Guía de buenas prácticas para la gestión de contenidos multilingües y en lengua no inglesa de los repositorios

Créditos de la imagen | Tommaso D'Incalci | Ikon Images CC BY-NC

30 de octubre de 2023

Elaborada por el grupo de trabajo de la COAR para el Apoyo al Multilingüismo y los Contenidos en Lengua no Inglesa de los Repositorios

Cita recomendada:

Grupo de trabajo de la COAR para el Apoyo al Multilingüismo y los Contenidos en Lengua no Inglesa en los Repositorios. Octubre de 2023. *Guía de buenas prácticas para la gestión de contenidos multilingües y en lengua no inglesa de los repositorios, versión 2*. Confederation of Open Access Repositories (COAR). DOI: 10.5281/zenodo.10053918

Agradecimientos

Miembros colaboradores del grupo de trabajo

Iryna Kuchma, EIFL (presidenta), Ucrania

Jagadish Aryal, Social Science Baha, Nepal

Andreas Czerniak, Universidad de Bielefeld - Biblioteca, Alemania

Christophe Dony, Biblioteca de la Universidad de Lieja, Bélgica

Joe Cera, Biblioteca de Derecho de Berkeley, Universidad de California, Estados Unidos

Sebastiano Giorgi-Scalari, Universitat Oberta de Catalunya, España

Gussun Gunes, Departamento de Gestión de Información y Archivos de la Universidad de Mármara, Turquía

Gultekin Gurdal, Instituto de Tecnología de Izmir İYTE, Turquía

Johanna Havemann, AfricArXiv, Alemania

Libio Huaroto Pajuelo, Universidad Peruana de Ciencias Aplicadas, Perú

Alan Ku (Gu Liping), Biblioteca Nacional de Ciencias, Academia China de Ciencias, China

Pierre Lasou, Biblioteca de la Universidad Laval, Canadá

Norma Aída Manzanera Silva, Centro de Investigaciones sobre América del Norte, Universidad Nacional Autónoma de México, México

Lautaro Matas, LA Referencia, España/América Latina

Ayako Mikami, Universidad de Hokkaido, Japón

Tomoki Nagase, Instituto Nacional de Informática, Japón

Tomasz Neugebauer, Universidad Concordia, Canadá

Jean-François Nominé, INIST, Francia

Milica Sevkusic, ITS SASA/EIFL, Serbia

Kathleen Shearer, COAR, Canadá/Internacional

Freddy Sumba, CEDIA, Ecuador

Nos gustaría dar las gracias a las siguientes personas por su aportación en estas recomendaciones:

Ginny Barbour (en nombre de Open Access Australia), Susanna Fiorini, Raina Heaton, JPCOAR (Japan Consortium for Open Access Repositories), Susan Kung, Cheng-Jen Lee, Devon Murphy, David Quispe, François Renaville, Sadie Roosa, Joan Spanne y Kelly Stathis.

Introducción **4**
Resumen de las recomendaciones **6**
Recomendaciones detalladas **7**
1. Declarar el idioma del recurso en cada elemento 7
2. Declarar el idioma de los metadatos (p. ej., atributo xml:lang) 10
3. Utilizar códigos de idioma estándar (de dos o tres letras) (ISO 639) 12
3.1. Introducción a las etiquetas y códigos de idiomas 12
3.2. Resumen del árbol de decisión para elegir una etiqueta de idioma 13
4. Habilitar el formato de codificación UTF-8 en el repositorio y utilizar el alfabeto/sistema de escritura original siempre que sea posible. Si es necesario transliterar metadatos, utilizar normas reconocidas (p. ej., ISO) 15
4.1. Transliteración frente a transcripción 15
5. Si el software del repositorio admite varios idiomas de interfaz, configurar la interfaz de usuario en el idioma o idiomas nativos del grupo destinatario, junto con una opción en inglés 16
6. Escribir el nombre o nombres de personas utilizando el sistema de escritura empleado en el documento depositado y proporcionar un identificador persistente que permita una identificación inequívoca 17
7. Incluir palabras clave en muchos idiomas, utilizar vocabularios y tesauros multilingües si es posible 20
7.1. Vocabularios y tesauros multilingües 20
8. Recomendaciones para gestores de repositorios sobre contenidos traducidos 25
Anexo 1. Casos de uso y retos **30**
Anexo 2. Declarar el idioma del recurso en cada elemento: ejemplos de implementación siguiendo las normas/directrices sobre metadatos **38**
Anexo 3. Declarar el idioma de los metadatos (atributo xml:lang): ejemplos de implementación siguiendo las normas/directrices sobre metadatos **40**
Anexo 4. Ejemplos de implementación de las normas ISO 639-1, ISO 639-2 e ISO 639-3 **43**
Anexo 5. Corrección de inconsistencias del código de idioma en registros de repositorios DSpace **45**
Anexo 6. Corrección de la falta del idioma de documento en registros de repositorios EPrints **46**
Anexo 7. Herramientas de tratamiento de textos **48**

Introducción

El multilingüismo es un rasgo fundamental en entornos comunicativos de investigación positivos, inclusivos y diversos. Publicar en una lengua local garantiza que el público de diferentes países tenga acceso a la investigación que dichos países financian y, a su vez, iguala las condiciones para los investigadores que hablan varias lenguas. La iniciativa de Helsinki sobre multilingüismo en la comunicación académica sostiene que la exclusión de las lenguas locales o nacionales para la publicación académica es el factor más importante —y, a menudo, olvidado— que impide a las sociedades utilizar y aprovechar la investigación realizada en el lugar donde viven. Aunque la posición dominante de una *lingua franca* —el inglés— sea útil para extender la divulgación de ideas a todo el mundo, también impide el uso de los resultados de la investigación en el ámbito local.

Tras décadas de políticas que han llevado a los investigadores a publicar en inglés, estamos empezando a apreciar un giro en esta tendencia. En Europa, Asia y muchas otras áreas de actuación, los responsables de políticas están introduciendo nuevas medidas para animar a los investigadores a publicar en lenguas locales y autóctonas. La recomendación de la UNESCO sobre la ciencia abierta, por ejemplo, pide a los estados miembros que fomenten "el multilingüismo en la práctica de la ciencia, en las publicaciones científicas y en las comunicaciones académicas". Esto reafirma y está en consonancia con afirmaciones no tan recientes, como las de la Declaración Universal de los Derechos Humanos, que instan a no discriminar a los investigadores por su idioma, y la recomendación de la UNESCO sobre la promoción y el uso del plurilingüismo y el acceso universal al ciberespacio, que insta a la comunidad a "tomar las medidas necesarias para reducir las barreras lingüísticas y [...] garantizar que todas las culturas puedan expresarse y tener acceso al ciberespacio en todas las lenguas, incluidas las indígenas".

El multilingüismo plantea un desafío particular a la hora de descubrir recursos. Si el idioma de un recurso académico no está etiquetado adecuadamente, no será indexado correctamente por las herramientas de descubrimiento. Ello se debe a que la indexación implica prácticas de análisis de texto, como el *stemming*, la lematización (agrupación de las formas flexionadas de una palabra para que puedan analizarse como un único elemento) y el tratamiento adecuado de las palabras vacías. Todas estas técnicas de análisis textual son muy específicas de cada lengua. La inclusión de etiquetas de idioma y la adopción de otras prácticas similares permite a los buscadores, agregadores e indexadores de información y a los servicios de descubrimiento identificar correctamente el idioma de todo el texto y procesar cada elemento en consecuencia. Además, es posible que los investigadores y otros buscadores de información que solo sepan leer en una o dos lenguas deseen conocer toda la investigación relevante en su área, independientemente de la lengua en que esté publicada. La correcta designación del idioma del recurso es importante para apoyar esta necesidad y ofrecer una mejor consulta multilingüe.

En agosto de 2022, la COAR puso en marcha el grupo de trabajo de la COAR para el Apoyo al Multilingüismo y los Contenidos en Lengua no Inglesa de los Repositorios con el objetivo de desarrollar y promover buenas prácticas en la gestión de contenidos multilingües y en lengua no inglesa en los repositorios. Basándose en diecisiete casos de uso aportados por distintas

comunidades de interesados (gestores y usuarios de repositorios, autores y traductores, agregadores y sistemas de descubrimiento), el grupo de trabajo identificó tres áreas relevantes para su labor: la mejora de la descubribilidad de contenido en lengua no inglesa, la curación de contenidos multilingües en un repositorio y la admisión de traducciones. Los casos de uso se documentan en el anexo 1.

En junio de 2023, el grupo de trabajo publicó un primer conjunto de proyectos de recomendaciones para su revisión por parte de la comunidad. La consulta dio lugar a una amplia variedad de aportaciones, las cuales fueron revisadas por el grupo de trabajo e incorporadas a una segunda versión de recomendaciones. El presente documento presenta las recomendaciones actualizadas a partir de las aportaciones de la comunidad.

Las recomendaciones establecen una serie de buenas prácticas para gestores de repositorios y desarrolladores de software para repositorios, y abordan cuestiones relativas a metadatos, palabras clave multilingües, interfaces de usuario, formatos y licencias que mejorarán la visibilidad, el descubrimiento y la reutilización del contenido de los repositorios en una amplia variedad de lenguas.

Es nuestro deseo que estas recomendaciones sean ampliamente adoptadas por los repositorios de todo el mundo. Algunas de las recomendaciones pueden ser inmediatamente adoptadas por los gestores de repositorios, mientras que otras llevarán algo más de tiempo y su plena aplicación requerirá esfuerzos colectivos por parte de gestores de repositorios, agregadores, investigadores y desarrolladores de software. En los próximos meses, la COAR y el grupo de trabajo difundirán ampliamente las recomendaciones y trabajarán para avanzar en su adopción en los repositorios de todo el mundo.

Resumen de las recomendaciones

Creadores y curadores de metadatos

Declarar el idioma del recurso en cada elemento.

Declarar el idioma de los metadatos (p. ej., atributo xml:lang).

Utilizar códigos de idioma estándar (de dos o tres letras) (ISO 639).

Habilitar el formato de codificación UTF-8 en el repositorio y utilizar el alfabeto/sistema de escritura original siempre que sea posible. Si es necesario transliterar metadatos, utilizar normas reconocidas (p. ej., ISO).

Si el software del repositorio admite varios idiomas de interfaz, configurar la interfaz de usuario en el idioma o idiomas nativos del grupo destinatario, junto con una opción en inglés.

Escribir el nombre o nombres de personas utilizando el sistema de escritura empleado en el documento depositado y proporcionar un identificador persistente que permita una identificación inequívoca (p. ej., ORCID).

Incluir palabras clave en muchos idiomas, utilizar vocabularios y tesauros multilingües si es posible.

Recomendaciones para gestores de repositorios sobre contenidos traducidos.

Desarrolladores de software y plataformas de repositorio

Garantizar que los códigos de idioma puedan utilizarse sistemáticamente en todas las colecciones del depósito y que sean compatibles.

Exponer el idioma de los metadatos mediante un protocolo de intercambio de metadatos, por ejemplo: OAI-PMH, GraphQL API, etc.

Mejorar la compatibilidad con los códigos de idioma ISO (por ejemplo, con los códigos de tres letras necesarios para algunos idiomas).

Garantizar que se admitan varios idiomas de interfaz.

Garantizar que los identificadores persistentes queden expuestos por OAI-PMH. El grupo de trabajo PIDs in Dublin Core™ ha desarrollado recomendaciones para hacer posible la exposición de identificadores persistentes (PID), incluyendo ORCID, a través de OAI-PMH.

Facilitar palabras clave en varios idiomas con el fin de aumentar la descubribilidad de contenido multilingüe en el repositorio. Por ejemplo, habilitando la integración en tiempo real con Wikidata (un caso sería, cuando un usuario empieza a escribir en el campo de

metadatos apropiado, la aparición de los términos relevantes de Wikidata en una lista desplegable para que el usuario los seleccione).

Permitir una asignación automática de términos controlados basada en los metadatos existentes.

Recomendaciones detalladas

1. Declarar el idioma del recurso en cada elemento

Recomendación

Es preceptivo declarar el idioma principal del documento. Los metadatos de idiomas deben codificarse utilizando el código ISO 639 (para más detalles, véase el punto 2.3 Utilizar códigos de lengua estándar (de dos o tres letras) (ISO 639)).

Directrices

Si el documento tiene un único idioma, los metadatos de idioma identificarán la lengua principal del recurso. La atribución de la lengua principal del recurso debe llevarse a cabo en el elemento.

Ejemplo 1. Idioma en XML de Dublin Core sencillo con codificación ISO 639-1

```
<dc:language>en</dc:language>
```

Ejemplo 2. Idioma en MODS con codificación ISO 639-2

```
<language>
<languageTerm authority="iso639-2b" type="code"
authorityURI="http://id.loc.gov/vocabulary/iso639-2"
valueURI="http://id.loc.gov/vocabulary/iso639-2/eng">eng</languageTerm>
</language>
```

Si todo el documento (por ejemplo, un volumen editado) tiene secciones importantes del texto en varios idiomas, los metadatos se repetirán para mencionar cada idioma.

Ejemplo 3. Documento bilingüe (francés/inglés) en XML de Dublin Core sencillo con codificación ISO-639-1

```
<dc:language>en</dc:language>
<dc:language>fr</dc:language>
```

Ejemplo 4. Documento bilingüe (francés/inglés) en MODS con codificación ISO 639-2

```
<language>
```

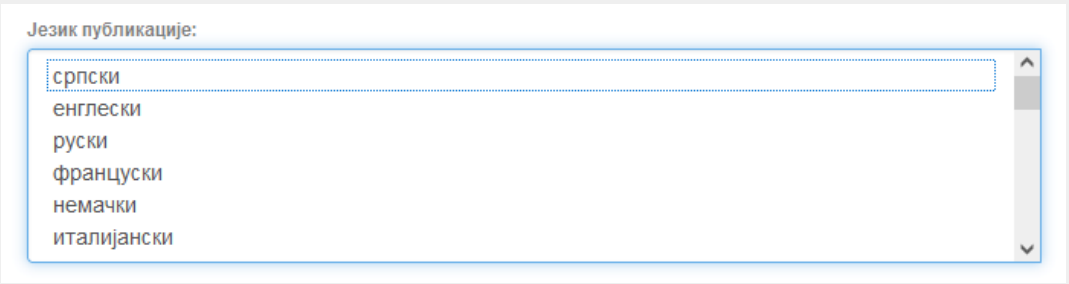
```
<languageTerm authority="iso639-2b" type="code"
authorityURI="http://id.loc.gov/vocabulary/iso639-2"
valueURI="http://id.loc.gov/vocabulary/iso639-2/eng">eng</languageTerm>
</language>
<language>
<languageTerm authority="iso639-2b" type="code"
authorityURI="http://id.loc.gov/vocabulary/iso639-2"
valueURI="http://id.loc.gov/vocabulary/iso639-2/fre">fre</languageTerm>

</language>
```

Ejemplo 5. EPrints
EPrints puede ampliarse para declarar información de idioma en el elemento o en el archivo, pero esto no sucede por defecto en EPrints. Del mismo modo, los *plug-ins* de exportación XML de EPrints, los metadatos incrustados y el código de la interfaz OAI-PMH podrían ampliarse para definir atributos xml:lang, pero esta ampliación no tiene lugar por defecto.

Ejemplo 6. Open Science Framework, OSF (marco de ciencia abierta)
Las nuevas mejoras en los metadatos del OSF para todos los proyectos, registros y prepublicaciones del OSF incluyen ahora el idioma de los materiales (más información en Nuevos metadatos del OSF para apoyar el cumplimiento de la política de intercambio de datos).

Ejemplo 7. Repositorio del Archivo Digital de la Academia Serbia de Ciencias y Artes (DAIS)
El contenido del repositorio está en más de quince idiomas. El filtrado por idioma no está habilitado y la búsqueda no es fácil de usar. El idioma se declara en el momento del envío, seleccionándolo de una lista desplegable (campo obligatorio). Pueden seleccionarse varios idiomas, pero "multilingüe" no existe como valor. Recomendaciones de las directrices de envío:

- El cuerpo principal del texto está en un idioma, y el título, los resúmenes y las palabras clave están en otro: solo declare el idioma principal (pero hay que proporcionar los metadatos en ambas lenguas).
- Publicación (por ejemplo, un volumen editado) con secciones importantes del texto en diferentes idiomas: declare todos los idiomas.
- Texto completo proporcionado en paralelo en varios idiomas: declare todos los idiomas.



En los metadatos, el idioma seleccionado se muestra como un código ISO de dos letras (pero se ofrece en formato legible por humanos en la lista desplegable).

dc.language.iso	sr

Ejemplo 8. Metadatos de preservación digital (PREMIS y METS)
METS (Metadata Encoding and Transmission Standard, o norma de codificación y transmisión de metadatos) y PREMIS (Preservation Metadata: Implementation Strategies, o metadatos para la preservación: estrategias de aplicación) son dos normas de metadatos que suelen utilizarse conjuntamente para proporcionar una amplia compatibilidad de metadatos en la preservación y gestión de objetos digitales. METS se centra principalmente en la codificación de metadatos descriptivos, administrativos y estructurales, y proporciona un marco para organizar y vincular varios tipos de metadatos dentro de un documento XML estructurado. PREMIS, por su parte, se centra en documentar las acciones, eventos y procesos implicados en la preservación a largo plazo de los objetos digitales. METS puede servir de contenedor para diversos metadatos, incluidos los metadatos PREMIS, lo que permite integrar la información específica de la preservación en el ámbito más amplio de la organización y descripción de objetos digitales.

El modelo de datos PREMIS no clasifica explícitamente el idioma como metadato técnico o descriptivo. La norma PREMIS tampoco define específicamente elementos o subelementos para capturar la información relativa al idioma. Sin embargo, el tipo de *propiedades significativas* y los componentes de la unidad semántica de valor en PREMIS bastan para capturar el idioma sin necesidad de un elemento XML específico del idioma.

El idioma puede considerarse un metadato técnico y significativo para la preservación utilizando el elemento <significantProperties> en PREMIS. Las propiedades significativas representan los aspectos de un objeto digital que influyen en su representación, comportamiento o interpretación, como el formato de archivo, el algoritmo de compresión, la versión de software, la resolución, el espacio de color y otras características técnicas que afectan a la representación y accesibilidad del objeto. Recomendamos que el idioma se codifique como una propiedad significativa en este sentido, utilizando PREMIS. Veamos un ejemplo donde el idioma especificado es el inglés:

<premis:significantProperties>

<premis:significantPropertiesType>language</premis:significantPropertiesType>
<premis:significantPropertiesValue>en</premis:significantPropertiesValue>
</premis:significantProperties>

Además, la información del idioma, si se considera una característica descriptiva del contenido intelectual (es decir, un metadato descriptivo), puede incrustarse en el documento METS mediante Dublin Core (utilizando la etiqueta <dc:language>) u otra sección de metadatos dentro del contenedor METS. El idioma también puede incrustarse en el METS como metadato técnico para documentos de texto utilizando TextMD[1] dentro del elemento PREMIS <objectCharacteristicsExtension>.

En el anexo 2 se muestran más ejemplos de implementación que siguen normas/directrices sobre metadatos.

2. Declarar el idioma de los metadatos (p. ej., atributo xml:lang)

Recomendación

Se recomienda utilizar el atributo xml:lang para indicar el idioma del campo de metadatos. Debido a la cardinalidad [0, 1], el atributo xml:lang podría describir el mismo elemento en diferentes idiomas, por lo que sería más preciso que el elemento dc:language.

Directrices

Independientemente de que se asuma mayoritariamente el inglés como estándar, el contenido debería ser expuesto con una referencia al idioma utilizado. Merece la pena hacerlo en el ámbito del repositorio, ya que ninguna otra parte interesada, como los agregadores (por ejemplo, BASE, OpenAIRE, etc.), puede deducir el idioma a partir del contenido de los metadatos.

En caso de que los elementos depositados tengan un título, u otros elementos de metadatos, en más de un idioma (por ejemplo, el título principal y el título de un resumen o *abstract*), habrá que asegurarse de que la información del idioma se indique mediante el atributo/subpropiedad xml:lang[2] y sea expuesto adecuadamente a través de protocolos de intercambio de metadatos, como OAI-PMH. Dado que algunos agregadores no pueden recoger la información completa ni todos los campos repetidos, se recomienda respetar el orden de introducción de metadatos de los títulos, es decir, proporcionar primero el título principal. Si es posible, se utilizará la alternativa dc.title para títulos adicionales.

[1] https://www.loc.gov/standards/textMD/
[2] https://www.w3.org/International/techniques/authoring-xml#natlang

Los agregadores como OpenAIRE[3] y BASE[4] identificarán correctamente el título principal basándose en la información proporcionada en el campo que indica el idioma del documento, independientemente del orden en que se haya introducido.

Sin embargo, en OAI-PMH el idioma de los metadatos no queda expuesto, por lo que solicitamos a los desarrolladores de software que lo tengan en cuenta en futuras versiones de sus plataformas.

Ejemplo 9. Cómo asignar el idioma cuando hay más de una lengua en los campos de metadatos
Se utiliza el atributo xml:lang para indicar el idioma del campo de metadatos.

<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>

He aquí un ejemplo de MODS de AILLA, en el que utilizamos los códigos de idioma ISO 639-3:
<titleInfo lang=”eng”>
<title>Iskonawa Oral Tradition</title>
</titleInfo>
<titleInfo lang=”spa”>
<title>Tradición Oral Iskonawa</title>

</titleInfo>
Véase el anexo 3 para más ejemplos de implementación siguiendo las normas/directrices sobre metadatos.

Ejemplo 10. DSpace
En DSpace 7, los pares de valores de idioma pueden incluir los idiomas e identificadores de idioma que se desee. Por defecto, DSpace ofrece diez pares de valores de idioma: inglés de Estados Unidos (en_US), inglés (en), español (es), alemán (de), francés (fr), italiano (it), japonés (ja), chino (zh), portugués (pt) y turco (tr). Pero esto se puede personalizar completamente en el archivo submission-forms.xml. Este archivo puede incluir identificadores de tres letras en caso de que en el material de destino de una colección haya idiomas con identificadores de tres letras. Durante los envíos, los valores de idioma aparecen en forma de lista desplegable, mientras que, en el modo de edición, el idioma es un campo de texto libre.

[3] https://www.openaire.eu
[4] https://www.base-search.net

Véase el anexo 5 para obtener información acerca de cómo se solucionan las inconsistencias del código de idioma en los repositorios que funcionan con versiones de DSpace anteriores.

Ejemplo 11. TIND IR
TIND IR es un repositorio basado en MARC. Esto significa que la forma más fácil de incluir información sobre contenidos multilingües es a través del campo 041 y los subcampos relevantes.[5]

Ejemplo 12. WEKO 3
WEKO3 es un software de repositorios desarrollado por el Instituto Nacional de Informática de Japón (NII) basado en el software INVENIO del CERN. Este software funciona con JAIRO Cloud, un sistema de repositorios basado en la nube que cuenta con el apoyo del Consorcio Japonés para Repositorios de Acceso Abierto (JPCOAR) y el NII. En WEKO 3, el esquema de metadatos JPCOAR es compatible por defecto y puede añadirse un atributo de idioma para cualquier metadato siempre que esté permitido en el esquema. En concreto, se acepta la codificación ISO 639-3 para el idioma del texto, mientras que para un atributo de idioma de otros elementos de metadatos se acepta ISO 639-1. Con cada campo puede añadirse una etiqueta de idioma con codificación ISO de dos caracteres utilizando el menú desplegable, la casilla de verificación y el botón de selección.

3. Utilizar códigos de idioma estándar (de dos o tres letras) (ISO 639)

3.1. Introducción a las etiquetas y códigos de idiomas

Identificar los idiomas de forma inequívoca es esencial para la interpretación, agregación y reutilización de los contenidos de una investigación. Las normas para las etiquetas de idioma se han ido actualizando y ampliando desde los comienzos de internet, en la década de 1990. La última norma de etiquetado de idiomas está definida por la BCP 47 del IETF (RFC 5646) en combinación con la ISO 639-3.

Las etiquetas de idioma son indispensables en los formatos HTML, XML y RDF para identificar un idioma natural. El código de idioma, en formato de dos o tres caracteres (como "en" para inglés), es el componente principal de una etiqueta de idioma y lo establece la norma ISO 639 (partes 1-3). El código de idioma puede ir seguido de subetiquetas destinadas a precisar o acotar el rango del idioma codificado de la siguiente manera:

idioma-extensión de idioma-sistema de escritura-región-extensión de la variante-uso privado.

[5] https://www.loc.gov/marc/bibliographic/bd041.html

La práctica del etiquetado de idiomas no supone ninguna complicación para un gran número de lenguas conocidas; la norma ISO 639 incluye códigos para más de 7.900 lenguas (a enero de 2023). Sin embargo, es importante tener en cuenta que las lenguas menos conocidas y las variedades regionales o las etapas históricas de las lenguas pueden no estar suficientemente representadas en la norma ISO 639. Las subetiquetas opcionales de la codificación BCP 47 ofrecen varias posibilidades para una identificación más precisa. La subetiqueta *private-use* "x" definida en BCP 47 puede utilizarse para identificar modalidades lingüísticas[6]. Además, ISO 639 es una norma que ha cambiado con el tiempo y ahora ofrece la oportunidad de proponer cambios:

> "El conocimiento de las lenguas humanas en cualquier momento histórico nunca es completo ni perfecto, sino que está en constante expansión. Dado el carácter exhaustivo de la norma ISO 639-3, los cambios en el conjunto de códigos son inevitables, sobre todo en lo que respecta a las lenguas menos conocidas o recientemente identificadas."[7]

Es importante recordar que el objetivo principal del etiquetado de idiomas es identificar y representar con exactitud, en función del contexto lingüístico y tecnológico, la lengua en uso. Si un código de 2 letras (ISO 639, parte 1) no es adecuado en un contexto específico, deberá utilizarse un código de 3 letras (ISO 639, partes 2 y 3) u otras subetiquetas (como para escritura, región o uso privado) con el fin de garantizar la interoperabilidad y la precisión en la identificación de la lengua. El conjunto de lenguas incluidas en la parte 1 de la norma ISO 639 se considera un subconjunto de la parte 2 y cualquier código de 2 letras de la parte 1 con un código de 3 letras correspondiente en la parte 2 o 3 se considera sinónimo con la misma extensión. Por ejemplo, los identificadores "fra", "fre" y "fr" designan la misma lengua. BCP 47 recomienda utilizar códigos de 2 letras siempre que existan, pero ISO 639 establece que debe permitirse la libre elección entre sinónimos siempre que sea posible. En este informe, recomendamos seguir esta parte de la recomendación de BCP 47 y utilizar códigos de 2 letras siempre que existan, pero, dependiendo del contexto específico de uso, puede ser adecuado utilizar códigos de 3 letras.

3.2. Resumen del árbol de decisión para elegir una etiqueta de idioma

A continuación presentamos un árbol de decisión resumido para elegir una etiqueta de idioma:

1. Busque el código de idioma en ISO 639.
2. Si encuentra un código ISO 639, parte 1, de dos letras para el idioma, utilícelo. Vaya al punto 5.
3. Si encuentra un código ISO 639, partes 2 o 3, de tres letras para el idioma, utilícelo. Vaya al punto 5.
4. Utilice la subetiqueta "x", reservada para uso privado, para definir un código de idioma personalizado. Vaya al punto 5.

[6] Tal como se describe en https://aclanthology.org/2020.lrec-1.408.pdf, por ejemplo.
[7] https://iso639-3.sil.org/code_changes/introduction

5. Decida si es necesaria y pertinente una subetiqueta para identificar el idioma. Por ejemplo, si el hecho de que se trate de una variante regional o un dialecto es importante en el contexto, considere la posibilidad de utilizar códigos de país ISO 3166 como subetiquetas (por ejemplo, "en-US" para el inglés americano). Si es necesario identificar variantes del sistema de escritura que sean relevantes, considere la posibilidad de utilizar como subetiquetas los códigos de escrituras ISO 15924 (por ejemplo, "sr-Latn" para el serbio en escritura latina).

Nota: la codificación de ISO 639 2 y 3 ha estandarizado algunas situaciones especiales:
* mis (de "miscellaneous"): se aplica cuando hay "idiomas no codificados".
* mul (de "multiple languages"): se utiliza cuando aparecen varios idiomas y no resulta práctico especificar todos los códigos de idioma correspondientes.
* und (de "undetermined"): se utiliza cuando hay que indicar un idioma pero no es posible identificarlo.
* zxx: figura en la lista de códigos como "sin contenido lingüístico", como los sonidos de animales (código añadido el 11 de enero de 2006).

El uso de códigos de idioma también puede resultar práctico en el caso de lenguas históricas o locales, regionales o clásicas (como el latín, el valón, etc.).[8]

Hallará más información sobre los códigos ISO 639-1, ISO 639-2 e ISO 639-3, y sobre las etiquetas de idiomas en el anexo 4.

Ejemplo 13. Lingüística y estudios lingüísticos
En lingüística y estudios lingüísticos, los códigos ISO 639-3 (de 3 letras) son un estándar. En primer lugar, la mayoría de las lenguas no tienen códigos de dos letras y, cuando los tienen, suelen prestarse a confusión porque no representan idiomas propiamente dichos (por ejemplo, cr para "cree", ms para "malayo" o zh para "chino"). Esto enturbia precisamente el tipo de diversidad que deseamos promover. Los lingüistas y los archivos lingüísticos también utilizan cada vez más los glottocodes (códigos de la base de datos Glottolog) para "languoides", ya que lo que llega a "contar" como lengua responde en gran medida a una visión política. Considere la posibilidad de disponer de un campo opcional para incluirlos.

Ejemplo 14. Repositorio institucional MiCISAN
El repositorio institucional MiCISAN utiliza la norma ISO 639-3.

[8] Ejemplos en valón: https://orbi.uliege.be/handle/2268/28421 y https://orbi.uliege.be/handle/2268/28419

4. Habilitar el formato de codificación UTF-8 en el repositorio y utilizar el alfabeto/sistema de escritura original siempre que sea posible. Si es necesario transliterar metadatos, utilizar normas reconocidas (p. ej., ISO)

Directrices y debate

UTF-8 es el formato de codificación de caracteres más extendido en la red (y en las tecnologías de internet). A fecha de 2023, está presente en el 98,0 % de todas las páginas web, y llega al 100 % en muchos idiomas.[9] Prácticamente todos los países y lenguas utilizan las codificaciones UTF-8 en la red en un 95 % de los casos.[10]

La mayoría de softwares de repositorios admite UTF-8 por defecto, como ocurre con DSpace 7, pero hay pasos en el proceso de instalación en los que es necesario asegurarse de que Tomcat utiliza UTF-8 por defecto o de forma similar.[11]

4.1. Transliteración frente a transcripción

La transliteración es la conversión de un texto de un sistema de escritura a otro (por ejemplo, del alfabeto griego al alfabeto latino) y se basa en la asignación de los grafemas de un sistema a otro de forma normalizada para que los lectores puedan reconstruir la ortografía original utilizando tablas de transliteración normalizadas o herramientas informáticas. Algunos países cuentan con normas de transliteración.

La transcripción es un tipo de conversión en el que el texto de la lengua de llegada captura el sonido en lugar de la ortografía.

A veces, la transliteración es inevitable. En las bases de datos bibliográficas y los catálogos de las bibliotecas encontramos enormes cantidades de metadatos transliterados o transcritos. En algunas comunidades de investigación, transliterar nombres e incluso títulos es una práctica habitual. Aunque ahora es común que se admita la codificación UTF-8, estas prácticas persisten. Si un repositorio ya contiene metadatos transliterados o su comunidad designada requiere que los metadatos sean transliterados, se recomienda hacer lo siguiente:

- Utilizar normas de transliteración reconocidas.
- Si es posible, elegir una norma y declararla en las páginas de preguntas frecuentes / manual de usuario / "acerca de" incluidas en el repositorio.
- Si ello no es posible, declarar todos los estándares utilizados en las páginas de preguntas frecuentes / manual del usuario / "acerca de".
- Proporcionar enlaces a las directrices de transliteración pertinentes (por ejemplo,. Biblioteca del Congreso) o herramientas[12] en las páginas de preguntas frecuentes / manual de usuario / "acerca de" para garantizar que los lectores puedan reconstruir la ortografía original.

[9] Encuesta de uso de las codificaciones de caracteres desglosadas por clasificación, *w3techs.com*. Consultado el 23-08-2023.
[10] https://en.wikipedia.org/wiki/UTF-8#cite_note-W3TechsWebEncoding-10
[11] https://wiki.lyrasis.org/display/DSDOC7x/Installing+DSpace
[12] Por ejemplo: https://alittlehebrew.com/transliterate/, https://www.translitteration.com

- Si los nombres de los autores están transliterados, utilizar identificadores como ORCID para conectar las diferentes variantes de los nombres.
- Utilizar códigos de idioma para los metadatos transliterados (por ejemplo, este recurso recomienda el-Latn para indicar texto en griego transliterado al alfabeto latino).

Si hay estándares de transliteración, debe evitarse la transcripción porque las normas no siempre son claras y se dificultaría la reconstrucción de la ortografía original. Si la transcripción es inevitable, habrá que seguir las reglas y normas de las lenguas.

Ejemplo 15. DataCite
DataCite requiere la transliteración de caracteres no latinos:

contributorName
Ocurrencias: 1
Definición: nombre completo de la persona que ha hecho la contribución.
Valores permitidos, ejemplos, otras restricciones: si se utiliza persona contribuidora, entonces contributorName es obligatorio.
Ejemplos: Patel, Emily; ABC Foundation
El formato para el nombre de persona puede ser: apellido, nombre de pila.
Los nombres en alfabeto no latino deben transliterarse según los esquemas ALA-LC.

5. Si el software del repositorio admite varios idiomas de interfaz, configurar la interfaz de usuario en el idioma o idiomas nativos del grupo destinatario, junto con una opción en inglés

Directrices

Una interfaz de usuario en varios idiomas facilita la navegación por el repositorio a usuarios de distintas comunidades. Por ejemplo, una interfaz en la lengua o lenguas maternas facilita a los usuarios locales la comprensión de los campos de metadatos al depositar contenidos, y a los usuarios internacionales la navegación y la búsqueda de contenidos.

Ejemplo 16. Dataverse
Dataverse admite interfaces de usuario en varios idiomas y depende de las traducciones de la comunidad realizadas por voluntarios. En el marco del proyecto Social Sciences and Humanities Open Cloud (SSHOC), se realizaron importantes avances hacia la creación de un directorio de paquetes de idiomas y se diseñó la herramienta en línea Weblate para facilitar nuevas traducciones. También está disponible una guía de usuario para Weblate.[13]

Ejemplo 17. DSpace

[13] https://doi.org/10.5281/zenodo.4807371

DSpace admite varios idiomas de interfaz. El texto que aparece en la interfaz se denomina "mensajes" y los archivos de mensajes (paquetes de idiomas) son aportados y gestionados por la comunidad al margen del núcleo del proyecto DSpace para permitir actualizaciones y lanzamientos más regulares. Los usuarios pueden modificar las traducciones de la comunidad o crear las suyas propias y enviarlas al proyecto dspace-api-lang en Github. Aparte de los mensajes, es posible localizar otros elementos, como las páginas de ayuda, los formularios de entrada y las plantillas de correo electrónico. En la documentación de DSpace pueden encontrarse instrucciones sobre cómo habilitar la interfaz en varios idiomas. DSpace 7 da un paso importante para facilitar las traducciones de la interfaz de usuario: https://wiki.lyrasis.org/pages/viewpage.a: las directrices para el soporte multilingüe en el *front-end* (UI) están disponibles.[14]

Ejemplo 18. EPrints
EPrints admite varios idiomas de interfaz y utiliza carpetas de "frases" y otros archivos específicos para cada idioma. Por defecto, EPrints solo viene empaquetado con frases en inglés, pero la comunidad ha compartido muchas traducciones a través de EPrints Bazaar y EPrints Files. EPrints utiliza el estándar de idioma ISO de dos letras para especificar subdirectorios de frases y otros tipos de directorios específicos de cada lengua, como por ejemplo:

- lib/lang/en/phrases/
- lib/lang/fr/static/
- lib/lang/de/templates/

Los metadatos de temas de EPrints están diseñados para dar cabida a etiquetas multilingües, con lo que las etiquetas de temas pueden mostrarse en función del idioma que el usuario haya definido para la interfaz.

EPrints está diseñado para utilizar por defecto frases en inglés; y si le faltan frases para otro idioma de interfaz declarado, utilizará las frases en inglés hasta que se añadan las frases que faltan. Hay una página wiki técnica sobre traducciones, pero puede que esté obsoleta, ya que ha sido editada muy pocas veces en los últimos años.

6. Escribir el nombre o nombres de personas utilizando el sistema de escritura empleado en el documento depositado y proporcionar un identificador persistente que permita una identificación inequívoca

Recomendación

[14] https://wiki.lyrasis.org/display/DSDOC7x/Multilingual+Support

Se recomienda escribir el nombre o nombres de persona tal y como aparecen en el documento depositado y proporcionar un identificador persistente que permita una identificación inequívoca, como ORCID.

Directrices y debate

Hay dos criterios principales a la hora de tratar los nombres de persona en los repositorios:

- Utilizando una forma preferida unificada que esté definida en un archivo de autoridad.
- Capturando los nombres tal y como aparecen en el documento depositado.

El primer criterio es el típico de los catálogos de las bibliotecas, donde se utiliza la forma unificada como encabezamiento del catálogo. Dependiendo del país, los nombres escritos originalmente en un alfabeto no latino se latinizarán o, a la inversa, se transcribirán/transliterarán según las normas utilizadas en un país concreto. Si un repositorio ofrece metadatos integrados que pueden importarse desde gestores de referencias y citas recomendadas preformateadas, este criterio puede no ser el adecuado, porque el formato del nombre en el repositorio diferirá del de la publicación.

Si los nombres se capturan tal y como aparecen en las publicaciones depositadas, el nombre de una misma persona aparecerá en el repositorio en varios formatos. En este caso, es importante utilizar identificadores persistentes, como ORCID, para garantizar la correcta identificación y conectar distintas versiones del nombre.

Ejemplo 19. DSpace
En la versión anterior de DSpace se necesitaba una solución provisional para mostrar las distintas versiones de los nombres de una manera sencilla para el usuario (por ejemplo, mediante una aplicación interna adicional). Ahora, DSpace CRIS y DSpace 7 no solo admiten una integración bidireccional con ORCID, sino que también tratan a las personas como entidades (entidades CRIS y entidades configurables, respectivamente).[15]

Ejemplo 20. Exposición de identificadores persistentes a través de OAI-PMH
También es importante garantizar que los identificadores persistentes se expongan a través de OAI-PMH. El grupo de trabajo PIDs in Dublin Core™ ha desarrollado recomendaciones para hacer posible la exposición de identificadores persistentes, incluyendo ORCID, a través de OAI-PMH. Se proponen dos soluciones, y ambas cubren varios casos de uso:
Opción 1: utilizar un atributo "id" con propiedades Dublin Core
Tanto el PID (identificador persistente) como la etiqueta son conocidos:
<dc:creator id="https://orcid.org/0000-0003-1541-5631">Walk, Paul</dc:creator>

Se conoce la etiqueta, pero no el PID:

[15] Por ejemplo: https://scholars.lib.ntu.edu.tw/cris/rp/rp00095 (DSpace CRIS)

<dc:creator id="">Walk, Paul</dc:creator>

Se conoce el PID, pero no la etiqueta:
<dc:creator id="https://orcid.org/0000-0003-1541-5631"></dc:creator> o
<dc:creator id="https://orcid.org/0000-0003-1541-5631"/>
Esta opción no es adecuada si es necesario incluir más de un PID.

Opción 2: utilizar propiedades anidadas para los identificadores
El PID y la etiqueta son conocidos:
<dc:creator>
<dc:identifier>https://orcid.org/0000-0003-1541-5631</dc:identifier>
<foaf:name>Walk, Paul</foaf:name>

Se conoce la etiqueta, pero no el PID:
<dc:creator>
<foaf:name>Walk, Paul</foaf:name>
</dc:creator>
o
<dc:creator>Walk, Paul</dc:creator>

Se conoce el PID, pero no la etiqueta:
<dc:creator>
<dc:identifier>https://orcid.org/0000-0003-1541-5631</dc:identifier>
</dc:creator>

En esta opción pueden proporcionarse varios PID (identificadores persistentes) para una misma propiedad:
<dc:creator>
<dc:identifier>https://orcid.org/0000-0003-1541-5631</dc:identifier>
<dc:identifier>http://paulwalk.net</dc:identifier>
<foaf:name>Walk, Paul</foaf:name>
</dc:creator>

Ejemplo 21. Esquema de metadatos JPCOAR
El esquema de metadatos JPCOAR también incluye un elemento para el identificador del investigador: jpcoar:nameIdentifier.[16] Este elemento puede utilizarse repetidamente con diferentes tipos de PID (KAKEN ID, ORCID, identificador del investigador y otros) y quedar expuesto para la Institutional Repositories Database (base de datos de repositorios institucionales del Japón) a través de OAI-PMH.

[16] https://schema.irdb.nii.ac.jp/en/schema/3-1

7. Incluir palabras clave en muchos idiomas, utilizar vocabularios y tesauros multilingües si es posible

Directrices y debate

La inclusión de palabras clave en muchos idiomas aumenta la descubribilidad de contenido en los repositorios. A este respecto, es importante distinguir entre palabras clave de texto libre (o "etiquetas") y términos controlados derivados de un vocabulario multilingüe controlado o tesauro. En el primer caso, las palabras clave en varios idiomas se introducen en el campo dc:subject, asegurándose de que el idioma está correctamente codificado.

Ejemplo 22. Repositorio institucional MiCISAN
El repositorio institucional MiCISAN siempre respeta el idioma del recurso. Si, por ejemplo, el recurso está en español, en el caso de las palabras clave se utilizan calificadores para diferenciar los metadatos en distintos idiomas:[17]

dc.subject.keywordseng
institutional repository

dc.subject.keywordseng
interoperability

dc.subject.keywordsspa
metadatos

dc.subject.keywordsspa
repositorio institucional

dc.subject.keywordsspa
interoperabilidad

Es importante señalar que el uso de palabras clave de texto libre no garantiza la coherencia ni revela relaciones jerárquicas entre los términos. El problema puede mitigarse seleccionando manualmente los términos que se añadirán como palabras clave desde vocabularios controlados. Sin embargo, una solución óptima contempla la integración de vocabularios controlados multilingües en el repositorio.

7.1. Vocabularios y tesauros multilingües

El uso de vocabularios controlados o tesauros[18] para los metadatos bibliográficos garantiza que un mismo concepto se describa de forma coherente. Junto con el uso de términos controlados para indicar el tipo de recurso, versión o derechos de uso, también se pueden utilizar vocabularios controlados para describir el contenido temático del recurso. En los vocabularios controlados multilingües, lo ideal es que cada término tenga un único equivalente en cada idioma y que las relaciones entre los términos sean las mismas. En un

[17] https://ru.micisan.unam.mx/handle/123456789/22232?show=full
[18] Registro de vocabularios controlados: https://bartoc.org/

entorno digital, a los términos del vocabulario se les asignan identificadores persistentes que pueden resolverse fácilmente.

Sin embargo, el uso de vocabularios controlados o tesauros conlleva ciertos desafíos.

- Para poder integrarse con los repositorios, los vocabularios controlados deben expresarse como datos interpretables por el sistema.
- Equivalencia forzada: no siempre es posible encontrar equivalentes verdaderos en todas las lenguas, por lo que el significado de los términos y las relaciones entre ellos en un idioma no se reflejarán con exactitud en sus equivalentes de otros idiomas.
- El proceso de asignación de términos controlados puede llevar mucho tiempo.
- Los investigadores no suelen estar familiarizados con el concepto de vocabularios controlados. Si los bibliotecarios no tienen los conocimientos especializados necesarios, los términos pueden ser demasiado generales e imprecisos.
- Existen muchos vocabularios controlados específicos de una disciplina y no es posible aplicarlos todos en repositorios multidisciplinares. Por otro lado, cabe la posibilidad de que los vocabularios generales no sean capaces de describir el contenido con precisión.
- Los vocabularios controlados más utilizados (por ejemplo, los encabezamientos de materia de la Biblioteca del Congreso de Estados Unidos o los vocabularios Getty) no incluyen por igual los distintos contextos culturales y grupos sociales.

En general, las plataformas de software de repositorios admiten la implementación de vocabularios controlados, aunque las soluciones de integración no siempre son óptimas.

Ejemplo 23. Dataverse
Dataverse es el repositorio de datos de código abierto desarrollado por el IQSS de la Universidad Harvard. La sólida comunidad de Dataverse está ayudando a mejorar la funcionalidad básica y a seguir desarrollándola. DANS-KNAW entregó el repositorio Dataverse listo para producción (Docker/k8s) a las comunidades de la Nube Europea de Ciencia Abierta (EOSC) CESSDA, CLARIN y DARIAH. Para hacer frente a los retos de integración de conjuntos de datos heterogéneos y multilingües, DANS-KNAW introdujo el soporte de vocabularios controlados externos (modelo de metadatos CESSDA conectado al marco Skosmos; soporte para la infraestructura de metadatos de componentes CLARIN y el Tesauro Europeo de Lenguas de las Ciencias Sociales (ELSST) alojado por CESSDA y ODISSEI en Skosmos; CESSDA tiene una versión actualizada con más propiedades de idioma).

Ejemplo 24. DSpace
DSpace ofrece tres formas de integrar vocabularios controlados:[19]

- Pares de valores en forma de lista controlada.

[19] https://wiki.lyrasis.org/display/DSDOC7x/Authority+Control+of+Metadata+Values

- Archivo XML con los términos (por ejemplo, para permitir la integración del sistema de clasificación decimal Dewey o el tesauro de términos griegos en los repositorios).[20]
- SolR Authority (se utilizó para la integración de ORCID antes de DSpace 7).[21]

Las entidades configurables de DSpace 7, aunque no se diseñaron inicialmente para este uso, podrían ser otra forma de implementar vocabularios controlados.

Ejemplo 25. TRIPLE
Se han dado varios intentos de superar las limitaciones de los vocabularios controlados existentes. El proyecto TRIPLE desarrolló un nuevo vocabulario controlado multilingüe (en nueve idiomas) para ciencias sociales y humanidades partiendo de vocabularios existentes.

Ejemplo 26. El vocabulario RVM Web
El vocabulario RVM Web, gestionado por la Universidad de Laval y utilizado por bibliotecas de todo Canadá, es un ejemplo de vocabulario controlado que intenta eliminar los sesgos culturales, históricos y coloniales:

- Es bilingüe (inglés y francés), pero no para todos los términos.
- En sus comienzos (hacia 1970) se creó traduciendo los encabezamientos de materias de la Biblioteca del Congreso (LCSH), y ahora ya es un producto independiente.
- La versión inglesa utiliza MeSH, AAT (tesauro Getty), HOMOsaurus (como novedad) y LCSH.
- Las relaciones entre los distintos términos del tesauro o del vocabulario se establecen manualmente. No es un proceso automatizado.
- La versión abierta RVM FAST no contiene AAT MeSH ni HOMOsaurus, solo LCSH (existe un plan para hacerla compatible con Linked Open Data con el fin de incluirla en DBpedia a corto plazo). (Véase un ejemplo aquí)
- Está incluido en WebDewey.
- Existe un identificador único para cada término (todavía no es público).
- Retos:
  - Sincronización entre los distintos productos (LCSH, RAMEAU, AAT, etc.). Se espera que mejore con el uso de las identificaciones.
  - ¿Cómo impulsar las actualizaciones de los términos utilizados en los sistemas?

Ejemplo 27. Wikidata
La integración de Wikidata en los repositorios, ya implantada en Europeana, puede ser una solución ampliamente aplicable para proporcionar palabras clave multilingües. Wikidata se basa tanto en el *crowdsourcing* como en los archivos de autoridades existentes y ya contiene un gran número de datos en varios idiomas.

[20] La primera integración del vocabulario de tipos de recursos de la COAR se llevó a cabo utilizando pares de valores o archivos XML: http://repositorium.sdum.uminho.pt/handle/1822/46066?mode=full
[21] https://wiki.lyrasis.org/display/DSDOC7x/ORCID+Authority

La importación de términos de varios vocabularios está habilitada a través de la herramienta Mix'n'match.

Wikidata como palabras clave

Wikidata es una base de conocimientos libre con más de 100 millones de elementos de datos. Actúa como almacenamiento central de datos estructurados generales de conceptos, e incluye etiquetas/traducciones de los conceptos en muchos idiomas. Por ello, el uso de conceptos de Wikidata como vocabulario controlado de palabras clave es especialmente prometedor, ya que puede proporcionar una mayor interoperatividad multilingüe con una menor inversión de tiempo.

Por ejemplo, el repositorio de datos de investigación basado en CKAN Depositar reutiliza Wikidata como fuente de palabras clave (más información aquí). Cabe señalar que las etiquetas de los conceptos de Wikidata seguirían cambiando. Por ello, Depositar solo almacena y expone el propio identificador (por ejemplo, "Q11030"). Después, consulta la API de MediaWiki para obtener las últimas etiquetas multilingües de un vocabulario de Wikidata. Sería mejor almacenar y exponer tanto 1) la etiqueta más reciente como 2) la etiqueta (antigua) en el momento de asignar una palabra clave.
Los conceptos de WikiData y otros términos de vocabulario controlado pueden codificarse utilizando las etiquetas JATS[22] <kwd-group> y <kwd> y añadiendo los atributos ***vocab***, ***vocab-identifier*** y ***vocab-term-identifier*** definidos en la Standards Tag Suite (STS) de la NISO:

- el nombre del vocabulario controlado ("wikidata") en el atributo ***vocab***[23]
- el identificador del vocabulario ("https://www.wikidata.org/") en el atributo ***vocab-identifier***[24]

[22] El Journal Article Tag Suite (JATS) es un formato XML utilizado para describir bibliografía científica publicada en línea. Se trata de una norma técnica elaborada por la Organización Nacional de Normas de Información de Estados Unidos (NISO) y aprobada por el Instituto Nacional Estadounidense de Normas con el código Z39.96-2012. El proyecto NISO fue una continuación del trabajo realizado por NLM/NCBI y popularizado por el repositorio PubMed Central de la NLM como norma de facto para el archivo e intercambio de revistas científicas de acceso abierto y sus contenidos con XML. Con la normalización de la NISO, la iniciativa de la NLM ha adquirido un mayor alcance, y otros repositorios, como SciELO y Redalyc, adoptaron el formato XML para los artículos científicos: https://en.wikipedia.org/wiki/Journal_Article_Tag_Suite
En JATS (Journal Article Tag Suite), cualquier campo de metadatos podía etiquetarse con un idioma. En el formato DTD del esquema JATS, el atributo xml:lang puede aplicarse a prácticamente cualquier elemento; véase: https://jats.nlm.nih.gov/articleauthoring/tag-library/1.2/attribute/xml-lang.html.
Ejemplos: títulos traducidos de PubMed Central https://www.ncbi.nlm.nih.gov/pmc/pmcdoc/tagging-guidelines/article/dobs.html#dob-at-transtitle.
Al utilizar el esquema JATS, el idioma de las palabras clave se registra mediante el atributo *xml:lang* de la etiqueta <kwd-group> (véase: https://jats.nlm.nih.gov/articleauthoring/tag-library/1.2/element/kwd-group.html). JATS agrupa las palabras clave por idioma con una serie de etiquetas <kwd> justo debajo de la etiqueta <kwd-group> de cada idioma.

[23] https://www.niso-sts.org/TagLibrary/niso-sts-TL-1-2-html/attribute/vocab.html

[24] https://www.niso-sts.org/TagLibrary/niso-sts-TL-1-2-html/attribute/vocab-identifier.html

- el identificador/URL de cada palabra clave en el atributo ***vocab-term-identifier*** (por ejemplo, "Q11030").[25] Para Wikidata, se trata del identificador del concepto, no de la etiqueta específica del idioma del concepto.

Hay varias formas de hacerlo. La norma JATS agrupa las palabras clave por idioma utilizando la etiqueta <kwd-group>. A continuación se muestra un ejemplo de etiquetado de metadatos de conceptos de Wikidata sobre fotografía (Q11633) y periodismo (Q11030) con las etiquetas de concepto en inglés (photography, journalism) y polaco (fotografia, dziennikarstwo) utilizando XML JATS:

```
<kwd-group xml:lang="en" vocab="wikidata" vocab-identifier="https://www.wikidata.org/">
  <kwd vocab-term-identifier="Q11633">photography</kwd>
  <kwd vocab-term-identifier="Q11030">journalism</kwd>
</kwd-group>
<kwd-group xml:lang="pl" vocab="wikidata" vocab-identifier="https://www.wikidata.org/">
  <kwd vocab-term-identifier="Q11633">fotografia</kwd>
  <kwd vocab-term-identifier="Q11030">dziennikarstwo</kwd>
</kwd-group>
```

Es posible que las actuales tecnologías de repositorios presenten limitaciones para esto.

Recomendación: añada todos los atributos descritos en el ejemplo: ***vocab***, ***vocab-identifier*** y ***vocab-term-identifier***

Recomendaciones para desarrolladores de software/plataformas de repositorios

- Se recomienda habilitar una integración en tiempo real con Wikidata (por ejemplo, cuando un usuario empieza a escribir en el campo de metadatos apropiado, los términos relevantes de Wikidata aparecen en una lista desplegable para que el usuario los seleccione).

- Se recomienda permitir una asignación automática de términos controlados basada en los metadatos existentes.

La indexación automática de contenidos podría hacer más eficiente el proceso de asignación de términos controlados. Este planteamiento, que se ha probado en repositorios institucionales individuales, ya es utilizado por los agregadores. Por ejemplo, Europeana enriquece automáticamente los metadatos a partir de vocabularios y conjuntos de datos externos, como GeoNames y DBpedia, y utiliza las relaciones semánticas y traducciones que ofrecen esos vocabularios. BASE asigna términos calculados del sistema de clasificación decimal Dewey basándose en los metadatos disponibles. El mismo planteamiento se utiliza en la plataforma de descubrimiento multilingüe GoTriple, donde los contenidos recogidos de diversas fuentes se anotan automáticamente utilizando términos controlados, gracias a lo cual es posible realizar búsquedas en varios idiomas en GoTriple.

[25] https://www.niso-sts.org/TagLibrary/niso-sts-TL-1-2-html/attribute/vocab-term-identifier.html

Otros avances podrían incluir la asignación de términos controlados a partir del texto completo de los documentos depositados y permitir una importación automatizada de los términos controlados asignados por los agregadores.

8. Recomendaciones para gestores de repositorios sobre contenidos traducidos

El multilingüismo y la traducción están estrechamente vinculados y se complementan entre sí. Las traducciones y los contenidos traducidos deben ser reconocidos como contribuciones válidas al ecosistema de la investigación y, como tales, apoyados y reconocidos como un valioso producto académico. Además, es necesario promover la diversidad lingüística en la cultura de la investigación. Para ello hay que fomentar y acreditar adecuadamente la traducción como práctica y como resultado. Esto puede lograrse en parte poniendo en práctica las siguientes ocho recomendaciones específicas:

1. **Incluir un campo específico para la función del traductor o traductores en los formularios de depósito de los archivos y repositorios en línea para dar cabida a los créditos del traductor (p. ej., dc.contributor.translator)**

Véanse algunas directrices en Rivero, Monica, Robert Estep, y Lorena Gauthereau-Bryson, "Digitization Practices for Translations: Lessons Learned from the Our Americas Archive Partnership Project", D-Lib Magazine, 17 (2011) doi:10.1045/september2011-rivero.

2. **Si es posible, dar cabida a la identificación del traductor con otros campos, como ORCID u otros identificadores interoperables similares; también organización o afiliación, si las hay**

3. **Incluir (sub)campo(s) específico(s) para el estado de la traducción del documento, lengua o lenguas utilizadas para el contenido traducido y lengua o lenguas del documento de origen, designándolas preferiblemente con códigos de idioma internacionales normalizados**

Ejemplo 28. Directrices CRIS v1.2 actualizadas
Las directrices CRIS v1.2 actualizadas ya incluyen el multilingüismo y las traducciones automáticas. Véase un ejemplo en la página 13 del tutorial CERIF y un ejemplo de CRIS en las directrices OpenAIRE actualizadas con un atributo adicional "trans=" en el elemento:
Versión:
https://github.com/openaire/guidelines-cris-managers/releases/tag/v1.2.0 con los valores:

- h := humana
- m := automática

- o := original

Véase también el CERIF XSD en la búsqueda cfTrans_Type.

Animamos también a llevar a cabo desarrollos similares en otros estándares y plataformas.

4. Permitir a los usuarios apuntar a otros registros relacionados del contenido traducido añadiendo campos de relación, como dc.relation

Las opciones de etiquetado en este campo de relación podrían incluir la información:

- "Es una traducción de"
- "Se ha traducido desde" (Esta segunda opción se podría utilizar más bien en caso de traducción parcial, por ejemplo, de un capítulo o una sección de un libro.)

Ejemplo 29. Crossref
Crossref trata el idioma como un atributo basado en códigos de dos letras que pueden utilizarse en varios elementos y gestiona las traducciones mediante atributos de relación específicos: isTranslationOf; hasTranslation.[26] Pero el problema es que un proveedor puede no utilizarlos al registrar sus contenidos.

Referencia del esquema:
El idioma existe en el esquema común como atributo:
https://data.crossref.org/schemas/common5.3.1.xsd
<xsd:attributeGroup name="language.atts">
<xsd:annotation>
<xsd:documentation>Los atributos de idioma se basan en la norma ISO 639</xsd:documentation>
</xsd:annotation>
<xsd:attribute name="language" use="optional">

La relación de traducción también es posible; véase el esquema de relación:
https://data.crossref.org/schemas/relations.xsd
<xsd:element name="intra_work_relation">
<xsd:complexType mixed="true">
<xsd:attribute name="relationship-type" use="required">
<xsd:annotation>
<xsd:documentation>Se utiliza para definir relaciones entre elementos que son esencialmente el mismo trabajo, pero que pueden diferir en algún aspecto que afecte a la cita, por ejemplo una diferencia de formato, idioma o revisión. Asignar identificadores diferentes a exactamente el mismo elemento disponible en un lugar o como copias en varios lugares puede ser problemático y debe evitarse.
</xsd:documentation>
</xsd:annotation>

[26] Véase la documentación: https://www.crossref.org/documentation/schema-library/markup-guide-metadata-segments/multi-language/; https://www.crossref.org/documentation/schema-library/metadata-deposit-schema-5-3-1/ y https://www.crossref.org/documentation/schema-library/markup-guide-metadata-segments/relationships/

```
<xsd:simpleType>
<xsd:restriction base="xsd:string">
<!--      Crossref   -->
<xsd:enumeration value="isTranslationOf"/>
<!--  hasTranslation  -->
<xsd:enumeration value="hasTranslation"/>
<!--  isTranslationOf  -->
```

5. **Dar cabida a este campo de relación con otros campos de identificación que apunten al documento original**

Puede utilizarse un DOI u otro PID del documento original, o un identificador o una URL si no existe un resolvedor interoperable.

Directrices y ejemplos

Las opciones de exportación de registros de contenidos traducidos deberían incluir idealmente toda la información anterior, con especificidades relativas al tipo de traducción cuando sea necesario, para lo cual se proporcionan más detalles de contexto a continuación.

Un ejemplo de registro para un contenido traducido o posteditado por un humano tendría el siguiente aspecto:

"Este material titulado '[título traducido]' es una traducción íntegra/parcial desde el [idioma - código de idioma estándar] con fecha [DD-MM-AAAA] a cargo de [nombre(s) de traductor(es) de 'título original'], escrito por [nombre(s) autor(es)] en [idioma - código de idioma estándar], publicado en [datos editoriales]/recuperado de [DOI, otro resolvedor de PID o URL].”

Traducción automática

En el transcurso de nuestra investigación, el tema de la traducción automática (TA) ha suscitado un acalorado debate en el seno del grupo de trabajo y hemos compartido la naturaleza de esta discusión en una entrada de blog: ¿Es posible aceptar en los repositorios contenidos académicos traducidos automáticamente?

Dada la complejidad de la cuestión, así como las implicaciones éticas, el grupo de trabajo ha optado por recomendar que los repositorios no acepten contenidos traducidos exclusivamente con motores de traducción. Esto coincide también con las recomendaciones del "Informe del grupo de trabajo Traducción y Ciencia Abierta" (2020). La TA debería ser percibida y utilizada como una tecnología de apoyo, etiquetada de forma transparente e inequívoca como asistencia automática, y habría que permitir que cambie dinámicamente en tiempo real en vez de mantenerla y conservarla en el repositorio como un recurso primario.

El grupo de trabajo seguirá analizando de cerca este escenario en rápida evolución y continuará estudiando las cuestiones y, posiblemente, publicando nuevas recomendaciones relacionadas con la traducción automática (TA) de textos académicos, la traducción asistida por TA y la TA de resúmenes y metadatos en repositorios.

Sin embargo, dos estudios exploratorios llevados a cabo como parte del proyecto francés Traducciones y ciencia abierta (cartografía y recopilación de corpus científicos bilingües y evaluación de la traducción automática en el contexto de los estudios de comunicación académica) han constatado que los investigadores han estado utilizando ampliamente la TA para traducir sus propias investigaciones y los metadatos relacionados, así como cargando contenidos multilingües en repositorios, incluso sin notificar que utilizaban TA. En estos casos, la TA puede ser, con mayor o menor precisión, posteditada, pero sin ningún grado de certeza en cuanto a la calidad a gran escala para el propietario o la entidad gestora del repositorio. Es posible que los repositorios no tengan la capacidad de detectar y revisar este material, ya que su volumen crece rápidamente. Además, esta es una práctica que los repositorios, en general, difícilmente pueden controlar debido a los costes y recursos. Por eso sería útil poner en marcha un sistema de aviso que permitiera a los investigadores proporcionar información sobre la naturaleza de la traducción subida al repositorio. Lo ideal es que este sistema de aviso distinguiera entre el contenido traducido y posteditado por humanos, y la traducción automática en bruto.

Este sistema de aviso también podría ser útil para repositorios con capacidad para mostrar una TA instantánea del contenido recuperado (de manera automática o bajo petición).

El aviso, que se mostraría como una advertencia para el usuario con el fin de concienciarlo sobre posibles errores y anticiparse a posibles reclamaciones, podría decir lo siguiente:

"Este documento/material es una traducción automática no revisada de [cita del original] realizada el día [DD-MM-AAAA] del [código del idioma de origen] al [código del idioma de destino] tal y como se publicó en [detalles de la publicación] / se recuperó de [DOI, otro resolvedor de PID o URL] utilizando [nombre de la herramienta de TA]. Esta traducción automática no ha sido revisada ni editada y se proporciona "tal cual" con el único propósito de ayudar a los usuarios a comprender al menos parte del contenido original expresado en [idioma de origen]. Esta cláusula no garantiza la corrección y exactitud de la citada traducción automática [en la lengua de destino] por parte de ninguna persona física o jurídica en ninguna parte de esta traducción. [En consecuencia, disponer de esta traducción no dará lugar a ninguna responsabilidad por parte de ninguna persona hacia ninguna otra persona en el caso de que se haga uso de ella, sea cual sea el propósito]. Se invita expresamente a los usuarios de esta traducción automática a que la hagan revisar, corregir o editar por un traductor profesional o un experto en la materia."

6. A no ser que el documento lo justifique (por ejemplo, traducción paralela, traducción comentada, versiones replicadas bilingües o multilingües), cargar las traducciones de los documentos como registros separados

Esto es especialmente apropiado en el caso de prefacios, introducciones u otras colaboraciones publicadas en volúmenes multilingües con múltiples colaboradores.

7. Promover el uso de licencias favorables a la (re)traducción para fomentar la traducción de contenidos de nueva producción y la

retraducción, así como promover los créditos de traducción (por ejemplo, CC-BY)

Encontrará más información a este respecto en: Susanna Fiorini, Franck Barbin, Martine Garnier-Rizet, Katell Hernandez Morin, Franziska Humphreys, et al., Rapport du groupe de travail "Traductions et science ouverte", [Rapport Technique] Comité pour la science ouverte. 2020, 44 p. ⟨hal-03640511⟩.

8. **Asegurarse de que se proporciona suficiente información y recomendaciones a los depositantes en un apartado de preguntas frecuentes u otra forma de implementar lo anterior**

Anexo 1. Casos de uso y retos

Estos son algunos casos de uso que impulsan las prácticas recomendadas:

1. Como miembro de una institución no inglesa, recibo en mi depósito documentos en inglés que tengo que describir.

Cuando se envía un documento nuevo en inglés al repositorio, es necesario describirlo con distintos campos de metadatos en diferentes idiomas (por ejemplo, resúmenes, títulos, palabras clave, tipo de documento) y utilizando vocabularios controlados no ingleses.
Ejemplo: la Universidad de Hokkaido utiliza el esquema de metadatos JPCOAR (los metadatos en varios idiomas se introducen en el mismo campo de metadatos, pero se diferencian mediante el atributo de idioma, por ejemplo, dc.description.abstract y dc.subject[27]). En dicho esquema, una columna de idioma en la parte derecha de la página muestra el código de idioma ISO de los metadatos. Cuando se depositan artículos de revistas, se incluyen todos los metadatos de la versión publicada (sin traducción del original; en las revistas en japonés, los resúmenes y las palabras clave suelen estar escritos también en inglés y el texto completo, en japonés); los resúmenes están en los metadatos y el atributo de idioma está incrustado; los nombres de los autores están en el idioma del artículo. Como mínimo hay un esquema para marcar metadatos para varios idiomas, pero surgen dudas sobre la descubribilidad de contenido y qué metadatos son más adecuados.

2. Como responsable de repositorio, a menudo gestiono artículos, tesis o disertaciones que están escritos en más de un idioma.

Todas las tesis y disertaciones llegan en francés, pero muchas contienen artículos insertados en forma de capítulos en la lengua en la que han sido escritos.
Ejemplo: en la Universidad de Lieja, si un documento está disponible en varios idiomas, cada versión en un idioma se pone a disposición como un registro diferente con metadatos en diferentes idiomas. Es el caso de un mismo documento en dos lenguas distintas para el que existen dos registros diferentes,[28] pero solo hay un atributo de idioma para el registro.

3. Como autor o autora, me gustaría ver mis artículos escritos en diferentes idiomas en un solo registro (para fines estadísticos e informativos).

Todos los artículos en varios idiomas se depositan en un único artículo y deben ser descritos adecuadamente.
Ejemplo: antes, la Universitat Oberta de Catalunya disponía de dos registros separados para los artículos en varios idiomas. Actualmente, a petición de los autores, las traducciones se unen en un único registro o incluso en el mismo documento de archivo, lo que simplifica el seguimiento de las citas y aumenta la visibilidad. Pero puede haber problemas para los agregadores de contenidos y los servicios de indexación.

[27] https://eprints.lib.hokudai.ac.jp/dspace/handle/2115/79104?mode=full&submit_simple=Show+full+item+record

[28] https://orbi.uliege.be/handle/2268/170862 y https://orbi.uliege.be/handle/2268/170863

4. Como responsable de repositorio, quiero ofrecer campos de envío en varios idiomas.

[ESTO PUEDE SER ESPECÍFICO DE DSPACE]. Al configurar los formularios de envío, las etiquetas y la ayuda/instrucciones de cada campo solo pueden estar escritas en un idioma. El multilingüismo solo puede conseguirse escribiendo la etiqueta en cada idioma en el mismo campo (Autor/Auteur).

5. Como responsable de repositorio, quiero tener el nombre y la descripción de una colección en más de un idioma.

Actualmente solo se permite un idioma para el nombre y la descripción de una colección.

Ejemplo: [ESTO PUEDE SER ESPECÍFICO DE DSPACE]. Estaría bien que los textos introductorios (en HTML), etc., de las comunidades/colecciones pudieran presentarse en varios idiomas. Esto podría lograrse fácilmente utilizando CSS y etiquetas div con nombre. Pero, lamentablemente, los atributos de HTML, como id y style, parecen eliminarse en la salida HTML. Es decir: texto se transforma en texto en la interfaz de usuario.

Como las colecciones y las comunidades son elementos en DSpace (y, por tanto, tienen sus propios metadatos), una forma de resolver este problema sería permitir seleccionar el idioma de los metadatos, como ya se puede hacer para los metadatos de objetos (es decir, los resúmenes).

Una solución rápida y sencilla al problema bilingüe de las colecciones/comunidades en DSpace es utilizar un delimitador (por ejemplo, la barra |) entre dos textos que describan estas entidades y sus campos de metadatos, según sea necesario. Basta con dividir el texto en el momento de la visualización para que solo se muestre el texto que esté activo en cada momento. Aquí puede ver la versión árabe de la lista de comunidades/colecciones. Al cambiar el idioma de la interfaz al inglés utilizando el icono del mundo situado en la parte superior, verá que aparecen todas en inglés. El mismo planteamiento se ha aplicado a los elementos de facetas, donde ahora verá valores controlados (como nombres de formatos/tipos, universidades/institutos superiores/departamentos, entidades, etc.) en varios idiomas.

6. Como responsable de repositorio, quiero poder gestionar las etiquetas en mi idioma de forma eficaz.

En los programas multilingües de código abierto (OJS, DSpace, EPrints, etc.), las etiquetas en inglés son las obligatorias a la hora de desarrollar nuevas funciones. Las actualizaciones de otros idiomas suelen quedarse atrás y son gestionadas posteriormente por la comunidad o, a veces, localmente. La traducción de las nuevas funcionalidades del software supone un gran reto.

Ejemplos: en el repositorio EPrints de ZORA (Zurich Open Repository and Archive) existe una versión alemana de la interfaz.

CSpace, en China, incluye un esquema de metadatos y una interfaz en diferentes idiomas, pero los gestores de repositorios siguen teniendo dificultades para describir el contenido de los repositorios.

Normalmente son los usuarios quienes seleccionan las etiquetas de idioma y también reciben formación sobre cómo depositar contenido multilingüe.

Los idiomas de interfaz de los repositorios desarrollados por el Centro Informático de la Universidad de Belgrado (Serbia) incluyen el inglés y el serbio (en ambos alfabetos: latino y cirílico).[29] Como los usuarios no estaban satisfechos con las traducciones disponibles, el equipo de desarrollo ideó una aplicación web propia para facilitar la traducción. La aplicación permite añadir, eliminar y cambiar las etiquetas seleccionadas en repositorios individuales o en todos los repositorios. Los cambios se propagan a los repositorios en 24 horas.

7. Como responsable de repositorio, quiero ofrecer la traducción de metadatos (por ejemplo, resúmenes, títulos y temas) al inglés.

Algunos metadatos deben traducirse al inglés utilizando herramientas de traducción automática.

Ejemplos: se utiliza una API de traducción de Google para traducir resúmenes, títulos y temas.

Esto también se podría conseguir recomendando o exigiendo en las directrices para el usuario una cantidad mínima de metadatos en inglés. En el archivo digital de la Academia Serbia de las Ciencias y las Artes se recomienda proporcionar al menos una breve descripción y palabras clave en inglés, ya que ello mejora la descubribilidad del contenido.

8. Como responsable de un repositorio nacional, debo depositar artículos en todas las lenguas de mi país.

Los contenidos están disponibles en los idiomas locales, pero algunos no disponen de código de lengua, no están en Unicode y no existen vocabularios controlados en esos idiomas.

Ejemplo: en Nepal, solo los títulos se introducen en lengua nepalí y el resto de metadatos están en inglés. No hay estandarización de palabras clave en lengua nepalí ni hay vocabularios controlados. Muchas lenguas locales no están en Unicode y a veces se utilizan palabras latinizadas. **Por ejemplo: किताब kitaba (latinizado) y book (la forma traducida al inglés). Esto genera problemas para la indexación de Google Scholar, que prefiere ver los metadatos en la lengua del artículo.**

9. Como responsable de repositorio, me gustaría exponer el idioma de los metadatos en OAI-PMH.

Actualmente no existe exposición del idioma de los metadatos en OAI-PMH.

[29] Por ejemplo: https://dais.sanu.ac.rs

Objetivo deseable: los repositorios deberían utilizar de forma coherente y consciente las etiquetas de idioma de los metadatos para garantizar que no se exponga información incorrecta sobre el idioma. Y un atributo de idioma debería ser exportable, incluyendo OAI-PMH. Otra opción podría ser un planteamiento proactivo por parte de los repositorios. Por ejemplo: descargar mensualmente una extracción de las hojas de referencia de metadatos y ponerlas a disposición pública para exponer los valores de idiomas.

10. Como agregador de contenido y responsable del sistema de descubrimiento, 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.

Existen problemas de indexación de contenidos con respecto al agregador (Solr, VuFind, etc.) porque no hay forma de separar los índices por idioma y utilizar herramientas específicas de cada idioma para enriquecer las experiencias de búsqueda.

La mayoría de los metadatos de los repositorios regionales no separan adecuadamente la información multilingüe. Incluso pueden encontrarse lenguas mezcladas en campos de metadatos textuales individuales.

Las palabras clave y los descriptores están en varios idiomas sin la identificación adecuada; cientos de repositorios utilizan vocabularios diferentes incluso en el mismo idioma. Se ha debatido en torno a la implementación de clasificadores automáticos para etiquetar los metadatos de los repositorios con vocabularios normalizados para la región.

Ejemplos: la red latinoamericana de repositorios de acceso abierto LA Referencia está desarrollando una herramienta de detección de idiomas (utilizando diferentes bibliotecas python para el procesamiento del lenguaje natural) para separar las lenguas en los campos textuales de metadatos con el fin de mejorar los metadatos para los agregadores. La idea es añadir etiquetas xml:lang adecuadas a cada campo textual de metadatos. Este etiquetado sería utilizado por el proceso de indexación con el fin de generar índices separados, pero aún así el problema de enfrentarse a varios idiomas en la interfaz de usuario de búsqueda es complejo de resolver.

CORE parece que utiliza una herramienta de detección de idiomas. Distinguir entre bosnio, croata, montenegrino y serbio es todo un reto, ya que se trata de lenguas muy similares. Debido a ello, las etiquetas de idiomas en CORE suelen ser incorrectas cuando se trata de estos idiomas. El uso de la etiqueta aglutinadora BCMS para los cuatro idiomas sería una solución a este problema.

11. Como agregador, me gustaría indexar correctamente los contenidos y ayudar a los usuarios a encontrarlos en sus idiomas.

Las directrices institucionales y temáticas de los repositorios de OpenAIRE (para la agregación de contenidos de repositorios) animan a utilizar el

atributo xml:lang para indicar el idioma de los metadatos. El agregador OpenAIRE admite la etiqueta xml language.
Ejemplo: **<dc:description>** Foreword [by] Hazel Anderson; Introduction; The scientific heresy: transformation of a society; Consciousness as causal reality [etc] **</dc:description>** **<dc:description** xml:lang="en-US"**>** A number of problems in quantum state and system identification are addressed. **</dc:description>** OpenAIRE admite la etiqueta de idioma xml y el agregador realiza comprobaciones de metadatos para el idioma, por ejemplo, en temas, títulos y resúmenes/descripciones, pero no en nombres. Se recomienda ORCID para los nombres; OpenAIRE I+T: título, descripción OpenAIRE también permite varios idiomas, que se indican en cada recurso de contenido.

12. Como investigador, quiero saber qué investigaciones hay en otros idiomas. Podría ser también un caso de uso para un paciente, etc.

Traducir los resúmenes y ponerlos a disposición, u ofrecer una opción de búsqueda por palabras clave en muchos idiomas podrían ser algunas de las soluciones. Las herramientas de aprendizaje profundo han empezado a hacerlo, por ejemplo, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
Ejemplos: en BASE, búsqueda multilingüe (el concepto de búsqueda está incluido en el tesaruro EuroVoc o el tesauro Agrovoc; ejemplo de búsqueda para climatología). Wikidata y Abstract Wikipedia proporcionan información independiente del idioma.

13. Como bibliotecario o archivero de preservación digital, necesito saber cómo debo incluir información en lenguaje natural en los metadatos técnicos y descriptivos para que los documentos de archivo digitales puedan indexarse de forma eficaz para su correcta recuperación y acceso.

Las mejores prácticas documentadas para la inclusión de información en lenguaje natural mediante normas de metadatos de preservación digital como METS y PREMIS facilitan una mayor accesibilidad, inclusión y diversidad de los archivos digitales.
Ejemplos: el idioma es una información necesaria para indexar con eficacia y recuperar texto (*stemming* —reducción de palabras a su raíz—, palabras vacías), contenidos de vídeo y audio (la conversión de voz a texto favorece la recuperación/indexación, el subtitulado de audio y la accesibilidad al vídeo). Pueden incluirse metadatos de idioma utilizando la etiqueta <dc:language> de Dublin Core como parte de los metadatos descriptivos internos (mdWrap) de un archivo METS.

Pueden incluirse metadatos de idioma como una de las propiedades significativas (<significantProperties>) de las unidades semánticas en PREMIS.

En el caso de documentos textuales, los metadatos de idioma se pueden incluir utilizando textMD, normalmente en forma de esquema de ampliación dentro de la sección de metadatos administrativos del estándar METS. El idioma también puede incluirse como parte de un documento textMD independiente dentro del elemento de PREMIS <objectCharacteristicsExtension>.

14. Como usuario, quiero poder utilizar una interfaz en mi propio idioma para enviar o consultar contenidos.

La interfaz del repositorio está disponible en varios idiomas.
Ejemplos: el repositorio de la Universitat Oberta de Catalunya pone a disposición del usuario final tres interfaces en sendos idiomas. Cada interfaz de idioma tiene los nombres de los campos de metadatos en su propio idioma, por ejemplo: *autor*, en catalán y español; *author*, en inglés. En todos los repositorios institucionales desarrollados por el Centro Informático de la Universidad de Belgrado, la interfaz de usuario final está disponible en inglés y serbio (tanto en alfabeto cirílico como latino). Sin embargo, las etiquetas y la ayuda en el formulario de entrada solo están disponibles en serbio porque no es posible alinearlas con el idioma de la interfaz en DSpace.

15. Como miembro de una institución en lengua inglesa, utilizo un catálogo para describir el contenido de mi repositorio, tanto en inglés como en otros idiomas.

El contenido se introduce en la lengua materna, pero puede haber problemas de encontrabilidad.
Ejemplo: en Berkeley Law (facultad de Derecho de la Universidad de Berkeley) se utiliza un sistema basado en MARC para describir contenidos. Al tratarse de menos del 1-3 % del contenido, no cabe esperar que la búsqueda con términos no ingleses devuelva ningún resultado a menos que el usuario esté buscando algo específico. No se utilizan términos temáticos en el repositorio, pero parece una forma fácil de aumentar la accesibilidad en otros idiomas. El catálogo y el repositorio están vinculados, y la búsqueda está disponible en muchos idiomas. Los catalogadores hablan muchos idiomas y son capaces de catalogar en lenguas no inglesas, pero aún así la mayor parte de la catalogación se realiza en inglés para hablantes monolingües.

16. Como institución que admite muchas traducciones, me gustaría que se reconociera el mérito de los traductores al depositar elementos traducidos en el repositorio.

Los traductores se pueden citar utilizando taxonomías, como por ejemplo, la taxonomía CREDIT, que actualmente solo está disponible en inglés y sería bueno contar con una traducción oficial a otros idiomas. Existen dos traducciones "no oficiales" al francés.[30] Los traductores son reconocidos en el repositorio institucional (por ejemplo, como colaboradores con nombres y funciones), pero no es el caso de otros archivos, como los de prepublicaciones.
Ejemplo: el repositorio de la Universidad de Lieja dispone de un campo de metadatos para el traductor (véase aquí).

17. Como traductor, me gustaría saber si existe una determinada traducción.

Como traductor, quiero saber si existe una traducción: • **Para una cita incluida en un documento fuente en el mismo idioma, pero necesito comprobar si existe una versión en el idioma de destino (original o traducida) del texto citado (con una referencia en las notas o la bibliografía del documento fuente) antes de decidir si traduzco la cita yo mismo o reutilizo la cita traducida existente en mi traducción.** • **Para utilizar textos sobre el mismo tema que la traducción que me han encargado, puede que necesite construir un corpus de documentos similares en las lenguas de partida y de llegada de mi encargo para utilizarlos en un software de concordancia que permita buscar cadenas de texto (palabras, términos, frases) en una lengua y recuperarlas en dos lenguas. Puedo buscar a través de una investigación documental una colección de documentos con su correspondiente traducción en la lengua de destino y, después, procesarlo con un software de alineación para obtener archivos con palabras y frases alineadas.** • **Para alinear traducciones existen dos opciones:** **a) Alimentar un sistema CAT (traducción asistida por ordenador).** **b) Alimentar los módulos de aprendizaje de un sistema de TA (traducción automática).**
Ejemplo: en todos estos casos, el hecho de que los documentos se registren con metadatos adecuados para designar la condición de original/traducción y apunten al equivalente o equivalentes en cuestión, podría ayudar a las búsquedas de escritorio mencionadas si los metadatos fueran interoperables con los motores de búsqueda, los catálogos de bibliotecas, los repositorios y los sistemas CRIS. Esto también será relevante en el campo de la edición de revistas, terminología, minería de textos y tecnologías lingüísticas. Para facilitar el trabajo en estos ámbitos necesitamos interoperabilidad e interconexiones entre los distintos sistemas.

[30] Véase https://coop-ist.cirad.fr/etre-auteur/reconnaitre-tous-les-contributeurs/3-la-taxonomie-credit-pour-identifier-toutes-les-contributions y https://www.redactionmedicale.fr/2018/03/la-taxonomie-credit-devrait-etre-utilisee-par-les-revues-francaises-pour-decrire-la-contribution-des

Translate Science está construyendo una herramienta de este tipo y por ello necesitamos buenos metadatos lingüísticos en los repositorios.

Anexo 2. Declarar el idioma del recurso en cada elemento: ejemplos de implementación siguiendo las normas/directrices sobre metadatos

Esquema Datacite 4.4	9 idiomas Uso: opcional Ocurrencias: 0-1 (no repetible) Codificación recomendada: IETF BCP 47 o códigos de idioma ISO 639-1
Dublin Core (DC)	Nombre del término: language Uso: opcional Ocurrencias: repetible La práctica recomendada es utilizar un valor no literal que represente un idioma de un vocabulario controlado como ISO 639-2 o ISO 639-3, o un valor literal consistente en una etiqueta de idioma de la mejor práctica actual 47 del IETF [IETF-BCP47].
Norma de metadatos para tesis y disertaciones electrónicas (ETDMS)	dc.language Uso: opcional Ocurrencias: 0-N (repetible) Los nombres de los idiomas deben registrarse utilizando la norma ISO 639-2 (o RFC 1766). Si no se especifica el idioma, se asume que es el inglés (en).
Esquema de descripción de objetos de metadatos (MODS)	Elemento de nivel superior: <language> Uso: opcional Ocurrencias: 0-N (repetible) Este recurso contiene textos en inglés y francés: <language> <languageTerm type="code" authority="iso639-2b">eng<languageTerm> </language> <language> <languageTerm type="code" authority="iso639-2b">fre<languageTerm> </language> Este recurso contiene texto en árabe egipcio, que está codificado como lengua individual en la norma ISO 639-3: <language> <languageTerm type="code" authority="rfc4646" >zh-Hans</languageTerm> </language> <language> <languageTerm type="code" authority="iso639-3" >arz</languageTerm> </language>

Directrices de OpenAIRE para repositorios bibliográficos, institucionales y temáticos	dc:language Uso: obligatorio si es aplicable (Mandatory if Applicable, MA) Ocurrencias: 0-N (repetible) Recomendación: tomar valores de una de las listas siguientes: ● IETF BCP 47, registro de subetiquetas lingüísticas de la IANA ● ISO 639-x, donde x puede ser 1, 2 o 3. Mejor práctica: utilizar la norma ISO 639-3, con lo que seguimos el estándar http://www.sil.org/iso639-3/ Si es necesario, repetir este elemento para indicar varias lenguas. Si las normas ISO 639-2 y 639-1 son suficientes para el contenido de un repositorio, pueden utilizarse alternativamente. Como hay un único mapeo, puede hacerse durante un proceso de agregación.
Japan Consortium for Open Access Repository (JPCOAR)	dc:language Uso: recomendado (R) Ocurrencias: 0-N (repetible: excepto término obligatorio) Instrucciones de uso Introducir las lenguas principales que se utilizan en el texto principal del recurso. Utilizar los códigos de idioma ISO 639-3. El uso de la macrolengua de ISO 639-3 es opcional. Notas No introducir nombres de idiomas. No introducir nombres de países. Introducir en orden de prioridad de idioma. Ejemplos recomendados El texto principal del recurso está en inglés. <dc:language>eng</dc:language> El texto principal del recurso está en inglés y japonés. <dc:language>eng</dc:language> <dc:language>jpn</dc:language> Ejemplos no recomendados No se recomienda ISO 639-1. <dc:language>ja</dc:language> No introducir varios idiomas en un mismo elemento. <dc:language>engjpn</dc:language> No utilizar mayúsculas ni caracteres de doble byte. <dc:language>JPN</dc:language> <dc:language> e n g </dc:language> No introducir nombres de idiomas. <dc:language>日本語</dc:language> No introducir nombres de países. <dc:language>US</dc:language> No introducir códigos de idioma distintos de ISO 639. <dc:language>en_US</dc:language>

Anexo 3. Declarar el idioma de los metadatos (atributo xml:lang): ejemplos de implementación siguiendo las normas/directrices sobre metadatos

Esquema Datacite 4.4	xml:lang="EN", por ejemplo <xs:element name="**title**" maxOccurs="unbounded"> <xs:annotation> <xs:documentation>Nombre o título por el que se conoce un recurso.</xs:documentation> </xs:annotation> <xs:complexType> <xs:simpleContent> <xs:extension base="xs:string"> <xs:attribute name="titleType" type="titleType" use="optional"/> <xs:attribute ref="xml:lang"/> Igualmente, para xs:element name="**creatorName**">, <xs:element name="**publisher**">, <xs:element name="**subjects**" minOccurs="0">, <xs:element name="**contributorName**">, <xs:element name="**rightsList**" minOccurs="0">, <xs:element name="**descriptions**" minOccurs="0">, <xs:element name="**language**" type="xs:language" minOccurs="0">, <xs:annotation> <xs:documentation>Primary language of the resource. Allowed values are taken from IETF BCP 47, ISO 639-1 language codes.</xs:documentation>
Dublin Core (DC)	Cuando se indique el idioma del valor, deberá codificarse utilizando el atributo "xml:lang". Por ejemplo: <dc:subject xml:lang="en">seafood</dc:subject> <dc:subject xml:lang="fr">fruits de mer</dc:subject>
Norma de metadatos para tesis y disertaciones electrónicas (ETDMS)	El idioma es un calificador global que puede utilizarse en cualquier elemento: https://ndltd.org/wp-content/uploads/2021/04/etd-ms-v1.1.html#qualifiers
Esquema de descripción de objetos de metadatos (MODS)	Hay atributos relacionados con el idioma: https://www.loc.gov/standards/mods/userguide/attributes.html#list ISO 639-2/b
Directrices del repositorio institucional y temático OpenAIRE	Uso del atributo xml:lang para indicar el idioma de los metadatos. Ejemplo: **<dc:description>** Foreword [by] Hazel Anderson; Introduction; The scientific heresy: transformation of a society; Consciousness as causal reality [etc] **</dc:description>** **<dc:description** xml:lang="en-US"**>** A number of problems in quantum state and system identification are addressed. **</dc:description>**

	OpenAIRE admite la etiqueta de idioma xml y el agregador realiza comprobaciones de metadatos para el idioma, por ejemplo, en temas, títulos y resúmenes/descripciones, pero no en nombres. Se recomienda ORCID para los nombres. OpenAIRE I+T: título https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/latest/field_title.html#dci-title. Descripción: https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/latest/field_description.html#attribute-lang-o OpenAIRE también permite varios idiomas https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/v4.0.0/field_language.html, el recurso de contenido tiene este idioma.
JPCOAR 1.0.2 https://schema.irdb.nii.ac.jp/ja/schema	El atributo xml:lang puede utilizarse para cada elemento. En principio se utilizará el código de idioma de dos dígitos de la norma ISO 639-1 (por ejemplo, "ja" para japonés, "en" para inglés). Para yomi japonés, utilizar “ja-Kana“. Cuando introduzca "yomi", deberá indicar información original (es decir, en kanji) con la descripción de que "xml:lang is 'ja'". En el caso del chino, es conveniente introducir por separado el chino simplificado como "zh-ch" y el chino tradicional como "zh-tw".
Esquema de metadatos JPCOAR 2.0 https://schema.irdb.nii.ac.jp/en/schema/2.0/14 https://schema.irdb.nii.ac.jp/en/schema/2.0/1	Cambio desde 1.0.2 : admite además "ja-Latn". Los extractos de la parte actualizada: La información de idioma para katakana yomi es xml:lang="ja-Kana", y para romaji yomi es xml:lang="ja-Latn". Allí donde introduzca un yomi, la información en xml:lang="ja" debe introducirse por separado del yomi.
Akdeniz, Esra, & Moilanen, Katja. (2023). Modelo de metadatos CMM CESSDA (3.0). Zenodo. https://doi.org/10.5281/zenodo.7528240	1.1.3.1. Idioma del título del estudio El idioma del contenido del elemento. M (Mandatory) ISO 639-1 Ocurrencias: 1 ddi:DDIInstance/s:StudyUnit/r:Citation/r:Title/r:String/@xml:lang Lo mismo para el idioma de subtitulación; Idioma del título alternativo; Idioma del motivo de la versión; Idioma del resumen; Idioma del tema de estudio (descriptivo); Idioma de la palabra clave (descriptivo); Idioma de la disciplina (texto libre); Idioma del tipo de fuente de datos (descriptivo); Idioma del modo de la colección de datos (descriptivo); Idioma de las condiciones de acceso a los datos; Idioma de las condiciones de acceso a los metadatos (estudio); Idioma del nombre completo de la organización; Idioma del nombre, abreviatura o acrónimo de la organización; Idioma de la descripción de la organización; Idioma de la descripción de la versión del conjunto de datos; Idioma del conjunto de datos; Idioma de la descripción del archivo del conjunto de datos; Idioma del nombre del archivo; Idioma del título del documento; Idioma del título de la publicación; Idioma

	del nombre de la revista/serie - 75 campos de metadatos en total para indicar el idioma; También hay campos de metadatos para indicar traducciones, por ejemplo 1.1.3.2 Estado de la traducción del título del estudio ¿Está traducido el contenido del elemento? R: verdadero, falso Ocurrencias: 0-1 ddi:DDIInstance/s:StudyUnit/r:Citation/r:Title/r:String/@isTranslated; y 28 campos de metadatos que mencionan la traducción

Anexo 4. Ejemplos de implementación de las normas ISO 639-1, ISO 639-2 e ISO 639-3

ISO 639-1 e ISO 639-2
Propiedad de idioma en el Data Catalog Vocabulary (DCAT) - Versión 2 (recomendación del W3C del 4 de febrero de 2020):

Alcance:	Se debería utilizar los recursos definidos por la Biblioteca del Congreso (ISO 639-1, ISO 639-2). Si se define un código ISO 639-1 (de dos letras) para el idioma, entonces se debería utilizar su IRI correspondiente; si no se define ningún código ISO 639-1, entonces se debería utilizar el IRI correspondiente del código ISO 639-2 (de tres letras).
Nota de uso:	Repita esta propiedad si el recurso está disponible en varios idiomas.

En el Data Catalog Vocabulary (DCAT) - Versión 3 se incluye el mismo redactado.
Borrador de trabajo del W3C del 10 de mayo de 2022.

Códigos para la representación de nombres de idiomas ordenados alfabéticamente por el código alfa-3/ISO 639-2: http://www.loc.gov/standards/iso639-2/php/code_list.php.

ISO 639-3
ISO 639-3 amplía los códigos ISO 639-2 con el objetivo de cubrir todas las lenguas naturales y funciona mejor para lenguas como el cebuano, el montenegrino o el quechua (que tiene variantes dependientes de la región del país). Por ejemplo, se recomienda en la Guía del repositorio ALICIA (también una videoguía, Perú).

Las recomendaciones de metadatos para el material de texto almacenado en repositorios de publicaciones finlandesas recomiendan la norma ISO 639-X para dc.language.iso. Es preferible utilizar los códigos de idioma de 3 caracteres de ISO 639-2 o ISO 639-3, según corresponda.

Sigue habiendo algunos problemas de implementación para un código de tres letras, ya que no todos los repositorios podrían admitirlo ahora (debido a problemas de software y lenguaje XML) y podría haber problemas similares con los agregadores (por ejemplo, OpenAIRE las recomendaciones https://www.w3.org/TR/xml/ y https://www.w3.org/TR/xml/#RFC1766).

Más información sobre las etiquetas de idiomas

Un artículo útil y más descriptivo sobre [etiquetas de idioma en HTML y XML] publicado en 2014 por W3 con ejemplos:

Examples:

Code	Language	Subtags
en	English	language
mas	Masai	language
fr-CA	French as used in Canada	language+region
es-419	Spanish as used in Latin America	language+region
zh-Hans	Chinese written with Simplified script	language+script

y una propuesta de uso.

language-extlang-script-region-variant-extension-privateuse

Para muchas lenguas menos conocidas, habladas por grupos minoritarios, y también para periodos históricos de las lenguas, simplemente no se dispone de códigos de idioma, que son la base de las etiquetas; véase a este respecto "[The Shortcomings of Language Tags for Linked Data When Modeling Lesser-Known Languages]", con recomendaciones para mejorar o desarrollar códigos de idioma ISO.

Anexo 5. Corrección de inconsistencias del código de idioma en registros de repositorios DSpace

Si la lengua de destino utiliza caracteres únicos, puede ser posible establecer automáticamente el valor de los metadatos de idioma.

He aquí un ejemplo en SQL para que DSpace especifique elementos utilizando una lengua de destino y les asigne un valor de idioma suponiendo que la lengua de destino no está representada por caracteres de 2 bytes:

```
update metadatavalue set text_lang='/*indicar aquí el código ISO del idioma de destino*/'
 where metadata_field_id in (/*indicar aquí los números de cada metadata_field_id cuyos metadatos acepten algún valor de cadena */)
 and length(text_value)!=octet_length(text_value)
 and text_value ~ '^[/*indicar aquí todos los caracteres específicos de uso exclusivo en la lengua de llegada† */].*'
 and (text_lang is null or text_lang != '');
```

† Se puede utilizar una expresión regular que abarque todos los caracteres de la lengua. Por ejemplo, para la escritura japonesa: [ぁ－ん ァ－ヿ 亜－腕]; y para la cirílica: [a－ԥA－ԤҤ－Ҿҥ－ҿ].

Con una tarea cron por la noche puede añadirse rápidamente el código "en" en cualquier metadato que no disponga del código de idioma. Véase cómo crear una consulta o función SQL para cambiar text_lang a "en".

Las herramientas Atmire CSV Power Tools pueden utilizarse para editar metadatos exportados (en y en_US, así como paréntesis, y otras cuestiones de idiomas).

Anexo 6. Corrección de la falta del idioma de documento en registros de repositorios EPrints

REAL es un repositorio de gestión de EPrints encargado en 2008 y que actualmente contiene más de 220.000 artículos repartidos en ocho colecciones. El contenido es diverso y está formado por artículos de investigación actuales subidos por investigadores y material digitalizado por la institución matriz, la Biblioteca y Centro de Información de la Academia Húngara de Ciencias. La versión actual del software REAL es la 3.3.15.

El campo de idioma de los documentos siempre ha estado presente, pero hasta ahora no había estado visible en los formularios de carga de documentos de la web, ni en ninguna de las vistas de un artículo, por lo que los depositantes o bibliotecarios no podían configurarlo ni comprobar su contenido.

<documents>
<document id="http://real.mtak.hu/id/document/xxxxx">
<files>
<file id="http://real.mtak.hu/id/file/yyyyy">
<filename>zzzzz.pdf</filename>
</file>
</files>
<eprintid>wwwwwwww</eprintid>
<format>text</format>
<language>hu</language>
<security>public</security>
</document>
</documents>

Recientemente hemos expuesto el campo y hemos descubierto que EPrints fijaba el contenido en función de la configuración de idioma utilizada en el navegador en el momento del depósito, es decir, que los valores que contiene son más o menos aleatorios. Para averiguar (y establecer) los valores correctos para cientos de miles de artículos, elaboramos una lista de identificadores de los artículos que deseábamos comprobar, descargamos los metadatos en formato DC, extrajimos el título e intentamos adivinar el idioma del documento basándonos en el idioma del título.

Nuestro script comenzaba con una hipótesis (la primera hipótesis era que el idioma del documento es el húngaro), las palabras del título se introducían en un corrector ortográfico, y si más de la mitad de las palabras eran
reconocidas, aceptábamos la hipótesis como verdadera. En la siguiente ejecución se comprobaron los elementos restantes con la hipótesis "el idioma es el inglés", y luego se probaron otros idiomas.

El fragmento de script C-shell que aparece a continuación muestra la prueba del título frente a la hipótesis "el idioma es el italiano" utilizando el corrector ortográfico Hunspell.
@ den = `grep ^title: $3-eprint-$item.txt |tr -d '{}[]'| awk -F':' '{print $2,$3}' | awk -F'=' '{print $1}' | hunspell -d it_IT -l | wc -l`

```
@ enu = `grep ^title: $3-eprint-$item.txt |tr -d '{}[]'| awk -F':' '{print $2,$3}' | awk -F='
'{print $1}' | wc -w`

@ discr = `echo $den $enu | ~/unixstat/stat/bin/dm "floor (x1/x2+0.49)"`
```

La experiencia con este método demuestra que, con algunos filtros, la tasa de error podría reducirse al 1-2 %, que es mucho mejor que el porcentaje de error actual del 40-50 %. Hay que tener en cuenta que existen documentos complicados, multilingües o muy técnicos (por ejemplo, de matemáticas) que suponen un desafío. No sabemos cómo etiquetar los documentos bilingües / multilingües.

Anexo 7. Herramientas de tratamiento de textos

Siempre que sea posible, habrá que especificar el idioma o idiomas del documento, de párrafos individuales y de frases mientras se escribe en la herramienta de tratamiento de textos.

Para especificar el idioma de determinados párrafos y frases en MS Word, OpenOffice, LibreOffice y herramientas similares, utilizaremos la configuración de idioma y los teclados adecuados mientras escribimos. Para especificar el idioma o idiomas en un documento existente, seleccionaremos el texto y definiremos el idioma utilizando la herramienta de idioma de la barra de herramientas o del menú. Para conservar esta información tras la conversión a PDF, el documento debe exportarse como PDF etiquetado. Sin embargo, dependiendo de la extensión PDF incorporada en el procesador de textos, esta información puede perderse durante la conversión a PDF.

W3C da recomendaciones sobre cómo especificar el idioma de un párrafo o frase con la entrada Lang en documentos PDF. Sin embargo, para aplicar estas recomendaciones en los archivos PDF se necesita el software comercial Adobe Acrobat.

LaTeX también permite trabajar en varios idiomas. Existen varios paquetes que permiten la composición tipográfica en distintos idiomas (como, babel o polyglossia), y esta función también está disponible en Overleaf, el editor colaborativo en línea de LaTeX.

Sin embargo, la interoperabilidad de las distintas herramientas de edición de texto y los formatos utilizados sigue siendo una asignatura pendiente. Se necesitan normas claras y la colaboración con los fabricantes de software para garantizar no solo que el texto creado en diversas herramientas de software siga siendo legible para humanos y máquinas, sino también que las diversas características y funcionalidades que estaban disponibles en el documento original (codificación, etiquetas, anotaciones, etc.) lo sigan estando tras la conversión a otros formatos.