Guia de bones pràctiques per a la gestió de continguts multilingües i en llengua no anglesa dels repositoris

Crèdits de la imatge | Tommaso D'Incalci | Ikon Images CC BY-NC

30 d'octubre de 2023

Elaborada pel grup de treball de la COAR per al Suport al Multilingüisme i els Continguts en Llengua no Anglesa en els Repositoris

COAR
Confederation of Open Access Repositories

Citació recomanada:

Grup de treball de la COAR per al Suport al Multilingüisme i els Continguts en Llengua no Anglesa en els Repositoris. Octubre de 2023. *Guia de bones pràctiques per a la gestió de continguts multilingües i en llengua no anglesa dels repositoris, versió 2*. Confederation of Open Access Repositories (COAR). DOI: 10.5281/zenodo.10053918

Agraïments:

Membres col·laboradors del grup de treball

Iryna Kuchma, EIFL (presidenta), Ucraïna

Jagadish Aryal, Social Science Baha, Nepal

Andreas Czerniak, Universitat de Bielefeld - Biblioteca, Alemanya

Christophe Dony, Biblioteca de la Universitat de Lieja, Bèlgica

Joe Cera, Biblioteca de Dret de Berkeley, Universitat de Califòrnia, Estats Units

Sebastiano Giorgi-Scalari, Universitat Oberta de Catalunya, Espanya

Gussun Gunes, Departament de Gestió d'Informació i Arxius de la Universitat de Marmara, Turquia

Gultekin Gurdal, Institut de Tecnologia d'Izmir İYTE, Turquia

Johanna Havemann, AfricArXiv, Alemanya

Libio Huaroto Pajuelo, Universitat Peruana de Ciències Aplicades, Perú

Alan Ku (Gu Liping), Biblioteca Nacional de Ciències, Acadèmia Xinesa de Ciències, la Xina

Pierre Lasou, Biblioteca de la Universitat Laval, Canadà

Norma Aída Manzanera Silva, Centre de Recerques sobre Amèrica del Nord, Universitat Nacional Autònoma de Mèxic, Mèxic

Lautaro Matas, LA Referencia, Espanya/Amèrica Llatina

Ayako Mikami, Universitat de Hokkaido, Japó

Tomoki Nagase, Institut Nacional d'Informàtica, Japó

Tomasz Neugebauer, Universitat Concordia, Canadà

Jean-François Nominé, INIST, França

Milica Sevkusic, ITS SASA/EIFL, Sèrbia

Kathleen Shearer, COAR, el Canadà/Internacional

Freddy Sumba, CEDIA, Equador

Ens agradaria donar les gràcies a les persones següents per la seva aportació en aquestes recomanacions:

Ginny Barbour (en nom d'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 i Kelly Stathis.

Introducció

El multilingüisme és un tret fonamental en entorns comunicatius de recerca positius, inclusius i diversos. Publicar en una llengua local garanteix que el públic de diferents països tingui accés a la recerca que aquests països financen i, al seu torn, iguala les condicions per als investigadors que parlen diverses llengües. La iniciativa de Hèlsinki sobre multilingüisme en la comunicació acadèmica sosté que l'exclusió de les llengües locals o nacionals per a la publicació acadèmica és el factor més important —i, sovint, oblidat— que impedeix a les societats utilitzar i aprofitar la recerca que es porta a terme al lloc on viuen. Encara que la posició dominant d'una *lingua franca* —l'anglès— sigui útil per estendre la divulgació d'idees arreu del món, també impedeix l'ús dels resultats de la recerca en l'àmbit local.

Després de dècades de polítiques que han portat els investigadors a publicar en anglès, comencem a veure un gir en aquesta tendència. A Europa, Àsia i moltes altres àrees d'actuació, els responsables de polítiques van introduint noves mesures per animar els investigadors a publicar en llengües locals i autòctones. La recomanació de la UNESCO sobre la ciència oberta, per exemple, demana als estats membres que fomentin "el multilingüisme en la pràctica de la ciència, en les publicacions científiques i en les comunicacions acadèmiques". Això reforça —i va en la mateixa línia— afirmacions no tan recents, com les de la Declaració universal dels drets humans, que insten a no discriminar els investigadors pel seu idioma, i la recomanació de la UNESCO sobre la promoció i l'ús del plurilingüisme i l'accés universal al ciberespai, que insta la comunitat a "prendre les mesures necessàries per reduir les barreres lingüístiques i [...] garantir que totes les cultures puguin expressar-se i tenir accés al ciberespai en totes les llengües, incloent-hi les indígenes".

El multilingüisme planteja un desafiament particular a l'hora de descobrir recursos. Si l'idioma d'un recurs acadèmic no està etiquetat adequadament, no serà indexat correctament per les eines de descobriment. Això es deu al fet que la indexació implica pràctiques d'anàlisis de text, com l'*stemming*, la lematització (agrupació de les formes flexionades d'una paraula perquè es puguin analitzar com un únic element) i el tractament adequat de les paraules buides. Totes aquestes tècniques d'anàlisi textual són molt específiques de cada llengua. La inclusió d'etiquetes d'idioma i l'adopció d'altres pràctiques similars permet als cercadors, agregadors i indexadors d'informació, i als serveis de descobriment identificar correctament l'idioma de tot el text i processar cada element en conseqüència. A més, pot ser que els investigadors i altres cercadors d'informació que només sàpiguen llegir en una o dues llengües vulguin conèixer tota la recerca rellevant en la seva àrea, independentment de la llengua en què estigui publicada. La designació correcta de l'idioma del recurs és important per donar suport a aquesta necessitat i oferir una consulta multilingüe de més qualitat.

L'agost del 2022, la COAR va posar en marxa el grup de treball de la COAR per al Suport al Multilingüisme i els Continguts en Llengua no Anglesa dels Repositoris amb l'objectiu de desenvolupar i promoure bones pràctiques en la gestió de continguts multilingües i en llengua no anglesa en els repositoris. Basant-se en disset casos d'ús aportats per diferents comunitats d'interessats (gestors i usuaris de repositoris, autors i traductors, agregadors i sistemes de descobriment), el grup de treball va identificar tres àrees rellevants per a la tasca que feia: la millora de la descobribilitat de contingut en llengua no anglesa, la cura de

continguts multilingües en un repositori i l'admissió de traduccions. Els casos d'ús es documenten en l'annex 1.

El juny del 2023, el grup de treball va publicar un primer conjunt de recomanacions preliminars perquè el revisés la comunitat. La consulta va donar lloc a una varietat àmplia d'aportacions, les quals van ser revisades pel grup de treball i incorporades a una segona versió de recomanacions. Aquest document presenta les recomanacions actualitzades a partir de les aportacions de la comunitat.

Les recomanacions estableixen una sèrie de bones pràctiques per a gestors de repositoris i desenvolupadors de programari per a repositoris, i aborden qüestions relatives a metadades, paraules clau multilingües, interfícies d'usuari, formats i llicències que milloraran la visibilitat, el descobriment i la reutilització del contingut dels repositoris en diverses llengües.

La nostra voluntat és que aquestes recomanacions siguin adoptades àmpliament pels repositoris de tot el món. Algunes de les recomanacions poden ser adoptades immediatament pels gestors de repositoris, mentre que altres necessitaran una mica més de temps i per aplicar-les de manera completa caldran esforços col·lectius per part de gestors de repositoris, agregadors, investigadors i desenvolupadors de programari. Els mesos vinents, la COAR i el grup de treball difondran arreu les recomanacions i treballaran per avançar perquè s'adoptin en els repositoris de tot el món.

Resum de les recomanacions

Creadors i curadors de metadades

Declarar l'idioma del recurs en cada element.

Declarar l'idioma de les metadades (p. ex., atribut xml:lang).

Utilitzar codis d'idioma estàndard (de dues o tres lletres) (ISO 639).

Habilitar el format de codificació UTF-8 en el repositori i utilitzar l'alfabet/sistema d'escriptura original sempre que es pugui. Si és necessari transliterar metadades, utilitzar normes reconegudes (p. ex., ISO).

Si el programari del repositori admet diversos idiomes d'interfície, configurar la interfície d'usuari en l'idioma o idiomes natius del grup destinatari, juntament amb una opció en anglès.

Escriure el nom o noms de persones utilitzant el sistema d'escriptura emprat en el document dipositat i proporcionar un identificador persistent que permeti una identificació inequívoca (p. ex., ORCID).

Incloure paraules clau en molts idiomes, fer servir vocabularis i tesaurus multilingües si és possible.

Recomanacions per a gestors de repositoris sobre continguts traduïts.

Desenvolupadors de programari i plataformes de repositori

Garantir que els codis d'idioma es puguin utilitzar sistemàticament en totes les col·leccions del dipòsit i que siguin compatibles.

Exposar l'idioma de les metadades mitjançant un protocol d'intercanvi de metadades, per exemple: OAI-PMH, GraphQL API, etc.

Millorar la compatibilitat amb els codis d'idioma ISO (per exemple, amb els codis de tres lletres necessaris per a alguns idiomes).

Garantir que s'admeten diversos idiomes d'interfície.

Garantir que els identificadors persistents queden exposats per OAI-PMH. El grup de treball PIDs in Dublin Core™ ha desenvolupat recomanacions per fer possible l'exposició d'identificadors persistents (PID), incloent-hi ORCID, a través d'OAI-PMH.

Facilitar paraules clau en diversos idiomes amb la finalitat d'augmentar la descobribilitat de contingut multilingüe en el repositori. Per exemple, habilitant la integració en temps real amb Wikidata (un cas seria, quan un usuari comença a escriure en el camp de metadades apropiat, l'aparició dels termes rellevants de Wikidata en una llista desplegable perquè l'usuari els seleccioni).

Permetre una assignació automàtica de termes controlats basada en les metadades existents.

Recomanacions detallades

1. Declarar l'idioma del recurs en cada element

Recomanació

És preceptiu declarar l'idioma principal del document. Les metadades d'idiomes s'han de codificar utilitzant el codi ISO 639 (per a més detalls, vegeu el punt 2.3 Utilitzar codis de llengua estàndard (de dues o tres lletres) (ISO 639)).

Directrius

Si el document té un únic idioma, les metadades d'idioma identificaran la llengua principal del recurs. L'atribució de la llengua principal del recurs s'ha de dur a terme en l'element.

Exemple 1. Idioma en XML de Dublin Core senzill amb codificació ISO 639-1

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

Exemple 2. Idioma en MODS amb codificació 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 tot el document (per exemple, un volum editat) té seccions importants del text en diversos idiomes, les metadades es repetiran per esmentar cada idioma.

Exemple 3. Document bilingüe (francès/anglès) en XML de Dublin Core senzill amb codificació ISO-639-1

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

Exemple 4. Document bilingüe (francès/anglès) en MODS amb codificació 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>
```

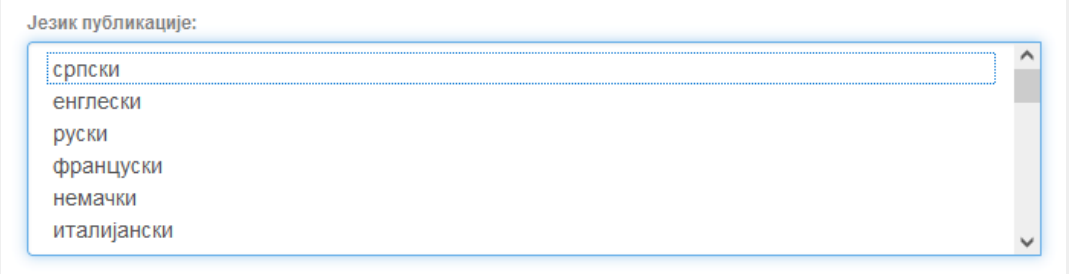
```
<language>
<languageTerm authority="iso639-2b" type="code"
authorityURI="http://id.loc.gov/vocabulary/iso639-2"
valueURI="http://id.loc.gov/vocabulary/iso639-2/fre">fre</languageTerm>

</language>
```

Exemple 5. EPrints
EPrints es pot ampliar per declarar informació d'idioma en l'element o en l'arxiu, tot i que aquesta declaració no es fa per defecte a EPrints. De la mateixa manera, els connectors (*plug-ins*) d'exportació XML d'EPrints, les metadades incrustades i el codi de la interfície OAI-PMH es podrien ampliar per definir atributs xml:lang, però aquesta ampliació no té lloc per defecte.

Exemple 6. Open Science Framework, OSF (marc de ciència oberta)
Les noves millores en les metadades de l'OSF per a tots els projectes, registres i prepublicacions de l'OSF inclouen ara l'idioma dels materials (més informació en Noves metadades de l'OSF per donar suport al compliment de la política d'intercanvi de dades).

Exemple 7. Repositori de l'Arxiu Digital de l'Acadèmia Sèrbia de Ciències i Arts (DAIS)
El contingut del repositori és en més de quinze idiomes. El filtratge per idioma no està habilitat i l'ús de la cerca no és fàcil. L'idioma es declara en el moment de l'enviament, seleccionant-lo d'una llista desplegable (camp obligatori). Es poden seleccionar diversos idiomes, però "multilingüe" no existeix com a valor. Recomanacions de les directrius d'enviament:

- El cos principal del text és en un idioma, i el títol, els resums i les paraules clau són en un altre: només declareu l'idioma principal (però cal proporcionar les metadades en totes dues llengües).
- Publicació (per exemple, un volum editat) amb seccions importants del text en diferents idiomes: declareu tots els idiomes.
- Text complet proporcionat en paral·lel en diversos idiomes: declareu tots els idiomes.



En les metadades, l'idioma seleccionat es mostra com un codi ISO de dues lletres (però s'ofereix en format llegible per a humans en la llista desplegable).

dc.language.iso	sr

Exemple 8. Metadades de preservació digital (PREMIS i METS)
METS (Metadata Encoding and Transmission Standard, o norma de codificació i transmissió de metadades) i PREMIS (Preservation Metadata: Implementation Strategies, o metadades per la conservació: estratègies d'aplicació) són dues normes de metadades que se solen fer servir conjuntament per proporcionar una compatibilitat àmplia de metadades en la preservació i la gestió d'objectes digitals. METS se centra principalment en la codificació de metadades descriptives, administratives i estructurals, i proporciona un marc per organitzar i vincular diversos tipus de metadades dins d'un document XML estructurat. PREMIS, al seu torn, se centra a documentar les accions, els esdeveniments i els processos implicats en la preservació a llarg termini dels objectes digitals. METS pot servir de contenidor per a diverses metadades, incloent-hi les metadades PREMIS, la qual cosa permet integrar la informació específica de la preservació en l'àmbit més ampli de l'organització i la descripció d'objectes digitals.

El model de dades PREMIS no classifica explícitament l'idioma com a metadada tècnica o descriptiva. La norma PREMIS tampoc no defineix específicament elements o subelements per capturar la informació relativa a l'idioma. Ara bé, amb el tipus de *propietats significatives* i els components de la unitat semàntica de valor a PREMIS n'hi ha prou per capturar l'idioma sense que calgui un element XML específic de l'idioma.

L'idioma es pot considerar una metadada tècnica i significativa per a la preservació utilitzant l'element <significantProperties> a PREMIS. Les propietats significatives representen els aspectes d'un objecte digital que influeixen en com es representa, es comporta o s'interpreta, com ara el format d'arxiu, l'algorisme de compressió, la versió de programari, la resolució, l'espai de color i altres característiques tècniques que afecten la representació i l'accessibilitat de l'objecte. En aquest sentit, recomanem que l'idioma es codifiqui com una propietat significativa, utilitzant PREMIS. Vegem un exemple en què l'idioma especificat és l'anglès:

```
<premis:significantProperties>

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

A més, la informació de l'idioma, si es considera una característica descriptiva del contingut intel·lectual (és a dir, una metadada descriptiva), es pot incrustar en el document METS mitjançant Dublin Core (utilitzant l'etiqueta <dc:language>) o

una altra secció de metadades dins del contenidor METS. L'idioma també es pot incrustar en el METS com a metadada tècnica per a documents de text utilitzant TextMD[1] a dins de l'element PREMIS <objectCharacteristicsExtension>.

En l'annex 2 es mostren més exemples d'implementació que segueixen normes/directrius sobre metadades.

2. Declarar l'idioma de les metadades (p. ex., atribut xml:lang)

Recomanació

Es recomana utilitzar l'atribut xml:lang per indicar l'idioma del camp de metadades. A causa de la cardinalitat [0, 1], l'atribut xml:lang podria descriure el mateix element en diferents idiomes, per la qual cosa seria més precís que l'element dc:language.

Directrius

Independentment que s'assumeixi majoritàriament l'anglès com a estàndard, el contingut hauria de ser exposat amb una referència a l'idioma utilitzat. Val la pena fer-ho en l'àmbit del repositori, ja que cap altra part interessada, com els agregadors (per exemple, BASE, OpenAIRE, etc.), pot deduir l'idioma a partir del contingut de les metadades.

En cas que els elements dipositats tinguin un títol, o altres elements de metadades, en més d'un idioma (per exemple, el títol principal i el títol d'un resum o *abstract*), caldrà assegurar-se que la informació de l'idioma s'indiqui mitjançant l'atribut/subpropietat xml:lang[2] i sigui exposat adequadament a través de protocols d'intercanvi de metadades, com OAI-PMH. Atès que alguns agregadors no poden recollir la informació completa ni tots els camps repetits, es recomana respectar l'ordre d'introducció de metadades dels títols, és a dir, proporcionar primer el títol principal. Si és possible, s'ha d'utilitzar l'alternativa dc.title per a títols addicionals.

Els agregadors com OpenAIRE[3] i BASE[4] identificaran correctament el títol principal basant-se en la informació proporcionada en el camp que indica l'idioma del document, independentment de l'ordre en què s'hagi introduït.

No obstant això, a OAI-PMH l'idioma de les metadades no queda exposat, per la qual cosa sol·licitem als desenvolupadors de programari que ho tinguin en compte en futures versions de les seves plataformes.

Exemple 9. Com s'assigna l'idioma quan hi ha més d'una llengua en els camps de metadades
Es fa servir l'atribut xml:lang per indicar l'idioma del camp de metadades.

<datacite:titles>

[1] https://www.loc.gov/standards/textMD/
[2] https://www.w3.org/international/techniques/authoring-xml#natlang
[3] https://www.openaire.eu
[4] https://www.base-search.net

<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>

Heus aquí un exemple de MODS d'AILLA, en el qual utilitzem els codis d'idioma ISO 639-3:
<titleInfo lang=”eng”>
<title>Iskonawa Oral Tradition</title>
</titleInfo>
<titleInfo lang=”spa”>
<title>Tradició Oral Iskonawa</title>

</titleInfo>
Vegeu l'annex 3 per a més exemples d'implementació seguint les normes/directrius sobre metadades.

Exemple 10. DSpace
A DSpace 7, els parells de valors d'idioma poden incloure els idiomes i identificadors d'idioma que es vulgui. Per defecte, DSpace ofereix deu parells de valors d'idioma: anglès dels Estats Units (en_US), anglès (en), espanyol (es), alemany (de), francès (fr), italià (it), japonès (ja), xinès (zh), portuguès (pt) i turc (tr). Però això es pot personalitzar completament en el fitxer submission-forms.xml. Aquest arxiu pot incloure identificadors de tres lletres en cas que en el material de destinació d'una col·lecció hi hagi idiomes amb identificadors de tres lletres. Durant els enviaments, els valors d'idioma apareixen en forma de llista desplegable, mentre que, en el mode d'edició, l'idioma és un camp de text lliure.

Vegeu l'annex 5 per obtenir informació sobre com se solucionen les inconsistències del codi d'idioma en els repositoris que funcionen amb versions de DSpace anteriors.

Exemple 11. TIND IR
TIND IR és un repositori basat en MARC. Això significa que la forma més fàcil d'incloure informació sobre continguts multilingües és a través del camp 041 i els subcamps rellevants.[5]

Exemple 12. WEKO 3
WEKO3 és un programari de repositoris desenvolupat per l'Institut Nacional d'Informàtica del Japó (NII) basat en el programari INVENIO del CERN. Aquest

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

programari funciona amb JAIRO Cloud, un sistema de repositoris basat en el núvol que compta amb el suport del Consorci Japonès per a Repositoris d'Accés Obert (JPCOAR) i el NII. A WEKO 3, l'esquema de metadades JPCOAR és compatible per defecte i s'hi pot afegir un atribut d'idioma per a qualsevol metadada sempre que estigui permès en l'esquema. En concret, s'accepta la codificació ISO 639-3 per a l'idioma del text, mentre que per a un atribut d'idioma d'altres elements de metadades s'accepta ISO 639-1. Amb cada camp es pot afegir una etiqueta d'idioma amb codificació ISO de dos caràcters utilitzant el menú desplegable, la casella de verificació i el botó de selecció.

3. Utilitzar codis d'idioma estàndard (de dues o tres lletres) (ISO 639)

3.1. Introducció a les etiquetes i codis d'idiomes

Identificar els idiomes de manera inequívoca és essencial per a la interpretació, l'agregació i la reutilització dels continguts d'una recerca. Les normes per a les etiquetes d'idioma s'han anat actualitzant i ampliant des del començament d'internet, la dècada de 1990. La darrera norma d'etiquetatge d'idiomes és definida per la BCP 47 de l'IETF (RFC 5646) en combinació amb l'ISO 639-3.

Les etiquetes d'idioma són indispensables en els formats HTML, XML i RDF per identificar un idioma natural. El codi d'idioma, en format de dos o tres caràcters (com "en" per a l'anglès), és el component principal d'una etiqueta d'idioma i ho estableix la norma ISO 639 (parts 1-3). El codi d'idioma pot anar seguit de subetiquetes destinades a precisar o limitar el rang de l'idioma codificat de la manera següent:

idioma-extensió d'idioma-sistema d'escriptura-regió-extensió de la variant-ús privat.

La pràctica de l'etiquetatge d'idiomes no representa cap complicació per a un gran nombre de llengües conegudes; la norma ISO 639 inclou codis per a més de 7.900 llengües (el gener del 2023). No obstant això, és important tenir en compte que les llengües menys conegudes i les varietats regionals o les etapes històriques de les llengües poden no estar prou representades en la norma ISO 639. Les subetiquetes opcionals de la codificació BCP 47 ofereixen diverses possibilitats per a una identificació més precisa. La subetiqueta *private-use* "x" definida en BCP 47 es pot utilitzar per identificar modalitats lingüístiques.[6] A més, ISO 639 és una norma que ha canviat amb el temps i ara ofereix l'oportunitat de proposar canvis:

> "El coneixement de les llengües humanes en qualsevol moment històric mai és complet ni perfecte, sinó que s'amplia constantment. Donat el caràcter exhaustiu de la norma ISO 639-3, els canvis en el conjunt de codis són inevitables, sobretot pel que fa a les llengües menys conegudes o identificades recentment."[7]

[6] Tal com es descriu a https://aclanthology.org/2020.lrec-1.408.pdf, per exemple.
[7] https://iso639-3.sil.org/code_changes/introduction

És important recordar que l'objectiu principal de l'etiquetatge d'idiomes és identificar i representar amb exactitud, en funció del context lingüístic i tecnològic, la llengua en ús. Si un codi de 2 lletres (ISO 639, part 1) no és adequat en un context específic, s'haurà d'utilitzar un codi de 3 lletres (ISO 639, parts 2 i 3) o altres subetiquetes (com ara per a escriptura, regió o ús privat) amb la finalitat de garantir la interoperabilitat i la precisió en la identificació de la llengua. El conjunt de llengües incloses en la part 1 de la norma ISO 639 es considera un subconjunt de la part 2 i qualsevol codi de 2 lletres de la part 1 amb un codi de 3 lletres corresponent en la part 2 o 3 es considera sinònim amb la mateixa extensió. Per exemple, els identificadors "fra", "fre" i "fr" designen la mateixa llengua. BCP 47 recomana utilitzar codis de 2 lletres sempre que existeixin, però ISO 639 estableix que ha de permetre's la lliure elecció entre sinònims sempre que sigui possible. En aquest informe, recomanem seguir aquesta part de la recomanació de BCP 47 i utilitzar codis de 2 lletres sempre que existeixin, però, depenent del context específic d'ús, pot ser adequat utilitzar codis de 3 lletres.

3.2. Resum de l'arbre de decisió per triar una etiqueta d'idioma

A continuació presentem un arbre de decisió resumit per triar una etiqueta d'idioma:

1. Busqueu el codi d'idioma a ISO 639.
2. Si trobeu un codi ISO 639, part 1, de dues lletres per a l'idioma, utilitzeu-lo. Aneu al punt 5.
3. Si trobeu un codi ISO 639, parts 2 o 3, de tres lletres per a l'idioma, utilitzeu-lo. Aneu al punt 5.
4. Utilitzeu la subetiqueta "x", reservada per a ús privat, per definir un codi d'idioma personalitzat. Aneu al punt 5.
5. Decidiu si és necessària i pertinent una subetiqueta per identificar l'idioma. Per exemple, si el fet que es tracti d'una variant regional o un dialecte és important en el context, considereu la possibilitat d'utilitzar codis de país ISO 3166 com a subetiquetes (per exemple, "en-US" per a l'anglès americà). Si és necessari identificar variants del sistema d'escriptura que siguin rellevants, considereu la possibilitat d'utilitzar com a subetiquetes els codis d'escriptures ISO 15924 (per exemple, "sr-Latn" per al serbi en escriptura llatina).

Nota: la codificació d'ISO 639 2 i 3 ha estandarditzat algunes situacions especials:
* mis (de "miscellaneous"): s'aplica quan hi ha "idiomes no codificats".
* mul (de "multiple languages"): s'utilitza quan apareixen diversos idiomes i no resulta pràctic especificar tots els codis d'idioma corresponents.
* und (d'"undetermined"): s'utilitza quan cal indicar un idioma però no és possible identificar-lo.
* zxx: figura en la llista de codis com a "sense contingut lingüístic", com els sons d'animals (codi afegit l'11 de gener de 2006).

L'ús de codis d'idioma també pot resultar pràctic en el cas de llengües històriques o locals, regionals o clàssiques (com el llatí, el való, etc.).[8]

[8]Exemples en való: https://orbi.uliege.be/handle/2268/28421 i https://orbi.uliege.be/handle/2268/28419

Trobareu més informació sobre els codis ISO 639-1, ISO 639-2 i ISO 639-3, i sobre les etiquetes d'idiomes en l'annex 4.

Exemple 13. Lingüística i estudis lingüístics
En lingüística i estudis lingüístics, els codis ISO 639-3 (de 3 lletres) són un estàndard. En primer lloc, la majoria de les llengües no tenen codis de dues lletres i, quan els tenen, solen prestar-se a confusió perquè no representen idiomes pròpiament dits (per exemple, cr per a "cree", ms per a "malai" o zh per a "xinès"). Això enterboleix precisament el tipus de diversitat que volem promoure. Els lingüistes i els arxius lingüístics també utilitzen cada vegada més els glottocodes (codis de la base de dades Glottolog) per a "languoides", ja que el que arriba a "comptar" com a llengua respon en gran manera a una visió política. Considereu la possibilitat de disposar d'un camp opcional per incloure'ls.

Exemple 14. Repositori institucional MiCISAN
El repositori institucional MiCISAN utilitza la norma ISO 639-3.

4. Habilitar el format de codificació UTF-8 en el repositori i utilitzar l'alfabet/sistema d'escriptura original sempre que sigui possible. Si és necessari transliterar metadades, utilitzar normes reconegudes (p. ex., ISO)

Directrius i debat

UTF-8 és el format de codificació de caràcters més estès a la xarxa (i en les tecnologies d'internet). El 2023, és present en el 98,0 % de totes les pàgines web, i arriba al 100 % en molts idiomes.[9] Pràcticament tots els països i llengües utilitzen les codificacions UTF-8 a la xarxa en un 95 % dels casos.[10]

La majoria de programaris de repositoris admet UTF-8 per defecte, com passa amb DSpace 7, però hi ha passos en el procés d'instal·lació en què és necessari assegurar-se que Tomcat fa servir UTF-8 per defecte o de manera semblant.[11]

4.1. Transliteració enfront de transcripció

La transliteració és la conversió d'un text d'un sistema d'escriptura a un altre (per exemple, de l'alfabet grec a l'alfabet llatí) i es basa en l'assignació dels grafemes d'un sistema a un altre de manera normalitzada perquè els lectors puguin reconstruir l'ortografia original utilitzant taules de transliteració normalitzades o eines informàtiques. Alguns països compten amb normes de transliteració.

[9] Enquesta d'ús de les codificacions de caràcters desglossades per classificació, *w3techs.com*. Consultat 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

La transcripció és un tipus de conversió en el qual el text de la llengua d'arribada captura el so en lloc de l'ortografia.

A vegades, la transliteració és inevitable. En les bases de dades bibliogràfiques i els catàlegs de les biblioteques trobem enormes quantitats de metadades transliterades o transcrites. En algunes comunitats de recerca, transliterar noms i fins i tot títols és una pràctica habitual. Encara que ara és comú que s'admeti la codificació UTF-8, aquestes pràctiques persisteixen. Si un repositori ja conté metadades transliterades o la seva comunitat designada necessita que les metadades siguin transliterades, es recomana fer el següent:

- Utilitzar normes de transliteració reconegudes.
- Si és possible, triar una norma i declarar-la en les pàgines de preguntes freqüents / manual d'usuari / "sobre" incloses en el repositori.
- Si això no és possible, declarar tots els estàndards utilitzats en les pàgines de preguntes freqüents / manual de l'usuari / "sobre".
- Proporcionar enllaços a les directrius de transliteració pertinents (per exemple, Biblioteca del Congrés) o eines[12] en les pàgines de preguntes freqüents / manual d'usuari / "sobre" per garantir que els lectors puguin reconstruir l'ortografia original.
- Si els noms dels autors estan transliterats, utilitzar identificadors com ORCID per connectar les diferents variants dels noms.
- Utilitzar codis d'idioma per a les metadades transliterades (per exemple, aquest recurs recomana el-Latn per indicar text en grec transliterat a l'alfabet llatí).

Si hi ha estàndards de transliteració, cal evitar la transcripció, perquè les normes no sempre són clares i es dificultaria la reconstrucció de l'ortografia original. Si la transcripció és inevitable, cal seguir les regles i normes de les llengües.

Exemple 15. DataCite
DataCite requereix la transliteració de caràcters no llatins:

contributorName
Ocurrències: 1
Definició: nom complet de la persona que ha fet la contribució.
Valors permesos, exemples, altres restriccions: si s'utilitza persona contribuïdora, llavors contributorName és obligatori.
Exemples: Patel, Emily; ABC Foundation
El format per al nom de persona pot ser: cognom, nom de pila. Els noms en alfabet no llatí s'han de transliterar segons els esquemes ALA-LC.

5. Si el programari del repositori admet diversos idiomes d'interfície, configurar la interfície d'usuari en l'idioma o idiomes natius del grup destinatari, juntament amb una opció en anglès

Directrius

[12]Per exemple: https://alittlehebrew.com/transliterate/, https://www.translitteration.com

Una interfície d'usuari en diversos idiomes facilita la navegació pel repositori a usuaris de diferents comunitats. Per exemple, una interfície en la llengua o llengües maternes facilita als usuaris locals comprendre els camps de metadades quan hi dipositen continguts, i als usuaris internacionals navegar-hi i cercar-hi continguts.

Exemple 16. Dataverse
Dataverse admet interfícies d'usuari en diversos idiomes i depèn de les traduccions de la comunitat fetes per voluntaris. En el marc del projecte Social Sciences and Humanities Open Cloud (SSHOC), es van fer avenços importants cap a la creació d'un directori de paquets d'idiomes i es va dissenyar l'eina en línia Weblate per facilitar noves traduccions. També hi ha disponible una guia d'usuari per a Weblate.[13]

Exemple 17. DSpace
DSpace admet diversos idiomes d'interfície. El text que apareix en la interfície es denomina "missatges" i els arxius de missatges (paquets d'idiomes) són aportats i gestionats per la comunitat al marge del nucli del projecte DSpace per permetre actualitzacions i llançaments més regulars. Els usuaris poden modificar les traduccions de la comunitat o crear les seves pròpies traduccions i enviar-les al projecte dspace-api-lang a Github. A part dels missatges, és possible localitzar altres elements, com les pàgines d'ajuda, els formularis d'entrada i les plantilles de correu electrònic. En la documentació de DSpace es poden trobar instruccions sobre com s'ha d'habilitar la interfície en diversos idiomes. DSpace 7 fa un pas important per facilitar les traduccions de la interfície d'usuari: https://wiki.lyrasis.org/pages/viewpage.a: les directrius per al suport multilingüe en el *front-end* (UI) estan disponibles.[14]

Exemple 18. EPrints
EPrints admet diversos idiomes d'interfície i utilitza carpetes de "frases" i altres arxius específics per a cada idioma. Per defecte, EPrints només ve empaquetat amb frases en anglès, però la comunitat ha compartit moltes traduccions a través d'EPrints Bazaar i EPrints Files. EPrints utilitza l'estàndard d'idioma ISO de dues lletres per especificar subdirectoris de frases i altres tipus de directoris específics de cada llengua, com per exemple:

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

Les metadades de temes d'EPrints estan dissenyades per donar cabuda a etiquetes multilingües, amb la qual cosa les etiquetes de temes es poden mostrar en funció de l'idioma que l'usuari hagi definit per a la interfície.

[13] https://doi.org/10.5281/zenodo.4807371
[14] https://wiki.lyrasis.org/display/dsdoc7x/multilingual+Support

EPrints està dissenyat per utilitzar per defecte frases en anglès; i si li falten frases per a un altre idioma d'interfície declarat, utilitzarà les frases en anglès fins que s'afegeixin les frases que falten. Hi ha una pàgina wiki tècnica sobre traduccions, però pot ser que estigui obsoleta, perquè ha estat editada molt poques vegades els darrers anys.

6. Escriure el nom o noms de persones utilitzant el sistema d'escriptura emprat en el document dipositat i proporcionar un identificador persistent que permeti una identificació inequívoca

Recomanació

Es recomana escriure el nom o noms de persona tal com apareixen en el document dipositat i proporcionar un identificador persistent que permeti una identificació inequívoca, com ara ORCID.

Directrius i debat

Hi ha dos criteris principals a l'hora de tractar els noms de persona en els repositoris:

- Utilitzant una forma preferida unificada que estigui definida en un arxiu d'autoritat.
- Capturant els noms tal com apareixen en el document dipositat.

El primer criteri és el típic dels catàlegs de les biblioteques, en què s'utilitza la forma unificada com a encapçalament del catàleg. Depenent del país, els noms escrits originalment en un alfabet no llatí es llatinitzaran o, al revés, es transcriuran/transliteraran segons les normes utilitzades en un país concret. Si un repositori ofereix metadades integrades que es poden importar des de gestors de referències i citacions recomanades preformatades, aquest criteri pot no ser l'adequat, perquè el format del nom en el repositori diferirà del de la publicació.

Si els noms es capturen tal com apareixen en les publicacions dipositades, el nom d'una mateixa persona apareixerà en el repositori en diversos formats. En aquest cas, és important utilitzar identificadors persistents, com ORCID, per garantir la identificació correcta i connectar diferents versions del nom.

Exemple 19. DSpace

En la versió anterior de DSpace es necessitava una solució provisional per mostrar les diferents versions dels noms d'una manera senzilla per a l'usuari (per exemple, mitjançant una aplicació interna addicional). Ara, DSpace CRIS i DSpace 7 no sols admeten una integració bidireccional amb ORCID, sinó que també tracten les persones com a entitats (entitats CRIS i entitats configurables, respectivament).[15]

Exemple 20. Exposició d'identificadors persistents a través d'OAI-PMH

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

També és important garantir que els identificadors persistents s'exposin a través d'OAI-PMH. El grup de treball PIDs in Dublin Core™ ha desenvolupat recomanacions per fer possible l'exposició d'identificadors persistents, incloent-hi ORCID, a través d'OAI-PMH. Es proposen dues solucions, i totes dues cobreixen diversos casos d'ús:

Opció 1: utilitzar un atribut "id" amb propietats Dublin Core

Tant el PID (identificador persistent) com l'etiqueta són coneguts:

```
<dc:creator id="https://orcid.org/0000-0003-1541-5631">Walk, Paul</dc:creator>
```

Es coneix l'etiqueta, però no el PID:

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

Es coneix el PID, però no l'etiqueta:

```
<dc:creator id="https://orcid.org/0000-0003-1541-5631"></dc:creator> o
<dc:creator id="https://orcid.org/0000-0003-1541-5631"/>
```

Aquesta opció no és adequada si és necessari incloure més d'un PID.

Opció 2: utilitzar propietats niades per als identificadors

El PID i l'etiqueta són coneguts:

```
<dc:creator>
        <dc:identifier>https://orcid.org/0000-0003-1541-5631</dc:identifier>
        <foaf:name>Walk, Paul</foaf:name>
```

Es coneix l'etiqueta, però no el PID:

```
<dc:creator>
<foaf:name>Walk, Paul</foaf:name>
</dc:creator>
o
<dc:creator>Walk, Paul</dc:creator>
```

Es coneix el PID, però no l'etiqueta:

```
<dc:creator>
        <dc:identifier>https://orcid.org/0000-0003-1541-5631</dc:identifier>
</dc:creator>
```

En aquesta opció es poden proporcionar diversos PID (identificadors persistents) per a una mateixa propietat:

```
<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>
```

Exemple 21. Esquema de metadades JPCOAR

L'esquema de metadades JPCOAR també inclou un element per a l'identificador de l'investigador: jpcoar:nameIdentifier.[16] Aquest element es pot utilitzar repetidament amb diferents tipus de PID (KAKEN ID, ORCID, identificador de l'investigador i altres) i quedar exposat per a la Institutional Repositories Database (base de dades de repositoris institucionals del Japó) a través d'OAI-PMH.

7. Incloure paraules clau en molts idiomes, utilitzar vocabularis i tesaurus multilingües si és possible

Directrius i debat

La inclusió de paraules clau en molts idiomes augmenta la descobribilitat de contingut en els repositoris. En referència a això darrer, és important distingir entre paraules clau de text lliure (o "etiquetes") i termes controlats derivats d'un vocabulari multilingüe controlat o tesaurus. En el primer cas, les paraules clau en diversos idiomes s'introdueixen en el camp dc:subject, assegurant-se que l'idioma està codificat correctament.

Exemple 22. Repositori institucional MiCISAN
El repositori institucional MiCISAN sempre respecta l'idioma del recurs. Si, per exemple, el recurs està en espanyol, en el cas de les paraules clau s'utilitzen qualificadors per diferenciar les metadades en diferents idiomes:[17]

dc.subject.keywordseng
institutional repository

dc.subject.keywordseng
interoperability

dc.subject.keywordsspa
metadatos

dc.subject.keywordsspa
repositorio institucional

dc.subject.keywordsspa
interoperabilidad

És important assenyalar que l'ús de paraules clau de text lliure no garanteix la coherència ni revela relacions jeràrquiques entre els termes. El problema es pot mitigar seleccionant manualment els termes que s'afegiran com a paraules clau des de vocabularis controlats. No obstant això, una solució òptima comporta la integració de vocabularis controlats multilingües en el repositori.

[16] https://schema.irdb.nii.ac.jp/en/schema/3-1
[17] https://ru.micisan.unam.mx/handle/123456789/22232?show=full

7.1. Vocabularis i tesaurus multilingües

L'ús de vocabularis controlats o tesaurus[18] per a les metadades bibliogràfiques garanteix que un mateix concepte es descrigui de manera coherent. Juntament amb l'ús de termes controlats per indicar el tipus de recurs, versió o drets d'ús, també es poden utilitzar vocabularis controlats per descriure el contingut temàtic del recurs. En els vocabularis controlats multilingües, l'ideal és que cada terme tingui un únic equivalent en cada idioma i que les relacions entre els termes siguin les mateixes. En un entorn digital, als termes del vocabulari se'ls assignen identificadors persistents que poden resoldre's fàcilment.

No obstant això, l'ús de vocabularis controlats o tesaurus comporta certs desafiaments.

- Per poder-se integrar amb els repositoris, els vocabularis controlats s'han d'expressar com a dades interpretables pel sistema.
- Equivalència forçada: no sempre és possible trobar equivalents veritables en totes les llengües, per la qual cosa el significat dels termes i les relacions entre ells en un idioma no es reflectiran amb exactitud en els seus equivalents d'altres idiomes.
- El procés d'assignació de termes controlats pot comportar molt de temps.
- Els investigadors no solen estar familiaritzats amb el concepte de vocabularis controlats. Si els bibliotecaris no tenen els coneixements especialitzats necessaris, els termes poden ser massa generals i imprecisos.
- Existeixen molts vocabularis controlats específics d'una disciplina i no és possible aplicar-los tots en repositoris multidisciplinaris. D'altra banda, hi ha la possibilitat que els vocabularis generals no siguin capaços de descriure el contingut amb precisió.
- Els vocabularis controlats més utilitzats (per exemple, els encapçalaments de matèria de la Biblioteca del Congrés dels Estats Units o els vocabularis Getty) no inclouen de la mateixa manera els diferents contextos culturals i grups socials.

En general, les plataformes de programari de repositoris admeten la implementació de vocabularis controlats, encara que les solucions d'integració no sempre són òptimes.

Exemple 23. Dataverse
Dataverse és el repositori de dades de codi obert desenvolupat per l'IQSS de la Universitat Harvard. La sòlida comunitat de Dataverse està ajudant a millorar la funcionalitat bàsica i a continuar desenvolupant-la. DANS-KNAW va lliurar el repositori Dataverse a punt per a la producció (Docker/k8s) a les comunitats del Núvol Europeu de Ciència Oberta (EOSC) CESSDA, CLARIN i DARIAH. Per afrontar els reptes d'integració de conjunts de dades heterogènies i multilingües, DANS-KNAW va introduir el suport de vocabularis controlats externs (model de metadades CESSDA connectat al marc Skosmos; suport per a la infraestructura de metadades de components CLARIN i el Tesaurus Europeu de Llengües de les Ciències Socials (ELSST) allotjat per CESSDA i ODISSEI a Skosmos; CESSDA té una versió actualitzada amb més propietats d'idioma).

Exemple 24. DSpace

[18] Registre de vocabularis controlats: https://bartoc.org/

DSpace ofereix tres maneres d'integrar vocabularis controlats:[19]

- Parells de valors en forma de llista controlada.
- Arxiu XML amb els termes (per exemple, per permetre la integració del sistema de classificació decimal Dewey o el tesaurus de termes grecs en els repositoris).[20]
- SolR Authority (es va utilitzar per a la integració d'ORCID abans de DSpace 7).[21]

Les entitats configurables de DSpace 7, encara que no es van dissenyar inicialment per a aquest ús, podrien ser una altra manera d'implementar vocabularis controlats.

Exemple 25. TRIPLE

S'han donat diversos intents de superar les limitacions dels vocabularis controlats existents. El projecte TRIPLE va desenvolupar un nou vocabulari controlat multilingüe (en nou idiomes) per a ciències socials i humanitats partint de vocabularis existents.

Exemple 26. El vocabulari RVM Web

El vocabulari RVM Web, gestionat per la Universitat de Laval i utilitzat per biblioteques de tot el Canadà, és un exemple de vocabulari controlat que intenta eliminar els biaixos culturals, històrics i colonials:

- És bilingüe (anglès i francès), però no per a tots els termes.
- Al començament (cap al 1970) es va crear traduint els encapçalaments de matèries de la Biblioteca del Congrés (LCSH), i ara ja és un producte independent.
- La versió anglesa utilitza MeSH, AAT (tesaurus Getty), HOMOsaurus (com a novetat) i LCSH.
- Les relacions entre els diferents termes del tesaurus o del vocabulari s'estableixen manualment. No és un procés automatitzat.
- La versió oberta RVM FAST no conté AAT MeSH ni HOMOsaurus, només LCSH (existeix un pla per fer-la compatible amb Linked Open Data amb la finalitat d'incloure-la a DBpedia a curt termini). (Vegeu un exemple aquí)
- Està inclòs a WebDewey.
- Hi ha un identificador únic per a cada terme (encara no és públic).
- Reptes:
  - Sincronització entre els diferents productes (LCSH, RAMEAU, AAT, etc.). S'espera que millori amb l'ús de les identificacions.
  - Com es poden impulsar les actualitzacions dels termes utilitzats en els sistemes?

Exemple 27. Wikidata

[19] https://wiki.lyrasis.org/display/dsdoc7x/authority+Control+of+Metadata+Values
[20] La primera integració del vocabulari de tipus de recursos de la COAR es va dur a terme utilitzant parells de valors o arxius XML: http://repositorium.sdum.uminho.pt/handle/1822/46066?mode=full
[21] https://wiki.lyrasis.org/display/DSDOC7x/ORCID+Authority

La integració de Wikidata en els repositoris, ja implantada a Europeana, pot ser una solució àmpliament aplicable per proporcionar paraules clau multilingües. Wikidata es basa tant en el *crowdsourcing* com en els arxius d'autoritats existents i ja conté un gran nombre de dades en diversos idiomes. La importació de termes de diversos vocabularis està habilitada a través de l'eina Mix'n'match.

Wikidata com a paraules clau

Wikidata és una base de coneixements lliure amb més de 100 milions d'elements de dades. Actua com a emmagatzematge central de dades estructurades generals de conceptes, i inclou etiquetes/traduccions dels conceptes en molts idiomes. Per això, l'ús de conceptes de Wikidata com a vocabulari controlat de paraules clau és especialment prometedor, ja que pot proporcionar més interoperativitat multilingüe amb menys inversió de temps.

Per exemple, el repositori de dades de recerca basat en CKAN Depositar reutilitza Wikidata com a font de paraules clau (més informació aquí). Cal assenyalar que les etiquetes dels conceptes de Wikidata continuarien canviant. Per això, Depositar només emmagatzema i exposa l'identificador propi (per exemple, "Q11030"). Després, consulta l'API de MediaWiki per obtenir les últimes etiquetes multilingües d'un vocabulari de Wikidata. Seria millor emmagatzemar i exposar tant 1) l'etiqueta més recent com 2) l'etiqueta (antiga) en el moment d'assignar una paraula clau.

Els conceptes de WikiData i altres termes de vocabulari controlat es poden codificar utilitzant les etiquetes JATS[22] <kwd-group> i <kwd> i afegint els atributs ***vocab***, ***vocab-identifier*** i ***vocab-term-identifier*** definits en la Standards Tag Suite (STS) de la NISO:

- el nom del vocabulari controlat ("wikidata") en l'atribut ***vocab***[23]
- l'identificador del vocabulari ("https://www.wikidata.org/") en l'atribut ***vocab-identifier***[24]

[22]El Journal Article Tag Suite (JATS) és un format XML utilitzat per descriure bibliografia científica publicada en línia. Es tracta d'una norma tècnica elaborada per l'Organització Nacional de Normes d'Informació (NISO) dels Estats Units i aprovada per l'Institut Nacional Estatunidenc de Normes amb el codi Z39.96-2012. El projecte NISO va ser una continuació del treball portat a terme per NLM/NCBI i popularitzat pel repositori PubMed Central de la NLM com a norma *de facto* per a l'arxivament i l'intercanvi de revistes científiques d'accés obert i els seus continguts amb XML. Amb la normalització de la NISO, la iniciativa de l'NLM ha adquirit un abast més gran, i altres repositoris, com ara SciELO i Redalyc, van adoptar el format XML per als articles científics: https://en.wikipedia.org/wiki/journal_article_tag_suite
A JATS (Journal Article Tag Suite), qualsevol camp de metadades es podia etiquetar amb un idioma. En el format DTD de l'esquema JATS, l'atribut xml:lang es pot aplicar a pràcticament qualsevol element; vegeu: https://jats.nlm.nih.gov/articleauthoring/tag-library/1.2/attribute/xml-lang.html. Exemples: títols traduïts de PubMed Central https://www.ncbi.nlm.nih.gov/pmc/pmcdoc/tagging-guidelines/article/dobs.html#dob-at-transtitle. En utilitzar l'esquema JATS, l'idioma de les paraules clau es registra mitjançant l'atribut *xml:lang* de l'etiqueta <kwd-group> (vegeu: https://jats.nlm.nih.gov/articleauthoring/tag-library/1.2/element/kwd-group.html). JATS agrupa les paraules clau per idioma amb una sèrie d'etiquetes <kwd> just sota l'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

- l'identificador/URL de cada paraula clau en l'atribut ***vocab-term-identifier*** (per exemple, "Q11030").[25] Per a Wikidata, es tracta de l'identificador del concepte, no de l'etiqueta específica de l'idioma del concepte.

Hi ha diverses maneres de fer-ho. La norma JATS agrupa les paraules clau per idioma utilitzant l'etiqueta <kwd-group>. A continuació es mostra un exemple d'etiquetatge de metadades de conceptes de Wikidata sobre fotografia (Q11633) i periodisme (Q11030) amb les etiquetes de concepte en anglès (photography, journalism) i polonès (fotografia, dziennikarstwo) utilitzant 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="Q11633">fotografia</kwd>
</kwd-group>
```

Pot ser que les tecnologies actuals de repositoris tinguin limitacions en aquest sentit.

Recomanació: afegiu tots els atributs descrits en l'exemple: ***vocab***, ***vocab-identifier*** i ***vocab-term-identifier***

Recomanacions per a desenvolupadors de programari/plataformes de repositoris

- Es recomana habilitar una integració en temps real amb Wikidata (per exemple, quan un usuari comença a escriure en el camp de metadades apropiat, els termes rellevants de Wikidata apareixen en una llista desplegable perquè l'usuari els seleccioni).

- Es recomana permetre una assignació automàtica de termes controlats basada en les metadades existents.

La indexació automàtica de continguts podria fer més eficient el procés d'assignació de termes controlats. Aquest plantejament, que s'ha provat en repositoris institucionals individuals, ja el fan servir els agregadors. Per exemple, Europeana enriqueix automàticament les metadades a partir de vocabularis i conjunts de dades externes, com GeoNames i DBpedia, i utilitza les relacions semàntiques i traduccions que ofereixen aquests vocabularis. BASE assigna termes calculats del sistema de classificació decimal Dewey basant-se en les metadades disponibles. El mateix plantejament s'utilitza en la plataforma de descobriment multilingüe GoTriple, en què els continguts recollits de diverses fonts s'anoten automàticament utilitzant termes controlats, gràcies a la qual cosa és possible fer cerques en diversos idiomes en GoTriple.

Altres avenços podrien incloure l'assignació de termes controlats a partir del text complet dels documents dipositats i permetre una importació automatitzada dels termes controlats assignats pels agregadors.

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

8. Recomanacions per a gestors de repositoris sobre continguts traduïts

El multilingüisme i la traducció estan estretament vinculats i es complementen entre ells. Les traduccions i els continguts traduïts han de ser reconeguts com a contribucions vàlides en l'ecosistema de la recerca i, com a tals, secundats i reconeguts com un producte acadèmic valuós. A més, cal promoure la diversitat lingüística en la cultura de la recerca. Per a això és necessari fomentar i acreditar adequadament la traducció com a pràctica i com a resultat. Això es pot aconseguir en part posant en pràctica les vuit recomanacions específiques següents:

1. **Incloure un camp específic per a la funció del traductor o traductors en els formularis de dipòsit dels arxius i repositoris en línia per donar cabuda als crèdits del traductor (p. ex., dc.contributor.translator)**

Vegeu algunes directrius en Rivero, Monica, Robert Estep, i Lorena Gauthereau-Bryson, "Digitization Practices for Translations: Lessons Learned from the Our Americas Arxivi Partenariat Project", D-Lib Magazine, 17 (2011) doi:10.1045/september2011-rivero.

2. **Si és possible, donar cabuda a la identificació del traductor amb altres camps, com ORCID o altres identificadors interoperables similars; també organització o afiliació, si n'hi ha**

3. **Incloure (sub)camp(s) específic(s) per a l'estat de la traducció del document, llengua o llengües utilitzades per al contingut traduït i llengua o llengües del document d'origen, designant-les preferiblement amb codis d'idioma internacionals normalitzats**

Exemple 28. Directrius CRIS v1.2 actualitzades
Les directrius CRIS v1.2 actualitzades ja inclouen el multilingüisme i les traduccions automàtiques. Vegeu un exemple en la pàgina 13 del tutorial CERIF i un exemple de CRIS en les directrius OpenAIRE actualitzades amb un atribut addicional "trans=" en l'element:
Versió:
https://github.com/openaire/guidelines-cris-managers/releases/tag/v1.2.0 amb els valors:

- h := humana
- m := automàtica
- o := original

Vegeu també el CERIF XSD en la cerca cfTrans_Type.

Animem també a dur a terme desenvolupaments semblants en altres estàndards i plataformes.

4. Permetre als usuaris apuntar a altres registres del contingut traduït que hi estiguin relacionats afegint camps de relació, com ara dc.relation

Les opcions d'etiquetatge en aquest camp de relació podrien incloure la informació:

- "És una traducció de"
- "S'ha traduït des de" (Aquesta segona opció es podria utilitzar més aviat en cas d'una traducció parcial, per exemple, d'un capítol o una secció d'un llibre.)

Exemple 29. Crossref
Crossref tracta l'idioma com un atribut basat en codis de dues lletres que es poden utilitzar en diversos elements i gestiona les traduccions mitjançant atributs de relació específics: isTranslationOf; hasTranslation.[26] Però el problema és que un proveïdor pot no utilitzar-los en registrar els seus continguts.

Referència de l'esquema:
L'idioma existeix en l'esquema comú com a atribut:
https://data.crossref.org/schemas/common5.3.1.xsd
<xsd:attributeGroup name="language.atts">
<xsd:annotation>
<xsd:documentation>Els atributs d'idioma es basen en la norma ISO 639</xsd:documentation>
</xsd:annotation>
<xsd:attribute name="language" use="optional">

La relació de traducció també és possible; vegeu l'esquema de relació:
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>S'utilitza per definir relacions entre elements que són essencialment el mateix treball, però que poden diferir en algun aspecte que afecti la citació, per exemple una diferència de format, idioma o revisió. Assignar identificadors diferents a exactament el mateix element disponible en un lloc o com a còpies en diversos llocs pot ser problemàtic i s'ha d'evitar.
</xsd:documentation>
</xsd:annotation>
<xsd:simpleType>
<xsd:restriction base="xsd:string">
<!-- Crossref -->
<xsd:enumeration value="isTranslationOf"/>
<!-- hasTranslation -->
<xsd:enumeration value="hasTranslation"/>
<!-- isTranslationOf -->

[26] Vegeu la documentació: 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/ i https://www.crossref.org/documentation/schema-library/markup-guide-metadata-segments/relationships/

5. Donar cabuda a aquest camp de relació amb altres camps d'identificació que apuntin al document original

Es pot utilitzar un DOI o un altre PID del document original, o un identificador o un URL si no existeix cap sistema de resolució interoperable.

Directrius i exemples

Les opcions d'exportació de registres de continguts traduïts haurien d'incloure idealment tota la informació anterior, amb especificitats relatives al tipus de traducció quan sigui necessari, per a la qual cosa a continuació es proporcionen més detalls de context.

Un exemple de registre per a un contingut traduït o posteditat per un humà tindria l'aspecte següent:

"Aquest material titulat '[títol traduït]' és una traducció íntegra/parcial des del [idioma - codi d'idioma estàndard] amb data [DD-MM-AAAA] a càrrec de [nom(s) de traductor(es) de 'títol original'], escrit per [nom(s) autor(es)] en [idioma - codi d'idioma estàndard], publicat en [dades editorials]/recuperat de [DOI, un altre sistema de resolució de PID o URL].”

Traducció automàtica

En el transcurs de la nostra recerca, el tema de la traducció automàtica (TA) ha suscitat un debat acalorat en el si del grup de treball i hem compartit la naturalesa d'aquesta discussió en una entrada de blog: És possible acceptar en els repositoris continguts acadèmics traduïts automàticament?

Donada la complexitat de la qüestió, així com les implicacions ètiques, el grup de treball ha optat per recomanar que els repositoris no acceptin continguts traduïts exclusivament amb motors de traducció. Això coincideix també amb les recomanacions de l'Informe del grup de treball Traducció i Ciència Oberta (2020). La TA hauria de ser percebuda i utilitzada com una tecnologia de suport, etiquetada de manera transparent i inequívoca com a assistència automàtica, i caldria permetre que canviï dinàmicament en temps real en comptes de mantenir-la i conservar-la en el repositori com un recurs primari.

El grup de treball continuarà analitzant de prop aquest escenari en ràpida evolució i continuarà estudiant les qüestions i, possiblement, publicant noves recomanacions relacionades amb la traducció automàtica (TA) de textos acadèmics, la traducció assistida per TA i la TA de resums i metadades en repositoris.

Sigui com sigui, dos estudis exploratoris duts a terme com a part del projecte francès Traduccions i ciència oberta (cartografia i recopilació de corpus científics bilingües i avaluació de la traducció automàtica en el context dels estudis de comunicació acadèmica) han constatat que els investigadors han estat utilitzant àmpliament la TA per traduir les seves pròpies recerques i les metadades relacionades, així com carregant continguts multilingües en repositoris, fins i tot sense notificar que utilitzaven TA. En aquests casos, la TA pot ser, d'una manera més o menys acurada, posteditada, però sense cap grau de certesa quant a la qualitat a gran escala per al propietari o l'entitat gestora del repositori. És possible

que els repositoris no tinguin la capacitat de detectar i revisar aquest material, ja que el seu volum creix ràpidament. A més, aquesta és una pràctica que els repositoris, en general, difícilment poden controlar a causa dels costos i recursos. Per això seria útil posar en marxa un sistema d'avís que permetés als investigadors proporcionar informació sobre la naturalesa de la traducció pujada al repositori. Idealment, aquest sistema d'avís faria possible distingir el contingut traduït i posteditat per humans de la traducció automàtica en brut.

Aquest sistema d'avís també podria ser útil per a repositoris amb capacitat per mostrar una TA instantània del contingut recuperat (de manera automàtica o a demanda).

L'avís, que es mostraria com un advertiment per a l'usuari amb la finalitat de conscienciar-lo sobre possibles errors i anticipar-se a possibles reclamacions, podria dir el següent:

"Aquest document/material és una traducció automàtica no revisada de [citació de l'original] feta el dia [DD-MM-AAAA] del [codi de l'idioma d'origen] al [codi de l'idioma de destinació] tal com es va publicar en [detalls de la publicació] / es va recuperar de [DOI, un altre sistema de resolució de PID o URL] utilitzant [nom de l'eina de TA]. Aquesta traducció automàtica no ha estat revisada ni editada i es proporciona "tal qual" amb l'únic propòsit d'ajudar els usuaris a comprendre almenys part del contingut original expressat en [idioma d'origen]. Aquesta clàusula no garanteix la correcció i l'exactitud de la traducció automàtica [en la llengua de destinació] per part de cap persona física o jurídica en cap part d'aquesta traducció. [En conseqüència, disposar d'aquesta traducció no comporta cap responsabilitat per part de cap persona envers cap altra persona en cas que se'n faci ús, sigui quin sigui el propòsit]. Es convida expressament els usuaris d'aquesta traducció automàtica que la facin revisar, corregir o editar per un traductor professional o un expert en la matèria."

6. Tret que el document ho justifiqui (per exemple, traducció paral·lela, traducció comentada, versions replicades bilingües o multilingües), carregar les traduccions dels documents com a registres separats

Això és especialment apropiat en el cas de prefacis, introduccions o altres col·laboracions publicades en volums multilingües amb col·laboradors múltiples.

7. Promoure l'ús de llicències favorables a la (re)traducció per fomentar la traducció de continguts de nova producció i la retraducció, així com promoure els crèdits de traducció (per exemple, CC-BY)

Trobareu més informació referent a això 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. **Assegurar-se que es proporciona prou informació i recomanacions als dipositants en un apartat de preguntes freqüents o una altra manera d'implementar el que s'ha esmentat**

Annex 1. Casos d'ús i reptes

Aquests són alguns casos d'ús que impulsen les pràctiques recomanades:

1. Com a membre d'una institució no anglesa, en el repositori rebo documents en anglès que he de descriure.

Quan s'envia un document nou en anglès al repositori, és necessari descriure'l amb diferents camps de metadades en diferents idiomes (per exemple, resums, títols, paraules clau, tipus de document) i utilitzant vocabularis controlats no anglesos.
Exemple: la Universitat de Hokkaido utilitza l'esquema de metadades JPCOAR (les metadades en diversos idiomes s'introdueixen en el mateix camp de metadades, però es diferencien mitjançant l'atribut d'idioma, per exemple, dc.description.abstract i dc.subject[27]). En aquest esquema, una columna d'idioma en la part dreta de la pàgina mostra el codi d'idioma ISO de les metadades. Quan es dipositen articles de revistes, s'inclouen totes les metadades de la versió publicada (sense traducció de l'original; en les revistes en japonès, els resums i les paraules clau solen estar escrits també en anglès i el text complet, en japonès); els resums són en les metadades i l'atribut d'idioma hi està incrustat; els noms dels autors són en l'idioma de l'article. Com a mínim hi ha un esquema per marcar metadades per a diversos idiomes, però sorgeixen dubtes sobre la descobribilitat de contingut i quines metadades són més adequades.

2. Com a responsable de repositori, sovint gestiono articles, tesis o dissertacions que estan escrits en més d'un idioma.

Totes les tesis i dissertacions arriben en francès, però moltes contenen articles inserits en forma de capítols en la llengua en la qual han estat escrits.
Exemple: a la Universitat de Lieja, si un document està disponible en diversos idiomes, cada versió en un idioma es posa a disposició com un registre diferent amb metadades en diferents idiomes. És el cas d'un mateix document en dues llengües diferents per al qual existeixen dos registres diferents,[28] però només hi ha un atribut d'idioma per al registre.

3. Com a autor o autora, m'agradaria veure els meus articles escrits en diferents idiomes en un sol registre (per a finalitats estadístiques i informatives).

Tots els articles en diversos idiomes es dipositen en un únic article i han de ser descrits adequadament.

[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/2268/170862ihttps://orbi.uliege.be/handle/2268/170863

Exemple: abans, la Universitat Oberta de Catalunya disposava de dos registres separats per als articles en diversos idiomes. Actualment, a petició dels autors, les traduccions s'uneixen en un únic registre o fins i tot en el mateix document d'arxiu, la qual cosa simplifica el seguiment de les citacions i augmenta la visibilitat. Però hi pot haver problemes per als agregadors de continguts i els serveis d'indexació.

4. Com a responsable de repositori, vull oferir camps d'enviament en diversos idiomes.

[AIXÒ POT SER ESPECÍFIC DE DSPACE]. En configurar els formularis d'enviament, les etiquetes i l'ajuda/instruccions de cada camp només poden estar escrites en un idioma. El multilingüisme només es pot aconseguir escrivint l'etiqueta en cada idioma en el mateix camp (Autor/Auteur).

5. Com a responsable de repositori, vull tenir el nom i la descripció d'una col·lecció en més d'un idioma.

Actualment només es permet un idioma per al nom i la descripció d'una col·lecció.

Exemple: [AIXÒ POT SER ESPECÍFIC DE DSPACE]. Estaria bé que els textos introductoris (en HTML), etc., de les comunitats/col·leccionis es poguessin presentar en diversos idiomes. Això es podria aconseguir fàcilment utilitzant CSS i etiquetes div amb nom. Però, lamentablement, els atributs d'HTML, com id i style, semblen eliminar-se en la sortida HTML. És a dir: text es transforma en text en la interfície d'usuari.

Com que les col·leccions i les comunitats són elements a DSpace (i, per tant, tenen les seves pròpies metadades), una manera de resoldre aquest problema seria permetre seleccionar l'idioma de les metadades, com ja es pot fer per a les metadades d'objectes (és a dir, els resums).

Una solució ràpida i senzilla al problema bilingüe de les col·leccions/comunitats a DSpace és utilitzar un delimitador (per exemple, la barra |) entre dos textos que descriguin aquestes entitats i els seus camps de metadades, segons sigui necessari. N'hi ha prou amb dividir el text en el moment de la visualització perquè només es mostri el text que estigui actiu en cada moment. Aquí podeu veure la versió àrab de la llista de comunitats/col·leccions. En canviar l'idioma de la interfície a l'anglès utilitzant la icona del món situat en la part superior, veureu que apareixen totes en anglès. El mateix plantejament s'ha aplicat als elements de facetes, on ara veureu valors controlats (com ara noms de formats/tipus, universitats/instituts superiors/departaments, entitats, etc.) en diversos idiomes.

6. Com a responsable de repositori, vull poder gestionar les etiquetes en el meu idioma de manera eficaç.

En els programes multilingües de codi obert (OJS, DSpace, EPrints, etc.), les etiquetes en anglès són les obligatòries a l'hora de desenvolupar noves funcions. Les actualitzacions d'altres idiomes se solen quedar enrere i són gestionades posteriorment per la comunitat o, a vegades, localment. La

traducció de les noves funcionalitats del programari representa un repte important.
Exemples: en el repositori EPrints de ZORA (Zurich Open Repository and Archive) hi ha una versió alemanya de la interfície. CSpace, a la Xina, inclou un esquema de metadades i una interfície en diferents idiomes, però els gestors de repositoris continuen tenint dificultats per descriure el contingut dels repositoris. Normalment són els usuaris els qui seleccionen les etiquetes d'idioma i també reben formació sobre com es diposita el contingut multilingüe. Els idiomes d'interfície dels repositoris desenvolupats pel Centre Informàtic de la Universitat de Belgrad (Sèrbia) inclouen l'anglès i el serbi (en tots dos alfabets: llatí i ciríl·lic).[29] Com que els usuaris no estaven satisfets amb les traduccions disponibles, l'equip de desenvolupament va idear una aplicació web pròpia per facilitar la traducció. L'aplicació permet afegir, eliminar i canviar les etiquetes seleccionades en repositoris individuals o en tots els repositoris. Els canvis es propaguen als repositoris en 24 hores.

7. Com a responsable de repositori, vull oferir la traducció de metadades (per exemple, resums, títols i temes) a l'anglès.

Algunes metadades s'han de traduir a l'anglès utilitzant eines de traducció automàtica.
Exemples: s'utilitza una API de traducció de Google per traduir resums, títols i temes. Això també es podria aconseguir recomanant o exigint en les directrius per a l'usuari una quantitat mínima de metadades en anglès. En l'arxiu digital de l'Acadèmia Sèrbia de les Ciències i les Arts es recomana proporcionar almenys una descripció breu i paraules clau en anglès, ja que això millora la descobribilitat del contingut.

8. Com a responsable d'un repositori nacional, he de dipositar articles en totes les llengües del meu país.

Els continguts estan disponibles en els idiomes locals, però alguns no disposen de codi de llengua, no estan en Unicode i no existeixen vocabularis controlats en aquests idiomes.
Exemple: al Nepal, només els títols s'introdueixen en llengua nepalesa i la resta de metadades són en anglès. No hi ha estandardització de paraules clau en llengua nepalesa ni hi ha vocabularis controlats. Moltes llengües locals no estan en Unicode i a vegades s'utilitzen paraules llatinitzades. **Per exemple:** किताब **kitaba (llatinitzat) i book (la forma**

[29]Per exemple: https://dais.sanu.ac.rs

traduïda a l'anglès). Això genera problemes per a la indexació de Google Scholar, que prefereix veure les metadades en la llengua de l'article.

9. Com a responsable de repositori, m'agradaria exposar l'idioma de les metadades a OAI-PMH.

Actualment no existeix l'exposició de l'idioma de les metadades a OAI-PMH.
Objectiu desitjable: els repositoris haurien d'utilitzar de manera coherent i conscient les etiquetes d'idioma de les metadades per garantir que no s'exposi informació incorrecta sobre l'idioma. I un atribut d'idioma hauria de ser exportable, incloent-hi OAI-PMH. Una altra opció podria ser un plantejament proactiu per part dels repositoris. Per exemple: descarregar mensualment una extracció dels fulls de referència de metadades i posar-los a disposició pública per exposar els valors d'idiomes.

10. Com a agregador de contingut i responsable del sistema de descobriment, vull saber quin és l'idioma del document de text complet que indexo per poder ajudar els usuaris a trobar el contingut en el seu idioma preferit.

Hi ha problemes d'indexació de continguts respecte a l'agregador (Solr, VuFind, etc.) perquè no hi ha manera de separar els índexs per idioma i utilitzar eines específiques de cada idioma per enriquir les experiències de cerca. **La majoria de les metadades dels repositoris regionals no separen adequadament la informació multilingüe. Fins i tot es poden trobar llengües barrejades en camps de metadades textuals individuals.** **Les paraules clau i els descriptors són en diversos idiomes sense la identificació adequada; centenars de repositoris utilitzen vocabularis diferents fins i tot en el mateix idioma. S'ha debatut entorn de la implementació de classificadors automàtics per etiquetar les metadades dels repositoris amb vocabularis normalitzats per a cada regió.**
Exemples: la xarxa llatinoamericana de repositoris d'accés obert LA Referencia desenvolupa una eina de detecció d'idiomes (utilitzant diferents biblioteques python per al processament del llenguatge natural) per separar les llengües en els camps textuals de metadades amb la finalitat de millorar les metadades per als agregadors. La idea és afegir etiquetes xml:lang adequades a cada camp textual de metadades. Aquest etiquetatge seria utilitzat pel procés d'indexació amb la finalitat de generar índexs separats, però encara així el problema d'enfrontar-se a diversos idiomes en la interfície d'usuari de cerca és complex de resoldre. CORE sembla que utilitza una eina de detecció d'idiomes. Distingir entre bosnià, croat, montenegrí i serbi és tot un repte, ja que es tracta de llengües molt semblants. A causa d'això, quan es tracta d'aquests idiomes, les etiquetes d'idiomes a CORE solen ser

incorrectes. L'ús de l'etiqueta aglutinadora BCMS per als quatre idiomes seria una solució a aquest problema.

11. Com a agregador, m'agradaria indexar correctament els continguts i ajudar els usuaris a trobar-los en els seus idiomes.

Les directrius institucionals i temàtiques dels repositoris d'OpenAIRE (per a l'agregació de continguts de repositoris) animen a utilitzar l'atribut xml:lang per indicar l'idioma de les metadades. L'agregador OpenAIRE admet l'etiqueta xml language.

Exemple: **<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 admet l'etiqueta d'idioma xml i l'agregador fa comprovacions de metadades per a l'idioma, per exemple, en temes, títols i resums/descripcions, però no en noms. Es recomana ORCID per als noms; OpenAIRE I+T: títol, descripció
OpenAIRE també permet diversos idiomes, que s'indiquen en cada recurs de contingut.

12. Com a investigador, vull saber quines recerques hi ha en altres idiomes. Podria ser també un cas d'ús per a un pacient, etc.

Traduir els resums i posar-los a disposició, o oferir una opció de cerca per paraules clau en molts idiomes podrien ser algunes de les solucions. Les eines d'aprenentatge profund ho han començat a fer, per exemple, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Exemples: a BASE, cerca multilingüe (el concepte de cerca està inclòs en el tesarurus EuroVoc o el tesaurus Agrovoc; exemple de cerca per a climatologia). Wikidata i Abstract Wikipedia proporcionen informació independent de l'idioma.

13. Com a bibliotecari o arxiver de preservació digital, necessito saber com he d'incloure informació en llenguatge natural en les metadades tècniques i descriptives perquè els documents d'arxiu digitals es puguin indexar de manera eficaç amb la finalitat que es puguin recuperar i s'hi pugui accedir de manera correcta.

Les millors pràctiques documentades per incloure informació en llenguatge natural mitjançant normes de metadades de preservació digital com METS i PREMIS faciliten més accessibilitat, inclusió i diversitat dels arxius digitals.

Exemples: l'idioma és una informació necessària per indexar amb eficàcia i recuperar text (*stemming* —reducció de paraules a la seva arrel—, paraules buides), continguts de vídeo i àudio (la conversió de veu a text afavoreix la recuperació/indexació, la subtitulació d'àudio i l'accessibilitat al vídeo).

Es poden incloure metadades d'idioma utilitzant l'etiqueta <dc:language> de Dublin Core com a part de les metadades descriptives internes (mdWrap) d'un arxiu METS.

Es poden incloure metadades d'idioma com una de les propietats significatives (<significantProperties>) de les unitats semàntiques a PREMIS.

En el cas de documents textuals, les metadades d'idioma es poden incloure utilitzant textMD, normalment en forma d'esquema d'ampliació a dins de la secció de metadades administratives de l'estàndard METS. L'idioma també es pot incloure com a part d'un document textMD independent a dins de l'element de PREMIS <objectCharacteristicsExtension>.

14. Com a usuari, vull poder utilitzar una interfície en el meu propi idioma per enviar o consultar continguts.

La interfície del repositori està disponible en diversos idiomes.

Exemples: el repositori de la Universitat Oberta de Catalunya posa a la disposició de l'usuari final tres interfícies perquè pugui triar el seu idioma. Cada interfície d'idioma té els noms dels camps de metadades en l'idioma corresponent, per exemple: *autor*, en català i espanyol; *author*, en anglès.
En tots els repositoris institucionals desenvolupats pel Centre Informàtic de la Universitat de Belgrad, la interfície d'usuari final està disponible en anglès i serbi (tant en alfabet ciríl·lic com llatí). No obstant això, les etiquetes i l'ajuda en el formulari d'entrada només estan disponibles en serbi perquè no és possible alinear-les amb l'idioma de la interfície en DSpace.

15. Com a membre d'una institució en llengua anglesa, utilitzo un catàleg per descriure el contingut del meu repositori, tant en anglès com en altres idiomes.

El contingut s'introdueix en la llengua materna, però hi pot haver problemes de trobabilitat.

Exemple: a Berkeley Law (facultat de Dret de la Universitat de Berkeley) s'utilitza un sistema basat en MARC per descriure continguts. Com que es tracta de menys de l'1-3 % del contingut, no cal esperar que la cerca amb termes no anglesos retorni cap resultat tret que l'usuari cerqui una cosa específica. No s'utilitzen termes temàtics en el repositori, però sembla una forma fàcil d'augmentar l'accessibilitat en altres idiomes.
El catàleg i el repositori estan vinculats, i la cerca està disponible en molts idiomes. Els catalogadors parlen molts idiomes i són capaços de catalogar en llengües no angleses, però tot i això la major part de la catalogació es fa en anglès per a parlants monolingües.

16. Com a institució que admet moltes traduccions, m'agradaria que es reconegués el mèrit dels traductors en dipositar elements traduïts en el repositori.

Els traductors es poden esmentar utilitzant taxonomies, com ara la taxonomia CREDIT, que actualment només està disponible en anglès, i seria bo comptar amb una traducció oficial a altres idiomes. Existeixen dues traduccions "no oficials" al francès.[30] Els traductors són reconeguts en el repositori institucional (per exemple, com a col·laboradors amb noms i funcions), però no és el cas d'altres arxius, com els de prepublicaciones.
Exemple: el repositori de la Universitat de Lieja disposa d'un camp de metadades per al traductor (vegeu-ho aquí).

17. Com a traductor, m'agradaria saber si existeix una traducció determinada.

Com a traductor, vull saber si existeix una traducció: • **Per a una citació inclosa en un document font en el mateix idioma, però necessito comprovar si existeix cap versió en l'idioma de destinació (original o traduïda) del text citat (amb una referència en les notes o la bibliografia del document font) abans de decidir si tradueixo la citació jo mateix o reutilitzo la citació traduïda existent en la meva traducció.** • **Per utilitzar textos sobre el mateix tema que la traducció que m'han encarregat, pot ser que necessiti construir un corpus de documents semblants en les llengües de partida i d'arribada del meu encàrrec per utilitzar-los en un programari de concordança que permeti buscar cadenes de text (paraules, termes, frases) en una llengua i recuperar-les en dues llengües. Puc cercar a través d'una recerca documental una col·lecció de documents amb la seva traducció corresponent en la llengua de destinació i, després, processar-ho amb un programari d'alineació per obtenir arxius amb paraules i frases alineades.** • **Per alinear traduccions hi ha dues opcions:** **a) Alimentar un sistema CAT (traducció assistida per ordinador).** **b) Alimentar els mòduls d'aprenentatge d'un sistema de TA (traducció automàtica).**
Exemple: en tots aquests casos, el fet que els documents es registrin amb metadades adequades per designar la condició d'original/traducció i apuntin a l'equivalent o equivalents en qüestió, podria ajudar les cerques d'escriptori esmentades si les metadades fossin interoperables amb els motors de cerca, els catàlegs de biblioteques, els repositoris i els sistemes CRIS. Això també pot ser rellevant en el camp de l'edició de revistes, terminologia, mineria de textos i tecnologies lingüístiques. Per facilitar el treball en aquests àmbits necessitem interoperabilitat i interconnexions entre els diferents sistemes.

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

Translate Science está construint una eina d'aquest tipus i per això necessitem bones metadades lingüístiques en els repositoris.

Annex 2. Declarar l'idioma del recurs en cada element: exemples d'implementació seguint les normes/directrius sobre metadades

Esquema Datacite 4.4	9 idiomes Ús: opcional Ocurrències: 0-1 (no repetible) Codificació recomanada: IETF BCP 47 o codis d'idioma ISO 639-1
Dublin Core (DC)	Nom del terme: language Ús: opcional Ocurrències: repetible La pràctica recomanada és utilitzar un valor no literal que representi un idioma d'un vocabulari controlat com a ISO 639-2 o ISO 639-3, o un valor literal consistent en una etiqueta d'idioma de la millor pràctica actual 47 de l'IETF [IETF-BCP47].
Norma de metadades per a tesis i dissertacions electròniques (ETDMS)	dc.language Ús: opcional Ocurrències: 0-N (repetible) Els noms dels idiomes s'han de registrar utilitzant la norma ISO 639-2 (o RFC 1766). Si no s'especifica l'idioma, s'assumeix que és l'anglès (en).
Esquema de descripció d'objectes de metadades (MODS)	Element de nivell superior: <language> Ús: opcional Ocurrències: 0-N (repetible) Aquest recurs conté textos en anglès i francès: <language> <languageTerm type="code" authority="iso639-2b">eng<languageTerm> </language> <language> <languageTerm type="code" authority="iso639-2b">fre<languageTerm> </language> Aquest recurs conté text en àrab egipci, que està codificat com a llengua 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>
Directrius d'OpenAIRE per a repositoris	dc:language Ús: obligatori si és aplicable (Mandatory if Applicable, DT.) Ocurrències: 0-N (repetible) Recomanació: prendre valors d'una de les llistes següents:

bibliogràfics, institucionals i temàtics	• IETF BCP 47, registre de subetiquetes lingüístiques de la IANA • ISO 639-x, on x pot ser 1, 2 o 3. Millor pràctica: utilitzar la norma ISO 639-3, amb la qual cosa seguim l'estàndard http://www.sil.org/iso639-3/ Si és necessari, repetir aquest element per indicar diverses llengües. Si les normes ISO 639-2 i 639-1 són suficients per al contingut d'un repositori, es poden utilitzar alternativament. Com que hi ha un mapatge únic, es pot fer durant un procés d'agregació.
Japan Consortium for Open Access Repository (JPCOAR)	dc:language Ús: recomanat (R) Ocurrències: 0-N (repetible: excepte terme obligatori) Instruccions d'ús Introduir les llengües principals que s'utilitzen en el text principal del recurs. Utilitzar els codis d'idioma ISO 639-3. L'ús de la macrollengua d'ISO 639-3 és opcional. Notes No introduir noms d'idiomes. No introduir noms de països. Introduir en ordre de prioritat d'idioma. Exemples recomanats El text principal del recurs és en anglès. <dc:language>eng</dc:language> El text principal del recurs és en anglès i japonès. <dc:language>eng</dc:language> <dc:language>jpn</dc:language> Exemples no recomanats No es recomana ISO 639-1. <dc:language>ja</dc:language> No introduir diversos idiomes en un mateix element. <dc:language>engjpn</dc:language> No utilitzar majúscules ni caràcters de doble byte. <dc:language>JPN</dc:language> **<dc:language> e n g </dc:language>** No introduir noms d'idiomes. **<dc:language>日本語</dc:language>** No introduir noms de països. <dc:language>US</dc:language> No introduir codis d'idioma diferents d'ISO 639. <dc:language>en_US</dc:language>

Annex 3. Declarar l'idioma de les metadades (atribut xml:lang): exemples d'implementació seguint les normes/directrius sobre metadades

Esquema Datacite 4.4	xml:lang="EN", per exemple <xs:element name="**title**" maxOccurs="unbounded"> <xs:annotation> <xs:documentation>Nom o títol pel qual es coneix un recurs.</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"/> Igualment, per a 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)	Quan s'indiqui l'idioma del valor, s'haurà de codificar utilitzant l'atribut "xml:lang". Per exemple: <dc:subject xml:lang="en">seafood</dc:subject> <dc:subject xml:lang="fr">fruits de mer</dc:subject>
Norma de metadades per a tesis i dissertacions electròniques (ETDMS)	L'idioma és un qualificador global que es pot utilitzar en qualsevol element: https://ndltd.org/wp-content/uploads/2021/04/etd-ms-v1.1.html#qualifiers
Esquema de descripció d'objectes de metadades (MODS)	Hi ha atributs relacionats amb l'idioma: https://www.loc.gov/standards/mods/userguide/attributes.html#list ISO 639-2/b
Directrius del repositori institucional i temàtic OpenAIRE	Ús de l'atribut xml:lang per indicar l'idioma de les metadades. Exemple: **<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 admet l'etiqueta d'idioma xml i l'agregador fa comprovacions de metadades per a l'idioma, per exemple, en temes, títols i resums/descripcions, però no en noms. Es recomana ORCID per als noms. OpenAIRE I+T: títol https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/latest/field_title.html#dci-title. Descripció: https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/latest/field_description.html#attribute-lang-o OpenAIRE també permet diversos idiomes https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/v4.0.0/field_language.html, el recurs de contingut té aquest idioma.
JPCOAR 1.0.2 https://schema.irdb.nii.ac.jp/ja/schema	L'atribut xml:lang es pot utilitzar per a cada element. En principi s'utilitzarà el codi d'idioma de dos dígits de la norma ISO 639-1 (per exemple, "ja" per al japonès, "en" per a l'anglès). Per a yomi japonès, utilitzeu “ja-Kana“. Quan introduïu "yomi", heu d'indicar informació original (és a dir, en kanji) amb la descripció que "xml:lang is 'ja'". En el cas del xinès, és convenient introduir per separat el xinès simplificat com a "zh-ch" i el xinès tradicional com a "zh-tw".
Esquema de metadades 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	Canvi des d'1.0.2 : admet a més "ja-Latn". Els extractes de la part actualitzada: La informació d'idioma per a katakana yomi (lectura en katakana) és xml:lang="ja-Kana", i per a romaji yomi (lectura en romaji) és xml:lang="ja-Latn". Allí on introdueixi un yomi, la informació en xml:lang="ja" s'ha d'introduir per separat del yomi.
Akdeniz, Esra, & Moilanen, Katja. (2023). Model de metadades CMM CESSDA (3.0). Zenodo. https://doi.org/10.5281/zenodo.7528240	1.1.3.1. Idioma del títol de l'estudi L'idioma del contingut de l'element. M (Mandatory) ISO 639-1 Ocurrències: 1 ddi:DDIInstance/s:StudyUnit/r:Citation/r:Title/r:String/@xml:lang El mateix per a l'idioma de subtitulació; Idioma del títol alternatiu; Idioma del motiu de la versió; Idioma del resum; Idioma del tema d'estudi (descriptiu); Idioma de la paraula clau (descriptiu); Idioma de la disciplina (text lliure); Idioma del tipus de font de dades (descriptiu); Idioma de la manera de la col·lecció de dades (descriptiu); Idioma de les condicions d'accés a les dades; Idioma de les condicions d'accés a les metadades (estudi); Idioma del nom complet de l'organització; Idioma del nom, abreviatura o acrònim de l'organització; Idioma de la descripció de l'organització; Idioma de la descripció de la versió del conjunt de dades; Idioma del conjunt de dades; Idioma de la descripció de l'arxiu del conjunt de dades; Idioma del nom de l'arxiu; Idioma del títol del document; Idioma del títol de

	la publicació; Idioma del nom de la revista/sèrie - 75 camps de metadades en total per indicar l'idioma; També hi ha camps de metadades per indicar traduccions, per exemple 1.1.3.2 Estat de la traducció del títol de l'estudi El contingut de l'element està traduït? R: vertader, fals Ocurrències: 0-1 ddi:DDIInstance/s:StudyUnit/r:Citation/r:Title/r:String/@isTranslated; i 28 camps de metadades que esmenten la traducció

Annex 4. Exemples d'implementació de les normes ISO 639-1, ISO 639-2 i ISO 639-3

ISO 639-1 i ISO 639-2
Propietat d'idioma en el Data Catalog Vocabulary (DCAT) - Versió 2 (recomanació del W3C del 4 de febrer de 2020):

Abast:	S'haurien de fer servir els recursos definits per la Biblioteca del Congrés (ISO 639-1, ISO 639-2). Si es defineix un codi ISO 639-1 (de dues lletres) per a l'idioma, llavors s'hauria d'utilitzar el seu IRI corresponent; si no es defineix cap codi ISO 639-1, llavors s'hauria d'utilitzar l'IRI corresponent del codi ISO 639-2 (de tres lletres).
Nota d'ús:	Repetiu aquesta propietat si el recurs està disponible en diversos idiomes.

En el Data Catalog Vocabulary (DCAT) - Versió 3 s'inclou la mateixa redacció.
Esborrany de treball del W3C del 10 de maig de 2022.

Codis per a la representació de noms d'idiomes ordenats alfabèticament pel codi alfa-3/ISO 639-2: http://www.loc.gov/standards/iso639-2/php/code_list.php.

ISO 639-3
ISO 639-3 amplia els codis ISO 639-2 amb l'objectiu de cobrir totes les llengües naturals i funciona millor per a llengües com el cebuà, el montenegrí o el quítxua (que té variants dependents de la regió del país). Per exemple, es recomana en la Guia del repositori ALICIA (també una videoguia, el Perú).

Les recomanacions de metadades per al material de text emmagatzemat en repositoris de publicacions finlandeses recomanen la norma ISO 639-X per a dc.language.iso. És preferible utilitzar els codis d'idioma de 3 caràcters d'ISO 639-2 o ISO 639-3, segons correspongui.

Continua havent-hi alguns problemes d'implementació per a un codi de tres lletres, ja que no tots els repositoris el podrien admetre ara (a causa de problemes de programari i llenguatge XML) i hi podria haver problemes similars amb els agregadors (per exemple, OpenAIRE les recomanacions https://www.w3.org/tr/xml/ i https://www.w3.org/tr/xml/#RFC1766).

Més informació sobre les etiquetes d'idiomes

Un article útil i més descriptiu sobre etiquetes d'idioma en HTML i XML publicat el 2014 per W3 amb exemples:

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

i una proposta d'ús.

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

Per a moltes llengües menys conegudes, parlades per grups minoritaris, i també per a períodes històrics de les llengües, simplement no es disposa de codis d'idioma, que són la base de les etiquetes; en referència a aquest aspecte, vegeu "The Shortcomings of Language Tags for Linked Data When Modeling Lesser-Known Languages", amb recomanacions per millorar o desenvolupar codis d'idioma ISO.

Annex 5. Correcció d'inconsistències del codi d'idioma en registres de repositoris DSpace

Si la llengua de destinació utilitza caràcters únics, pot ser possible establir automàticament el valor de les metadades d'idioma.

Heus aquí un exemple en SQL perquè DSpace especifiqui elements utilitzant una llengua de destinació i els assigni un valor d'idioma suposant que la llengua de destinació no està representada per caràcters de 2 bytes:

```
update metadatavalue set text_lang='/*indiqueu aquí el codi ISO de l'idioma de destinació*/'
 where metadata_field_id in (/*indiqueu aquí els números de cada metadata_field_id les metadades del qual acceptin algun valor de cadena */)
 and length(text_value)!=octet_length(text_value)
and text_value '^[/*indiqueu aquí tots els caràcters específics d'ús exclusiu en la llengua d'arribada† */].*'
 and (text_lang is null or text_lang != '');
```

† Es pot utilitzar una expressió regular que comprengui tots els caràcters de la llengua. Per exemple, per a l'escriptura japonesa: [ぁ－んァ－ヶ亜－腕]; i per a la ciríl·lica: [а-яА-ЯЂ-Ҝђ-ҝ].

Amb una tasca cron de nit es pot afegir ràpidament el codi "en" en qualsevol metadada que no disposi del codi d'idioma. Vegeu com es crea una consulta o funció SQL per canviar text_lang a "en".

Les eines Atmire CSV Power Tools es poden utilitzar per editar metadades exportades (en i en_US, així com parèntesis, i altres qüestions d'idiomes).

Annex 6. Correcció de la manca de l'idioma de document en registres de repositoris EPrints

REAL és un repositori de gestió d'EPrints encarregat el 2008 i que actualment conté més de 220.000 articles repartits en vuit col·leccions. El contingut és divers i està format per articles de recerca actuals pujats per investigadors i material digitalitzat per la institució matriu, la Biblioteca i Centre d'Informació de l'Acadèmia Hongaresa de Ciències. La versió actual del programari REAL és la 3.3.15.

El camp d'idioma dels documents sempre ha estat present, però fins ara no havia estat visible en els formularis de càrrega de documents de la web, ni en cap de les visualitzacions d'un article, per la qual cosa els dipositants o bibliotecaris no ho podien configurar ni podien comprovar-ne el contingut.

<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>

Recentment hem exposat el camp i hem descobert que EPrints fixava el contingut en funció de la configuració d'idioma utilitzada en el navegador en el moment del dipòsit, és a dir, que els valors que conté són més o menys aleatoris. Per esbrinar (i establir) els valors correctes per a centenars de milers d'articles, vam elaborar una llista d'identificadors dels articles que volíem comprovar, vam baixar les metadades en format DC, vam extreure el títol i vam intentar endevinar l'idioma del document basant-nos en l'idioma del títol.

El nostre script començava amb una hipòtesi (la primera hipòtesi era que l'idioma del document és l'hongarès), les paraules del títol s'introduïen en un corrector ortogràfic, i si més de la meitat de les paraules eren
reconegudes, acceptàvem la hipòtesi com a veritable. En l'execució següent es van comprovar els elements restants amb la hipòtesi "l'idioma és l'anglès", i després es van provar altres idiomes.

El fragment de script C-shell que apareix a continuació mostra la prova del títol enfront de la hipòtesi "l'idioma és l'italià" utilitzant el corrector ortogràfic 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)"`
```

L'experiència amb aquest mètode demostra que, amb alguns filtres, la taxa d'error es podria reduir a l'1-2 %, que és molt millor que el percentatge d'error actual del 40-50 %. Cal tenir en compte que hi ha documents complicats, multilingües o molt tècnics (per exemple, de matemàtiques) que representen un desafiament. No sabem com s'han d'etiquetar els documents bilingües / multilingües.

Annex 7. Eines de tractament de textos

Sempre que sigui possible, cal especificar l'idioma o idiomes del document, de paràgrafs individuals i de frases mentre s'escriu en l'eina de tractament de textos.

Per especificar l'idioma de determinats paràgrafs i frases en MS Word, OpenOffice, LibreOffice i eines similars, hem d'utilitzar la configuració d'idioma i els teclats adequats mentre escrivim. Per especificar l'idioma o idiomes en un document existent, cal seleccionar el text i definir l'idioma utilitzant l'eina d'idioma de la barra d'eines o del menú. Per conservar aquesta informació després de la conversió a PDF, el document s'ha d'exportar com a PDF etiquetat. Malgrat això, depenent de l'extensió PDF incorporada en el processador de textos, aquesta informació es pot perdre durant la conversió a PDF.

W3C dona recomanacions sobre com s'ha d'especificar l'idioma d'un paràgraf o frase amb l'entrada Lang en documents PDF. Ara bé, per aplicar aquestes recomanacions en els arxius PDF es necessita el programari comercial Adobe Acrobat.

LaTeX també permet treballar en diversos idiomes. Hi ha diversos paquets que permeten la composició tipogràfica en diferents idiomes (com ara babel o polyglossia), i aquesta funció també està disponible a Overleaf, l'editor col·laboratiu en línia de LaTeX.

No obstant això, la interoperabilitat de les diferents eines d'edició de text i els formats utilitzats continua essent una assignatura pendent. Calen normes clares i la col·laboració amb els fabricants de programari per garantir no sols que el text creat en diverses eines de programari continuï essent llegible per a humans i màquines, sinó també que les diverses característiques i funcionalitats que estaven disponibles en el document original (codificació, etiquetes, anotacions, etc.) ho continuïn estant després de la conversió a altres formats.