Good Practice Advice for Managing Multilingual and non-English Language Content in Repositories

Image credit | Tommaso D'Incalci | Ikon Images CC BY-NC

October 30, 2023

Prepared by the COAR Task Force on Supporting Multilingualism and non-English Content in Repositories

COAR
Confederation of Open Access Repositories

Cite as:

COAR Task Force on Supporting Multilingualism and non-English Content in Repositories. October 2023. *Good Practice Advice for Managing Multilingual and non-English Language Content in Repositories, Version 2*. Confederation of Open Access Repositories (COAR). DOI: 10.5281/zenodo.10053918

Acknowledgements

Contributing Task Force Members

Iryna Kuchma, EIFL (chair), Ukraine

Jagadish Aryal, Social Science Baha, Nepal

Andreas Czerniak, Bielefeld University – Library, Germany

Christophe Dony, ULiège Library, Belgium

Joe Cera, Berkeley Law Library, University of California, USA

Sebastiano Giorgi-Scalari, Open University of Catalonia, Spain

Gussun Gunes, Marmara University Information and Records Management Department, Türkiye

Gultekin Gurdal, Izmir Institute of Technology İYTE, Türkiye

Johanna Havemann, AfricArXiv, Germany

Libio Huaroto Pajuelo, Universidad Peruana de Ciencias Aplicadas, Peru

Alan Ku (Gu Liping), National Science Library, Chinese Academy of Sciences, China

Pierre Lasou, Bibliothèque de l'Université Laval, Canada

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

Lautaro Matas, LA Referencia, Spain/Latin America

Ayako Mikami, Hokkaido University, Japan

Tomoki Nagase, National Institute of Informatics, Japan

Tomasz Neugebauer, Concordia University, Canada

Jean-Francois Nomine, INIST, France

Milica Sevkusic, ITS SASA/EIFL, Serbia

Kathleen Shearer, COAR, Canada/International

Freddy Sumba, CEDIA, Ecuador

We would like to thank the following people for their input to these recommendations:

Ginny Barbour (on behalf of 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 and Kelly Stathis.

Introduction

Multilingualism is a critical characteristic of a healthy, inclusive, and diverse research communications landscape. Publishing in a local language ensures that the public in different countries has access to the research they fund, and also levels the playing field for researchers who speak different languages. The Helsinki Initiative on Multilingualism in Scholarly Communication asserts that the disqualification of local or national languages in academic publishing is the most important – and often forgotten – factor that prevents societies from using and taking advantage of the research done where they live. While the dominant position of a *lingua franca* – English – is useful for the widespread dissemination of ideas across the world, it also impedes the use of research results at the local level.

After decades of policies that have directed researchers to publish in English, we are starting to see a reversal of this trend. In Europe, Asia, and many other jurisdictions, policy makers are introducing new measures that encourage researchers to publish in local and indigenous languages. The UNESCO Recommendation on Open Science, for example, calls on member states to encourage "multilingualism in the practice of science, in scientific publications and in academic communications". This aligns and reinforces less recent statements such as mentions in the Universal Declaration of Human Rights, which urges that research now be discriminated against based on their language and the UNESCO Recommendation concerning the Promotion and Use of Multilingualism and Universal Access to Cyberspace, which calls on the community to *"take the necessary measures to alleviate language barriers and ...ensure that all cultures can express themselves and have access to cyberspace in all languages, including indigenous ones."*

Multilingualism presents a particular challenge for the discovery of resources. If the language of a scholarly resource is not labelled properly it will not be correctly indexed by discovery services. That is because indexing involves text analysis practices such as stemming, lemmatization (grouping together the inflected forms of a word so they can be analysed as a single item), and the appropriate treatment of stop-words. All of these text analysis techniques are very language specific. Including the language tag and adopting other such practices enables information seekers, aggregators, indexers, and discovery services to correctly identify the language of the full text and process items accordingly. Moreover, while researchers and other information seekers may only be able to read in one or two languages, they want to know about all the relevant research in their area, regardless of the language in which it is published. Proper language attribution of the resource is important to support this need and offer better multilingual retrieval.

In August 2022, COAR launched the COAR Task Force on Supporting Multilingualism and non-English Content in Repositories to develop and promote good practices for repositories in managing multilingual and non-English content. Based on 17 use cases contributed by different stakeholders communities (repository managers and users, authors and translators, aggregators and discovery systems), the Task Force identified three relevant areas for their work: enhancing discoverability of non-English content; curating multilingual content in a repository; and supporting translations. The use cases are documented in Appendix 1.

In June 2023, the Task Force published an initial set of draft recommendations for community review. The consultation resulted in a wide range of input, which was reviewed by the Task Force and incorporated into a second version of the recommendations. This document presents the updated recommendations based on this community input.

The recommendations identify good practices for repository managers and repository software developers, and focus on the topics of metadata, multilingual keywords, user interfaces, formats, and licences that will improve the visibility, discovery and reuse of repository content in a variety of languages.

We very much hope that these recommendations will be widely adopted by repositories world wide. Some of the recommendations can be immediately adopted by repository managers, while others will take more time and fully implementing the recommendations will require collective efforts on the part of repository managers, aggregators, researchers, and software developers. In the coming months, COAR and the Task Force will disseminate the recommendations widely and work to advance their adoption in repositories across the globe

Summary Recommendations

Creators and curators of metadata

Declare the language of the resource at the item level

Declare the language of the metadata (e.g. xml:lang attribute)

Use standard (two-letter or three-letter) language codes (ISO 639)

Enable UTF-8 support in your repository and use the original alphabet / the writing system whenever possible. If it is necessary to transliterate metadata, use recognised standards (e.g. ISO)

If the repository software supports multiple interface languages, set up the user interface in the native language(s) of the target group, along with as English option

Write personal name/s using the writing system used in the deposited document and provide a persistent identifier enabling unambiguous identification, such as ORCID

Include keywords in many languages, use multilingual vocabularies and thesauri if possible

Recommendations for repository managers on translated content

Repository software/platforms developers

Ensure that language codes can consistently be used across the repository collections and are well supported

Expose the language of metadata via metadata exchange protocol, e.g. OAI-PMH, GraphQL API, etc.

Improve support for ISO language codes, e.g. three-letter codes needed for some languages

Ensure that multiple interface languages are well supported

Ensure that persistent identifiers are exposed via OAI-PMH. PIDs in Dublin Core™ Working Group has developed recommendations to make it possible to expose persistent identifiers including ORCID, via OAI-PMH

Provide support for multilingual keywords to increase the discoverability of multilingual repository content. For example, enable a real-time integration of Wikidata – e.g. when a user starts typing in the appropriate metadata field, relevant Wikidata terms appear in a drop-down list for the user to select

Enable automatic assignment of controlled terms based on the existing metadata

Detailed Recommendations

1. Declare the language of the resource at the item level

Recommendation

Declaring the primary language of the document is considered mandatory. The language metadata must be encoded using the ISO-639 language code (see 2.3 Use standard (two-letter or three-letter) language codes (ISO 639) for more details.

Guidelines

If the document has only one language, language metadata identifies the primary language of the resource. Attribution of the primary language of the resource must be done at the item level.

Example 1 language in simple Dublin Core XML with ISO-639-1 encoding

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

Example 2 language in MODS with ISO 639-2 encoding

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

If the whole document (e.g. edited volume) has important sections of the text in different languages, repeat the language metadata to mention each language.

Example 3: bilingual (French/English) document in simple Dublin Core XML with ISO-639-1 encoding

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

Example 4: bilingual (French/English) document in MODS with ISO 639-2 encoding

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

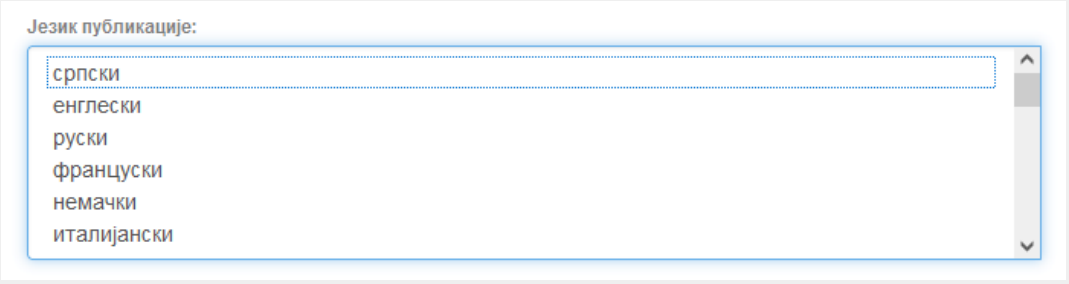
```
<languageTerm authority="iso639-2b" type="code"
uthorityURI="http://id.loc.gov/vocabulary/iso639-2"
valueURI="http://id.loc.gov/vocabulary/iso639-2/fre">fre</languageTerm>
</language>
```

Example 5: EPrints
EPrints can be extended to declare language information at the item or file level but this is not in place on EPrints by default. Similarly EPrints XML export plugins, embedded metadata and OAI-PMH interface code could be extended to define xml:lang attributes but it does not do this by default.

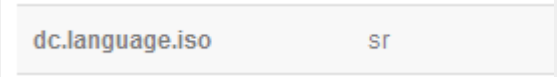
Example 6: OSF – Open Science Framework
New metadata enhancements on OSF for all OSF Projects, Registrations, and Preprints now includes the language of materials, more details in New OSF Metadata to Support Data Sharing Policy Compliance.

Example 7: Digital Archive of the Serbian Academy of Sciences and Arts (DAIS) Repository
The repository content is in more than 15 languages. Filtering by language is not enabled, and searching is not user friendly. Language is declared at submission, by selecting from a drop-down list (mandatory field). Multiple languages can be selected, but "Multilingual" doesn't exist as a value. Recommendations in the submission guidelines:

- The main body of the text in one language and title, abstracts and keywords in a different language: declare just the main language (but provide the metadata in both languages).
- Publication (e.g. edited volume) with important sections of the text in different languages: declare all the languages.
- Entire text provided in parallel in multiple languages: declare all the languages.



In the metadata, the selected language is displayed as a two-letter ISO code (but is given in a human-readable format in the drop-down)



Example 8: Digital Preservation Metadata (PREMIS & METS)
METS (Metadata Encoding and Transmission Standard) and PREMIS (Preservation Metadata: Implementation Strategies) are two metadata standards commonly used together to provide comprehensive metadata support for preserving and managing digital objects. METS is primarily focused on encoding descriptive, administrative, and structural metadata, providing a framework for organising and linking various types of metadata within a structured XML document. PREMIS, on the other hand, focuses on documenting the actions, events, and processes involved in the long-term preservation of digital objects. METS can serve as a container for various metadata, including PREMIS metadata, allowing for the integration of preservation-specific information within the broader context of digital object organisation and description.

The PREMIS data model does not explicitly classify language as technical or descriptive metadata. Neither does the PREMIS standard specifically define elements or sub-elements for capturing language information. However, the *significant properties* type and value semantic unit's components in PREMIS are sufficient for capturing language without a need for a dedicated language XML element.

Language can be considered to be technical metadata, significant for preservation, using the <significantProperties> element in PREMIS. Significant properties represent the aspects of a digital object that impact its rendering, behaviour, or interpretation, such as file format, compression algorithm, software version, resolution, colour space, and other technical features that affect the object's rendering and accessibility. We recommend that language is encoded as a significant property in that sense, using PREMIS, for example (in this example, the language specified is english):

```
<premis:significantProperties>

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

In addition, language information, if considered as a descriptive characteristic of the intellectual content (i.e. descriptive metadata), can be embedded into the METS document using Dublin Core (using the <dc:language> tag) or another metadata section within the METS container. Language can also be embedded into the METS as technical metadata for text documents using TextMD[1] within the PREMIS <objectCharacteristicsExtension> element.

See more implementation examples following metadata standards/guidelines in the Appendix 2.

[1] https://www.loc.gov/standards/textMD/

2. Declare the language of the metadata (e.g. xml:lang attribute)

Recommendation

Use the xml:lang attribute to indicate the language of the metadata field. Because of the cardinality [0, 1], the xml:lang attribute could describe the same element in different languages, so this would be more accurate than the dc:language element.

Guidelines

Regardless of the fact that English is mainly assumed to be the standard, the content should be exposed with a reference to the language used. It is worth doing it at the repository level as no other stakeholder, such as aggregators (i.e BASE, OpenAIRE, etc.), can infer language from the content of the metadata.

In case the deposited items has a title, or other metadata elements, in more than one language (e.g. the main title and the title of a summary or abstract), make sure that language information is indicated using the xml:lang[2] attribute/subproperty and properly exposed via metadata exchange protocols, such as OAI-PMH. As some aggregators are unable to harvest full information and all repeated fields, it is recommended to respect the order of titles metadata input, i.e. to provide the main title first. If possible, use dc.title alternative for additional titles.

Aggregators such as OpenAIRE[3] and BASE[4] will correctly identify the main title based on the information provided in the field indicating language of the document, regardless of the order at input.

However, there is no exposure of the language of metadata in OAI-PMH and therefore we request that software developers consider this in future versions of their platforms.

Example 9: How to attribute language when there is more than one language in the metadata fields

Use of the xml:lang attribute to indicate the language of the metadata field.

```
<datacite:titles>
<datacite:title xml:lang="en">Open Access</datacite:title>
<datacite:title xml:lang="pl">Otwarty Dostęp</datacite:title>
</datacite:titles>

<dc:title xml:lang="en">Open Access</dc:title>
<dc:title xml:lang="fr">Libre Accès</dc:title>
```

Here's a MODS example from AILLA, where we use ISO 639-3 language codes:

```
<titleInfo lang="eng">
<title>Iskonawa Oral Tradition</title>
</titleInfo>
```

[2] https://www.w3.org/International/techniques/authoring-xml#natlang
[3] https://www.openaire.eu
[4] https://www.base-search.net

<titleInfo lang=”spa”>
<title>Tradición Oral Iskonawa</title>

</titleInfo>
See more implementation examples following metadata standards/guidelines in the Appendix 3.

Example 10: DSpace
In DSpace 7, the value-pairs set for languages can include whatever languages and language identifiers desired. By default, DSpace provides 10 languages value-pairs: English (United States) (en_US), English (en), Spanish (es), German (de), French (fr), Italian (it), Japanese (ja) and Chinese (zh), Portuguese (pt), Turkish (tr). However, it is fully customizable in the submission-forms.xml file. It can include three letter identifiers, if there are languages with three-letter identifiers that are present in the target material for a collection. During submissions, language values are displayed as a dropdown list while in the edition mode, language is a free text field.

See Appendix 5 on how to fix language code inconsistencies in repositories running on previous versions of DSpace.

Example 11: TIND IR
The TIND IR is a MARC-based repository. That means that the easiest way to include information about multilingual content is through the 041 field and relevant subfields[5].

Example 12: WEKO 3
WEKO3 is a repository software developed by NII (National Institute of Informatics, Japan) based on INVENIO by CERN. This software operates JAIRO Cloud, a cloud-based repository system, which is supported by JPCOAR (Japan Consortium for Open Access Repositories) and NII. In WEKO 3, JPCOAR metadata schema is supported by default and a language attribute can be added for any metadata as long as it is allowed in the schema. Specifically, ISO-639-3 is acceptable as the language of the text, and for a language attribute of other metadata elements, ISO-639-1 is acceptable. With each field, you can add a language tag in the form of a two-character ISO using the dropdown menu, checkbox and radio button.

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

3. Use standard (two-letter or three-letter) language codes (ISO 639)

3.1 Introduction to language tags and codes

The identification of languages in an unambiguous way is essential for the interpretation, aggregation and re-use of research content. The standards for language tags have been updated and extended since the early years of the Internet in the 1990s. The latest language tag standard is defined by IETF's BCP 47 (RFC 5646) in combination with ISO 639-3.

A language tag is a requirement in HTML, XML and RDF, as a means to identify a natural language. A language code, in the form of a two- or three-character identification such as 'en' for English, is the main constituent of a language tag and is provided by the ISO 639 standard (parts 1-3). The language code can be followed by sub-tags refining or narrowing the range of the encoded language in the following form:

language-extlang-script-region-variantextension-privateuse.

The practice of language tagging is straightforward for a large number of well-known languages; ISO 639 includes codes for over 7900 languages (as of January 2023). However, it is important to note that lesser-known languages and regional varieties or historical stages of languages may not be sufficiently represented in ISO 639. The optional sub-tags compliant with BCP 47 offer some possibilities for more fine grained identification. The "x" *private-use* sub-tag defined in BCP 47 can be used for the identification of language variations[6]. In addition, ISO 639 is a standard that has changed over time, and offers opportunities for the submission of change requests:

> "Knowledge of human languages at any point in time will never be complete or perfect, but is always expanding. Given the comprehensive nature of ISO 639-3, changes to the code set are inevitable, especially in respect to lesser-known or newly identified languages."[7]

It is important to remember that the primary objective of language tagging is to accurately identify and represent a language being used, depending on the language and technology context of use. If a 2-letter code (ISO 639 Part 1) is not appropriate in a specific context, then a 3-letter code (ISO 639 Part 2 and 3) or other subtags (such as for script, region or private use) should be used to ensure interoperability and precision in language identification. The set of languages included in Part 1 of ISO 639 is considered a subset of Part 2 and any single 2-letter code in Part 1 with a corresponding 3-letter code in Part 2 or 3 are considered to be synonyms with the same extension. For example, the identifiers "fra", "fre" and "fr" designate the same language. BCP 47 recommends the use of the 2-letter codes whenever they exist, but ISO 639 states that free choice between synonyms should be allowed whenever possible. In this report, we recommend following this part of the BCP 47 recommendation and using 2-letter codes whenever they exist, but it may be appropriate to use 3-letter codes depending on the specific context of use.

[6] As described in https://aclanthology.org/2020.lrec-1.408.pdf, for example
[7] https://iso639-3.sil.org/code_changes/introduction

3.2 Summary decision tree to choose a language tag

The following is a summary decision tree for how to determine a language tag:

1. Search for the language code in ISO 639
2. If a 2-letter 639 Part 1 code is found for the language, use it. Go to 5.
3. If a 3-letter 639 Part 2 or 3 code is found for the language, use it. Go to 5.
4. Use the "x" subtag, reserved for private use, to define a custom language code. Go to 5.
5. Determine if a sub-tag is required and relevant to identify the language. For example, if the fact that this is a regional variation or dialect is important in the context, consider using ISO 3166 country codes as subtags (e.g., "en-US" for American English). If there are relevant writing system variations to identify, consider using ISO 15924 script codes as subtags (e.g., "sr-Latn" for Serbian in Latin script).

Note: The ISO 639 2 and 3 have standardised some special situations:
* mis is listed as "uncoded languages" (originally an abbreviation for "miscellaneous")
* mul (for "multiple languages") is applied when several languages are used and it is not practical to specify all the appropriate language codes
und (for "undetermined") is used in situations in which a language or languages must be indicated but the language cannot be identified.
* zxx is listed in the code list as "no linguistic content", e.g. animal sounds (code added on 11 January 2006).

Using language codes can also be practical for historical or local, regional or classical languages (e.g. Latin, Walloon, etc.)[8].

See more about ISO 639-1, ISO 639-2 and ISO 639-3 and language tags in Appendix 4.

Example 13: Linguistics and language studies
In linguistics and language studies, ISO-639-3 (3-letter) codes are a standard. First, most languages don't have 2-letter codes, and when they do they are often confusing because they don't represent languages (e.g. cr for 'Cree', ms for 'Malay', or zh for 'Chinese). This obscures exactly the type of diversity we hope to promote. Linguists and language archives are also increasingly using glottocodes for "languoids", since what gets to "count" as a language is largely political. Consider having an optional field to include those as well.

Example 14: MiCISAN institutional repository
MiCISAN institutional repository uses ISO 639-3.

[8] Examples in Walloon: https://orbi.uliege.be/handle/2268/28421 and https://orbi.uliege.be/handle/2268/28419

4. Enable UTF-8 support in your repository and use the original alphabet/writing system whenever possible. If it is necessary to transliterate metadata, use recognised standards (e.g. ISO)

Guidelines and discussion

UTF-8 is the dominant encoding for the World Wide Web (and internet technologies), accounting for 98.0% of all web pages, and up to 100% for many languages, as of 2023[9]. Virtually all countries and languages have 95% or more use of UTF-8 encodings on the web.[10]

Most repository software supports UTF-8 by default, for example, DSpace 7, but there are steps in the installation process where it is necessary to ensure that Tomcat uses UTF-8 as default and the like.[11]

4.1 Transliteration vs transcription

Transliteration is the conversion of text from one system of writing to another (e.g. from the Greek alphabet to the Latin alphabet) that relies on mapping graphemes from one writing system to those in another in a standardised way, so that readers can reconstruct the original spelling using standardised transliteration tables or software tools. Some countries have transliteration standards.

Transcription is the type of conversion where the text in the target language captures sound rather than spelling.

Transliteration is sometimes unavoidable. Huge amounts of transliterated or transcribed metadata can be found in bibliographic databases and library catalogues. In some research communities transliterating names and even titles is a common practice. Although support for UTF-8 is now common, these practices persist. If a repository already contains transliterated metadata or its designated community requires that metadata be transliterated, the following recommendations should be followed:

- Use recognised transliteration standards.
- If possible, choose one standard and declare it in the repository's FAQ / user manual / about pages.
- If this is not possible, declare all used standards in the FAQ / user manual / about pages.
- To ensure that readers can reconstruct the original spelling, provide links to relevant transliteration guidelines (e.g. Library of Congress) and/or tools[12] in FAQ / user manual / about pages.
- If author names are transliterated, identifiers such as ORCID should be used to connect different name variants.
- Use language codes for transliterated metadata (e.g. this resource recommends e.g. el-Latn to indicate text in Greek transliterated to Roman alphabet).

---

[9] "Usage Survey of Character Encodings broken down by Ranking". *w3techs.com*. Retrieved 2023-08-23.
[10] https://en.wikipedia.org/wiki/UTF-8#cite_note-W3TechsWebEncoding-10
[11] https://wiki.lyrasis.org/display/DSDOC7x/Installing+DSpace
[12] E.g. https://alittlehebrew.com/transliterate/, https://www.translitteration.com

If there are transliteration standards, transcription should be avoided because rules are not always clear, which makes it difficult to reconstruct the original spelling. If transcription is unavoidable, follow the rules and standards for your languages.

Example 15: DataCite
DataCite requires transliteration of non-Roman characters:
contributorName
Occurrence: 1
Definition: The full name of the contributor.
Allowed values, examples, other constraints: If Contributor is used, then contributorName is mandatory.
Examples: Patel, Emily; ABC Foundation
The personal name format may be: family, given. Non- roman names should be transliterated according to the ALA-LC schemas.

5. If the repository software supports multiple interface languages, set up the user interface in the native language(s) of the target group, along with an English option

Guidelines

A user interface in the multiple languages makes it easier for users from different communities to navigate through the repository. For example, an interface in the native language(s) makes it easier for local users to both understand the metadata fields when depositing content; at the same time, an interface in English makes it easier for international users to browse and search content.

Example 16: Dataverse
Dataverse supports multilingual user interfaces and relies on community translations done by volunteers. Major progress towards creating a directory of language packs was made within the Social Sciences and Humanities Open Cloud (SSHOC) project and the online tool Weblate was designed to facilitate new translations. A user guide for Weblate is also available[13].

Example 17: DSpace
DSpace provides support for multiple interface languages. The text displayed on the interface is called "messages" and the messages files (language packs) are contributed and managed by the community outside the core DSpace project to allow more regular updates and releases. Users can modify community translations or create their own and commit them to the dspace-api-lang project on Github. Apart from messages, it is possible to localise other elements, such as help pages, input forms and email templates. Instructions on how to enable the interface in multiple languages is available in the DSpace documentation. DSpace

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

7 makes a step forward towards facilitating UI translations: https://wiki.lyrasis.org/pages/viewpage.a: Guidelines for multilingual support on the frontend (UI) are available[14].

Example 18: EPrints
EPrints provides support for multiple interface languages, using language-specific folders of "phrases" and other files. By default, EPrints only comes packaged with English language phrases, but the community has shared many translations in the EPrints Bazaar and on EPrints Files. EPrints uses the two letter ISO language standard to specify sub-directories of phrases and other types of language specific directories, for example:

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

EPrints subject metadata is designed to accommodate multilingual labels, so subject labels can be displayed according to what language the user has set for the interface.

EPrints is designed to default to English phrases; if it has missing phrases for another declared interface language, it will use the English language phrases until the missing phrases are added. There is a technical wiki page about translations but it may be out of date as it has only been edited a few times in the last few years.

6. Write personal name(s) using the writing system used in the deposited document and provide a persistent identifier enabling unambiguous identification

Recommendation

Write personal name/s as displayed in the deposited document and provide a persistent identifier enabling unambiguous identification, such as ORCID.

Guidelines and discussion

There are two main approaches to handling personal names in repositories:

- using a unified preferred form, as defined in an authority file;
- capturing the names as they are rendered in the deposited document.

The former approach is typical of library catalogues, where the unified form is used as a catalogue heading. Depending on the country, names that are originally written in a non-Roman alphabet will be Romanised, or, conversely, transcribed/transliterated according to the rules used in a particular country. If a repository offers embedded metadata that can

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

be imported into reference managers and preformatted recommended citations, this approach may not be optimal because the format of the name in the repository will differ from that in the publication.

If names are captured as they are displayed on deposited publications, the name of the same person will appear in the repository in various formats. In this case, it is important to use persistent identifiers, such as ORCID, to ensure proper identification and connect various name versions.

Example 19: DSpace
While in previous version of DSpace a workaround was required to display various name versions in a user friendly way (e.g. by means of an additional in-house application), DSpace CRIS and DSpace 7 not only support bidirectional integration with ORCID, but also treat persons as entities (CRIS entities and configurable entities, respectively)[15].

Example 20: Exposing persistent identifiers via OAI-PMH
It is also important to ensure that persistent identifiers are exposed via OAI-PMH. PIDs in Dublin Core™ Working Group has developed recommendations to make it possible to expose persistent identifiers including ORCID, via OAI-PMH. Two solutions are proposed and both cover several use cases.

Option 1: Using an 'id' attribute with Dublin Core properties
Both PID and label are known
<dc:creator id="https://orcid.org/0000-0003-1541-5631">Walk, Paul</dc:creator>

Label is known, but PID is not
<dc:creator id="">Walk, Paul</dc:creator>

PID is known, but label is not
<dc:creator id="https://orcid.org/0000-0003-1541-5631"></dc:creator> or
<dc:creator id="https://orcid.org/0000-0003-1541-5631"/>
This option is not suitable if it is necessary to include more than one PID.

Option 2: Using nested properties for identifiers
PID and label are known
<dc:creator>
<dc:identifier>https://orcid.org/0000-0003-1541-5631</dc:identifier>
<foaf:name>Walk, Paul</foaf:name>

Label is known, but PID is not
<dc:creator>
<foaf:name>Walk, Paul</foaf:name>
</dc:creator>

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

```
or:
<dc:creator>Walk, Paul</dc:creator>

PID is known, but label is not
<dc:creator>
        <dc:identifier>https://orcid.org/0000-0003-1541-5631</dc:identifier>
</dc:creator>

In this option, it is possible to provide multiple PIDs for the same property
<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>
```

Example 21: JPCOAR metadata schema
JPCOAR metadata schema also includes an element for researcher identifier, jpcoar:nameIdentifier[16]. This element can be used repeatedly with different types of PIDs (KAKEN ID, ORCID, researcher ID, and others) and exposed for the Institutional Repositories Database via OAI-PMH.

7. Include keywords in many languages, use multilingual vocabularies and thesauri if possible

Guidelines and discussion

The inclusion of keywords in many languages increases the discoverability of repository content. In this context, it is important to distinguish between free-text keywords (or "tags") and controlled terms derived from a controlled multilingual vocabulary or thesaurus. In the former case, keywords in several languages are provided in the dc:subject field, making sure that the language is properly encoded.

Example 22: MiCISAN institutional repository
MiCISAN institutional repository always respects the language of the resource. If, for example, the resource is in Spanish, in the case of keywords, qualifiers are used to differentiate the metadata in different languages[17]:

```
dc.subject.keywordseng
institutional repository

dc.subject.keywordseng
interoperability
```

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

dc.subject.keywordsspa
metadatos

dc.subject.keywordsspa
repositorio institucional

dc.subject.keywordsspa
interoperabilidad

It is important to note that using free-text keywords does not ensure consistency, nor does it reveal hierarchical relations among terms. The problem can be mitigated by selecting manually the terms to be added as keywords from controlled vocabularies. However, an optimal solution involves the integration of multilingual controlled vocabularies in the repository.

7.1 Multilingual vocabularies and thesauri

The use of controlled vocabularies or thesauri[18] for bibliographic metadata ensures that the same concept is described consistently. Along with using controlled terms to indicate resource type, version, or usage rights, controlled vocabularies can be used to describe the subject content of the resource. In multilingual controlled vocabularies, each term ideally has only one equivalent in every language and the relations among terms are the same. In a digital environment, the vocabulary terms are assigned persistent identifiers that can easily be resolved.

However, the use of controlled vocabularies or thesauri involves some challenges

- In order to be integrated with repositories, controlled vocabularies must be expressed as machine-readable data.
- Forced equivalency: it is not always possible to find true equivalents in all languages, due to which the meaning of terms and relations between them in one language will not be accurately reflected in their counterparts in other languages.
- The process of assigning controlled terms may be time-consuming.
- Researchers are usually not familiar with the concept of controlled vocabularies. If librarians do not have the required expert knowledge, the terms may be too general and inaccurate.
- There are many disciplinary specific controlled vocabularies and it is not possible to apply all of them in multidisciplinary repositories. On the other hand, general vocabularies may not be able to describe the content accurately.
- Widely used controlled vocabularies (e.g. Library of Congress Subject Headings, or Getty vocabularies) are not equally inclusive to various cultural contexts and social groups.

Generally speaking, repository software platforms support the implementation of controlled vocabularies, although integration solutions are not always optimal.

[18] A registry of controlled vocabularies: https://bartoc.org/

Example 23: Dataverse
Dataverse is the open source data repository developed by IQSS of Harvard University. A strong Dataverse community is helping to improve the basic functionality and develop it further. DANS-KNAW delivered production ready (Docker/k8s) Dataverse repository for the European Open Science Cloud (EOSC) communities CESSDA, CLARIN and DARIAH. To address the heterogeneous and multilingual datasets integration challenges, DANS-KNAW introduced external controlled vocabularies support (CESSDA Metadata Model connected to Skosmos framework; support for CLARIN Component MetaData Infrastructure and the European Language Social Science Thesaurus (ELSST) hosted by CESSDA and ODISSEI in Skosmos - CESSDA has an updated version with more language properties).

Example 24: DSpace
DSpace offers three ways to integrate controlled vocabularies[19]:

- Value pairs in a controlled list form;
- XML file containing the terms (e.g. to support the integration of Dewey Decimal Classification or the Thesaurus of Greek terms in repositories)[20];
- SolR Authority (was used for the ORCID integration before DSpace 7[21].

The DSpace 7 Configurable entities, though not initially designed for this usage, could be another way to implement controlled vocabularies.

Example 25: TRIPLE
There have been a number of attempts to overcome the limitations of the existing controlled vocabularies. The project TRIPLE developed a new multilingual (nine languages) controlled vocabulary for Social Sciences and Humanities by building upon existing vocabularies.

Example 26: The vocabulary RVM Web
The vocabulary RVM Web, maintained by Université Laval and used by libraries across Canada, is an example of a controlled vocabulary seeking to eliminate cultural, historical, and colonial biases:

- It's bilingual – in English and French, but not for all terms;
- Initially (around 1970) it was built by translating Library of Congress Subject Headings (LCSH) and is now an independent product;
- English version uses MeSH, AAT (Getty Thesaurus), HOMOsaurus (newly used) and LCSH;
- Relations between the different thesaurus or vocabulary terms are established manually. It is not an automated process;

[19] https://wiki.lyrasis.org/display/DSDOC7x/Authority+Control+of+Metadata+Values
[20] The first integration of COAR Resources Type Vocabulary was using either value pairs or XML files : http://repositorium.sdum.uminho.pt/handle/1822/46066?mode=full
[21] https://wiki.lyrasis.org/display/DSDOC7x/ORCID+Authority

- Open version RVM FAST does not contain AAT MeSH and HOMOsaurus, only LCSH (there is a plan to make it compliant with Linked Open Data in order include it in DBpedia, in the short term); example;
- Included in WebDewey;
- Unique identifier for each term (not yet public right now);
- Challenges:
    - Synchronisation between the different products (LCSH, RAMEAU, AAT, etc.). This will hopefully be improved with the use of IDs;
    - How to push updates of the terms used in systems?

Example 27: Wikidata
Integration of Wikidata into repositories, already implemented in Europeana, may be a widely applicable solution for providing multilingual keywords. Wikidata relies on both crowdsourcing and the existing authority files and it already contains a large number of data items in various languages. The import of terms from various vocabularies is enabled via the tool Mix'n'match.

Wikidata as keywords
Wikidata is a free knowledge base with more than 100 million data items. It acts as central storage for a general structured data of concepts, including the concept labels/translations in many languages. As a result, the use of Wikidata concepts as a controlled vocabulary of keywords is particularly promising as it can provide more multilingual interoperability with a lower time investment.

For example, Depositar – a research data repository based on CKAN – reuses Wikidata as the source of keywords, see more details here. It should be noted that the concept labels of Wikidata would keep changing. So depositar only stores and exposes the identifier (e.g. "Q11030") itself. Then it inquires the MediaWiki API to get the latest multilingual labels of a Wikidata vocabulary. Iit would be better to store and expose both (1) the latest label and (2) the (old) label at the time of the assignment of a keyword.

WikiData concepts and other controlled vocabulary terms can be encoded using JATS[22] <kwd-group> and <kwd> tags, with the addition of ***vocab***, ***vocab-identifier*** and ***vocab-term-identifier*** attributes defined in the NISO Standards Tag Suite (STS):

- the name of the controlled vocabulary ("wikidata") in the ***vocab*** attribute[23]
- the vocabulary identifier ("https://www.wikidata.org/") in the ***vocab-identifier*** attribute[24]
- the identifier/URL of each keyword in the ***vocab-term-identifier*** (e.g. "Q11030") attribute[25]. For WikiData, this is the identifier of the concept, not the language-specific label of the concept.

There is more than one way to do it, the JATS standard bundles the keywords by language using the <kwd-group> tag. The following is an example of metadata tagging of the wikidata concepts of photography (Q11633) and journalism (Q11030) with the concept labels in English (photography, journalism) and Polish (fotografia, dziennikarstwo) using JATS xml:

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

There might be limitations for this in the current repository technologies.

Recommendation: adding all the attributes described in the example – ***vocab***, ***vocab-identifier*** and ***vocab-term-identifier***

---

[22] The Journal Article Tag Suite (JATS) is an XML format used to describe scientific literature published online. It is a technical standard developed by the National Information Standards Organization (NISO) and approved by the American National Standards Institute with the code Z39.96-2012. The NISO project was a continuation of the work done by NLM/NCBI, and popularised by the NLM's PubMed Central as a de facto standard for archiving and interchange of scientific open-access journals and its contents with XML. With the NISO standardisation the NLM initiative has gained a wider reach, and several other repositories, such as SciELO and Redalyc, adopted the XML formatting for scientific articles: https://en.wikipedia.org/wiki/Journal_Article_Tag_Suite
In JATS (Journal Article Tag Suite), any metadata field could be tagged with a language. In the DTD format of the JATS schema, the xml:lang attribute can be applied to almost any element, see: https://jats.nlm.nih.gov/articleauthoring/tag-library/1.2/attribute/xml-lang.html. Examples: PubMed Central translated titles https://www.ncbi.nlm.nih.gov/pmc/pmcdoc/tagging-guidelines/article/dobs.html#dob-at-transtitle.
Using the JATS schema, the language of keywords is recorded using the *xml:lang* attribute of the <kwd-group> tag (see: https://jats.nlm.nih.gov/articleauthoring/tag-library/1.2/element/kwd-group.html). JATS groups the keywords by language, with a series of <kwd> tags immediately under each language's <kwd-goup> tag.

[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

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

Recommendations for repository software/platforms developers

- Enable a real-time integration of Wikidata – e.g. when a user starts typing in the appropriate metadata field, relevant Wikidata terms appear in a drop-down list for the user to select.
- Enable automatic assignment of controlled terms based on the existing metadata.

Automatic indexing of content could make the process of assigning controlled terms more efficient. This approach, which has been tested in individual institutional repositories, is already used by aggregators. For example, Europeana performs automatic metadata enrichment relying on external vocabularies and datasets such as GeoNames and DBpedia and uses the semantic relations and translations offered by these vocabularies. BASE assigns computed Dewey Decimal Classification terms based on available metadata. The same approach is used in the multilingual discovery platform GoTriple, where content harvested from various sources is automatically annotated using controlled terms, due to which it is possible to search GoTriple in multiple languages.

Additional steps forward could include the assignment of controlled terms based on the full text of deposited documents and enabling an automated import of the controlled terms assigned by aggregators.

8. Recommendations for repository managers on translated content

Multilingualism and translation are inevitably intertwined and can complement one another. Translations and translated content should be recognised as valid contributions to the research ecosystem and, as such, supported and acknowledged as a valuable scholarly output and promote linguistic diversity in research culture. To do so, it is necessary to encourage and properly credit translation both as a practice and output. This can be in part achieved with the implementation of the following eight specific recommendations:

1. **Include a specific field for the role of translator(s) in deposit forms of online archives and repositories to accommodate translator crediting (e.g. use dc.contributor.translator)**

See some guidelines in Rivero, Monica, Robert Estep, and Lorena Gauthereau-Bryson, 'Digitization Practices for Translations: Lessons Learned from the Our Americas Archive Partnership Project', D-Lib Magazine, 17 (2011) doi:10.1045/september2011-rivero.

2. **Accommodate translator identification with other fields such as ORCID or other similar interoperable identifiers if possible; organisation or affiliation if any**

3. **Include specific (sub)field(s) for the document's translation status, the language(s) used for the translated content, and the language(s) of the source document, preferably by designating the languages in international standard language codes**

Example 28: Updated CRIS Guidelines v1.2
Updated CRIS Guidelines v1.2 already includes multilingualism and machine translations. See an example on page 13 of the CERIF tutorial and a CRIS example in updated OpenAIRE guidelines with an additional attribute 'trans=' is at the element:
release:
https://github.com/openaire/guidelines-cris-managers/releases/tag/v1.2.0 with values:

- h := human
- m := machine
- o := origin

Also see the CERIF XSD at search cfTrans_Type.

We encourage similar developments in other standards and platforms.

4. Allow users to point to other related records of the translated content by adding relation fields such as the dc.relation field

Labelling options in this relation field could include:

- "Is a translation of"
- "Is translated from" (This second option could be best used in case of partial translation, e.g. of a book chapter or section).

Example 29: Crossref
Crossref is handling language as an attribute using 2 letters codes that can be used in multiple elements and manage translations using specific relation attributes: isTranslationOf; hasTranslation[26]. But the issue is that a provider may not be using it when registering their content.

Schema reference:
Language exists in the common schema as an attribute :
https://data.crossref.org/schemas/common5.3.1.xsd
<xsd:attributeGroup name="language.atts">
<xsd:annotation>
<xsd:documentation>Language attributes are based on ISO 639</xsd:documentation>
</xsd:annotation>
<xsd:attribute name="language" use="optional">

Translation relation is also possible, see relation schema:
https://data.crossref.org/schemas/relations.xsd

[26] See documentation: 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/ and https://www.crossref.org/documentation/schema-library/markup-guide-metadata-segments/relationships/

```
<xsd:element name="intra_work_relation">
<xsd:complexType mixed="true">
<xsd:attribute name="relationship-type" use="required">
<xsd:annotation>
<xsd:documentation>Used to define relations between items that are essentially the same work but may differ in some way that impacts citation, for example a difference in format, language, or revision. Assigning different identifiers to exactly the same item available in one place or as copies in multiple places can be problematic and should be avoided. </xsd:documentation>
</xsd:annotation>
<xsd:simpleType>
<xsd:restriction base="xsd:string">
<!--      Crossref   -->
<xsd:enumeration value="isTranslationOf"/>
<!--  hasTranslation  -->
<xsd:enumeration value="hasTranslation"/>
<!--  isTranslationOf  -->
```

5. Accommodate this relation field with other fields of identification pointing at the original document

Use a DOI or other PID of the original document or a handle or a URL if there is no interoperable resolver.

Guidelines and examples

Export options of records of translated content should ideally include all of the above information, with specificities regarding the type of translation when necessary, for which further context detail is provided below.

An example of record for a human translated or post-edited content could read something akin to:

“This material titled ‘[translated title]’ is an integral/partial translation in [language name - standard language code] dated [DD-MM-YYYY] by [translator(‘s)s name(s) of “original title” by [author(‘s)s name(s) in [language name - standard language code] as published in [publication details]/retrieved from [DOI, other PID resolver or URL].”

Machine translation

In the course of our work, the topic of machine translation (MT) has sparked a heated discussion within the Task Group and we shared the nature of this discussion in a blog post: Is there a case of accepting machine translated scholarly content in repositories?

Given the complexity of the issues as well as the ethical implications, the Task Group has chosen to recommend that repositories not accept exclusively MTed content. This is also in line with the “Report of the ‘Translation and open science’ working group” (2020). Rather, MT should be perceived and used as an assistive technology and allowed to change dynamically in real time, transparently and unambiguously labeled as machine assistance, rather than curated and preserved as a primary resource in the repository.

The Task Group will monitor this rapidly evolving landscape, and continue to consider the issues and possibly publish further recommendations related to machine translation for scholarly texts, MT-assisted translation, as well as MT of abstracts and metadata in repositories.

However, during two exploratory studies carried out as part of the French Translations and Open Science project (mapping and collection of scientific bilingual corpora and MT evaluation in the context of scholarly communication studies), it was observed that researchers have been widely using MT to translate their own research and related metadata and upload multilingual content in repositories, even without notifying that they used MT. In these cases, MT can be – more or less accurately – post-edited, but without any degree of certainty as to quality on a broad scale for a repository owner or managing entity. Repositories may not be able to detect and review this material as its volume is growing rapidly. In addition, this is a practice that repositories can hardly control in general, keeping the costs and resources issues in mind. This is why it would be useful to put in place a notice system allowing researchers to provide information about the nature of the translation uploaded. Ideally, this notice system would distinguish human translated and post-edited content from raw MT.

This notice system could be also useful for repositories that have capacities to display an instant MT of the retrieved content (automatic or on-demand).

The notice, displayed as a warning for the user in order to raise awareness on potential errors and anticipate any claims, could read as follows:

“This document/This material is an unrevised machine translation dated [DD-MM-YYYY] of [citation of original] from [source language code] into [target language code] as published in [publication details]/ retrieved from [DOI, other PID resolver or URL] using [the name of the MT tool].This machine translation has not been reviewed or edited and is provided “as is” for the sole purpose of assisting users in understanding at least part of the subject matter of the original content expressed in [source language]. This provision does not imply a guarantee of correctness and accuracy of the said machine translation [in target language] by any natural or legal person in any part of this translation. [Consequently, the provision of this translation shall not give rise to any liability on the part of any person to any other person in the event that this translation is used for any purpose whatsoever.] Users of this machine translation are expressly invited to have it checked, revised or edited by a professional translator or relevant expert.”

6. Unless the document justifies it (e.g. parallel translation, commented translation, mirrored bilingual or multilingual versions), upload translations of documents as separate records

This is especially true for prefaces, introductions, or other contributions published in multilingual multi-contributor volumes.

7. **Promote the use of (re)translation-friendly licences to encourage translation of newly produced content and retranslation as well as promote translation crediting (e.g CC-BY)**

See more details about this in 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. **Make sure to provide sufficient information and recommendations for depositors in the form of FAQ or another form to implement the above**

Appendix 1. Use cases and challenges

Some of the use cases that are driving the recommended practices are as follows:

1. As a non-English institution, I am receiving in my repository documents in English that I need to describe.

When a new English document is submitted to the repository, it needs to be described with different metadata fields in different languages (e.g. abstracts, titles, keywords, document type) and using non-English controlled vocabularies.
Example: Hokkaido University uses JPCOAR metadata schema – Metadata in different languages is put in the same metadata field but distinguished by the language attribute, e.g. dc.description.abstract and dc.subject[27]– a language column on the right side of the page shows the ISO language code of the metadata. When journal articles are deposited, every metadata on the published version is included (no translation from the original; in Japanese language journals typically abstracts and keywords are written in English as well and full-text - in Japanese); abstracts are in metadata and the language attribute is embedded; authors names in the language of the article. At least, there is a scheme to mark metadata for multi-language; but there are concerns about discoverability and what is more suitable metadata.

2. As a repository manager, I often deal with articles, thesis or dissertations that are written in more than one language.

All thesis and dissertation are submitted in French but many contain articles inserted as chapters in the language they were written in.
Example: At ULiège, if a document is available in different languages, each language version is made available as a different record with metadata in different languages. Example of the same document in two different languages, for which two different records exist[28]. But there is only one language attribute for the record.

3. As an author, I would like to see my articles written in different languages in one record - for statistics and for reporting

All articles in different languages are deposited in one item and need to be described properly.
Example: At Open University of Catalonia there were two separate records for articles in different languages in the past. Now, by request from authors, translations are together in one record or even in the same file document, which simplifies citations tracking and increases visibility. But there might be issues for content aggregators and indexing services.

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

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

4. As a repository manager, I want to provide submission fields in different languages

[THIS MAY BE SPECIFIC TO DSPACE]. When configuring submission forms, the labels and help/instructions for each field can only be written in one language. Multilingualism can only be achieved by typing the label in each language in the same field (Author/Auteur).

5. As a repository manager, I want to have a collection name and description in more than one language

Currently only one language is allowed for a collection name and description.
Example: [THIS MAY BE SPECIFIC TO DSPACE]. It would be nice if introductory texts (HTML) etc. of communities/collections could be presented in multiple languages. This could quite easily be accomplished by using CSS and named divs. But unfortunately html attributes, such as id and style, seem to be removed in the html output - i.e. text is transformed to text in the UI. As collections and communities are items in DSpace (and thus have their own metadata), maybe a way to solve this problem would be to allow language selections at the metadata level, like it could be done already for objects metadata (i.e abstracts). A simple and quick workaround to the bilingually issue of collections/communities in DSpace is to use a delimiter, like the bar | , in between two texts describing these entities and their metadata fields as needed. All is required is to split the text at viewing time so that only the text in the currently active is displayed. Here you will see the Arabic version of the communities/collections list. When switching the language to the English interface, using the world icon on top, you will see them all appear in English. The same approach has been applied to the facets elements, where you now see controlled values like names of formats/ types, universities/ colleges/ departments, entities, etc. in multiple languages.

6. As a repository manager, I want to be able to manage labels in my language efficiently.

In open source multilingual softwares (OJS, DSpace, Eprints, etc.), the English labels are the mandatory ones when developing new features. Other languages' updates are often lagging behind and managed afterwards by the community or sometimes locally. Translations for new software functionalities is a challenge.
Examples: At ZORA (Zurich Open Repository and Archive) EPrints repository there is a German version of the interface. CSpace in China includes a metadata schema and interface in different languages, but repository managers still have challenges describing content in repositories. It's usually up to the users to select language tags and users are trained on how to deposit multilingual content. The interface languages of the repositories developed by the University of Belgrade Computer Centre (Serbia) include English and Serbian (in two alphabets: Cyrillic and

Latin)[29]. As the users were not satisfied with the available translations, the development team devised an in-house web application to facilitate translation. The application allows adding, removing and changing selected labels in individual or in all repositories. Changes are propagated to the repositories within 24 hours.

7. As a repository manager, I want to offer metadata translation in English - e.g. abstracts, titles and subjects

Some metadata need to be translated in English using machine translation tools
Examples: A Google translation API is used for translating abstracts, titles and subjects. This could also be achieved by recommending or requiring at least minimum metadata in English in user guidelines. In the Digital Archive of the Serbian Academy of Sciences and Arts, providing at least a brief description and keywords in English is recommended, as this improves content discoverability.

8. As a national repository, I need to deposit items in all languages of the country.

Content is available in local languages, but some of them don't have the language code, aren't in Unicode and there are no controlled vocabularies in those languages.
Example: In Nepal, only titles are added in Nepali language and the rest of metadata are in English, There is no consistency for keywords standardisation in Nepali language and no controlled vocabularies. Many local languages aren't in Unicode and sometimes romanised words are used – e.g. किताब kitaba (romanised) and a book (in translated form). This creates issues for Google Scholar indexing that would like to see metadata in the language of the article.

9. As a repository manager, I would like to expose the language of the metadata in OAI-PMH.

Currently there is no exposure for the language of the metadata in OAI-PMH.
Wish list: Repositories should consistently and consciously use metadata language tags to ensure that incorrect language information isn't exposed. And a language attribute should be exportable, including OAI-PMH. Another option could be a proactive approach by repositories – downloading, e.g. on the monthly basis, the extraction of metadata reference sheets and making them openly available to expose the language values.

10. As an aggregator and discovery system, I want to know what is the language of the full text document I am indexing, so I can assist users in finding content in their preferred language

[29] E.g. https://dais.sanu.ac.rs

There are issues with indexing contents at aggregator level (Solr, VuFind, etc.) because there is no way to separate the indexes by language and use language specific tools to enrich the search experiences.

Most regional repositories metadata does not have proper separation of multilingual information. Even mixed languages can be found on single textual metadata fields.

Keywords and descriptors are in multiple languages without the proper identification, hundreds of repositories are using different vocabularies even in the same language. Some ideas were discussed around the implementation of automatic classifiers to tag repository metadata with normalised vocabularies for the region.

Examples: LA Referencia is developing a language detecting tool (using different python libraries for natural language processing) to separate languages in metadata textual fields in order to improve metadata at aggregator level. The idea is to add proper xml:lang tags to every textual metadata field. This tagging would be used by the indexing process in order to generate separated indexes, still the problem of dealing with different languages in the search UI is complex to solve.

CORE seems to use a language detection tool. Distinguishing among Bosnian, Croatian, Montenegrin and Serbian is a challenge, as these languages are very similar. Due to this, language tags in CORE are usually incorrect when it comes to these languages. Using the common tag BCMS languages would be a solution to this problem.

11. As an aggregator, I would like to index content correctly and assist users in finding content in their languages.

OpenAIRE Institutional and thematic Repository Guidelines (for aggregating repository content) encourage the use of the xml:lang attribute to indicate the language of the metadata. OpenAIRE aggregator supports the xml language tag

Example: **<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 supports the xml language tag and the aggregator conducts metadata checks for language, e.g. in subjects, titles and abstracts/descriptions; no names though – ORCID is recommended for names – OpenAIRE I+T: Title, Description
OpenAIRE also allows multiple languages – a content resource has this language.

12. As a researcher, I want to know what research is out there in other languages. Could also be a use case for a patient, etc.

Translating abstracts and making them available, offering an option to search by keywords in many languages could be some of the solutions and deep learning tools started offering this, e.g. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
Examples: BASE – a multilingual search, (search term is included in Eurovoc Thesaurus or Agrovoc Thesaurus. Example search for climatology). Wikidata and Abstract Wikipedia providing information independent of language.

13. As a digital preservation librarian or archivist, I need to know how to include natural language information in technical and descriptive metadata so that digital archival documents can be effectively indexed for retrieval and access.

Documented best practices for the inclusion of natural language information using digital preservation metadata standards such as METS and PREMIS facilitate increased accessibility, inclusion and diversity of digital archives.
Examples: Language is required information for effective indexing for retrieval of text (word stemming, stop words), video and audio content (speech-to-text allows for retrieval/indexing, subtitling of audio and video for accessibility). Language metadata can be included using Dublin Core's <dc:language> tag as a part of the Internal Descriptive Metadata (mdWrap) of a METS file. Language metadata can be included as one of the <significantProperties> of semantic units in PREMIS. For text documents, language metadata can be included using textMD, most commonly as an extension schema used within the METS administrative metadata section. Language can also be included as a part of standalone textMD document within the PREMIS element <objectCharacteristicsExtension>.

14. As a user, when submitting or browsing content, I want to be able to use an interface in my own language.

Repository interface is available in different languages.
Examples: At Open University of Catalonia, the repository has three language interfaces for the repository end-user . Each language interface has metadata fields names in the same language - e.g. Autor in Catalan and Spanish, Author in English. In all institutional repositories developed by the University of Belgrade Computer Centre, the end-user interface is available in English and Serbian (both Cyrillic and Latin) However, the labels and help in the input form are available only in Serbian because it is not possible to align them with the interface language in DSpace.

15. As an English language institution I use a catalogue to describe content in my repository - in English and other languages

Content is entered in native language, but findability might be an issue.
Example: At Berkeley Law, a MARC based system is used for describing content. Since this is < 1-3% of the content there is no expectation that searching in non-English terms will return any results unless the user is looking for something specific. Subject terms in the repository aren't used, but this seems like an easy way to increase accessibility in other languages. The catalogue and repository are linked and search is available in many languages. The catalogers speak many languages and are capable of cataloguing in non-English languages, but still most cataloguing is done in English aimed at single language speakers.

16. As an institution that supports a lot of translations, I would like to credit translators when depositing translated items in the repository.

Translators could be credited using taxonomies, e.g. CREDIT taxonomy, which is only available in English now, and it would be good to have an official translation into other languages. Two 'unofficial' French translations exist'[30]. Translators are acknowledged in the institutional repository (e.g. as contributors with names and roles), but it's not a case for some other archives, e.g. preprint archives.
Example: ULiège repository has a translator metadata field, e.g. see here.

17. As a translator, I would like to know whether a translation exists

As a translator, I need to know whether a translation exists:

- **For a quotation embedded in a source document in the same language, but I need to check if there's a target (original or translated) language version of the quoted text (with a reference in the notes or bibliography of the source document), before deciding whether to translate the quotation myself or reuse the existing translated quotation in my translation;**
- **To use text about the same topic as the translation I'm assigned, I may need to build a corpus of similar documents in the source and target languages of my assignment to use them in concordancing software which allow to search text strings (words, terms, phrases) in one language and retrieve in two languages. I may seek through a desktop research a collection of documents with their translation in the target language and then process them in an aligning software to obtain aligned files for words/phrases.**
- **To build alignments, either**
  a) To feed into a CAT (computer aided translation) systems, or

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

b) To feed into the learning modules of MT (Machine Translation) systems.
Example: In all those cases, having documents being recorded with proper metadata designating the original/translation status and pointing to the matching counterpart(s), might help the above desktop searches if the metadata were interoperable with search engines, library catalogues, repositories and CRIS systems. This will also be relevant for journal editors, terminologists, text miners and language technologists. To facilitate their work we need interoperability and interconnections between different systems. Translate Science is building such a tool and that is why we need good language metadata from repositories.

Appendix 2. Declare the language of the resource at the item level: Implementation examples following metadata standards/guidelines

Datacite Schema 4.4	9 Language Usage: optional Occurrence: 0-1 (Not repeatable) Recommended encoding IETF BCP 47 or ISO 639-1 language codes
Dublin Core (DC)	Term Name: language Usage: optional Occurrence: repeatable Recommended practice is to use either a non-literal value representing a language from a controlled vocabulary such as ISO 639-2 or ISO 639-3, or a literal value consisting of an IETF Best Current Practice 47 [IETF-BCP47] language tag.
Electronic thesis and dissertation metadata standard (ETDMS)	dc.language Usage : Optional, Occurrence: 0-N (Repeatable) Language names themselves should be recorded using ISO 639-2 (or RFC 1766). If the language is not specified, it is assumed to be English (en).
Metadata Object Description Schema (MODS)	Top Level Element <language> Usage : Optional Occurrence: 0-N (Repeatable) This resource contains both English and French text: <language> <languageTerm type="code" authority="iso639-2b">eng<languageTerm> </language> <language> <languageTerm type="code" authority="iso639-2b">fre<languageTerm> </language> This resource contains text in Egyptian Arabic, which is coded as an individual language in ISO 639-3: <language> <languageTerm type="code" authority="rfc4646" >zh-Hans</languageTerm> </language> <language> <languageTerm type="code" authority="iso639-3" >arz</languageTerm> </language>
OpenAIRE Guidelines for Literature,	dc:language Usage: Mandatory if Applicable (MA) Occurrence: 0-N (Repeatable) Recommendation: take values from one of the following lists:

institutional, and thematic Repositories	• IETF BCP 47, the IANA Language Subtag Registry • ISO 639-x, where x can be 1,2 or 3. Best Practice: we use ISO 639-3 and by doing so we follow: http://www.sil.org/iso639-3/ If necessary, repeat this element to indicate multiple languages. If ISO 639-2 and 639-1 are sufficient for the contents of a repository they can be used alternatively. Since there is a unique mapping this can be done during an aggregation process.
Japan Consortium for Open Access Repository (JPCOAR)	dc:language Usage: R (Recommended) Occurrence: 0-N (Repeatable: expect mandatory term) Usage Instructions Enter the main languages that are used in the main text of the resource. Use the ISO 639-3 language codes. It is optional to use the ISO 639-3 macrolanguage. Notes Do not enter language names. Do not enter country names. Enter in order of language priority. Recommended Examples The main text of the resource is in English. <dc:language>eng</dc:language> The main text of the resource is in English and Japanese. <dc:language>eng</dc:language> <dc:language>jpn</dc:language> Unrecommended Examples ISO 639-1 is not recommended. <dc:language>ja</dc:language> Do not enter multiple languages in one element. <dc:language>engjpn</dc:language> Do not use capital letters and double-byte characters. <dc:language>JPN</dc:language> <dc:language>eng</dc:language> Do not enter language names. <dc:language>日本語</dc:language> Do not enter country names. <dc:language>US</dc:language> Do not enter language codes other than ISO 639. <dc:language>en_US</dc:language>

Appendix 3. Declare the language of the metadata (xml:lang attribute): Implementation examples following metadata standards/guidelines

Datacite Schema 4.4	xml:lang="EN", for example <xs:element name="**title**" maxOccurs="unbounded"> <xs:annotation> <xs:documentation>A name or title by which a resource is known.</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"/> Similarly, for 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)	Where the language of the value is indicated, it should be encoded using the 'xml:lang' attribute. For example: <dc:subject xml:lang="en">seafood</dc:subject> <dc:subject xml:lang="fr">fruits de mer</dc:subject>
Electronic thesis and dissertation metadata standard (ETDMS)	Language is a global qualifier that can be used in any element: https://ndltd.org/wp-content/uploads/2021/04/etd-ms-v1.1.html#qualifiers
Metadata Object Description Schema (MODS)	There are Language-Related Attributes https://www.loc.gov/standards/mods/userguide/attributes.html#list ISO 639-2/b
OpenAIRE institutional and thematic repository Guidelines	The use of the xml:lang attribute to indicate the language of the metadata. Example: **<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 supports the xml language tag and the aggregator conducts metadata checks for language - e.g. in subjects, titles and abstracts/descriptions; no names though - ORCID is recommended

	for names - OpenAIRE I+T: Title https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/latest/field_title.html#dci-title , Description https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/latest/field_description.html#attribute-lang-o OpenAIRE also allows multiple languages https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/v4.0.0/field_language.html - content resource has this language.
JPCOAR 1.0.2 https://schema.irdb.nii.ac.jp/ja/schema	xml:lang attribute can be used for each element In principle, use the two-digit language code of ISO 639-1 (e.g. Japanese: "ja"; English: "en"). For Japanese 'yomi', use "ja-Kana". Where 'yomi' is entered, you must enter its original information (i.e. in kanji) with the description that 'xml:lang is "ja"'. For Chinese, it is desirable to separately enter simplified Chinese as "zh-ch" and traditional Chinese as "zh-tw".
JPCOAR Metadata Schema 2.0 https://schema.irdb.nii.ac.jp/en/schema/2.0/14 https://schema.irdb.nii.ac.jp/en/schema/2.0/1	Change from 1.0.2 : additionally support "ja-Latn". The Excerpts from updated part: The language information for Katakana Yomi is xml:lang="ja-Kana" and for Romaji Yomi, xml:lang="ja-Latn". Where a Yomi is entered, the information in xml:lang="ja" must be entered separately from the Yomi.
Akdeniz, Esra, & Moilanen, Katja. (2023). CMM CESSDA Metadata Model (3.0). Zenodo. https://doi.org/10.5281/zenodo.7528240	1.1.3.1 Language of Study Title The language of the content of the element. M ISO 639-1 Occurrence 1 ddi:DDIInstance/s:StudyUnit/r:Citation/r:Title/r:String/@xml:lang Similarly for Language of Subtitle; Language of Alternative Title; Language of Versioning Reason; Language of Abstract; Language of Study Topic (descriptive); Language of Keyword (descriptive); Language of discipline (free text); Language of Type of Data Source (descriptive); Language of Mode of Data Collection (descriptive); Language of Data Access Conditions; Language of Metadata Access Conditions (Study); Language of Full Name of Organization; Language of Organization Name Abbreviation/Acronym; Language of Description of the organisation; Language of Dataset Version Description; Dataset Language; Language of Dataset File Description; Language of File Name; Language of Document Title; Language of Publication Title; Language of Name of the Journal/Serial - 75 metadata fields overall to indicate the language; There are also metadata fields to indicating translations, e.g. 1.1.3.2 Translation Status of Study Title Is the content of the element translated? R

	true, false Occurrence 0-1 ddi:DDIInstance/s:StudyUnit/r:Citation/r:Title/r:String/@isTranslated; and 28 metadata fields mentioning translation

Appendix 4. ISO 639-1, ISO 639-2 and ISO 639-3 implementation examples

ISO 639-1 and ISO 639-2
Language property in the Data Catalog Vocabulary (DCAT) - Version 2 (W3C Recommendation 04 February 2020):

Range:	Resources defined by the Library of Congress (ISO 639-1, ISO 639-2) SHOULD be used. If a ISO 639-1 (two-letter) code is defined for language, then its corresponding IRI SHOULD be used; if no ISO 639-1 code is defined, then IRI corresponding to the ISO 639-2 (three-letter) code SHOULD be used.
Usage note:	Repeat this property if the resource is available in multiple languages.

The same wording is included in the Data Catalog Vocabulary (DCAT) - Version 3 W3C Working Draft 10 May 2022.

Codes for the Representation of Names of Languages arranged alphabetically by alpha-3/ISO 639-2 Code: http://www.loc.gov/standards/iso639-2/php/code_list.php.

ISO 639-3
ISO 639-3 extends the ISO 639-2 alpha-3 codes with an aim to cover all known natural languages and works better for such languages as Cebuano, Montenegrin, Quechua (which has variations by region of the country) languages. For example, it's recommended in the ALICIA repository Guide (also a video guide, Peru).

Metadata recommendations for text material stored in Finnish publication repositories recommend the ISO 639-X standard for dc.language.iso. It is preferable to use the 3-character language codes of ISO 639-2 or ISO 639-3, as appropriate.

There are still some implementation issues for a three-letter code as not all repositories could support this now (software and XML language issues) and there might be similar issues with aggregators (for example, OpenAIRE follows https://www.w3.org/TR/xml/ and https://www.w3.org/TR/xml/#RFC1766).

More about language tags

A useful and more descriptive article on "Language tags in HTML and XML" (2014) from W3 with examples:

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

and a proposal to use

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

For many lesser-known languages spoken by minorities and also for historical stages of languages, language codes, the basis of language tags, are simply not available, see "The Shortcomings of Language Tags for Linked Data When Modeling Lesser-Known Languages" with recommendations to improve or develop ISO language codes.

Appendix 5: Fixing language code inconsistencies in DSpace repository records

If the target language uses unique characters, it may be possible to automatically set the value of the language metadata.

Here is a SQL example for DSpace to specify items using a target language and set language value to them under the assume that the target language is not represented by 2-byte characters:

```
update metadatavalue set text_lang='/*Insert here the ISO code of target language*/'
 where metadata_field_id in (/* Insert here each metadata_field_id numbers of
which metadata accept some string value */)
 and length(text_value)!=octet_length(text_value)
 and text_value ~ '^[/*Insert here all specific characters uniquely used in target
language† */].*'
 and (text_lang is null or text_lang != '');
```

† You can use a regular expression that covers all the characters of the language. To take some examples, for Japanese:[ぁ-んァ-ヿ亜-腕] and for Cyrillic Scripts:[a-тA-TЋ-Ӿћ-ҳ].

It's an overnight cron job to add 'en' to any metadata lacking a language code, see more in Creating a SQL query or function to change text_lang to 'en'.

The Atmire CSV Power Tools could be used for editing exported metadata (en and en_US, as well as brackets, and other languages issues).

Appendix 6: Fixing missing document language in EPrints repository records

REAL is a repository running EPrints, commissioned in 2008, which contains presently more than 220000 items in eight collections. The content is diverse, partly current research articles uploaded by researchers, partly material digitised by the parent institution, the Library and Information Centre of the Hungarian Academy of Sciences. The current REAL software version is 3.3.15.

The language field for documents - though present - was not, up till now, visible in the web document upload forms, nor in any views of an item, and thus depositors or librarians were unable to set it or check its content.

```
<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>
```

We have recently exposed the field, and found that its content was set by EPrints based on the language setting used in the browser at deposit - that is, the values contained are more or less random. To find out (and set) the correct values for hundreds of thousands of items, we produced a list of IDs for the items to check, downloaded metadata in DC format, extracted the title, and tried to guess the language of the document based on the language of the title.

Our script started with a hypothesis (the first hypothesis was that the language of the document is hungarian), the title words were fed to a spellchecker, and if more than half of the words were
recognised, we accepted the hypothesis as true. In the next run remaining items were checked against the “language is english” hypothesis, then further languages were tested.

The C-shell script excerpt below shows the test of the title against the “language is italian” hypothesis, using the spell checker 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)"`
```

Experience with this method shows that - with some filtering - the error rate could be reduced to 1-2%, which is much better then the present error of 40-50%. We have to note that there are complicated, multilingual or highly technical (e.g. mathematics) documents, which represent a challenge. We do not know how to label bilingual / multilingual documents.

Appendix 7: Text processing tools

Whenever possible, specify the language(s) of the document, of individual paragraphs and phrases while writing, in the text processing tool.

To specify the language of particular paragraphs and phrases in MS Word, OpenOffice, LibreOffice and similar tools, use appropriate language settings and keyboards while typing. To specify language(s) in an existing document, select the text and define the language using the Language tool in the toolbar or menu. To preserve this information after conversion to PDF, the document should be exported as a tagged PDF. However, depending on the PDF extension built in the text processor, this information may be lost during conversion to PDF.

W3C provides recommendations on how to specify the language for a passage or phrase with the Lang entry in PDF documents. However, in order to implement these recommendations in PDF files, commercial software, Adobe Acrobat is required.

Multilingual support is also provided for LaTeX. There are a number of packages enabling typesetting in different languages, e.g. babel or polyglossia, and this feature is also available in Overleaf, collaborative cloud-based LaTeX editor

However, the interoperability of various text editing tools and formats used remains an open issue. Clear standards and collaboration with software produces is necessary to ensure that text created in various software tools remains not only readable for humans and machines, but also that the various features and functionalities (e.g. encoding, tags, annotations) available in the original document remain available after conversion to other formats.