COAR Confederation of Open Access Repositories

Good practice advice on managing multilingual and non-English language content in repositories

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

Version for Community Consultation: June 1-30, 2023

Provide your feedback here

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. And after decades of policies that have directed researchers to publish in English, we are beginning to see a reversal of this trend. 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". In China, Europe, and other jurisdictions, policy makers are introducing new measures that encourage researchers to publish in local languages.

Multilingualism presents a particular challenge for the discovery of research outputs. Although 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. Yet, discovery systems such as Google Scholar and other scholarly indexes tend to provide access prevailingly to the content available in the language of the user. In addition, the language of a scholarly resource is often not labelled appropriately, meaning a large portion of non-English resources are excluded from search results. Furthermore, many scholarly communications infrastructures are sub-optimal in their support for a variety of languages since little attention was paid to this issue during their design process.

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. This document presents the results of the task force work focusing on identifying good practices for metadata, multilingual keywords, user interfaces, translations, formats, licenses, and indexing that will improve the visibility and discovery of repository content in a variety of languages along with implementation guidance for the repository community.

Implementing multilingual support requires a joint effort of repository managers and aggregators, researchers and software/tools developers to ensure smooth implementation.

We've identified three broad areas: enhancing discoverability of non-English content; curating multilingual content in a repository and promoting language diversity and supporting translations in the use cases analyzed. 17 use cases from the perspective of repository managers and users, authors and translators, aggregators and discovery systems are driving the recommended practices and are available in Appendix 1.

Recommendations

Recommendations for repository managers

- Declare the language of the resource at the item level
- Declare the language of the metadata (xml:lang attribute)
- Use two-letter language codes whenever they are available, and three-letter codes if necessary (ISO 639)
- Enable UTF-8 support in your repository and use the original alphabet / the writing system whenever possible
- Use Unicode UTF-8 for all metadata
- For metadata in different languages use repeatable fields and never mix metadata in different languages in the same metadata field
- Write personal name/s as displayed in the deposited document and provide a persistent identifier enabling unambiguous identification, such as ORCID
- If the repository software supports multiple interface languages, it is recommended to set up the user interface in the native language(s) of the target group, along with that in English
- Explore the possibilities of integrating Wikidata - free and collaborative knowledge base for multilingual keywords - in your repository
- Enable further recognition of translation and translated content as valid contributions to the research ecosystem to further support and acknowledge translation as a valuable scholarly output and promote linguistic diversity in research culture

Recommendations for repository software/platforms developers

- Expose the language of metadata via metadata exchange protocol, e.g. OAI-PMH, GraphQL API, etc.
- Provide support for multilingual keywords to increase the discoverability of multilingual repository content
- 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
- Improve support for ISO language codes, e.g. three-letter codes needed for some languages

Recommendations for content creators

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

Actions for repository managers: Correct labelling of languages

When the language of the resource is correctly attributed, it allows discovery and indexing services to properly process and parse the text. 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 which are language specific. Including the language tag enables information seekers, aggregators, and other discovery services to correctly identify the language of the full text and treat items accordingly.

Declare the language of the resource at the item level

Declaring the primary language of the document is considered mandatory. The language metadata must be encoded using the ISO-639 language code.

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 or parts of the documents contains more than one language, the language metadata should be repeated 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>
```

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

Declare the language of the metadata (xml:lang attribute)

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

Declare the language of the metadata even in English. It seems like an additional effort, but it's worth it, aggregators can't assume that.

The *xml:lang* attribute/subproperty is described at https://www.w3.org/TR/xml/#sec-lang-tag . The values of the attribute are language identifiers as defined by [IETF BCP 47], *Tags for the Identification of Languages*.

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>
```

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

Use language codes (ISO)

IANA recommends using two-letter codes whenever they are available, and three-letter codes if necessary (e.g. if no two letter code exists): http://www.iana.org/assignments/language-subtag-registry/language-subtag-registry and https://en.wikipedia.org/wiki/IETF_language_tag.

According to the "Internet Official Protocol Standards" IETF RFC 1766 : Tags for the Identification of Languages: “The language tag is composed of 1 or more parts: A primary language tag and a (possibly empty) series of subtags.

In the primary language tag all 2-letter tags are interpreted according to ISO standard 639, "Code for the representation of names of languages" [ISO 639].

The information in the subtag may for instance be:

- Country identification, such as en-US (this usage is described in ISO 639)
- Dialect or variant information, such as no-nynorsk or en-cockney
- Languages not listed in ISO 639 that are not variants of any listed language, which can be registered with the i-prefix, such as i-silbo
- Script variations, such as az-arabic and az-cyrillic"

Using language codes can also be practical for historical or non-official languages (e.g. Latin, Walloon, etc.). Examples in Walloon:
https://orbi.uliege.be/handle/2268/28421
https://orbi.uliege.be/handle/2268/28419

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

Repository software/platforms implementation examples

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).

DSpace

DSpace 7 allows using a language attribute for any metadata you want - a two-character ISO in the dropdown menu, but in the field you can write anything and there are different variations. E.g. you can add more than one language option for an item - for several fields, but you can't specify on which field you are referring to. For earlier versions of DSpace there was a need to find workarounds for this. See Appendix 5 on how to fix language code inconsistencies in repositories running on previous versions of DSpace.

However, there is no exposure of the language of metadata in OAI-PMH and it's a request to software developers.

Multiple repository interface languages

If the repository software supports multiple interface languages, it is recommended to set up the user interface in the native language(s) of the target group, along with that in English.

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 localize 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 7 makes a step forward towards facilitating UI translations: https://wiki.lyrasis.org/pages/viewpage.a:

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: https://doi.org/10.5281/zenodo.4807371.

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.

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.

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.

Samvera

TIND IR

https://www.tind.io/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 (https://www.loc.gov/marc/bibliographic/bd041.html). While the language codes used for cataloguing (https://www.loc.gov/marc/languages/language_name.html) do not conform to the recommendation of this group, the processes and details of repository entries are more flexible and should probably instead use the language code methods described in this recommendation. The use of subfields allows for granular declaration of item language, summary language, table of contents language, and more.

The XML might look something like the following:

```
<datafield tag="041" ind1="0" ind2=" ">
   <subfield code="a">it</subfield>
   <subfield code="a">en</subfield>
   <subfield code="a">fr</subfield>
</datafield>
```

WEKO 3

WEKO3 is a cloud-based repository system supported by JPCOAR (Japan Consortium for Open Access Repositories). It is developed based on INVENIO by CERN. In WEKO3, 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.

Recommendations on multilingual keywords

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. This approach does not ensure consistency, not 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.

Multilingual vocabularies and thesauri

The use of controlled vocabularies or thesauri[1] 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.

For example, DSpace offers three ways to integrate controlled vocabularies: https://wiki.lyrasis.org/display/DSDOC7x/Authority+Control+of+Metadata+Values

- 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)[2]
- SolR Authority (was used for the ORCID integration before DSpace 7: https://wiki.lyrasis.org/display/DSDOC7x/ORCID+Authority)

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

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.

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

[2] 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

The vocabulary RVM Web (https://rvmweb.bibl.ulaval.ca/), 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;
- It is not automated;
- Open version RVM FAST does not contain AAT MeSH and HOMOsaurus, only LCSH: https://rvmweb.bibl.ulaval.ca/rvmweb/recherche/init.do?repertoire=rvmfast (there is a plan to make it compliant with Linked Open Data in order include it in DBpedia, in the short term); example: https://rvm.bibl.ulaval.ca/rvmweb/lod/notice.do?noControle=RVMFAST-000315572&repertoire=RVMFAST
- Included in WebDewey;
- Unique identifier for each term (not yet public right now);
- Challenges:
    - Synchronization 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?

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.

WikiData concepts and other controlled vocabulary terms can be encoded using JATS[3] <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) https://www.niso-sts.org/ :

- the name of the controlled vocabulary ("wikidata") in the ***vocab*** attribute (https://www.niso-sts.org/TagLibrary/niso-sts-TL-1-2-html/attribute/vocab.html)
- the vocabulary identifier ("https://www.wikidata.org/") in the ***vocab-identifier*** attribute (https://www.niso-sts.org/TagLibrary/niso-sts-TL-1-2-html/attribute/vocab-identifier.html)
- the identifier/URL of each keyword in the ***vocab-term-identifier*** (e.g. "Q11030") attribute (https://www.niso-sts.org/TagLibrary/niso-sts-TL-1-2-html/attribute/vocab-term-identifier.html). 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***

Recommendations for repository software/platforms developers

[3] 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 popularized 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 standardization 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.

- 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.

Recommendations for repository managers on personal names

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

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 catalogs, where the unified form is used as a catalog heading. Depending on the country, names that are originally written in a non-Roman alphabet will be Romanized, or, conversely, transcribed/transliterated according to the rules used in a particular country. If a repository offers embedded metadata that can 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.

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) – e.g. https://scholars.lib.ntu.edu.tw/cris/rp/rp00095 (DSpace CRIS).

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>
```

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>
```

JPCOAR metadata schema also includes an element for researcher identifier, jpcoar:nameIdentifier (https://schema.irdb.nii.ac.jp/en/schema/3-1). It is exposed for

Institutional Repositories Database (IRDB:https://irdb.nii.ac.jp/) via OAI-PMH, however, different types of IDs (KAKEN ID, ORCID, researcher ID, and others) are not integrated.

Recommendations for repository managers on transliteration

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 recognized standards (e.g. ISO).

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 standardized way, so that readers can reconstruct the original spelling using standardized 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 catalogs. 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 recognized 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 (e.g. https://alittlehebrew.com/transliterate/, https://www.translitteration.com) 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 https://eidr.org/documents/Using_EIDR_Language_Codes.pdf)

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.

Recommendations for repository managers on translated content

Multilingualism and translation are inevitably intertwined and can complement one another. It is therefore important to enable further recognition of translation and translated content as valid contributions to the research ecosystem to further support and acknowledge translation 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 recommendations:

- 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)
- Accommodate translator identification with other fields such as:
    - ORCID or other similar interoperable identifiers if possible
    - organization or affiliation if any
- 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.
- Allow users to point to other related records of the translated content by adding relation fields such as the dc.relation field. Labeling 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).
- Accommodate this relation field with other fields of identification pointing at the original document with:
    - DOI or other PID of original document or a handle
    - URL if no interoperable resolver
- Export options of records of translated content should ideally include all of the above information, which could read like this:

  “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].”

- 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.
- 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 https://hal-lara.archives-ouvertes.fr/OUVRIR-LA-SCIENCE/hal-03640511

- Make sure to provide sufficient information and recommendations for depositors in the form of FAQ or another form to implement the above.

Recommendations on 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.

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, https://eprints.lib.hokudai.ac.jp/dspace/handle/2115/79104?mode=full&submit_simple=Show+full+item+record - 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: https://orbi.uliege.be/handle/2268/170862 and https://orbi.uliege.be/handle/2268/170863. 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.

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. <div id="swedish">text</div> is transformed to <div>text</div> 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 text 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: https://repo-nu.maktabat-online.com/community-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) https://www.zora.uzh.ch/ 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), e.g. https://dais.sanu.ac.rs/. As the users were not satisfied with the available translations, the development team devised an in-house web application to facilitate translation: https://trapist.rcub.bg.ac.rs/DESI/. 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 https://cloud.google.com/translate 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: https://repowiki.rcub.bg.ac.rs/index.php/DAIS_-_Digital_Archive_of_the_Serbian_Academy_of_Sciences_and_Arts:_Metadata.

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 standardization in Nepali language and no controlled vocabularies. Many local languages aren't in Unicode and sometimes romanized words are used - e.g. किताब kitaba (romanized) 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

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 normalized 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 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. Action: promote this to repositories

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 (https://arxiv.org/abs/1810.04805).

Examples: BASE - multilingual search https://www.base-search.net/Search/Advanced, (search term is included in Eurovoc Thesaurus or Agrovoc Thesaurus. Example search for climatology). Wikidata and Abstract Wikipedia providing information independent of language: https://meta.wikimedia.org/wiki/Abstract_Wikipedia.

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 (https://www.loc.gov/standards/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 https://openaccess.uoc.edu/. 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 catalog 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 catalog and repository are linked and search is available in many languages. The catalogers speak many languages and are capable of cataloging in non-English languages, but still most cataloging 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': 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.

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 https://orbi.uliege.be/handle/2268/290642.

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
  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 catalogs, 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	dc:language Usage: Mandatory if Applicable (MA)

Literature, institutional, and thematic Repositories	Occurrence: 0-N (Repeatable) Recommendation: take values from one of the following lists: ● 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 Draft https://schema.irdb.nii.ac.jp/ja/schema/2.0-draft/14 https://schema.irdb.nii.ac.jp/ja/schema/2.0-draft/1	Change from 1.0.2 : additionally support "ja-Latn". The Excerpts from updated part: For Japanese ‘yomi’with Katakana characters, use "ja-Kana" and For Japanese ‘yomi’ with Roman letters (alphabet), use "ja-Latn". Where ‘yomi’ is entered, you must enter its original information (i.e. in kanji) with the description that ‘xml:lang is "ja"’.
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 (freetext); 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 organization; 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

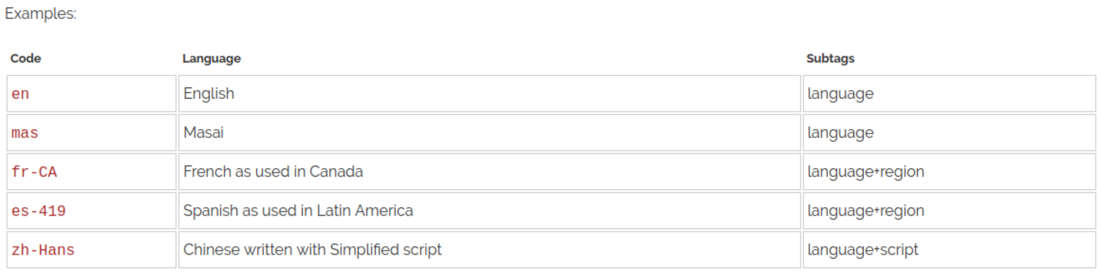
ISO 639-3 extends the ISO 639-2 alpha-3 codes with an aim to cover all known natural languages and works better for such languages as Cebuano, Montenegrin, Quechua (which has variations by region of the country) languages. For example, it's recommended in the ALICIA repository Guide (also a video guide,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

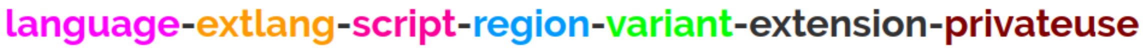
A useful and more descriptive article on “Language tags in HTML and XML” (2014) from W3 with examples:



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:[а-тА-ТЋ-Ӿћ-ӿ].
```

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 digitalised 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.