

(Draft) S-MDEG Multilingual Recommendations

Draft Release 1.0, October 4th 2024

Subgroup Members

Teresa Bascik
Jordan Pedersen
Caroline Sullivan
Mike Nason

Contents

[Summary](#)

[Recommendations](#)

[Recommended Practices for Stock DSpace 7](#)

[Recommendations that Require Further Development/Advocacy](#)

[Working Notes:](#)

[Creators and curators of metadata](#)

[Repository software/platforms developers](#)

Summary

Many Canadian institutions produce multilingual content. Some have policies for multilingual metadata where others might have policies for multilingual UI across institutional websites. Broadly speaking, multilingual metadata seems like an easier problem to solve than it actually is. Between platform capabilities and limitations of various metadata schemas, even those putting their best foot forward may find it difficult to embrace multilingualism in their repositories as fully as they may prefer.

Rather than reinvent the wheel, the S-MDEG Multilingualism subgroup reviewed the report released in October 2023 by the [COAR Task Force on Supporting Multilingualism and non-English Content in Repositories](#). The report, [Good Practice Advice for Managing Multilingual and non-English Language Content in Repositories](#), was intended for Creators and Curators of Metadata. S-MDEG worked from an assumption that moving toward these practices is advisable. We value the work of COAR because repositories

work best when they are interoperable, and the Task Force had already laid much of the foundational work for supporting multilingual content in this report.

When reviewing, we kept the following questions in mind:

- Can DSpace 7 meet a given recommendation out of the box?
- Are there foreseeable issues we may encounter in attempting to implement a given recommendation?

Recommendations

S-MDEG Recommendations are divided into two categories.

Recommended Practices for Stock DSpace 7 | COAR Multilingual recommendations that can and should be pursued with out-of-the-box DSpace7. While some configuration adjustments may be required by either repository managers or Scholars Portal, these recommendations have a lower barrier to entry.

Recommendations that Require Further Development/Advocacy | COAR Multilingual recommendations that we see value in but may currently be unachievable without advocacy or development resources to accommodate.

Recommended Practices for Stock DSpace 7

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

This is common practice for institutions using DSpace currently, where a user would input the item language in the submission form (ex. `dc.language.iso`). By default, DSpace submission form controlled vocabulary is configured for ~9 languages, including "other". From [DSpace Documentation](#):

language: If set to true a drop down menu will be shown, containing languages. The selected language will be used as language tag of the metadata field. A compulsory argument value-pairs-name must be given containing the name of the value pair that contains all the languages: e.g. `<language value-pairs-name="common_iso_languages">true</language>`.

S-MDEG believes it may be useful to expand this controlled vocabulary list to other languages identified by Scholaris client institutions. Specifically, institutions should

consider aligning with the language codes in the COAR Multilingual Task Force recommendation (detailed below), along with any additional languages that are active at a given institution.

There is, however, a small caveat with some *imported* metadata. DSpace can import metadata from a DOI leveraging the Crossref REST API. Language metadata is *not successfully retrieved* from the Crossref API. (The issue is currently open in GitHub: <https://github.com/DSpace/DSpace/issues/9046>) Additionally, metadata retrieved from open scholarly infrastructure sources such as Crossref, ORCID, Datacite, etc., may not include language data if not provided at the point of publication.

COAR Recommendation 3: Use standard (two-letter or three-letter) language codes (ISO 639)

DSpace recommends using the ISO 639 codes for interface development. In [DSpace 7 translation documentation](#), instructions are given that translations need to use the 2-letter ISO 639 code when creating translated content. For Scholaris users, the DSpace interface codebase should be sufficient (the default being en, with fr available). It would be beneficial to contribute to the French translation if issues arise; information on current volunteers are in the [DSpace translation documentation](#).

S-MDEG recommends adopting the summary decision tree provided in COAR recommendations whenever a decision about language of metadata is required (basically, if a 2-letter ISO 639-1 code is available and appropriate, use it, if not, use a 3-letter ISO 639-3 code if available, and determine if a sub tag is required and relevant). Broadly speaking, managing language metadata imported from external sources may introduce more challenges with consistency. However, we believe at minimum the following 2-letter codes will be most relevant for the language of metadata for Scholaris users:

- en = English
- fr = French
- ik = Inupiaq
- iu = Inuktitut
- cr = Cree
- oj = Ojibway

Indigenous languages are *somewhat* supported by ISO standards. More common languages or dialects are more likely to have language codes in at least *one* of the 639 subsets. For example, Michif is not present in the 2-letter codes; it appears in ISO 639-3,

and is likely to be relevant for some Scholaris users (crg = Michif). Generally, there is more granularity for First Nations and Inuit languages available in ISO 639-3 (ex. Cree could instead be crj = Southern East Cree, crk = Plains Cree, crl = Northern East Cree, crm = Moose Cree, csw = Swampy Cree, cwd = Woods Cree), and these 3 letter codes should be used if possible.

But, many languages and dialects of First Nations communities *aren't included* in the ISO standard. Mi'kmaq, Maliseet, Maliseet–Passamaquoddy, for example, are endangered languages throughout Maine, New Brunswick, and Nova Scotia. These aren't present in *any* ISO standard. Though ISO 639-3 is more in-depth regarding dialects, micro- and macro-languages, and other specifics, it remains *deficient* for describing many of the indigenous languages used by First Nations communities across North America.

If the whole document has important sections of the text in different languages, repeat the language metadata to mention each language, as per the COAR recommendations.

Additionally, it's worth noting that previous DSpace versions used `en_us` as the default. However, to align with the COAR recommendations, we suggest that any metadata stored as `en_us` be changed to `en` as part of pre- or mid-migration data cleaning.

COAR Recommendation 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).

S-MDEG feels that the UTF-8 Open Sans font that is the default with DSpace7 will accurately display non-Latin characters (based on testing by Scholars Portal).

- We support using the original script of the material in the metadata ([example from UOttawa](#)). However, this presents problems for [Recommendation 2](#).

Broadly, S-MDEG recommends that decisions about transliteration or translation be left to the individual institution. The labour to accommodate such work will be highly dependent on language skills, staff capacity, and language policy at each institution.

COAR Recommendation 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 that in English

DSpace 7 automatically supports a basic UI translation. Per [DSpace Multilingual Support Documentation](#), “DSpace will look at the user's browser language. If it has a language file in the user's language, it will render the interface in that language. If not, it will default to English or another default that you have configured.”

Locale translations for user interface are largely community dependent. Every software update that adds to – or modifies – a platform's user interface introduces potential conflicts in translations. S-MDEG would encourage Scholars Portal and Canada's DSpace user community *in general* to take shared responsibility in committing to the work of localizing DSpace for users globally. Many institutions in Quebec commit directly to translation efforts in GitHub. (eg. [dspace-api-lang repository for french locales in github](#))

Additionally, *some clients* may bring additional requirements to the table based on institutional policies. For example, the University of Ottawa requires multilingual issues across software to be tracked and have specific needs around URLs and file-paths. Tenable or not, these concerns should be evaluated.

Recommendations that Require Further Development/Advocacy

COAR Recommendation 2: Declare the language of the metadata

S-MDEG recognizes the value of correctly coded language metadata *without* post-submission cleanup. This plays an important role in search and discovery, especially when being harvested or if Scholaris pursues a federated search option. At minimum, title, subject, and abstract metadata should be the priority for declaring metadata language. These three elements have the largest footprint in downstream discovery. Institution metadata may also be of significant importance in locating materials downstream though, ideally, Research Organization Registry (ROR) implementation across DSpace would significantly alleviate this problem.

However, the S-MDEG multilingual working group has identified some challenges in implementing metadata *identifying the language of metadata* as a whole. In DSpace 7, a submission form can be hard-coded to identify the language of a specific metadata element (i.e. en instead of en_us) where users are not required to input the language of metadata themselves. But DSpace operates under an assumption that all deposited metadata matches the language of the form in general. For example, if one uploads a French abstract for an English thesis, you cannot assert the language of an individual metadata element from the deposit form.

It's possible to modify forms such that language is asserted element-by-element, but this raises a number of UI/UX concerns and limits form utility. It's also not possible to *require* such metadata (see GitHub issue: <https://github.com/DSpace/DSpace/issues/8590>), which presents potential for errors and/or incomplete data.

Because current workflows vary so widely across institutions, it is challenging to know how to best implement this recommendation. Some institutions have in-library staff support to review and improve submission metadata, others have minimal oversight and may even rely on external groups who don't have capacity to address metadata concerns (ex. an office of grad studies tasked to submit and approve a thesis collection). Some currently accept submissions with exclusively English or French metadata, and others require multilingual metadata (ex. UOttawa being fully bilingual, UdeM requiring some fields in the non-dominant language).

If no action is taken now, we might see some harm to discovery over time. There may be a need for extensive metadata clean-up projects. S-MDEG recommends that institutions develop plans that fit with their staffing capacity and language capacities locally. To provide a more sustainable improvement for all DSpace users, we also recommend that the Scholaris group support developing (or advocating for) a configuration option to make language a required attribute in the submission form.

If DSpace forms are improved to allow for element-specific language metadata (as per the open GitHub issue) for basic works metadata (title, keywords, abstract) S-MDEG would recommend leveraging this support. The Scholaris user community may need to work with Scholars Portal and the broader DSpace community to agree upon how it is implemented. For the time being, each institution should do their best to assert language metadata where possible.

COAR Recommendation 7: Include keywords in many languages, use multilingual vocabularies and thesauri if possible

The S-MDEG multilingual subgroup and others in Canada's repository community have spoken extensively about the desire to have wide-ranging, translated controlled vocabularies. However, even within the [COAR Recommendations](#), there is uncertainty and ambiguity on how this work can and should be done. Some examples are given for the DSpace context specifically:

DSpace offers three ways to integrate controlled vocabularies:

- 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](#));
 - SolR Authority (was used for the ORCID integration before DSpace 7).
- [DSpace 7 Configurable entities](#), though not initially designed for this usage, could be another way to implement controlled vocabularies.

As well as general recommendations for repository 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.

Wikidata or automatic assignment of controlled terms would likely be the most fruitful options due to the labour of maintenance required to update controlled vocabularies in the suggestions made specifically for DSpace. There are also significant challenges in localizing metadata in this space. We recognize that this would take significant development effort, but would likely improve the interoperability and usefulness of all metadata, not just multilingual content.

COAR Recommendation 6: Write personal name(s) using the writing system used in the deposited document and provide a persistent identifier enabling unambiguous identification, such as ORCID

In previous versions 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 allow for the management of virtual metadata through implementation of "entities" ([CRIS entities](#) and [configurable entities](#), respectively).

S-MDEG recommends recording names as they appear in submitted documents and enabling configurable entities for people to support the ORCID integration. It is worth noting, however, that the ORCID integration in DSpace 7/8 is under continued development and there are numerous open GitHub issues filed as part of an [ORCID Global Participation Fund](#) project ([26 open issues](#) at time of writing). Informally, we've heard that these are hopefully going to be addressed by the end of 2024.

COAR Recommendation 8: Recommendations for managing translated content

Broadly speaking, S-MDEG supports the recommendations of COAR's Multilingual Task Force. However, in cases where DSpace, Scholaris, or a Canadian-specific context may uniquely require adjustment or consideration, additional comments from S-MDEG are amended.

- A. 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`)

This is consistent with the S-MDEG Minimum Viable Metadata Recommendations.

- B. Accommodate translator identification with other fields such as ORCID or other similar interoperable identifiers if possible; organization or affiliation if any.

This would be great, but see [Recommendation 6](#) for more information on the challenges. Without enabling configurable entities, there is no easy, recommended path to associating ORCID values as child metadata or attributes against Dublin Core author elements.

- C. 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.

Dublin Core is not especially great for metadata extending beyond title- or item-level. Language codes affixed to specific elements are individualized. An item only has one language tag for the source language, `dc.language.iso`. There really isn't an obvious location for a DC record in DSpace to record languages of a source document. At best, the working group recommends using relational metadata and language codes for associated records.

- D. Allow users to point to other related records of the translated content by adding relation fields such as the `dc.relation` field.
- E. Accommodate this relation field with other fields of identification pointing at the original document.

- F. Unless the document justifies it (e.g. parallel translation, commented translation, mirrored bilingual or multilingual versions), upload translations of documents as separate records.
- G. 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).
- H. Make sure to provide sufficient information and recommendations for depositors in the form of FAQ or another form to implement the above.

This can be recommended to Scholaris client institutions on a case-by-case basis.