

リポジトリにおける 多言語・非英語コンテンツの 管理のための推奨事項 (解説)

北海道大学附属図書館/JPCOAR作業部会
(研究データ作業部会・コンテンツ流通促進作業部会)

三上 絢子

本日紹介する推奨事項の概要

<https://www.coar-repositories.org/news-updates/what-we-do/multilingual-and-non-english-content/>

COARの「リポジトリにおける多言語・非英語コンテンツタスクフォース」作成
2023年11月に公開された推奨事項

推奨事項の主な対象者

- メタデータの作成・管理を行う者
- リポジトリ管理者
- リポジトリソフトウェア作成者（例：Dspaceの開発コミュニティ）
- 基盤開発者（例：OpenAIREのようなアグリゲータ、メタデータスキーマの制定コミュニティ）



推奨事項の公開ページ

本日紹介する推奨事項の構成

要約版と詳細版が存在

- 要約版は8項目+6項目の全14項目
(Webページ)

メタデータの作成・管理を行う際の推奨事項が8項目、
リポジトリソフトウェア・基盤開発者への推奨事項が6項目

要約版ページ

- 詳細版は8章＋AppendixのPDF
(Zenodoで公開)

「メタデータの作成・管理を行う際の推奨事項」に沿った章立てだが、
「リポジトリソフトウェア・基盤開発者への推奨事項」6項目の内容も含む

詳細版ページ

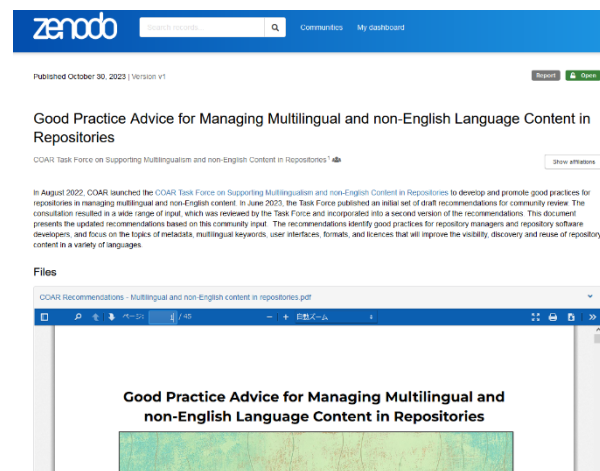
Summary of Recommendations

Creating and Curating Metadata

1. Declare the language of the resource at the item level
2. Declare the language of the metadata (e.g. xml:lang attribute)
3. Use standard (two-letter or three-letter) language codes (ISO 639)
4. 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)
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
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
7. Include keywords in many languages, use multilingual vocabularies and thesauri if possible
8. Recommendations for repository managers on translated content

Repository Software / Platform Developers

1. Ensure that language codes can consistently be used across the repository collections



メタデータの作成・管理を行う際の推奨事項 (要約版)

1. 資料の言語を表すメタデータをアイテム単位で記述する
2. メタデータの記述に使用した言語を、メタデータ中に付記する(記述例: xml:lang 属性の付与)
3. 言語表記には(ISO 639によって)標準化された(2文字または3文字の)コードを使用する
4. リポジトリにおいてUTF-8の使用をサポートし、可能な限り原文の表記体系の文字で記載する。翻字の必要がある場合には、周知された標準規格(例: ISO)に従う
5. リポジトリソフトウェアが複数言語のUIをサポートする場合には、英語と共に対象とする利用者の母語でも利用できるように設定する
6. 人名は資料中に記載されている表記方法を使って記載し、ORCIDのような、表記に曖昧さのない永続的な識別子を付与する
7. 多言語によるキーワードを記載する。可能であれば、多言語の語彙とシソーラスに対応する
8. 翻訳されたコンテンツを扱う際のリポジトリ管理者への勧告

リポジトリソフトウェア・基盤開発者への推奨事項 (要約版)

1. 収録対象のリポジトリ資料全体に対して、使用される言語コードの一貫性が保たれるようにする
2. メタデータ交換プロトコル(OAI-PMH、GraphQL APIなど)で交換される情報に、メタデータの記述に使われている言語を含める
3. ISO言語コードのサポートを改善する(例:三文字の言語コードが必要ないいくつかの言語への対応)
4. 永続的識別子がOAI-PMHを通じて公開されるようにする(PIDs in Dublin Core™ Working GroupはORCIDを含む永続的識別子をOAI-PMH経由で公開できるようにするための勧告を作成した)
5. 多言語のリポジトリ資料の発見性を高めるため、多言語のキーワードに関する支援機能を提供する。例えば、Wikidataとのリアルタイムでの連携(ユーザーがメタデータの入力を始めると、関連するWikidataの用語がドロップダウンリストに表示され、そこから選択できるようになる)など
6. 既存のメタデータに基づいた統制語の自動割り当てを可能にする

推奨事項(詳細版)の章立て

1. アイテムレベルで登録資料の言語を明示する
2. メタデータの記述言語を明示する
3. 標準化された言語コード(ISO)を使用する
4. UTF-8をサポートし、可能な限り原文の表記法を使用する
5. 英語に加えて母語のインターフェースを利用可能にする
6. 登録資料に沿った人名の表記とPIDの使用
7. 多言語のキーワード、語彙、シソーラスを使用する
8. 翻訳資料を扱う際の推奨事項

Appendix

メタデータの作成
管理を行う際の
推奨事項(要約版)

2.

1. 3.

リポジトリソフトウェア・
基盤開発者への推奨事項
(要約版)

4.

5. 6.

1. アイテムレベルで登録資料の言語を明示する

推奨事項

資料の主要な言語を明示することは必須とみなされる。
言語は、ISO 639の言語コードを用いて明示しなければならない(言語コードの詳細は後述)。

ガイドラインの内容

- 記述例(Dublin Core, MODS*)
- 複数の言語を含む資料についての記述例
EprintsやOpen Science Frameworkなどのフレームワークでの例

記述例(Dublin Core):

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

参考: JPCOARスキーマガイドラインの記述例

```
<dc:language>eng</dc:language>  
<dc:language>jpn</dc:language>
```

2. メタデータの記述言語を明示する

推奨事項

メタデータに入力された値の言語を示すにはxml:lang属性を使う。

1つの値に対してxml:lang属性が記述できるのは1回までであるため、異なる言語で書かれた値ごとに1:1対応で記述される

複数言語のタイトルに対する記述例:

<dc:title xml:lang="en">Open Access</dc:title>

<dc:title xml:lang="fr">Libre Accès</dc:title>

JPCOARスキーマでも
この記述例と同様に記述する

留意点

- アグリゲータによって、どのタイトルが本タイトルかを判別する方法は異なる
(dc:languageを元に判別してくれるアグリゲータもあれば、
記述の順番のみで判断するアグリゲータもある)
- 「メタデータの記述言語」の情報は現状、OAI-PMHの仕様自体には含まれない

3.標準化された言語コード(ISO)を使用する

言語表記の標準規格はBCP47(RFC 5646) で、ISO 639と下位タグの組み合わせによって定義される

ISO 639(言語コード):

2文字コード(639-1) 日本語:ja、英語:en、フランス語:fr、中国語:zn

3文字コード(639-2、-3) 日本語:jpn、英語:eng、フランス語:fre/fra、中国語:chi/zho

言語によっては3文字のコードのみを持つ(セブアノ語:ceb など)

下位タグ:

言語コード+ハイフンの後に続けて入力し、言語の地域やバリエーションなどを指定する。

地域の例:en-US(アメリカ英語)、zh-CN(中国語簡体字)、zh-TW(中国語繁体字)など

バリエーションの例:ja-Kana(日本語のヨミ)など

(3.続き)使用する言語コードを決定するための簡易フローチャート

まず、ISO 639中の言語コードを探し、該当するものがある場合は以下の優先順で使用する

1. 2文字のコード (ISO 639-1)
2. 3文字のコード (ISO 639-2, -3)
3. 下位タグ「x」を使用する (私的利用コード (Private Use) として RFC5649 に定義あり)

続けて、言語を特定するために下位タグが必要かつ関連性があるかどうかを判断する。

下位タグの使用例としては、

- 地域的なバリエーションや方言であることが文脈上重要である場合
ISO 3166の国コードを下位タグとして使用することを検討する (例えば、アメリカ英語には「en-US」)
- 言語を識別するために関連する文字体系のバリエーションがある場合
ISO 15924の文字コードを下位タグとして使用することを検討する
(例えば、ラテン文字のセルビア語には「sr-Latn」)

があげられる。

4. UTF-8をサポートし、 可能な限り原文の表記法を使用する

UTF-8は2023年時点で98%以上のWebサイトで使用されるエンコーディング
可能な限り、原文のアルファベット/表記方法を使用することを推奨する

翻字(Transliteration)が不可避であったり、翻字が一般的なコミュニティの場合は、
理解可能な翻字の規格に準拠した方法で行うことが望ましい。

翻字の例:

- ギリシャ文字で書かれたギリシャ語をアルファベットで表記しなす
- セルビア語(キリル文字-アルファベット)は併用されている

日本語では、ひらがな-ローマ字の対応付けが翻字にあたる
(この場合の翻字の規格は「日本式ローマ字」)

5. 英語に加えて母語のインターフェースを利用可能にする

対象となるユーザーの母語のサポート:

コンテンツのメタデータ要素を理解しやすくなる

英語のUIのサポート:

国外のユーザが資料を探しやすくなる

- 公用語が複数ある、またはある地域において公用語以外の言語が多く使用されている場合
例: カタルーニャ地域(スペイン)におけるカタルーニャ語など

6.登録資料に沿った人名の表記とPIDの使用

推奨事項

人名は資料中に記載されている表記方法を使って記載し、
加えて、ORCIDのように表記に曖昧さのないPID(永続的な識別子)を付与する

使用言語の違いによって、同一著者の資料が複数の名前(日本語とアルファベットなど)で
記載されるため、著者に紐づく永続的な識別子(ORCID等)の使用が重要になる。

7. 多言語のキーワード、語彙、シソーラスを使用する

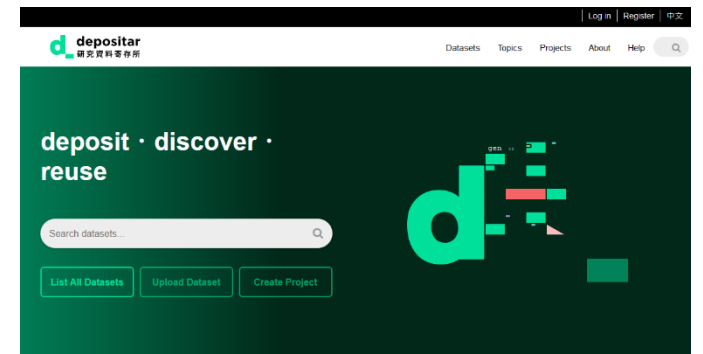
自由記述の場合は一貫性がなく、上下関係が明確でないことから、統制語彙を使用することが望ましい

ただし、統制語彙の使用においてもいくつかの課題がある

- 全ての言語に対して対応する語彙があるとは限らない
- 統制語彙の割り当てに時間がかかる場合がある
- 研究者は統制語彙の概念に慣れておらず、
図書館員では専門的な語彙を正確に割り当てられない可能性がある
- 専門分野に特有の語彙が多数あり、それらを網羅することは不可能である
- 一般的な語彙や広く使われている語彙では不十分な場合がある

多言語の統制語彙入力のため、
例えば以下のようなサポートを提供することが望ましい

- Wikidataを使用した統制語彙の付与
- 既存のメタデータを利用したレコメンド



Wikidataと連携した研究データ検索Webサービスの例
Depositar: <https://data.depositar.io/en/>

8. 翻訳資料を扱う際の勧告

1. 翻訳者のクレジットが記入できる項目を設けること
2. 翻訳者の識別子が記述可能なこと
3. 翻訳状態（機械/人的/オリジナル）、翻訳言語、原文の言語が記述可能な項目を有すること
4. 関連情報の項目を設け原本と翻訳資料が関連付けられること
5. 関連情報に、原文の情報やリンクを含めること
（特に、機械翻訳された資料は判別できるように権利情報の表記を行うこと）
6. 翻訳資料は特段の指定がない限り、原文とは別のレコードとして公開すること
7. 新たに刊行された資料の翻訳や再翻訳を促進するため、
適したライセンス（CC-BY等）の使用を促すこと
8. こうした取り組みを実施するため、FAQ等により投稿者に
十分な情報と推奨方針を確実に提供すること

Appendix

- 各研究機関のリポジトリでの取り組み例
- 各種メタデータスキーマでの記述例
- 言語コード (ISO 639) についての詳細情報
- DSpaceやEprintsでの既存レコードの修正例