リポジトリにおける
多言語・非英語コンテンツ管理のための
COAR グッドプラクティスアドバイス

オープンアクセスリポジトリ推進協会

コンテンツ流通促進作業部会

本資料「リポジトリにおける多言語・非英語コンテンツ管理のための COAR グッドプラクティスアドバイス」は、オープンアクセスリポジトリ連合（COAR）により英語(en)で作成された下記翻訳元資料" Good Practice Advice for Managing Multilingual and non-English Language Content in Repositories"を、JPCOAR コンテンツ流通促進作業部会が 2024/2/21 に日本語(ja)へ一部（メタデータ記述例、付録（Appendix）除く）翻訳したものです。

翻訳元資料：

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

DOI: 10.5281/zenodo.10053918

目次

序章

多言語使用は健全かつ包摂性のある多様な研究コミュニケーションの場において重要な特徴である。ローカル言語で出版することは、研究を助成する異なる国々の国民による研究へのアクセスを保証し、異なる言語を話す研究者にとって活動を行う場を平準化することを保証する。研究コミュニケーションにおける多言語使用のヘルシンキイニシアティブは、学術出版におけるローカル言語または母国語の排除は、社会からその地域で行われた研究を利用し、活用することを妨げる最も重要な要因であり、たびたび忘れられがちな要因であると主張している。共通言語としての英語の支配的な立場は、世界中にアイデアを流通させることには有用である一方、地域レベルでの研究成果の活用を妨げている。

何十年間にわたり研究者に英語での出版を指示する政策が行われてきたが、この傾向に転換が見られ始めている。ヨーロッパ、アジア、その他多くの地域において政策立案者が研究者にローカル言語や現地の言語での出版を奨励する新たな方策を導入している。例えば、ユネスコのオープンサイエンスに関する推奨勧告において、科学の実践、学術出版、学術コミュニケーションにおける多言語使用を参加国に推奨している。これは、世界人権宣言における研究において今や言語に基づく不公平な取り扱いを受けることがないようにすることといった言及や、ユネスコ推奨勧告における多言語使用およびサイバー空間への普遍的なアクセスを促進する言及といった最近の宣言に一致し、強化するものである。ユネスコによる宣言とはコミュニティに対して、言語の障壁を緩和することおよび全ての文化が自身を表現し、その地域の言語を含む全ての言語でサイバー空間にアクセスできるようにすることを保証するために必要な措置を講じることを呼びかけるものである。

多言語使用は、情報資源の発見に対して、ある種の挑戦であることを示している。学術情報資源の言語が適切に表記されていなかった場合、ディスカバリサービスに正確に索引付けされない。これは、索引付けが語幹解釈や見出し語解析（語形が変化した単語を集約することで一つの単語として分析可能となる）、ストップワードの適切な処理といったテキスト分析が含まれることが理由となる。これら全てのテキスト分析技法は、言語に大きく依存する。言語タグを含めることやその他の類似する対応により、情報検索者、アグリゲータ、索引作成者、ディスカバリサービスは本文言語を正確に特定し、適切な方法でアイテムを処理することができるようになる。さらに、研究者やその他の情報検索者は１つまたは２つの言語しか読むことができないかもしれないが、出版された言語に関わらず、自身の研究分野に関連する全ての研究について知りたいと考えている。情報資源の適切な言語属性はこのニーズを支援するために重要であり、より良い多言語使用の復権を提案するものである。

リポジトリにおける多言語および非英語コンテンツの管理に関するグッドプラクティス

を構築し促進するため、COAR は 2022 年 8 月にリポジトリにおける多言語使用および非英語コンテンツを支援する COAR タスクフォースを設置した。異なる利害関係者のコミュニティ（リポジトリ管理者、リポジトリユーザー、著者、翻訳者、アグリゲータ、ディスカバリシステム）から寄せられた 17 のユースケースに基づき、タスクフォースは 3 つの関連性のある領域（非英語コンテンツの発見可能性の強化、リポジトリへの多言語コンテンツの収集、翻訳支援）を特定した。これらのユースケースは付録 1 に記載している。

2023 年 6 月、タスクフォースはコミュニティのレビューを受け付けるための最初の推奨事項の草稿を発表した。協議の結果、様々な意見が寄せられ、タスクフォースによる検討を経て、推奨事項の第 2 版に盛り込まれた。この文書はコミュニティの意見に基づいて変更された推奨事項であることを示している。推奨事項はリポジトリ管理者、リポジトリソフトウェア開発者に向けたグッドプラクティスを明らかにするとともに、様々な言語のリポジトリコンテンツの視認性、発見可能性と再利用の向上につながるメタデータ、多言語キーワード、ユーザーインターフェース、形式、ライセンスに焦点を当てている。これらの推奨事項が世界中のリポジトリに広く採用されることを強く望んでいる。推奨事項の内いくつかはリポジトリ管理者により即時に採用できるものである一方、その他の推奨事項は採用に時間を要するものや完全に推奨事項を実装するためにはリポジトリ管理者、アグリゲータ、研究者、ソフトウェア開発者の共同的な努力が必要となるものもある。今後の数か月で COAR およびタスクフォースは推奨事項を広く普及させ、世界中のリポジトリでの実装が促進されるよう働きかけていく。

推奨事項（簡易版）

◆ メタデータの作成・管理を行う際の推奨事項：

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推奨事項

◆ リポジトリソフトウェア・基盤開発者への推奨事項：

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：3 文字の言語コードが必要な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）

1. アイテムレベルで資料の言語を明示する

◆ 推奨事項

資料の主要言語を明示することは必須であると考えられる。言語のメタデータは、ISO-639 の言語コードを用いて記載されなければならない（詳細は「3. 標準化された（2 文字または 3 文字の）言語コード（ISO 639）を使用する」を参照）。

◆ ガイドライン

資料中に一つの言語しか含まれない場合には、言語のメタデータは資料の主な言語を示す。主な言語の記載は、アイテム単位で行われなければならない。

論文集など、文書中に異なる言語により書かれた重要な章が含まれている場合は、言語メタデータを繰り返し記述して各言語について記載する。

メタデータの標準/ガイドラインに従った、より多くの実装例を付録 2 に記載する。

2. メタデータの記述言語を明記する（例：xml:lang 属性）

◆ 推奨事項

メタデータの記述に使われた言語を示すために xml:lang 属性を使用する。xml:lang 属性の繰り返し回数は[0, 1]のため、同じ要素について異なる言語で記述することができ、dc:language 要素よりも正確に言語を示すことができる。

◆ ガイドライン

主に英語が標準的に使用される言語であるとの仮定にかかわらず、資料の公開においては使用されている言語の情報を付記するべきである。

アグリゲータ（BASE、OpenAIRE など）のような他のステークホルダーはメタデータの内容から言語を推測することができないため、リポジトリ側において言語情報を付記することにも意義がある。

登録されたアイテムが、複数の言語によるタイトルまたは他のメタデータ要素を持っている場合（例えば、メインタイトルと要約や抄録のタイトル）、言語情報が xml:lang[1]属性を使用して示され、OAI-PMH のようなメタデータ交換プロトコルを介して適切に公開されるように留意する必要がある。アグリゲータによっては、全メタデータ要素や繰り返し入力されるフィールドをすべて取得することができない場合があるため、タイトルのメタデータ入力においては順序に気を付けること（すなわち、メインタイトルを最初に提供す

[1] https://www.w3.org/International/techniques/authoring-xml#natlang

ること）が推奨される。可能であれば、追加タイトルには dc.title alternative を使用する。

OpenAIRE[2]や BASE[3]などのアグリゲータは、入力時の順序に関係なく、文書の言語を示すフィールドに提供された情報に基づいてメインタイトルを正しく識別できる。しかし、OAI-PMH においてはメタデータの言語は出力されていないため、学術情報流通基盤のソフトウェア開発者に対しては将来のバージョンにおいてこの点について考慮することが望まれる。

3. 標準化された（2 文字または 3 文字の）言語コード（ISO 639）を使用する。

3.1 言語タグと言語コードについて

言語を一意に識別できる形で記載することは、研究内容の解釈、集約、再利用に不可欠である。

言語タグの標準規格は、1990 年代のインターネット初期から更新・拡張されてきた。最新の言語タグの標準規格は、IETF の BCP 47（RFC 5646）と ISO 639-3 の組み合わせで定義されている。

言語タグは、自然言語を識別する手段として、HTML、XML、RDF で要求されるものである。英語の'en'のような 2 文字または 3 文字の言語コードは言語タグの主な構成要素であり、ISO 639 標準規格（Part1-3）で定義されている。言語コードの後には、下記のような形で、言語の範囲を絞り込んだり、狭めたりする下位タグを付けることができる。

言語(language)-（言語に対応する）拡張(extlang)-文字体系(script)-地域(region)-異体(variant)-拡張(extension)-私的使用(privateuse)

言語へのタグ付けの試みは、多数にのぼる良く知られた言語に対して妥協無しに行われており、ISO639 には 7900 以上の言語のコードが含まれている（2023 年 1 月現在）。しかし、あまり知られていない言語や、地域的な変化や歴史的な言語は、ISO 639 では十分に表現されていない可能性があることに注意することが重要である。BCP47 に準拠したオプションのサブタグは、よりきめ細かい識別のための追加の選択肢を提供する。BCP47 で定義されている私的使用タグ "x "は、言語のバリエーション[4]を識別するために使うことが

[2] https://www.openaire.eu

[3] https://www.base-search.net/

[4] https://aclanthology.org/2020.lrec-1.408.pdf にて例が紹介されている

できる。

さらに、ISO639 は時代とともに変化してきた規格であり、変更のリクエストを送る機会も提供されている。

“どの時点においても、人間の言語に関する知識は決して完全でも完璧でもなく、常に拡大し続けている。ISO 639-3 の包括的な性質を考慮すると、特に少数言語や新しい言語を尊重する場合においては、コードセットへの変更は避けられない。”[5]

言語タグ付けの第一の目的が、使用されている言語と技術の文脈に応じて、使用されている言語を正確に識別し、表現することであると忘れないことが重要である。2 文字のコード（ISO 639 Part1）が特定の文脈で適切でない場合は、3 文字のコード（ISO 639 Part2 および 3）またはその他の下位タグ（文字体系、地域、私的使用など）を使用して、言語識別の相互運用性と精度を確保する必要がある。ISO 639 Part1 には、Part2 での対象となっている言語の一部が収録されていると考えられる。

また、Part1 の 2 文字コードに対する Part2 または Part3 の 3 文字コードも、同様に対象となる言語の範囲を拡張したものであるため、同義語とみなされる。例えば、"fra"、"fre"、"fr "という識別子は、同じ言語を表す。

BCP47 では、2 文字コードが存在する場合は常に 2 文字コードを使用することを推奨しているが、ISO639 では、可能な限り、同義語の中からの自由な選択を認めるべきであるとしている。本報告書では、BCP47 の勧告に従い、2 文字コードが存在する場合は常に 2 文字コードを使用することを推奨するが、特定の文脈においての使用においては、3 文字コードを使用することが適切な場合もある。

3.2 言語タグを選択するためのフローチャート

以下は、言語タグを決定する方法のフローチャートである。

1. ISO 639 で言語コードを検索
2. 言語に対応する 2 文字の ISO 639 Part1 のコードが見つかった場合は、それを使用する。5.に進む。
3. 言語に対応する 3 文字の ISO 639 Part2 または Part3 のコードが見つかった場合は、それを使用する。5.に進む。
4. 私的使用のために予約されている "x "サブタグを使用して、カスタム言語コードを定義

[5] https://iso639-3.sil.org/code_changes/introduction

する。5.に進む。

5. 言語を特定するために下位タグが必要かどうか、関連性があるかどうかを判断する。たとえば、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な場合は、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（たとえば、アメリカ英語の場合は「en-US」）。識別に関連する文字体系のバリエーションがある場合は、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：ラテン文字で書かれたセルビア語は「sr-Latn」）。

注：ISO 639 Part2 および Part3 では、いくつかの特殊な状況に対応した標準化がある。

- mis は "uncoded languages"（もともとは "miscellaneous "の略語）として定義されている。
- mul（"multiple languages "の略語）は、複数の言語が使用され、適切な言語コードをすべて指定することが現実的でない場合に適用される。
- und（"undetermined "の略語）は、言語を示す必要があるが、特定できない場合に使用される。
- zxx は、動物の鳴き声など、「言語的内容なし」としてコードリストに記載されている（2006 年 1 月 11 日に追加されたコード）。
- 言語コードの使用は、歴史的言語、地方固有の言語、古典的言語（ラテン語、ワロン語など）[6]にも有効である。
  ISO 639-1、ISO 639-2、ISO 639-3 および言語タグについては、付録 4 を参照のこと。

4. リポジトリで UTF-8 サポートを有効にし、可能な限り元のアルファベット/文字体系を使用する。メタデータの翻字が必要な場合は、認知された標準（ISO など）を使用する。

◆ ガイドラインと議論

UTF-8 はワールド・ワイド・ウェブ（およびインターネット技術）においてもっとも多くのシェアを持つエンコーディングであり、2023 年現在[7]、全ウェブページの 98.0%、多くの言語においては最大 100%を占めている。[8]

ほとんどのリポジトリソフトウェアはデフォルトで UTF-8 をサポートしているが、例えば DSpace 7 のように、Tomcat がデフォルトで UTF-8 を使用するようにインストールの際に確認する必要がある場合もある。[9]

---

[6] ワロン語での記述例：https://orbi.uliege.be/handle/2268/28421 、https://orbi.uliege.be/handle/2268/28419

[7] "Usage Survey of Character Encodings broken down by Ranking". *w3techs.com*. Retrieved 2023-08-23.

[8] https://en.wikipedia.org/wiki/UTF-8#cite_note-W3TechsWebEncoding-10

[9] https://wiki.lyrasis.org/display/DSDOC7x/Installing+DSpace

4.1 翻字と転写

翻字(transliteration)とは、ある文字体系から別の文字体系（たとえばギリシャ文字からラテン文字）へとテキストを置き換える方法のことである。翻字においては標準化された方法にて書記素を変換するため、翻字の対応表やソフトウェアを使うことで、読者は言語の綴りを再構成することができる。国によっては翻字の標準的な規格を設けているところもある（日本におけるローマ字のヘボン式など）。

転写(transcription)は、対象言語のテキストの綴りではなく、音を書き起こすことによる変換である（日本語におけるカタカナで表記される外来語も転写に当てはまる。）

場合によっては、翻字が避けられないことがある。書誌データベースや図書館目録には、膨大な量の翻字または転写されたメタデータが見られる。一部の研究コミュニティでは、人名やタイトルをも翻字するのが一般的な方法である。現在では UTF-8 のサポートが普及しているが、このような慣習が残っている。

リポジトリにすでに翻字されたメタデータが含まれている場合や、指定されたコミュニティがメタデータの翻字を要求している場合は、以下の推奨事項に従うべきである。

- 認知された翻字の標準を使用する。
- 可能であれば、リポジトリ全体で使用する翻字の標準を 1 つに選び、リポジトリの FAQ/利用ガイド/概要ページで宣言する。
- 可能な限り、リポジトリ内で使用されている翻字の標準を全てリポジトリの FAQ/ユーザーマニュアル/概要ページで宣言する。
- 読者が元の綴りを復元できるように、関連する翻字ガイドライン（例：米国議会図書館）や翻字ツール [10]へのリンクを FAQ／利用ガイド／概要ページで紹介する。
- 著者名が翻字されている場合は、ORCID のような識別子を使用して、異なる表記法による名前のバリエーション同士を関連付ける。
- 翻字されたメタデータには言語コードを使用する（例えば、ギリシャ文字からラテン文字への翻字に対して el-Latn の使用など）

もし翻字の標準がある場合には、規則が必ずしも明確でなく元の綴りを復元することが難しくなるため、転写は避けるべきである。やむを得ず転写を行う場合は、その言語の規則や基準に従うべきである。

[10] 例：https://alittlehebrew.com/transliterate/, https://www.translitteration.com

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

◆ ガイドライン

複数言語の UI があることで、異なるコミュニティのユーザー同士が、リポジトリを通したナビゲーションを容易に行えるようになる。

例えば、ローカルユーザーは、母語の UI があることでコンテンツの登録時にメタデータ要素を理解しやすくなり、同時に、英語の UI があることで、世界中のユーザーがコンテンツを閲覧・検索しやすくなる。

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

◆ 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 のような明白な個人の識別ができる PID（永続的な識別子）を付与する。

◆ ガイドラインと議論

リポジトリ上で人名を扱う方法は主に二つある。

- 典拠ファイルにおける定義のように、統一・推奨された形式を使用すること。
- 資料中で与えられた名前を取り込むこと。

一つ目の方法は、図書目録でよく見られ、統一された形式が目録の見出しに使用される。国によって、ローマ字以外のアルファベットで記載されていた名前がローマ字に変換されたり、反対に（ローマ字が）特定の国のルールに従って転写・字訳されたりする。もしリポジトリが、文献管理ソフトによるインポートを可能にする埋め込みメタデータや推奨される引用形式の例示を提供している場合、リポジトリ上の人名の形式が出版物上と異なる場合があるため、この方法は最適ではないかもしれない。

人名を資料中の表示のまま取り込んだ場合、同一人物の名前が、リポジトリ上で様々な形式で表示される。この場合、正しく個人識別を行い、様々なバージョンの名前を結びつけるために、ORCID のような PID（永続的識別子）を使用することが重要である。

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

◆ ガイドラインと議論

多言語でキーワードを記載することで、リポジトリ上のコンテンツの発見性が高まる。この文脈では、自由記述のキーワード（または「タグ」）と、統制語（統制された多言語の語彙やシソーラスから派生）とを区別することが重要である。前者の場合、複数の言語のキーワードが”dc:subject”要素に記述され、言語が適切にエンコードされていることが確認される。

自由記述のキーワードは一貫性がなく、各用語間の階層関係が不明瞭であることに注意が必要である。統制語彙からキーワードを手動で選択することで、この問題は軽減できるが、最適な解決方法は、リポジトリ上で多言語統制語彙と連携することである。

7.1 多言語語彙とシソーラス

書誌メタデータに、統制語彙やシソーラス[11]を使用することで、同じ概念を一貫して記述することができる。統制語は、資料の資源タイプ、バージョンやアクセス権を示すためだけでなく、主題を記述するためにも使用できる。多言語統制語彙において、各用語が、全ての言語で唯一の同義語を持つことが理想的である。各用語同士の関係性も同様である。デジタル環境において、語彙には簡単に処理できる永続的識別子が割り当てられる。

しかしながら、統制語彙やシソーラスの使用には幾つか課題がある。

- リポジトリと連携するには、統制語彙が機械可読なデータで表現されなければならない。
- 無理やりな同義性：ある言語における用語の意味や用語間の関係は、他の言語の対応する用語には正確に反映されない。そのため、全ての言語で正真正銘の同義語を見つけられるとは限らない。
- 統制語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馴染みがないことが多い。一方で、図書館員が必要な専門知識を持っていなければ、用語が一般的すぎて不正確になる可能性がある。
- 専門分野に特有の語彙が多数あり、全てを学際的なリポジトリに適用することは不可能である。一方で、一般的な語彙ではコンテンツを正確に記述できない。
- 広く使用される統制語彙（例：米国議会図書館件名標目表（LCSH）、ゲティ語彙）は、様々な文化的背景や社会集団に等しく浸透しているわけではない。

一般的にリポジトリのソフトウェアは、必ずしも最適化された組み込まれ方とは限らな

[11] 統制語彙のデータベース: https://bartoc.org/

いが、統制語彙の実装をサポートしている。

キーワードとしての Wikidata の使用

Wikidata は、1 億以上のデータ項目を持つ無料のナレッジベースである。概念の一般的な構造化データの中央ストレージとしての機能を果たしており、様々な言語の概念のラベルや概念の翻訳を含む。Wikidata の概念をキーワードの統制語彙として使用することは、より少ない時間投資でより多くの言語の相互運用性を提供できる、という意味で特に有望である。

例えば、 CKAN に基づいた研究データリポジトリである Depositar は、Wikidata をキーワードの情報源として再利用している。詳しくはこちらを参照いただきたい。Wikidata の概念のラベルは絶えず変遷することに注意が必要である。そのため、Depositar は、識別子（identifier）のみを保存・公開している（例：“Q11030”）。そして、Wikidata の語彙の最新の多言語ラベルを取得するためには、MediaWiki API が必要である。（1）最新のラベル（2）キーワードを割り当てた時点での（古い）ラベルの両方を保存・公開するとよりよいであろう。

Wikidata の概念や他の統制語彙は、JATS[12]の<kwd-group> タグや <kwd> タグを使い、NISO Standards Tag Suite (STS)で定義される、vocab 属性、vocab-identifier 属性、vocab-term-identifier 属性を加えることでエンコードできる。

---

[12] Journal Article Tag Suite (JATS)は、オンラインで出版される科学文献を記述するために使用される XML フォーマットである。このフォーマットは全米情報標準化機構（NISO）によって開発され、米国規格協会によって技術標準（Z39.96-2012）として承認されている。NISO のプロジェクトは、NLM/NCBI が行った研究を引き継いだもので、NLM の PubMed Central によって、科学的なオープンアクセスジャーナルとそのコンテンツのアーカイブを XML で交換するためのデファクトスタンダードとして普及した。NISO の標準化により、NLM のイニシアチブはより広範な範囲に及ぶようになり、SciELO や Redalyc など他のいくつかのリポジトリも科学論文のための XML フォーマット（https://en.wikipedia.org/wiki/Journal_Article_Tag_Suite）を採用した。JATS（ジャーナル・アーティクル・タグ・スイート）では、どのメタデータ・フィールドにも言語のタグを付けることができた。JATS スキームの DTD フォーマットでは、xml:lang 属性はほとんどすべての要素に適用できる（参照：https://jats.nlm.nih.gov/articleauthoring/tag-library/1.2/attribute/xml-lang.html）（例：PubMed Central の翻訳タイトル https://www.ncbi.nlm.nih.gov/pmc/pmcdoc/tagging-guidelines/article/dobs.html#dob-at-transtitle）。JATS スキーマでは、キーワードの言語は<kwd-group>タグの xml:lang 属性を用いて記述される（参照： https://jats.nlm.nih.gov/articleauthoring/tag-library/1.2/element/kwd-group.html）。JATS はキーワードを言語ごとにグループ化し、各言語の<kwd-goup>タグのすぐ下に一連の<kwd>タグを記述する。

- vocab 属性に統制語彙の名前（”wikidata”）[13]
- vocab-identifier 属性に語彙の識別子(“https://www.wikidata.org/”)[14]
- vocab-term-identifier 属性にそれぞれのキーワードの識別子/URL（例：“Q11030”）[15]

WikiData の場合、これは概念の言語固有のラベルではなく、識別子のことである。

方法は他にもある。JATS では<kwd-group>タグを使って、キーワードを言語ごとにまとめている。以下に示すのは、写真(Q11633)とジャーナリズム(Q11030)の Wikidata の概念に、英語(photography, journalism)及びポーランド語(fotografia, dziennikarstwo)の概念のラベルを付け、JATS xml を使ってメタデータのタグ付けを行った例である。

現在のリポジトリの技術では、上記を行うには限界があるかもしれない。

推奨事項：上記の例に記述した全ての属性（vocab 属性、vocab-identifier 属性、vocab-term-identifier 属性）を追加すること。

リポジトリソフトウェア/プラットフォーム開発者に向けた推奨事項

- Wikidata とのリアルタイムな連携の実現
  例：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。
- 既存のメタデータに基づいた統制語の自動割り当ての実現

コンテンツの自動索引により、統制語の割り当てのプロセスが、より効果的になる可能性がある。この個々の機関リポジトリでテストされてきたアプローチは、既にアリゲーターで使用されている。例えば Europeana は、GeoNames や DBpedia といった外部の語彙やデータセットに依存した自動的なメタデータの充実化を行っており、それらの語彙が提供する意味関係や翻訳を利用している。BASE は利用可能なメタデータに基づき、算出されたデューイ十進法の用語を割り当てている。多言語のディスカバリープラットフォームである GoTriple でも、同じアプローチが用いられている。GoTriple では、様々な情報源からハーベストされたコンテンツが、統制語を使って自動的にアノテーションされることで、多言語での検索が可能となっている。

将来に向けたさらなるステップとしては、資料に基づいた統制語の割り当てや、アグリゲータが割り当てた統制語の自動インポートの実現が考えられる。

---

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

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

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

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

多言語使用と翻訳は、必然的に関連し合い、互いに補完し合う。翻訳と翻訳されたコンテンツは、研究エコシステムへの確かな貢献として認識されるべきである。そしてだからこそ、価値ある学術成果として支援され、認められ、また研究文化における言語多様性を促進するべきである。そのためには、翻訳を実践と成果の両面から促進し、適切に評価することが必要である。こちらは次の 8 つの具体的な推奨事項を実行することで、ある程度達成できる。

1. 翻訳者のクレジットが記入できる項目を設けること(例：" dc.contributor.translator"を使用する。)以下のガイドラインを参照せよ。
   Rivero, Monica, Robert Estep, and Lorena Gauthereau-Bryson, 'Digitization Practices for Translations: Lessons Learned from the Our Americas Archive Partnership Project', D-Lib Magazine, 17 (2011) doi:10.1045/september2011-rivero.
2. 翻訳者の識別子が記述可能なこと（可能ならば ORCID か類似する相互運用可能な識別子、該当するものがあれば組織や所属も）
3. 翻訳状態（機械/人的/オリジナル）、翻訳言語、原文の言語が記述できる(サブ)項目を設けること。国際規格の言語コードで言語を明示することが望ましい。
   他の規格やプラットフォームでも、類似の開発を促進している。
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。
   この関連項目のラベルの選択肢としては、以下が挙げられる。
   - "Is a translation of"
   - "Is translated from"
     （こちらは、部分的な翻訳の場合に最適と思われる。例：図書の章や節）
5. 関連情報に、原文の情報やリンクを含めること。
   相互運用性のあるリゾルバがなければ、原文の DOI か他の PID(永続的識別子)、handle、URL を使用すること。

◆ ガイドラインと例

翻訳されたコンテンツの記録のエクスポートオプションには、上記の全ての情報を含むことが理想的であり、必要があれば、翻訳のタイプに関する特性も含むことが望ましい。こちらについては、以下に詳細を示す。

人間が翻訳または編集したコンテンツの記録の例としては、次のようなものがある。

「この[資料のタイトル]と題された資料は、 [出版物の詳細]に掲載された/[DOI、その他の PID リゾルバ、若しくは URL]から検索された、[(原文の)著者名]の[(原文の)言語名- 言語コード規格]の"[原文のタイトル]"を、[翻訳者名]が、[DD-MM-YYYY(年月日)]に、 [(翻訳物の)言語名- 言語コード規格]へ全て/一部翻訳したものである。

◆　機械翻訳

我々の活動の過程で、機械翻訳(MT)の話題は、タスクグループ内で白熱した議論を生んだ。この議論を次のブログで共有している。” Is there a case of accepting machine translated scholarly content in repositories?”

倫理的な意味合いだけでなく、問題の複雑さを考慮し、タスクグループは、リポジトリが機械翻訳されたコンテンツのみを引き受けないよう推奨することにした。これは”Report of the ‘Translation and open science’ working group” (2020)に沿ったものでもある。機械翻訳は、リポジトリの主要なリソースとしてキュレーション・保存されるよりも、むしろ、補助的な技術として認知・使用され、リアルタイムで絶えず変化することを許容され、分かりやすく明白に「機械による補助」とみなされるべきである。

タスクグループは、この急速に進化する情勢を追い、問題を審議し続ける。場合によっては、リポジトリ上の抄録・メタデータの機械翻訳と同様に、学術テキストの機械翻訳や、機械翻訳に補助された翻訳に関して、さらなる推奨事項を発表する可能性もある。

しかしながら、フランスの Translations and Open Science プロジェクト（学術コミュニケーション研究の文脈における、バイリンガルな科学文献と機械翻訳の評価のマッピングと収集）の一環として行われた 2 つの調査研究の間で、研究者は、公表していなくとも、自身の研究や関連メタデータを翻訳し、リポジトリに多言語のコンテンツをアップロードするために、幅広く機械翻訳を使用していることが分かった。

これらの場合において機械翻訳は、精度はともあれ後から編集できるが、リポジトリの所有者や管理者に対して、広範囲に渡って品質を保証できる仕組みはない。資料のボリュームが急速に増大しているため、リポジトリは、このような資料を検知し、評価することができない可能性がある。加えて、コストとリソースの問題を念頭に置くと、これは大抵リポジトリがコントロールできない業務である。

このため、研究者がアップロードされた翻訳の性質についての情報を提供できるような、掲示システムを導入することが有用であろう。この掲示システムで、人間が翻訳したコンテンツや編集後のコンテンツと、編集前の機械翻訳とを区別できることが理想的である。

この掲示システムは、（自動若しくはオンデマンドで）取得したコンテンツに対して即時に機械翻訳を表示できるリポジトリにとってもなお有用であろう。

この掲示は、潜在的な誤りへの認識を高め、あらゆるクレームに先手を打つためにユーザーに向けた警告として、次の例のように表示される：

“この文書/資料は、[出版の詳細]に掲載された/[DOI、その他の PID リゾルバ、若しくは URL]から検索された[原文の引用]の、[元の言語]から[翻訳言語]への、[DD-MM-YYYY(年月日)]付の、[翻訳機械のツール名]を使用した未修正の機械翻訳である。

この機械翻訳は校閲や編集をされることなく、ユーザーが、[元の言語]で表された原文の主題を最低限理解することのみを目的として、「そのまま」提供される。本但し書きは、この翻訳のいかなる部分についても、またいかなる自然人または法人による[翻訳言語]への機械翻訳についても、その正確性や精度を保証するものではない。[従って、いかなる目的で使用される場合であっても、この翻訳の提供により、何人に対しても責任は生じない。] 本機械翻訳のユーザーにおいては、プロの翻訳家または関連する専門家による確認、修正、編集を受けることを強く推奨する。”

6. 翻訳物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。（例：並行翻訳、コメント付き翻訳、バイリンガル版またはマルチリンガル版）
これは特に、多言語・多著者の出版物に収録された緒言や序文、その他の寄稿に当てはまる。

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 など）の使用を促すこと。
詳細については、以下を参照せよ：Susanna Fiorini, Franck Barbin, Martine Garnier-Rizet, Katell Hernandez Morin, Franziska Humphreys, et al.. Rapport du groupe de travail "Traductions et science ouverte". [Rapport Technique] Comité pour la science ouverte. 2020, 44 p. ⟨hal-03640511⟩

8. 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事項を確実に提供すること。