Vol.2020-IOT-51 No.9
Vol.2020-SPT-39 No.9
2020/9/4

マテリアル・インフォマティクスの実現に向けた材料データリポジトリの開発

田辺浩介[1]　松田朝彦[1]

概要： 物質・材料研究機構では、マテリアルズ・インフォマティクスの実現に向けた研究データ公開基盤として、材料データリポジトリ"Materials Data Repository"を開発した。Materials Data Repository は、既存の機関リポジトリの扱う文献情報に加えて、材料科学特有の研究データの保存と公開を行うためのアプリケーションである。本発表では、Materials Data Repository の備える機能、開発過程におけるメタデータ項目やデータ公開のワークフローの検討内容、ならびに本格的なデータ駆動型材料研究に向けた Materials Data Repository の将来像について述べる。

キーワード：データリポジトリ、マテリアルズ・インフォマティクス、オープンサイエンス

A development of Materials Data Repository for materials informatics

KOSUKE TANABE[†1]　ASAHIKO MATSUDA[†1]

Abstract: At the National Institute for Materials Science, we developed the "Materials Data Repository" as a platform for publishing research data to conduct materials informatics. The Materials Data Repository is an application to store and publish not only publications (as in traditional institutional repositories) but also research datasets specific to materials science. In this presentation, we will cover the functionalities of our repository, considerations regarding the metadata model and data publishing workflow, and a future vision of the Materials Data Repository for a more advanced data-driven materials research.

Keywords: Data repository, Materials informatics, Open science

1. はじめに

Materials Data Repository(MDR)[1]は、物質・材料研究機構（NIMS、以下「機構」）で 2020 年 6 月 15 日から運用を開始したデータリポジトリである。NIMS では、材料科学と情報学を融合させた材料研究・開発の手法である「マテリアルズ・インフォマティクス」、また材料科学におけるオープンサイエンスの実現に向けて、材料研究のデータ収集・解析・公開を行うための情報基盤「材料データプラットフォーム DICE」[2]の構築を行っている。

DICE 全体のアーキテクチャを図 1 に示す。DICE は、材料データベースやテキストデータマイニング、IoT ファイル転送システムなど複数のアプリケーションによって構成されており、MDR は DICE の中で、他のアプリケーションによって生成された研究データの共有と外部公開を行うためのアプリケーションとして位置づけられている。

2. MDR の開発

2.1 DICE のメタデータ

DICE は、材料科学に関するデータを「つくる」「ためる」「つかう」「公開する」ことをテーマとして構築を行っている[3]。このため、DICE を構成する各アプリケーションで

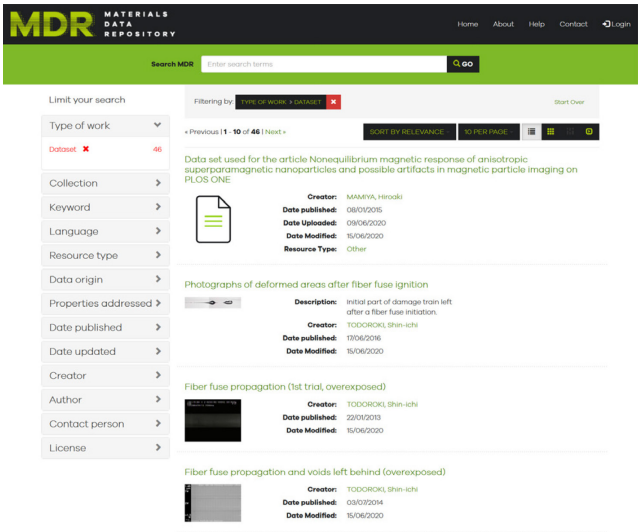

図 1　MDR のデータ検索結果表示画面

は、計測や実験によって得られた実際の研究データ（実データ）に対して、実験に用いた試料や計測に用いた装置の情報など、材料科学に特化したメタデータを付与することができ、かつそれらを解釈できることが要件として求められている。このため、DICE では材料メタデータを記述する

1 物質・材料研究機構 統合型材料開発・情報基盤部門
Research and Services Division of Materials Data and Integrated System, National Institute for Materials Science

Vol.2020-IOT-51 No.9
Vol.2020-SPT-39 No.9
2020/9/4

ためのスキーマ「共通メッセージ形式」を新たに設計し、MDR を含めた各アプリケーションでそのスキーマに従ったデータを送受信できるようにすることとしている。

材料メタデータの構造を図 2 に示す。材料メタデータは、材料科学共通の「共通メタデータ」と、観点別の「分野別メタデータ」によって構成されている。材料科学には物理的な試料とその計測、計算科学的アプローチをはじめ、データの性質により異なる観点による記述が必要なため、分野全体を横断するメタデータに加えて、観点ごとに細分化されたメタデータを設定できることを目的として設計されている。分野別メタデータには、「計測」「試料」「材料特性」「合成・プロセス」「計算」の各項目が定義されている。例として、MDR に実装されている共通メタデータ項目を表 1 に、試料メタデータ項目を表 2 に示す。

DICE で生成される材料データには、各システムでこれらの材料メタデータが付与され、データ公開時に実データとともに MDR に送付されることになる。

2.2 MDR のシステム実装

MDR は、DICE の各システムで生成された材料データを登録・保存し、検索可能な状態にして公開を行うデータリポジトリである。MDR への材料データ登録のユースケースには、他のシステムから機械的に送付されてくる場合に加えて、直接人手で Web ブラウザを用いて登録を行う場合も要求された。このため、MDR の開発にあたっては、メタデータ項目を柔軟に定義できるソフトウェアを選定する必要があった。

検討の結果、DICE では MDR のベースとするデータリポジトリソフトウェアとして、Hyrax[4] 2.6.0 を採用した。Hyrax は Samvera Community によって開発が行われているオープンソース・ソフトウェアである。Ruby on Rails によって記述され、機関リポジトリ用ミドルウェアの Fedora や検索エンジンの Solr と連携して動作する。Hyrax は主に北米圏の大学の機関リポジトリにおいて広く用いられているソフトウェアであり、データリポジトリとしてもデューク大学やミシガン大学などで採用例がある[5]。

Hyrax の特徴的な機能として、メタデータのスキーマ定義を RDF の書式で行い、またその定義をもとにメタデータ入力用のフォームを自動的に生成することが挙げられる。Fedora は RDF ストアとして動作するようになっており、Hyrax で作成されたメタデータは RDF として Fedora に保存されるようになっている。メタデータのスキーマは Hyrax 側でのみ行えばよく、データベース側(Fedora)であらかじめスキーマを定義しておく必要はない。

これらの利用実績や機能的な特長を評価し、MDR の構築用ソフトウェアとして Hyrax を採用することとした。

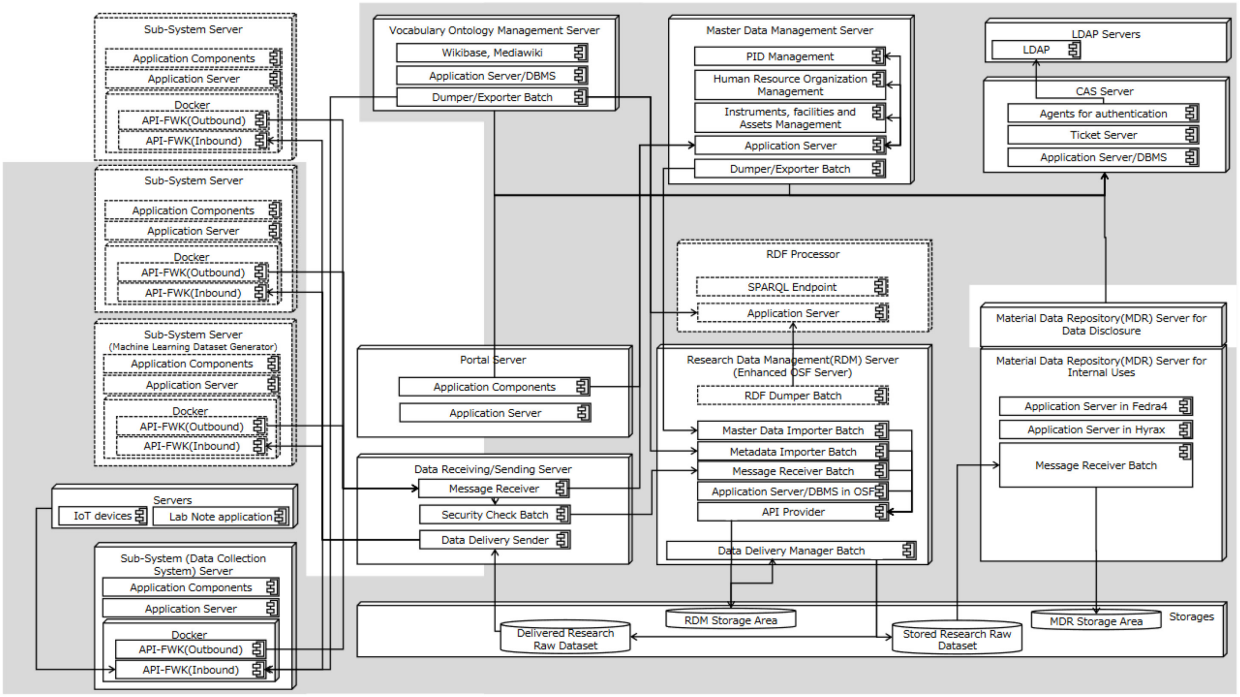

図 2　DICE 全体のアーキテクチャ[6]

Vol.2020-IOT-51 No.9
Vol.2020-SPT-39 No.9
2020/9/4

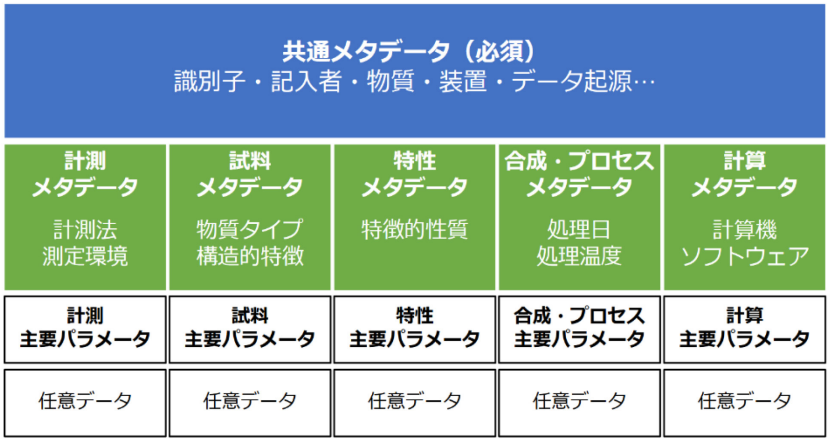

図 3　DICE 共通メッセージ形式のメタデータ構造[7]

Field	Required	想定用途	記入方式
First published at URL	System	公知であることのエビデンス申告	登録者記入
Supervisor approval	System	上長承認済申告	登録者記入
Title	System	題名	登録者記入
Alternative title	No	別の題名（日本語等）	登録者記入
Data origin	System	データの性質	登録者選択
Abstract or Summary	System	アブストラクト	登録者記入
Keyword	System	キーワード	登録者記入
Specimen	Policy	試料・対象物質	登録者記入
Surname	Policy	関係者姓（英字）	登録者記入
Given name	Policy	関係者名（英字）	登録者記入
Name	Policy	関係者 姓, 名 （英字）	登録者記入
Role	Policy	関係者が本Workにどう携わったか	登録者選択
Orcid	Policy	ORCID	登録者記入
Organization	Policy	著者所属組織	登録者記入
Sub organization	No	著者所属組織部門	登録者記入
choose type	No	著者に関する別IDがあれば	登録者選択
Identifier	No	著者に関する別IDがあれば	登録者記入
choose type	No	本Workに関する別IDがあれば	登録者選択
Identifier	No	本Workに関する別IDがあれば	登録者記入
choose type	Policy	日付の種類	登録者選択
Date	Policy	日付（作成日等）	登録者記入
Rights	Policy	ライセンス	登録者選択
Date	No	当該ライセンスのもと公開した日	登録者記入
Version	No	バージョン	登録者記入
Date	No	当該バージョンリリース日	登録者記入
Title	No	その他関係するリンク先のタイトル	登録者記入
Url	No	その他関係するリンク先のURL	登録者記入
Relationship	No	上記リンク先が本Workにどう関係するか	登録者選択
Label	No	カスタムメタデータの項目名	登録者記入
Description	No	カスタムメタデータの値	登録者記入
(NIMS DOI)	No	フォームには表示されず、運用側で付与	運用者記入

表 1　MDR での書誌メタデータ項目

Field	Required	想定用途	記入方式
Title	No	試料／対象物質（以下「試料」）の名前	登録者記入
choose type	No	試料の化学組成に関するIDの種類	登録者選択
Identifier	No	試料の化学組成に関するID	登録者記入
Description	No	試料の化学組成の説明・記述	登録者記入
choose type	No	試料の結晶構造に関するIDの種類	登録者選択
Identifier	No	試料の結晶構造に関するID	登録者記入
Description	No	試料の結晶構造の説明・記述	登録者記入
Description	No	試料の説明	登録者記入
choose type	No	試料のIDの種類	登録者選択
Identifier	No	試料のID	登録者記入
Material type	No	物質の分類	登録者記入
Material sub type	No	物質の副分類	登録者記入
Description	No	物質の分類の説明・記述	登録者記入
choose type	No	物質の分類のIDの種類	登録者選択
Identifier	No	物質の分類のID	登録者記入
Title	No	試料購入記録のタイトル	登録者記入
Date	No	試料購入記録に関する日付	登録者記入
choose type	No	試料購入記録のIDの種類	登録者選択
Identifier	No	試料購入記録のID	登録者記入
Organization	No	試料供給元組織	登録者記入
Sub organization	No	試料供給元組織部門	登録者記入
Organization	No	試料製造元組織	登録者記入
Sub organization	No	試料製造元組織部門	登録者記入
Purchase record item	No	試料購入記録データ	登録者記入
choose type	No	試料形状のIDの種類	登録者選択
Identifier	No	試料形状のID	登録者記入
Description	No	試料形状の説明・記述	登録者記入
choose type	No	試料の態のIDの種類	登録者選択
Identifier	No	試料の態のID	登録者記入
Description	No	試料の態の説明・記述	登録者記入
Category	No	試料の構造的特徴の分類	登録者記入
Sub category	No	試料の構造的特徴の副分類	登録者記入
Description	No	試料の構造的特徴の説明・記述	登録者記入
choose type	No	試料の構造的特徴のIDの種類	登録者選択
Identifier	No	試料の構造的特徴のID	登録者記入

表 2　MDR で実装されている試料メタデータの一覧

3. MDR の特長

MDR はマテリアルズ・インフォマティクスに資するためのデータリポジトリを目指して開発されており、MDR に登録された材料データに対する機械的なアクセスを当初から重視している。

先述のとおり、Hyrax はメタデータを RDF として定義・保存できるようになっており、ユーザはそれらのメタデータを RDF/Turtle と JSON の両方の形式で取得できる。また、Hyrax は実データの取得を行うための WebAPI として ResourceSync API[8]を備えている。これは、多くの機関リポジトリが備えている WebAPI である OAI-PMH 同様、リポジトリのデータの一括取得や同期に用いるための WebAPI であるが、OAI-PMH がメタデータのみを対象としているのに対し、ResourceSync は実データも一括取得や同期の対象としている。この機能は、大量のデータの収集と分析を必要とするマテリアルズ・インフォマティクスにおいて有用性の高いものである。DICE では、論文中の図表や画像の情報を検索するアプリケーション“FigResourceMiner”[9]が開発されているが、このアプリケーションは ResourceSync を

用いて、MDR に登録された材料データを収集するようになっている。

4. MDR の運用

4.1 データの登録方法

2020 年 8 月時点では、MDR にデータを登録できるのは機構の職員のみとなっており、ユーザ認証には機構の LDAP アカウントを用いている。しかし、DICE では機構外部のユーザへのサービス提供を想定しており、また DICE 上のアプリケーションに対するシングルサインインを実現するため、DICE の各アプリケーションはユーザ認証に CAS (Central Authentication Service)[10]を用いることが求められている。これに伴い、MDR でも 2020 年内にユーザ認証を CAS に切り替える予定としている。

MDR ではデータの登録方法について、バッチプログラムによる登録と、Web フォームによる手動登録の 2 種類の方法を想定している。バッチで登録を行う場合は、各システムで DICE の共通メッセージ形式によるメタデータを JSON 形式で記述し、実ファイルとともに zip ファイルに格納して、機構内の所定のファイルストレージに保存する（図 4）。MDR はバッチプログラムを起動して定期的にこの zip ファイルを読み込み、登録作業を行う。一方、Web フォームを用いて登録を行う場合、登録者は MDR にログインし、Web フォームからメタデータの入力と実ファイルのアップロードを行う。登録できるファイルサイズの上限は、バッチによる登録では特に制限を設けていないが、Web フォームによる登録では 1 ファイルあたり 100MB としている。

MDR でのデータ登録時には、書誌メタデータに対して、いくつかのメタデータの入力を必須としている。具体的な入力必須項目は、表 1 の Required 列を参照されたい。

4.2 データ公開の手順

MDR のデータ公開業務は、MDR の開発担当者と機構図書館の担当者を含めた 4 名の職員から成る運用担当チームによって行われている。

MDR に登録された直後のファイルは、公開範囲が「ログインユーザのみ」に設定され、外部からの参照が行えない状態になっている。ファイルが登録されると、MDR の運用担当チームがメタデータの入力内容をチェックし、必要に応じて不足しているメタデータについての記入を登録者に依頼する。メタデータの記入のチェックは、著者名や公知 URL など、データの公開に利用する部分に対してのみ行い、材料データに関するメタデータの内容のチェックは行わない。メタデータのチェックが完了すると、MDR 運用担当チームはデータの公開作業を行い、登録者にメールで公開完了の連絡を行う。公開作業で行われる具体的な作業内容は、DOI の付与と MDR のメタデータへの追加、ならびにデータの公開範囲の「全体公開」への変更である。

運用担当チームによるメタデータの入力内容のチェックで特に重要となるのが、「データ作成者の ORCID 番号の入力」「登録されたデータをリポジトリで公開する権利があるか」と「登録されたデータが公知であるかどうか」の 3 点である。1 点目の ORCID 番号は、データ作成者を世界中で一意に識別するために入力するものであり、ここで登録された ORCID 番号は、データへの DOI 付与の際にタイトルや作成日などのメタデータとあわせて、ジャパンリンクセンターや DataCite などの DOI 付与機関に送付され、世界的に流通することになる。2 点目のデータ登録の権利に対しては、例えば投稿論文の著者版ファイルが登録された場合、その論文掲載誌の出版社が著者版ファイルの機関リポジトリへの登録を許可しているかどうかを、出版社の著作権ポリシーを参照して確認することになる。3 点目の「データが公知であるか」という点については、MDR 運用担当チームが“First published URL”に入力されている URL にアクセスして、そのデータに関連する論文や資料が公開されているかどうかの確認を行う。この「公知であるか」という点の確認の必要性と詳細については、第 5.2 節において述べる。

公開が行われたデータは、第 3 節で述べた ResourceSync API などを通して、DICE の他のアプリケーションや外部の検索サービスに配信される。MDR へのデータ登録と公開 DICE アプリケーションとの連携の構成を図 4 に示す。

5. MDR の課題と将来

5.1 メタデータの記述に関する課題

前述のとおり、MDR において Hyrax を採用したのはメタデータの定義が柔軟に行えることが大きな理由であったが、MDR の開発を進めるにつれて、Hyrax でのデータ登録やメタデータの定義に、いくつかの機能的な制限が存在することがわかった。MDR で問題となった制限は、具体的には以下のものである。

- 複数の値をとる項目において、入力の順序が保証されない
- 識別子の種類」と「識別子」のようなキーと値をセットで持つメタデータが、デフォルトではサポートされていない。このため、このような構造を持つメタデータ項目については、別途自前でフォームや入力のバリデーション機能を作成しなければならない
- データのディレクトリの構造を保持したまま MDR にファイルをアップロードすることができない

Vol.2020-IOT-51 No.9
Vol.2020-SPT-39 No.9
2020/9/4

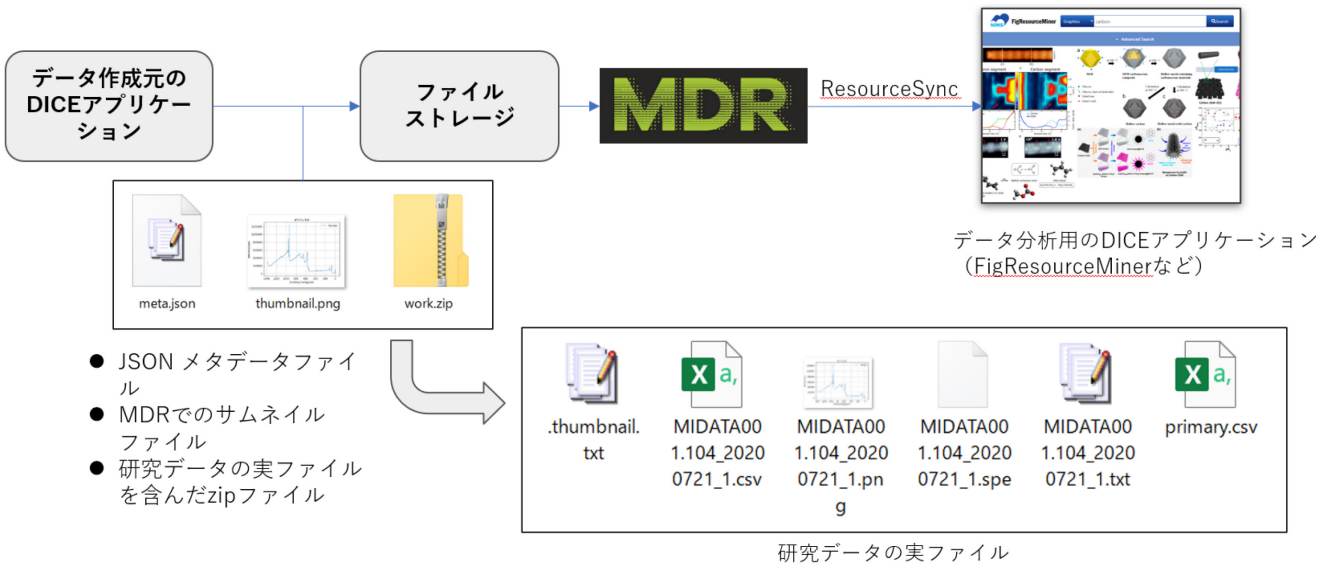

図 4　MDR でのデータ登録と他の DICE アプリケーションとの連携

特に入力の順序が保証されないという制約は、材料メタデータのみならず、著者順のような書誌メタデータにおいても大きな問題となる。この制約は Hyrax の開発コミュニティでも課題として認識されており[11]、Hyrax の次期バージョンとなる Hyrax 3.0 では、データの入力順序を保持できるように改善されることになっているが、2020 年 8 月時点では Hyrax 3.0 はまだリリース候補版の状態となっており、正式にはリリースされていない。著者名については、全ての著者名を著者順どおりに収録するメタデータ項目を別途作成することで対応したが、材料メタデータで複数の値を取る項目は多数存在するため、開発に必要な工数が大きく膨らむことが想定された。

また、DICE の共通メタデータ形式である「共通メッセージ形式」に起因する問題点も複数指摘された。「共通メッセージ形式」の設計にあたっては、機構内の複数の研究者によるヒアリングを実施したものの、実際に MDR の Web フォームによる研究データの登録を試行したところ、以下のような指摘が寄せられた。

- フォームが巨大かつ複雑すぎて、入力に非常に手間がかかる
- メタデータ定義の粒度が荒すぎて（もしくは細かすぎて）、どのようにメタデータを記述すればよいのかわからない

これらの制約を回避する方法として、MDR では共通メッセージ形式で記述しきれないメタデータについて、別途メタデータ記述用の CSV ファイルを用意し、実データとともに登録するという手法を検討し、部分的な実装を行っている。具体的には、“primary.csv”というファイル名で作成された CSV ファイルは、メタデータが記述されているファイルとみなし、MDR の詳細画面上にメタデータとして表示するという実装を行っている（図 3）。CSV によるメタデータの記述にあたっては、CSV on the Web[12]に準拠した、JSON-LD ファイルによる CSV ファイルへのアノテーションの付与も検討課題として挙がっているが、まだ具体的な実装には至っていない。

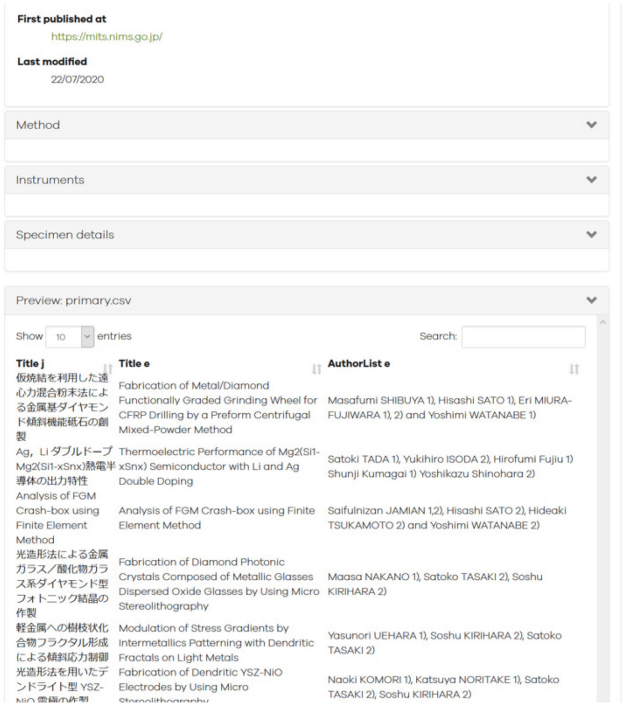

図 3 MDR 上で primary.csv 内のデータをメタデータとして表示している例

また、現在 MDR では、研究データパッケージングへの対応を進めている。ここでのデータパッケージングとは、メタデータと実ファイルをひとつのアーカイブファイルにまとめることを指す。データパッケージングの手法はすでにさまざまな分野で広く用いられており、例えば電子書籍で用いられる ePub や Microsoft Office で用いられる

OpenDocument、MDR へのバッチ登録の際に使用する zip ファイルも、データパッケージングの一例である。研究データを対象としたデータパッケージングのフォーマットも多数存在しているが[13]、汎用的なデータパッケージングのフォーマットとして RO-crate[14]という標準規格の開発がシドニー工科大学やマンチェスター大学の研究グループを中心に行われており、2019 年 11 月にバージョン 1.0 の仕様が公開された。RO-crate は、以下のファイル群をひとつの zip ファイルにまとめたものである。

- 研究データの実ファイル（ディレクトリを含む）
- JSON-LD によって記述したメタデータファイル
- 人間が参照するためのプレビュー用 HTML ファイル（オプション）

メタデータを JSON-LD によって記述するという仕様は、Hyrax がメタデータの定義や保存を RDF ベースで行うという特長を活用できるものである。また、アメリカ国立標準技術研究所(NIST)において、材料メタデータを Schema.org の拡張語彙を用いて表現する Materials Schemas[15]の開発が行われており、MDR においても JSON-LD によるメタデータ記述に Materials Schemas を用いることができるかどうかの検討を行っている。

5.2 データの公開に関する課題

現在の MDR は、第 4.2 節で述べたとおり、すでに公知になっているデータのみを登録・公開できるようになっている。また、MDR に登録されるデータは、一部の例外を除いてすべて認証なしでの公開となっており、「特定の機構外部のユーザ（共同研究者やジャーナルの査読者）に対してのみデータを限定的に公開する」という運用を行っていない。このような運用になっている理由は、MDR の持つ「材料科学におけるオープンサイエンスに資する」という目的によるものであるが、一方で研究活動にあたっては、研究が進行している途中のデータを共同研究者やジャーナル査読者と共有することも当然求められる。

しかし、MDR において未公知データを限定したユーザのみに公開するユースケースを検討した結果、輸出管理の観点から以下の点が問題として挙がった。

- 未公知データを登録する際、それが輸出管理の規制に該当するかどうかを判断することが困難である。論文の場合、論文誌に掲載された情報は公知となるため輸出管理の対象外となるが、公開対象を限定した場合は公知とみなされず、輸出管理の規制対象になる可能性が高い
- 機構外部のユーザに MDR のログインアカウントを発行してデータをダウンロードさせる場合、そのユーザが輸出管理上問題のない国や組織に所属しているかどうかを判断することが困難である

ユーザの所属組織の確認については、学認や eduGAIN, ORCID によるユーザ認証を行い、そのユーザ情報に付随する所属情報を用いるという方法も検討されたが、ユーザの所属する組織がこれらの認証サービスに参加していることが前提となり、特に企業における採用が困難であると予想された。また、これらの認証サービスを用いても、ユーザがその組織に所属していることを確実に保証できるものではないと判断されたため、採用に至らなかった。このため、機構外部からのデータ登録は、当面は限定された研究機関からのみ行う予定となっている。

6. おわりに

MDR は運用開始からまだ日が浅いが、その短期間の運用においても、データ登録や公開の障壁となる多くの課題を発見することとなった。本事例が、他分野におけるデータリポジトリの活用の参考となれば幸いである。

参考文献

[1] Materials Data Repository. https://mdr.nims.go.jp, (参照 2020-08-09)
[2] DICE. https://dice.nims.go.jp/, (参照 2020-08-09)
[3] 閃きを生む、物質・材料開発の“知能”「材料データプラットフォーム」とは. NIMS NOW. 2019, vol. 19, no.1, p.8-12. https://www.nims.go.jp/publicity/nimsnow/vol19/hdfqf100000aoslh-att/hdfqf100000aosp0.pdf, (参照 2020-08-09).
[4] Hyrax: a community-supported repository front-end. https://hyrax.samvera.org/, (参照: 2020-08-09)
[5] Samvera Implementations: In-production. https://wiki.lyrasis.org/display/samvera/Samvera+Implementations%3A+In-production, (参照 2020-0-8-09)
[6] 菊地伸治, 門平卓也, 鈴木峰晴, 内藤裕幸. 高付加価値科学データ創出を指向した研究データ管理プラットフォームのアーキテクチャ. 信学技報, vol. 119, no. 66, p. 7-17, https://mdr.nims.go.jp/concern/publications/m039k565z?locale=en, (参照 2020-08-09)
[7] 松田朝彦. 材料データリポジトリにおける共通メタデータ・分野別メタデータ. Japan Open Science Summit 2019. May 27-28th, 2019. https://doi.org/10.34968/nims.1359, (参照 2020-08-09).
[8] ResourceSync Framework Specification. https://www.openarchives.org/rs/1.1/resourcesync, (参照 2020-08-09)
[9] Yasuhiro Takada, FigResourceMiner: A search text in images, and graph visualization platform for academic articles. Poster presented at NIMS WEEK 2019. October 30, 2019. Tokyo. https://www.nims.go.jp/nimsweek/day2/poster.html#poster-list, (参照 2020-08-09)
[10] CAS. https://www.apereo.org/projects/cas, (参照 2020-08-09)
[11] Hyrax Metadata Ordering Working Group. https://wiki.lyrasis.org/display/samvera/Hyrax+Metadata+Ordering+Working+Group. (参照 2020-08-09)
[12] CSV on the Web: A Primer. https://www.w3.org/TR/tabular-data-primer/, (参照 2020-08-09)
[13] Approaches to Research Data Packaging - RDA 11th Plenary BoF meeting . https://www.rd-alliance.org/approaches-research-data-packaging-rda-11th-plenary-bof-meeting, (参照 2020-08-09)
[14] Research Object Crate (RO-Crate). https://www.researchobject.org/ro-crate/, (参照 2020-08-09)
[15] Materials Schemas https://pages.nist.gov/material-schema/, (参照 2020-08-09)