資料2

2023年9月29日

研究データ利活用協議会 研究資料・実験機器へのPID付与検討小委員会
議論の経緯と今後に向けた提言

研究データ利活用協議会（RDUF）
研究資料・実験機器へのPID付与検討小委員会

1. はじめに

「研究資料・実験機器へのPID付与検討小委員会」（以下、以下、「本小委員会」という。）は、研究データ利活用協議会（RDUF）の設置する小委員会として2022年4月から2023年9月までの期間に活動を行った。以下では、本小委員会における議論の経緯及び今後に向けた提言を報告する。

2. 背景

学術情報流通において、PID（Persistent Identifier、永続的識別子）は様々なオブジェクトを識別するとともに、オブジェクトへのアクセス向上やオブジェクト間の関係の明確化に対して重要な役割を果たしている。

これまでPIDが付与されてきたオブジェクトとしては、論文等の研究成果、研究データ、研究者、研究機関等が挙げられる。ジャーナル論文にはDOI（Digital Object Identifier、デジタルオブジェクト識別子）が付与されることが一般的となり、参考文献を記載する際の識別子としてDOIが広く用いられている。さらに近年、プレプリント、博士論文のほか、分野によっては学会発表資料、書籍等様々なドキュメントにもDOIが付与されている。また、研究データへのDOIも分野間の差異はあるがその範囲は広がりつつある。さらに、研究者についてはORCID[1]、研究機関や研究資金配分機関についてはROR[2]等、研究プロセスに関わった人や機関についてもPIDが整備されてきた。

一方で、研究に用いた試料、史資料、機材等の有体物としての研究資源については、機関や分野固有のID（Identifier、識別子）を付与されることはあるものの、PIDについては広く普及している状況ではなく、例えば計測機器については、RDA（Resarach Data Alliance）のPIDINST Working Group[3]において分野横断的な運用ソリューションについて議論され、ドキュメントが公開されている。また、ドイツではPIDの国家ロードマップを作るためのプロジェクトが2023年3月から実施され、その対象として機材や有体物も含まれる状況である[4]。

オープンサイエンスの対応を含め、研究DXの推進には、これらの有体物に関する情報もサイバー空間において参照できるようにすることが不可欠であるが、有体物の対象やそれを取り扱う分野が幅広く、ボーンデジタルのデータとは異なる特性を持っているため、PID付与を実現するには解決すべき課題が多いと考えられる。

本小委員会では、これら有体物としての研究資源に関するIDやメタデータの現状や課題について各委員から事例紹介を行うとともに意見交換を行った。また、RDA PIDINST Working Groupが作成・公開している「PIDINST Whitepaper」の日本語訳作業を通して有体物へのPID付与に関する状況や先進事例を理解するとともに、日本語訳ドキュメントを公開することにより国内における情報共有の活性化を図った。

---

[1] https://orcid.org
[2] https://ror.org/
[3] https://www.rd-alliance.org/groups/persistent-identification-instruments-wg
[4] https://doi.org/10.5438/vb6v-4m30

3. 議論の経緯

3.1. 有体物に対するPID付与に関する事例紹介・情報共有

2022年4月から8月にかけて開催した小委員会において、各委員の持つバックグラウンドや現在取り扱っている有体物に対するPID付与の状況や課題について事例紹介とPIDに関連した話題提供を行った。

3.1.1. 事例紹介の概要

具体的な事例として下記3件の紹介が行われた。

○事例1 「サンプルIDとPIDのギャップ」 福田委員

- 海洋研究開発機構が保有する船舶の調査航海で取得されたデータ・サンプルに関するPID付与の取り組み状況とともに、サンプルIDとPIDに対するギャップについて紹介した。
  - 取得されるサンプルは生物サンプル、岩石サンプル、堆積物コアサンプル、その他サンプル（海水、沈降粒子、懸濁粒子、大気、エアロゾル等）に大別している。
  - 航海及び潜航はIDで区別しており、航海IDを単位としてPIDとなるDOIを付与している。サンプルを識別するためのIDとしては、船上で付与するOnboard ID（航海IDや潜航IDの文字列が含まれることが多い）を使用しているが、生物サンプルについてはデータベースに登録した際に自動で採番されるJAMSTEC No.も付与している。なお、現状では、サンプルに対してのPID付与は行っていない。
  - サンプル情報のフローでは、航海IDの単位としたメタデータシートに採取したサンプルの基本情報の記述を求めており、メタデータシートが提出された後はサンプル管理データベースに登録している。公開可能な情報については、公開猶予期間終了後に機構のサイトで公開するとともに、生物、岩石サンプルについては、外部機関の関連データベースでも公開される。
  - 文献におけるサンプル引用には、JAMSTEC No.や Onboard ID が用いられており、現時点ではサンプルにPIDを付与しなくても、研究や成果の公開を行う際に大きな支障は生じていない。だが、一意性や永続性の保証がなく、分野間の共通性もない状況であること、ジャーナルによってはIGSNのようなPIDの引用を推奨されるケースも増えていることから、今後データや研究スタイルが変化していく中で、サンプルに対するPID付与への要望が高まる可能性がある。
  - 継続的にPIDの維持、管理を行うには、メタデータやデータベースを運用する体制の継続や、維持するための費用負担等が必要である。また、インターネット環境にアクセスできないことの多いフィールド調査において、サンプル採取とPID発番のタイムラグへの対処等も必要になることが想定される。

○事例2 「核融合研における実験データや実験装置へのPID付与の取り組みの現況」 中西委員

- 核融合科学研究所におけるオープンサイエンスの取り組みと核融合分野の実験データに関する国際動向を紹介した。
  - 核融合実験データは原則として、ボーンデジタル（計測機器からの数値データ）であり、自動的／大規模なデータ収集が行われ、大規模分析が標準的な研究の進め方である。
  - 核融合研では、現有のLHD（大型ヘリカル装置）実験以外に、大学も装置実験データも遠隔収集／保管し、国内で参照・解析プラットフォームを分野内に提供している。SNET遠隔集録と呼ばれている。
  - 「プラズマ・核融合クラウド」構想では核融合データの「超」分野利活用、来たるべきITER（次世代核融合実験装置）で取り扱う巨大データの技術基盤整備、データを鍵とした社会一般／産業界との連携が目標となっており、オープンサイエンス化が喫緊の課題である。
  - 同分野で先行しているのは欧州である。Horizon2020においてEUから予算、組織面で強力な支援のもと、“FAIR4Fusion”を掲げて、核融合実験データオープン化を実施中である。
  - 核融合研のPIDに関する現在の取り組みとしては、下記が挙げられる。

        - 大容量データの保存コスト削減のために、計測器、実験回の単位で圧縮／書庫化を行い、サイズ低減を図っている。LHD実験では、書庫化されたデータ総数が、およそ2～3千万個である。
        - なお、核融合実験データの参照は、計測（データ）名と実験番号の2つの主キーで検索／抽出するのが世界共通である。ただし、各サイトにプライベートPIDはあるが、グローバルPIDではない状況で、上記のFAIR4FusionでもePICやDOI登録が検討されたが、全データオブジェクトに対する登録発行には至っていない。
        - LHD+SNET実験データシステムでは、共同研究者向けに仮想専用線（SINET L2/L3-VPN）経由で各装置の実験データを共有してきたが、オープンデータ化を進めており、一般に公開を始めている。DOI登録も視野に入れている。
    - 核融合研におけるシステム化と今後の対応として、下記が挙げられる。
        - Gakunin RDM、Gakunin Cloudのフレームワークプラットフォームを最大限活用し、プライベートなオンプレミス実験データストレージとの透過的な接続を実現する。
        - 実験の運転等に関するメタデータのデータベースが複数分散して存在しているが、PIDランディングページにまとめて一般公開し、オープンサイエンス化を図る。
        - 計測・解析データオブジェクトの他に、例えば、LHD本体の核融合プラズマ発生装置周りに設置された約100種の計測機器に対しても、PID付与を検討している。

○事例3 「JASRI, SPring-8の試料PID付与に向けた取り組み」松本委員

- SPring-8での試料PID付与に向けた取り組みを紹介した。
    - 未知の材料の同定には、標準試料のデータベースで照会できる環境整備が必要である。NIMS（材料物質研究所）のMDR（Materials Data Repository）[5]と協力し、BL14B2（産業利用のBL）においてJASRIスタッフが測定とデータベース化を推進している。
    - XAFS（X線吸収微細構造）計測装置のメタデータは、YAML形式で整理されており、項目はモノクロメータの設定やイオンチャンバーの条件等で構成される。
    - 計測試料の準備には、試料を粉砕／ペレット化する等の作業が必要であり、試料メタデータもYAML形式で記述している。項目はサプライヤ、ロット番号、化学式等で構成される。
    - その他のメタデータとして課題に関する情報がある。
    - PIDは採用していないが、ローカルIDによる管理体制は一通り確立している。また、データベースも利用されている状況である。
    - 他機関によるXAFSデータの収集・整合性の維持のため、日本XAFS研究会でコミュニティでデータベース活動を推進しており、NIMSが中心となってMDRにデータを集約している。
    - NIMSのMDRはオープンデータで、GUIでのデータ閲覧も可能である。
    - NIMSがデザインするメタデータは3層に分かれており、第1層はDataCite等と互換性のある共通メタデータ、第2層は大分類、第3層はより詳細な情報が格納される。XAFSスペクトルはこの第3層に該当する。
    - 試料メタデータの課題として、メタデータの種類の統一、メタデータ作成の作業量、単体だけでなく化合物／混合物の取扱い方が挙げられる。
        - 材料の命名方法についてはばらつきが大きく、NIMS MDRでの材料辞書MatVocの整備や、オントロジーの整理が進められているほか、日本放射光学会データ構造化諮問委員会でもデータ記述の構造化について議論がなされている。
        - 化合物・混合物に関する試料表現については、NEDO燃料電池プロジェクト等で実践中である。

3.1.2. PIDとデータ品質保証

PIDに関連した話題提供を行った。

○「PID付与の観点からみたデータ品質保証」岡山委員

- データの品質の維持を考慮する上で、「不正」と「望ましくない」データ利用の違いに留意する必要がある。
- 研究開発におけるデータ管理フローとして、「生データ」→「データ可読化」→「データの高付加価値化」の流れを考慮する必要がある。

[5] https://mdr.nims.go.jp

- IDをどのように付与し、どのようにデータの品質を管理するかを実践するためには、実験機器のPIDとデータのPIDについて、総体としての管理方法の開発が必要である。
- データの価値の位置づけとして、組織の資産としても品質保証が求められる。
- データの品質保証の標準としてISO/IEC20512があるが、15の指標のうちデータの品質にかかわるものは5つある。
- データ管理においてデータ品質を維持するためには次の5つの視点が必要である。
  1. データ管理プロセスの視点
  2. トレーサビリティにおける品質保証の視点
  3. データキュレーションにおける品質保証の視点
  4. 改ざんされないデータ記録の視点
  5. データをどのように利用したいかの視点

3.1.3. 3.1.1及び3.1.2の情報共有と意見交換を踏まえた状況の整理

事例として紹介された対象物は異なるものの、以下が共通事項として挙げられる。

- どの研究機関でも「管理番号」を採番している。これは、実験史資料およびデジタルデータいずれにおいても共通である。
- 「管理番号」とともに、メタデータや情報をデータベースを用いて管理／整理している、あるいはデータベース化される下地が整った状態で情報が整理されている状態である。
- しかし、多くの資料については機関固有のID体系での管理にとどまっており、PIDの要件を満たしていないことが多い。このような管理体制下では、資料の一意性や所在の永続性についての保証が薄く、また分野間の共通性がない等、資料とデータの利活用に不便な情況となっている。
- 特に、有体物の資料については、その資料の由来(対象物が取得された、場所、時間、方法等)や所在(資料の保存場所、保存方法、利用方法等)を示す情報が、メタデータとして管理される必要がある。
- さらに、実験データの品質保証やデータ管理において、データ管理プロセス、トレーサビリティ、データキュレーション、改ざん防止のデータ記録などが重要となる。
- 資料の管理ID→PID化の具体的メリットが見えにくい。DOIなどPIDを新たに付番するだけでは、データ管理の工数削減にはつながらず、「PIDを通じて、データや資料等の所在が世界的・永続的に可視化され、その流通が保証される」ことに価値を見出すべきである。
- 有体物の史資料にPIDを付番し、より利便性の高いデータとして活用するためには、共有可能なメタデータスキーマを採用し、それを管理し流通させるための、データベース構築の構築と維持が必須である。しかし、これらの設計／構築／運用においては、計算機および人的双方に大きなリソースが必要とされている。
- 一貫性のあるメタデータの設計は、技術的な困難が多い。専門のスキルを持った人材育成も合わせて検討する必要がある。また、PIDに紐づくメタデータは必然的に世界標準となるため、世界的な動向を踏まえたキャッチアップと世界規模での利用展開を意識する必要がある。

3.2. PIDINSTに関するドキュメントの日本語訳

研究資源へのPIDとメタデータの付与・管理に関する世界的動向の調査の一環として、RDAのPIDINST Working Groupが公開している各種英文ドキュメント[6]の日本語訳作業を行うこととした。これらのドキュメントは、実験機器を特定するメタデータ標準に関する内容が中心となっており、有体物管理のよりよい方法を検討するための参考になると考えられる。

2022年8月からPIDINSTの日本語訳作業についてオンライン作業環境を整備し、同年10月の小委員会で日本語訳作業の分担や進め方を調整し、各委員による日本語訳作業を経て同年12月には本文日本語訳が完了、2023年1月にはクックブックの日本語訳作業が完了した。

[6] https://docs.pidinst.org/en/latest/

その後、完成版に向けた校正／編集を行うとともに、RDAのPIDINST WGに投稿し、本小委員会の活動内容と成果報告を行った。PIDINST WGからはこれらの活動が好意的に評価され、2023年7月に、日本語訳ドキュメント並びにソースコードが、PIDINST WG本体のWebページに公式に掲載された。ドキュメントは下記のURLから閲覧可能である。日本語訳版を公開することにより、国内で有体物のうち特に実験機器へのPID付与に関心がある機関やその担当者が自機関でのPIDを検討する上で役立てられることが期待される。

PIDINSTに関するドキュメント日本語訳版の公開URL：
https://docs.pidinst.org/ja

- PIDINSTホワイトペーパー（PIDINST Whitepaper の日本語訳）
- ePIC クックブック（ePIC Cookbookの日本語訳）
- DataCite クックブック（DataCite Cookbookの日本語訳）

上記日本語訳ドキュメントのソースコード(PIDINST WG 内のプロジェクト):
https://github.com/rdawg-pidinst/white-paper-ja

4. 今後に向けた提言

PIDとこれに付随するメタデータの重要性は、2010年代初頭に発表されたFAIR原則[7] の中でも明確に述べられている。本小委員会の活動では、非デジタルの研究史資料、試料、装置へのPID付番の課題を実務的な立場から総括し、改めてその意義を問い直すこととなった。今後の研究データ管理の方向性と課題について、研究者、研究機関、研究コミュニティそれぞれに対し、以下の提言を述べることとする。

4.1. 研究者への提言

- 研究成果、研究データに対し、PIDやPIDに紐づくメタデータを付与することは、このデータの存在を明らかにする強力な手段である。特に、PIDに紐づくメタデータはその研究分野に置ける共通言語を定義し、データのトレーサビリティを保証する。従って、データにPIDを付与し、これを公開することは、研究者のプレゼンスや信頼の向上につながる。
- 研究者は、自身のデータを整理する場合に、既に公開され、流通しているPIDやメタデータとの関係を考慮することが望ましい。これらにより自身の研究成果もまた、速やかにこのデータ流通網（データエコシステム）に参加させることができる。
- 研究者独力による「完璧な」PID体系とメタデータスキーマの開発は困難である。しかしながら、複数の研究者が、メタデータスキーマを共有し改善に取り組むことは、研究者間の相互理解を高めるとともに、研究パフォーマンスの向上に資するであろう。

4.2. 研究機関等、組織的対応への提言

- 研究機関／研究グループ等の研究組織がその活動を維持し、拡大するためには、PIDに基づく研究資源の管理が必須である。ここでいう研究資源とは、1次データから最終成果物となる研究データや史資料等、および実験装置やソフトウェア等の研究環境すべてを指し、さらにはデジタル、非デジタルを問わない。
- 管理対象となる研究資源は、今後も、量的／質的双方において、大幅に拡大することは間違いない。一方、PID及びメタデータの付番、管理には技術的／金銭的側面からの課題も大きい。研究組織は、組織内における研究資源のID管理、及びこれらのグローバルなPIDとの接続性の要否について、戦略的な知見を持つべきである。

[7] https://force11.org/info/the-fair-data-principles/

4.3. 分野間・組織間コミュニティへの提言

- 装置や資料といった有体物に対するPID付番は、従来より検討されてきたデジタル化された研究データと同様に、学術分野共通の課題である。さらに、付番されるPIDは、学術分野間をつなぐことで、新たな学際領域を発掘する大きな可能性を提供する。
- RDUFを始め、分野／組織横断的な研究データ利活用に関するコミュニティは、事例収集とその共有について、更なる活動強化を期待したい。

以上