

第1回 SPARC Japan セミナー2020

「研究データ公開:フルオープンと制限公開の境界線」

材料分野の研究データ公開における理想と課題 ～NIMSの事例から～

篠田 陽子

(物質・材料研究機構 統合型材料開発・情報基盤部門)

講演要旨



物質・材料研究機構では2017年に材料データプラットフォームセンター(DPFC)を設置し、材料分野におけるデータ駆動型研究の促進に貢献する材料データプラットフォームDICEの構築に取り組んでいる。DICEは材料データの入り口(収集)から出口(利活用)までを一貫して扱うシステムであり、取り扱う研究データの来歴や性質は多岐にわたるため、DPFCでは各サービスの在り方や運用について検討し議論を重ねてきた。本発表ではDICEでの事例を紹介しつつ、研究データ公開における理想と課題を考える機会としたい。



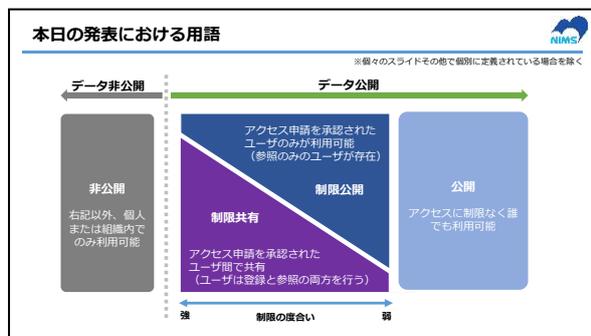
篠田 陽子

国立研究開発法人物質・材料研究機構 統合型材料開発・情報基盤部門 材料データプラットフォームセンター データサービスチーム所属。研究機関、研究費配分機関、大学研究室で研究支援業務に従事後、2016年に同機構に転職し、2019年6月より現職。研究データポリシーやサービス利用規約の策定に携わっている。

本日の発表における用語

本編に入る前に、本日の発表における用語を整理させていただきます(図1)。「公開」はアクセスに制限なく誰でも利用可能なこと、「制限公開」は承認ユーザーのみが利用可能なこと、「制限共有」は承認ユーザーの間で共有(ユーザーは登録と参照の両方を行う)すること、「非公開」は右記以外、個人または組織内でのみ利用可能なこととする。

ザラ間でデータを共有すること、そして、これらを合わせて「データ公開」、それ以外を「非公開」「データ非公開」とします。



(図1)

物質・材料研究機構(NIMS)について

最初に、NIMSについて簡単にご紹介します(図2)。NIMSは、その名が示すとおり、国内で唯一、物質・材料科学の研究を専門とする公的研究機関です。茨城県つくば市内の3地区と、兵庫県西播磨地区のSPring-8内に事業所があります。役職員は約1,600人で、このうち50~60%が研究職員、25%がエンジニア職員として働いています。研究組織としては8つの研究拠点・部門があり、各組織で特色ある物質・材料

の研究開発を行っています。

私が所属する統合型材料開発・情報基盤部門 (MaDIS) は、データとの融合で物質・材料の研究開発の在り方を革新することを目指して 2017 年 4 月に設置された新しい部門です。材料データプラットフォームセンターでは、データ駆動型の研究開発を支える基盤として、本日紹介する材料データプラットフォームの構築・運用に取り組んでいます。私はここで、主にデータの利用に関する枠組みやルールづくりに携わっています。

材料分野の研究データ

材料分野の研究データにはさまざまなものがあります。それらのデータをどのように取り扱うかを考えると、図 3 のようなイメージになります。左側の制限公開や非公開のところにあるものが多めになっています。実際には個々のデータで状況が異なるため、必ずしもこの限りではないのですが、材料分野の研究データは、例えば生物や化学、物理など他の基礎科学の分野に比べると、相対的にオープンにされにくい傾向にあると

言えます。

では、なぜオープンにされにくいのか。大きな理由の一つとして、材料は産業界や知財との関連が強いということが挙げられます。使われてこそその材料といわれるように、材料は社会や産業に役立つ物質であり、使える、そして使われる材料を作り出すための化合物組成や構造、成形プロセスなどは、研究戦略やノウハウに直結しやすい状況があります。

その一方で、近年のオープンサイエンスの流れや、マテリアルズ・インフォマティクス、マテリアルズインテグレーションといったデータ駆動による材料設計・開発への取り組みも加速的に進んでおり、データの利活用に対する期待や要望も高まっています。

こういった背景からも、材料分野の研究データにおいては、研究の新たな発展、そしてイノベーション促進のためにも、保護と利活用のバランスを取りながら戦略的にオープンにしていくことが重要になってきます。

NIMS の研究データポリシー

NIMS は国研として日本全体のイノベーションシステムをけん引する役割があり、また、NIMS の強みであり資産である材料データの利活用を図っていくという使命があります。このため NIMS では、研究データに対する基本的な考え方や利用ルールの整備と、データ利活用のためのデータインフラの構築・運用を両輪で進めてきました。

2018 年 8 月に研究データポリシーを制定しました。このポリシーの最初でも述べているのですが、研究活動を通じて得られる研究データは重要な知的資産であり、適切な管理と広く有効な利活用を図っていくために、研究データの収集、管理、利活用に関する基本的な方針を、研究データポリシーとして定めています (図 4)。基本的な考え方としては、NIMS の研究戦略や分野特性を考慮するとともに、研究データの性質や来歴等に留意しつつ取り扱うこととしています。

一口に研究データといっても、形式や種類はさまざま

1. 物質・材料研究機構 (NIMS) について

国立研究開発法人物質・材料研究機構
National Institute for Materials Science

国内で唯一、物質・材料科学 (Materials Science) の研究を専門とする公的研究機関

所在地: 茨城県つくば市内 3 地区、兵庫県西播磨地区 (SPRING-8内)

従業員数: 1,564名 (2020年4月1日現在。定年制職員以外も含む)

研究組織: 機能性材料研究拠点
エネルギー・環境材料研究拠点
磁性・スピントロニクス材料研究拠点
構造材料研究拠点 (RCSM)
国際ナノアーキテクトニクス研究拠点 (MANA)
先端材料解析研究拠点
複合材料開発・情報基盤部門 (MaDIS)
技術開発・共用部門 (RNFS)

MaDIS
エネルギー材料設計グループ
データ駆動分子設計グループ
データ駆動構造材料グループ
デバイス材料設計グループ
データ駆動無機材料グループ
SIP-MEラボ
材料データプラットフォームセンター

(図 2)

2. 材料分野の研究データ

データ非公開 ← データ公開

・ソフトウェア
・機械学習モデル
・性能
・プロセス
・シミュレーションされたデータ (商品データベース)

・実験データ
・計算データ
・シミュレーション
・化合物組成
・物性
・構造

・計測データ (第一原理計算など)
・標準物質のデータ
・文獻
・論文のバックデータ etc...

公開する? 公開しない?

材料分野の研究データは、生物や化学、物理など基礎科学の分野に比べると相対的にオープンにされにくい傾向にある。

(図 3)

ます。NIMSの研究データポリシーでは、対象とする研究データを電子データとしています(図5)。種類としては大きく3つ、①NIMSの研究成果物等のうち、一般に公表されたもの、②NIMSの研究成果物等で一般に公表されていないもののうち、NIMSが組織として収集・保管し、利活用を図るべきものとして選定したもの、③NIMSが構築する材料データプラットフォーム等を通じた利活用を図るため、NIMS外から提供を受けたものとしています。

これらの研究データを提供・公開するときの考え方として、まず研究データの提供・供用については、NIMS内外のデータの利活用によって発展的な成果を生み出すことを目指すとともに、必要に応じてアクセス制限や利用の限定などの措置も取り得ることとしています。研究データの公開(ポリシーでは、インターネット上で公表しアクセスできるようにすることを指す)については、公開することを原則としつつ、権利保護や情報保護の観点から配慮が必要なものは公開対象外としています。また、公開に際しては必要に応じて条件を付けることとしています。

(図4)

(図5)

材料データプラットフォーム DICE

この研究データポリシーと両輪となっていくデータインフラが、材料データプラットフォーム(DICE)です(図6)。材料科学の研究を推進し、材料開発を加速するためのプラットフォームであり、材料データの入り口から出口までを一貫して取り扱うシステムです。

DICEで取り扱うデータは、研究活動における日々の実験データや計測データ、データベース、論文や報告書、さらには利活用のためにNIMS外から提供されるデータなど、来歴や性質がさまざまです。このためプラットフォームでは、提供サービスごとにデータのオープン・クローズを検討する方法を取っています。具体的には、何をどこまで公開するのか、誰に使わせるのか、どう使わせるのかを細かく組み合わせることによって最適化を図っています。今回はDICEのデータ公開の事例として、MDR、MatNavi、共有データリポジトリの3つのサービスについて説明します。

Materials Data Repository (MDR)

MDRは、「公開」の事例です(図7)。MDRは材料研究データや文献等を収集・保存し、インターネットを介して広く利用できるようにするための公開データリポジトリで、2020年6月15日から公開を開始しています。名前に「材料データ」とあるように、材料に関するメタデータ項目をサポートしており、登録データにDOIを付けることによって引用を容易にしているのが特徴です。

MDRで公開するデータは主にNIMSの公知データ

(図6)

です。公開されたデータは誰でも利用することができ、検索・閲覧・ダウンロードが可能です。データ登録に関しては、現在はNIMSユーザーに限定しています。

データの利用条件としては、各データに示されているライセンス条件に従うこと、それから、公表時に出典を明示することとしています。

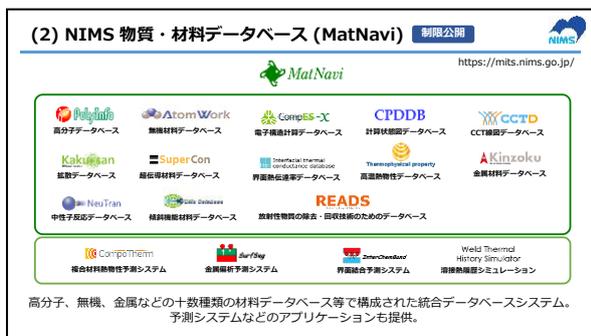
NIMS 物質・材料データベース (MatNavi)

MatNavi は、「制限公開」の事例です(図 8)。高分子、無機、金属などの十数種類の材料データベース等で構成された統合データベースシステムで、予測システムなどのアプリケーションも提供しています。これまでたくさんのユーザーにご利用、ご活用いただいています。

MatNavi では、NIMS が組織として収集したデータ(データベース)や、NIMS が利用および管理の権限を保有しているデータを公開しています。データベースの利用にはユーザー登録をお願いしています。利用者はデータの検索・閲覧ができますが、Web スクレイピング等によるデータの大量取得は禁止しています。



(図 7)



(図 8)

また、MatNavi のデータは、教育または研究開発目的での自己利用に限定して利用を許諾しています。

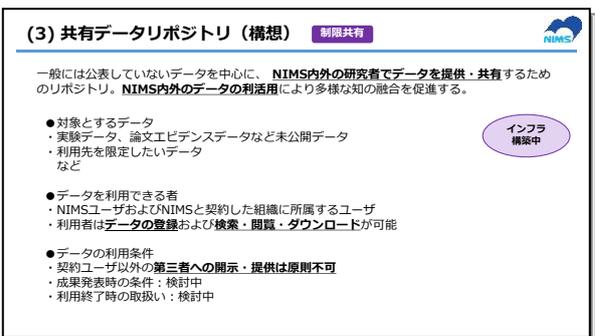
共有データリポジトリ (構想)

共有データリポジトリは、「制限共有」の事例です(図 9)。こちらはまだ構想段階ですが、NIMS 内外の研究者でデータを提供・共有し、これらのNIMS内外のデータの利活用によって多様な知の融合を促進することを目指すものです。

共有データリポジトリでは、実験データや論文エビデンスデータなどの未公開データ、あるいは利用先や利用者、利用する方法などを限定したいデータなど、一般には公表していないデータを扱うため、利用者はNIMSユーザーおよびNIMSと契約した組織に所属するユーザーに限定されます。そして、「共有」という名が示しているように、利用者は検索・閲覧・ダウンロードの他にデータの登録も行えることになります。また、このリポジトリで共有されるデータは第三者への開示・提供は原則不可となります。

DICE の事例のまとめ

紹介した3つのサービスを比較すると、図10の表のようになります。現状、NIMSでは、研究データの公開の仕方を基盤あるいはサービスごとに検討していますが、各サービスで扱えるデータの範囲が限定されており、これを広げていくことや、サービス間の連携をうまく図っていくことなど、まだ課題が残されている状況です。



(図 9)

研究データ公開における理想と課題

最後に、研究データ公開における理想と課題について、5つの項目を挙げて話題提供したいと思います。なお、これは主にプラットフォームのサービス担当者の視点から見たもので、発表者の私見も含まれていますので、組織としての考え方とは異なっている場合もあることをご承知おきいただければと思います。

1つ目は、本日のセミナータイトルにもなっている公開（フルオープン）と制限公開／制限共有の境界線です（図11）。明確な基準があり、一律に線引き・判断できるのが理想ですが、実際はデータの内容やデータ提供者の考え方によって公開可否の判断はさまざまであり、さらに時間軸が加わって状況が変化していくこともあります。ですので、結局はデータごとに細やかに判断していく必要があります。また、国研としての役割の中でのバランスという観点もあります。知の最大価値化の実現に向けて、どのポイントに比重を置くかが重要になってきますが、安全や保護を意識すると非公開に偏りがちになります。

2つ目は、データ（提供者）の権利の保護という点

サービス名称	公開	制限公開	制限共有
	MDR	MatNavi	共有データリポジトリ
対象データ	公知データ	利活用を図るため、NIMSが組織として収集したデータ	一般には公表していないデータ（未公開データ、利用先を限定するデータ、など）
データの提供者	主にNIMS	NIMS	NIMS内外の研究者（組織契約をしたユーザ）
データの利用者	制限なし	利用登録をした者	NIMS内外の研究者（組織契約をしたユーザ）
データ利用者ができること	検索・閲覧・ダウンロード	検索・閲覧	検索・閲覧・ダウンロード
データの利用条件*	- 各データ等のライセンス条件に従う - 出典の記載	- データの大量取得は禁止 - 教育又は研究開発目的での自己利用に限定 - 出典の記載	- 第三者への開示・提供は原則不可 - 成果の取扱いは検討中

*サービス利用に際しましては、規約、約款等で詳細をご確認ください。

（図10）

4. 研究データ公開における理想と課題

●公開（フルオープン）と制限公開／制限共有の境界線

理想：明確な基準があり、一律に線引き・判断ができる。

課題：データの内容やデータ提供者の考え方によって公開可否の判断は様々。時間経過とともに状況が変化することもある（例：論文発表や出願など）すばった線引きや一律に白黒つけることは難しい。

課題：国研としての役割の中での「バランス」。知の最大価値化の実現に向けて、どこに比重を置くか？「安全」「保護」を意識しすぎると非公開に偏りがち。



（図11）

です（図12）。理想としては、法律ですべてきれいに保護されて安心できると思いますが、実際にはデータは無体物であり、民法上も権利の対象とはなりません。また、著作権、特許権、営業秘密なども、ある一定の条件を満たしていないと保護されないということがあり限定的です。このためデータの保護においては、原則として契約を通じて保護を図る必要があります。この手続きで難しい部分が出てきます。

これと関連するのが、3つ目の利活用のシナリオです。データ利用者とデータ提供者の双方にとってうれしい使われ方がされるのが理想ですが、課題としては、例えば成果の取り扱いについて、データを使って得られた生成物・成果物は誰のものかということがあります。特に昨今のマテリアルズ・インフォマティクスやマテリアルズインテグレーション、AIで利用したときの成果の取り扱いなどは難しいところがあります。約款等であらかじめ扱いを定めておくことが必要になる場合もあります。また、データ（提供者）の貢献をどう評価していくか。データの引用・利活用状況の可視化や、その可視化された貢献を組織内で適正に評価に反映していく仕組みづくりも必要だと感じています。

4つ目はデータの提供方法です（図13）。いろいろなデータ公開のステータスがありますが、非公開から公開までワンストップサービスで提供できるのがやはり理想です。課題としては、データの形式や性質、権利関係などが一様ではなく、ステークホルダーも非常に多い状況になっているため、1つのシステムで多様なデータに対応することは、技術面あるいはコスト面の問題もあり、実際のところ困難ということがありま

4. 研究データ公開における理想と課題

●データ（提供者）の権利の保護

理想：法律で保護される。

課題：データは無体物であり、民法上、所有権や占有権、用益権、担保物権の対象とはならない。著作権、特許権、営業秘密などによる保護は限定的。データの保護は原則として契約を通じて図る必要がある

●利活用シナリオ

理想：データ利用者とデータ提供者にとって双方よし

課題：成果の取扱い。データを使って得られた生成物・成果物は誰のもの？（MIやAIで利用したとき）約款等で予め扱いを定めておくことが必要になる場合もあり。

課題：データ（提供者）の貢献の評価
データの引用・利活用状況の可視化（DOI付与や出典記載その他）や組織内で評価に反映する仕組みづくりが必要。

（図12）

す。

また、ルールを整備していくところでも、ステークホルダーが多くなればなるほど調整が難しくなります。現状としては、サービスごとに対応しながらそれらをシームレスにつなぐ仕組みをつくることで、ワンストップに近いサービスをつくっていくことが必要ではないかと思っています。

5つ目は費用です。ユーザー視点でいえば費用負担は低い方がいいですが、運用側の視点でいうと、インフラの維持・運用、セキュリティ確保などにおいては多大な費用がかかるという現実があります。持続可能なデータ公開のためには有料化を視野に入れることも必要であり、データを集めるにも公開するにも費用が発生する時代が来るのではないかと思っています。こういった費用が大きくなればなるほど、研究活動そのものにも負担が生じ、影響が出てくる恐れがあります。こういったところも視野に入れながら、うまいデータ公開の仕組みを考える必要があると思っています。

とソフト（契約・規則など）を両輪としてデータ公開の枠組みをつくっていくことが大切だと感じています。

それから、安心・安全なデータ公開を実現するためには、やはりお金がかかります。持続可能なデータ公開のためには、インフラの維持・運用、セキュリティ確保にかかる費用も考えていく必要があると思います。

最後に、当発表に当たり多くのアドバイスとお力添えを頂きました当センターの谷藤センター長、データサービスチームの田沼チームリーダー、桑島さんに御礼を申し上げます。

まとめ

材料分野は産業や知財との関連が強く、研究データはオープンにされにくい状況がありますが、研究の新たな発展やイノベーションを促進するためには、研究データを戦略的にオープンにしていくことが求められます。

また、研究データを保護しつつ利活用を図るためには、データ公開に際し、適切なアクセス制限や利用条件を付すことも重要です。研究データの使われ方や利活用シナリオを念頭に置いた上で、ハード（インフラ）

4. 研究データ公開における理想と課題 	
●データの提供方法	
理想：非公開～公開までワンストップサービスで提供	
課題：データの形式や性質は様々でなく、ステークホルダーも多い。 1つのシステムで多様なデータに対応することは、技術とコストの問題もあり困難。 ルールの面でも、ステークホルダーが多くなるほど調整は難しくなる。 サービスごとに対応しつつ、これらをシームレスにつなぐ仕組みが必要。	
●費用	
理想：ユーザー視点では、費用負担は低いほうがいい。	
課題：運用側の視点では、インフラの維持・運用、セキュリティ確保などには多大な費用がかかる。 持続可能なデータ公開のためには、有料化を視野にいれることも必要か。 データを集めるにも公開するにも、費用が発生する時代が来る？ データ公開・利用にかかる費用の増大は、研究活動そのものにも影響がでる恐れあり。	

(図 13)