SPARC Japan セミナー2024

「オープンアクセス義務化の先にあるもの：来るべき世界に向けて」

# ライフサイエンスにおけるオープンアクセスの歴史

川島 秀一

（情報・システム研究機構ライフサイエンス統合データベースセンター）

**講演要旨**

ライフサイエンス分野におけるオープンアクセスは、科学の進展と国際的な協力を促進するために重要な役割を果たしてきました。本講演では、ヒトゲノム計画を契機として発展したゲノムデータ共有の歴史を振り返り、GenBank や PubMed など初期のオープンデータベースの成功事例を通じて、オープンデータの現状を紹介します。また、FAIR 原則がオープンデータの適切な公開と利活用をどのように支えているかを明らかにします。さらに、技術的および倫理的課題を考察し、AI やビッグデータ解析が実現する未来の可能性を展望します。

**川島 秀一**

**情報・システム研究機構データサイエンス共同利用基盤施設ライフサイエンス統合データベースセンター特任准教授。博士（科学）。京都大学化学研究所、東京大学医科学研究所ヒトゲノム解析センターを経て2012年よりDBCLSにて、生命科学分野のデータベース構築、データベースの統合化技術開発などに取り組んでいる。**

私は 2024 年度より、前任者から引き継いで SPARC Japan セミナーの企画ワーキンググループに参加しています。ジャーナルのオープンアクセス（OA）に関しては無知なのですが、本日は仕事柄使うことが多い、ライフサイエンス分野のオープンデータについてご紹介いたします。

## OA の重要性

皆さまもご存知のとおり、OA とは、学術研究の成果を、無料で誰でもインターネット上でアクセスできるようにする公開モデルのことです。ライフサイエンス分野では、OA は非常に重要視されており、実際にそれがうまく回っています。

まず、オープンであることで研究が加速します。有名な例では、ヒトゲノムプロジェクトや 2024 年にノーベル化学賞を受賞した AlphaFold2（アルファフォールド 2）（タンパク質の構造を予測する AI モデルの開発）は、まさにオープンデータに依存した研究です。また、疾病研究・公衆衛生の分野の貢献として記憶に新しいところですが、COVID-19 のゲノム配列が即座に公開され、その結果、迅速なワクチン提供につながった例が挙げられます。また、研究においては再現性が重要視されますが、データがオープンになっていることで第三者が研究を再解析できることが保証されている点でも重要です。

ライフサイエンス分野では、論文を発表する際に、例えば遺伝子配列データの場合、国際塩基配列データベース（International Nucleotide Sequence Database：INSD）、タンパク質構造データは蛋白質構造データバンク（Protein Data Bank：PDB）など、公共データベ

ースへの登録が義務付けられているケースも多いです。そのようなことから、研究とデータのオープン化、公共化がうまく実現しています。

## 生命科学分野におけるオープンデータの歴史

ライフサイエンス分野におけるオープンデータの歴史を少しご紹介します（図 1）。この分野のオープンデータは、1879 年に創刊した米国の軍医総監局図書館の Index Medicus に始まります。これは医学文献をコレクションしたもので、当時はまだ紙媒体（書籍）でした。その後、1964 年に米国国立医学図書館（NLM）に移管され、MEDLARS と名前が変わります。このころ既にコンピュータ情報になっています。それが 1971 年には MEDLINE と名前を変えてオンライン化を実現しています。つまり、医学系論文は 50 年以上前に一応オンラインでの利用が可能になっているのです。これが後にインターネット（WWW）で公開され、現在は PubMed という形で利用されています。

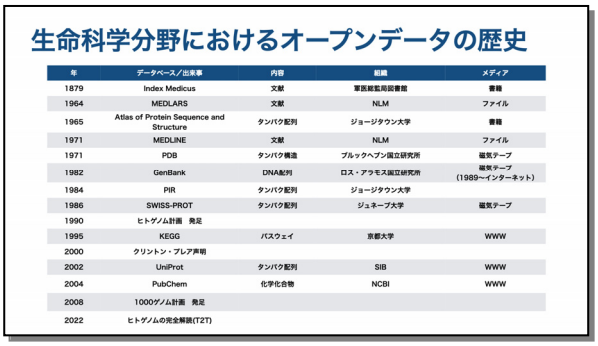
一方で、たんぱく質の分子データに関するオープンデータとしては、PDB が 1971 年に開始されました。これはタンパク質の立体構造を集めたデータベースで、現在に至るまで開発が続けられ、維持されています。また、遺伝子配列情報では、GenBank というデータベースが 1982 年に開発され、現在も続いています。ライフサイエンス分野では、このような長い歴史があります。時代ごとに必要になったさまざまな分子データや病気のデータがデータベース化され、基本的にはオープンデータとして誰でも閲覧可能となっています。

特に皆さまもよくご存知と思いますが、1990 年にはヒトゲノム計画が始まり、2000 年にクリントン・ブレア声明としてドラフトゲノムが公開、2003 年には完全ゲノムが公開されました。ゲノムプロジェクトは、特に国際協調やテクノロジーの進歩などと連動して、加速度的にデータを増やしました。ヒトゲノムは 2022 年に真の意味での完全解読がなされたと言われます。最初のヒトゲノムでは 10 年以上かけてヒト 1 人分のゲノムを解読しましたが、今では 1,000 人ゲノム、50 万人ゲノム、100 万人ゲノムというように、急速にデータが増えている状況です。

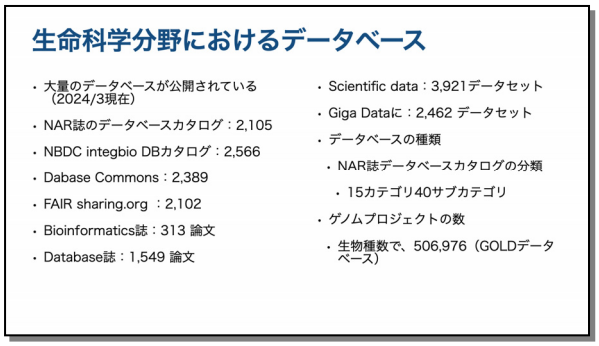
## 生命科学分野におけるデータベース

ライフサイエンス分野では、さまざまなデータがあり、それらが別々のデータベースで公開されています（図 2）。一説には 1 万以上のデータベースがあると聞いたことがありますが、根拠が見つけられませんでしたので、ここでは具体的な数字を示せる情報でご紹介します。

例えば、2024 年 3 月現在で、研究論文誌『Nucleic Acids Research』が作成しているバイオのデータベースカタログには約 2,100 のデータベースが公開されています。また、科学技術振興機構（JST）バイオサイエンスデータベースセンター（NBDC）の Integbio データベースカタログには約 2,500 のデータベースが公開されています。さらに、最近ではデータそのものを投稿するデータジャーナルも出てきています。代表的なものとして Scientific data や Giga Data があり、何千というデータが公開されています。




（図 1）




（図 2）

## ヒトゲノム計画とオープンデータの黎明期

この分野において、特に「ヒトゲノム計画」は、データの量を加速する意味でも、それを再利用する意味でも重要な存在です。最初のヒトゲノム計画は 1990 年に始まりました。私はアカデミアの人間として一般的にオープンデータ、オープンアクセスは素晴らしいと思っていますが、公的資金による研究でもあり、単純に必ずしも全員一致でオープンにしようとはなっていません。やはり特許の問題などがあり、当初は公開するか否かが大きな議論になりました。最終的に、バミューダ原則（1996 年 2 月）というものが宣言され、基本的には研究データを迅速に公開して誰でも利用できる形となり現在に至っています。

## バミューダ原則と FAIR 原則

バミューダ原則は三つの原則から成っています（図 3）。第一に、「ゲノムデータは、速やかに公開されるべき」ということです。シーケンシングされた DNA 配列は 24 時間以内に公開データベースに登録することが推奨されています。

次に、「すべての研究者がデータを自由に利用できるべき」であることです。あらゆる人がゲノムデータに無料でアクセスできることを保障しなければなりません。

三つ目は、「データ共有が科学の進歩を加速する」とうたっています。競争よりも国際協力を優先するということです。ゲノム情報を多くの研究者が利用できる環境を整えなければいけません。

この宣言が国際プロジェクトの原則として（参加国に）共有されることで、研究データの共有と再利用が盛んに行われているのではないかと思います。

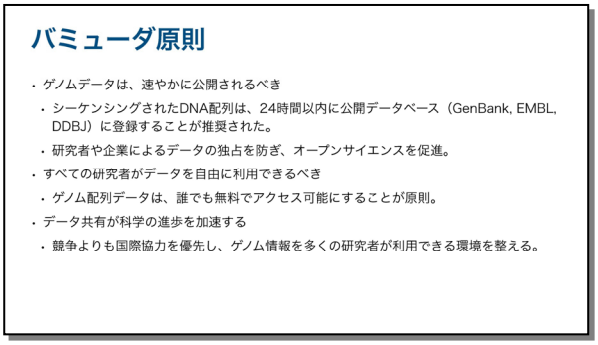
また、2016 年に公開された FAIR 原則は、データ集約型科学を推進するために必要なデータ共有の原則をまとめたもので、特に欧州では広く共有されています（図 4）。FAIR とは、データを「Findable（見つけられる）」「Accessible（アクセスできる）」「Interoperable（相互運用できる）」「Reusable（再利用できる）」という 4 原則の頭文字の略です。

例えば、Horizon Europe（欧州連合（EU）の研究・イノベーション枠組みプログラム）は予算規模 955 億ユーロほどの大型のプロジェクト（2021-2027 年）ですが、研究データの管理・公開の要件として FAIR 原則に従うことを規定しています。

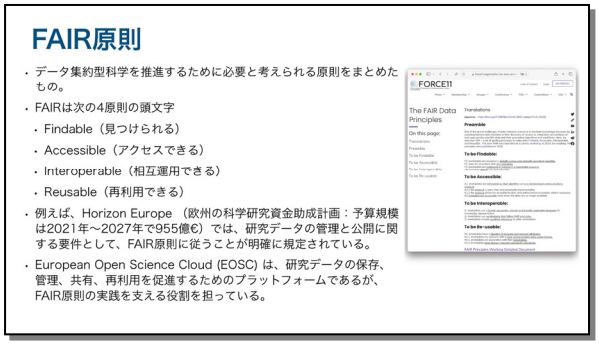
## 国際社会における対立とオープンデータの課題

一方で、よいことばかりではありません。今まで、生物から取得された遺伝情報をデジタル化したデータ（Digital Sequence Information：DSI）はオープンにされていますが、生き物を提供している側である途上国には、そのデータから得られる利益は配分されていません。データの商業利用に対する公平な利益共有に関する途上国からの主張は、生物多様性条約締約国会議（CBD-COP）で取り上げられ、南北問題に発展しています。

先進国は、イノベーションのためにはデータは共有すべきだと主張しています。一方で途上国は、その（商業利用に対する）利益を得ておらず、DSI 利用者が利益を公平に共有する仕組みを作るべきだと主張し




（図 3）

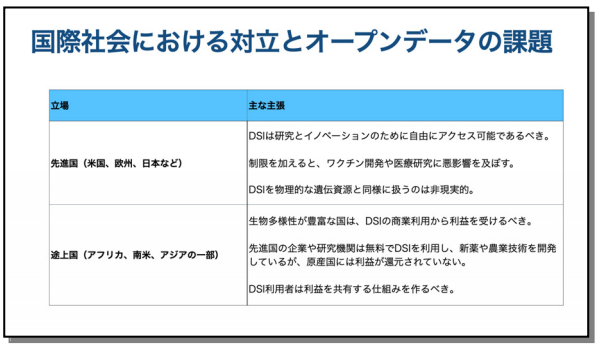



（図 4）

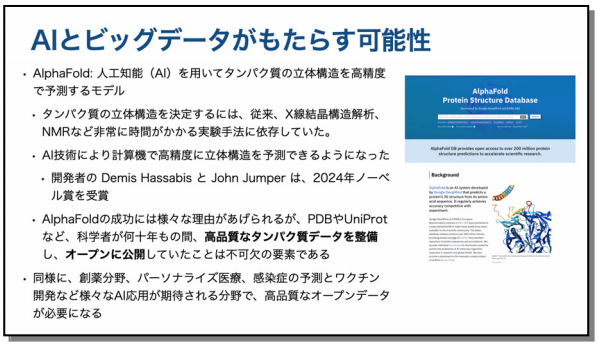
ています。現在も議論が続いているところですが、何かしらの対応が必要となるでしょう（図 5）。

## AI とビッグデータがもたらす可能性

最後に将来的な話をします。2024 年、AlphaFold2 という、AI を使ってタンパク質の立体構造を高精度で予測するモデルの開発者がノーベル化学賞を授賞しました。AlphaFold2 の成功にはさまざまな要因が挙げられますが、科学者が何十年にもわたって PDB や UniProt などの高品質なタンパク質データをメンテナンスし、オープンに公開していたことは不可欠な要素と言えます。やはりオープンデータは人類の科学の発展には必要不可欠な要素ですので、うまく落とし所を見つけて継続していくべきではないかと思います（図 6）。




（図 5）




（図 6）