生命科学分野におけるデータの共有の現状と課題

Data sharing in the life sciences: How far we have come and where we must go

高祖 歩美¹

KOSO Ayumi¹

.....

1 独立行政法人科学技術振興機構 バイオサイエンスデータベースセンター (〒102-8666 東京都千代田区四番町5-3)
Tel: 03-5214-8491 E-Mail: a2koso@jst.go.jp

1 National Bioscience Database Center (NBDC), Japan Science and Technology Agency (JST) (5-3 Yonbancho Chiyoda-ku, Tokyo 102-8666)

原稿受理 (2013-06-07)

情報管理 56(5), 294-301, doi: 10.1241/johokanri.56.294 (<http://dx.doi.org/10.1241/johokanri.56.294>)

著者抄録

実験や計測を通して得られたデータを広く研究者間で公開して、利活用できれば、重複したテーマに研究費が配分されることが減り、過去の研究結果の追試や再現が容易になり、科学技術の進歩が加速され、新しいビジネスや経済効果につながると考えられている。生命科学分野では、ヒトの全ゲノム情報を解読した国際的なヒトゲノムプロジェクトを契機として、研究者が保有するデータを公開して、皆で利活用するデータ共有が進められてきた。本稿では、科学技術振興機構 バイオサイエンスデータベースセンターが中心となって推進してきたわが国におけるデータ共有の取り組みと課題について概説する。各国のファンディング機関や政府が採っているデータ共有の方針や他人が利用できるデータが満たすべき3つの属性の課題（形式の課題、意味の課題、法上の課題）を中心に紹介する。

キーワード

生命科学, データ共有, 再利用, データベース, データ共有の方針, 公募要領, データ形式, メタデータ, 利用規約

1. はじめに

政府の保有するデータを、皆が利活用できるように利用条件を緩和して公開することによって、新しいサービスや雇用の創出、社会的な課題の解決に役立てようとする「オープンデータ」の取り組みがわが国でも始まりつつある。2013年1月に『情報管理』に掲載された「オープンデータ」の記事を記憶して

いる読者も多いことだろう¹⁾。データの公開と利活用は、科学の世界、とりわけ生命科学分野では、ヒトのゲノムが解読された2000年はじめ頃から盛んに取り組まれ、データ共有 (data sharing) とも称される。生命科学分野におけるデータ共有の1つの成功事例として、近年米国で始まった遺伝子解析の事業が挙げられる。インターネット上で、キットを注文して唾液を採取し返送するだけで、病気のかかりやすさ、

薬に対する感受性や身体的特徴、体質などを自分の遺伝子情報に基づいて分析してもらえる。このサービスの利用者は、世界的に25万人に及ぶ^{注1)}。実験や計測を通して得られたデータを広く研究者間で公開して利活用できれば、過去の研究結果の追試や再現が容易になり、重複したテーマに研究費が配分されることが減り、科学技術の進歩が加速され、新しいビジネスや経済効果につながると考えられている。

本稿では、生命科学分野におけるデータ共有を促進するためになされてきた取り組みを紹介する。わが国では、データ共有に必要な仕組みの整備は、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）が、その前身のプロジェクト（文部科学省統合データベースプロジェクト）の取り組みを引き継いでデータ共有に必要な仕組みの整備を進めてきた。NBDCは2011（平成23）年にJST内に設置され、生命科学分野のデータベースを使いやすくなるための研究開発やサービス提供を行っている。これらの活動やNBDC設立までの歴史的な経緯についてご興味のある方は、先行文献を参照されたい^{2), 3)}。

2. 国内外のデータ共有の方針

現在の生命科学分野におけるデータの共有と再利用を方向付けたプロジェクトとして、ヒトゲノムプロジェクトの果たした役割は大きい。ヒトゲノムプロジェクトは、欧米や日本を始めとする多数の国が参加して、ヒトの全遺伝子の配列を解読した国際的なプロジェクトである。本プロジェクトでは、解読された遺伝子配列はパブリックドメインに帰するものとされ、関係者はそのデータを解析後24時間以内に公開して、誰でも自由に利用できるようにすること（バミューダ原則）とした^{注2)}。

このバミューダ原則という取り決めは、データ共有の観点から非常に意義深い。バミューダ原則によって、一研究グループ、あるいは一企業に遺伝子配列の情報が独占されることはなく、インターネットに

接続することができれば誰でもヒトゲノムプロジェクトを通して得られたデータを利用することができることとなった。この精神は、現在の生命科学分野にも引き継がれている。例えば、米国のファンディング機関である国立衛生研究所（National Institutes of Health: NIH）や米国国立科学財団（National Science Foundation: NSF）が助成する研究費を申請する場合、その研究費を使って取得するデータをどのように共有し、管理するのかを記した計画書を提出しなければならない^{注3)}。これらの計画の進捗は研究費が執行されている間、確認されるほか、計画の実施が次の研究費の採択・不採択にも勘案される。加えて、2013年1月14日より、NSFのデータ共有の方針は強化され、計画書が添付されていない申請書は、不採択、あるいは審査せずに返却されることとなった^{注4)}。英国のウェルカム・トラスト（Wellcome Trust）や英国バイオテクノロジー・生物科学研究会議（Biotechnology and Biological Sciences Research Council: BBSRC）においても同様のデータ共有の方針が打ち出されている^{注5)}。

また、欧州において2014年から開始される研究とイノベーションに関する800億ユーロ規模のプログラム「Horizon2020」では、オープンデータを推進し、そのための情報基盤の開発や支援を行うことが発表されている^{注6)}。このように欧米の生命科学分野に関する主要なファンディング機関では、研究費の申請者に対して、公的な支援を受けて産出されるデータをどのように共有するのか明らかにすることを申請の段階から義務づけている。欧米ではこの制度的な要請によって、研究者がデータを共有することを間接的に後押ししている。

対するわが国はどうであろうか。わが国では、研究費の申請時にデータ共有に関する計画書の提出は義務づけられていない。しかし、文部科学省やJST、厚生労働省の助成を受けてなされた研究については、論文発表等で公表された成果にかかわる生データの複製物、または構築した公開用データベースの複製

物をNBDCに提供いただけるよう、それぞれの公募要領に記載している。このようなデータ提供への協力依頼は、2008（平成20年度）に実施された文部科学省の委託事業の公募要領にはじめて掲載された。その後、表1に示すように2013（平成25）年の4月までにその数や対象となる補助金や事業の種類が増えつつある。

3. 他人が再利用できるデータ

各国の政府やファンディング機関が推進するデータを共有すること、自分で提供したデータを他人が理解して再利用できることとは、具体的にはどのようなことなのか。例えば、数字の羅列されたファイル（データ）をインターネット上にアップロードしておけばよいというものではない。少なくともデータの3つの属性に関する課題を解決して、データを提供する必要がある。1つ目は、特定のアプリケーションやソフトウェアに依存しない形式でデータが提供されていること（形式の課題）、2つ目は、何のデータであるか、取得した実験条件や機械の情報、測定単位など、データの文脈が提供されていること（意味の課題）、そして、3つ目はどのようなライセンス（利用規約）のもとでデータが利用できるかが表明されていること（法上の課題）である。加えて、NBDC

ではこれらのデータの3つの属性を、人が人の言語を介して理解できるようにするだけでなく、機械も同様に判別して理解できるように、セマンティックWeb化する取り組みを行っている。そうすることによって、今よりも高度な検索や推論機能を機械で実現できるようになると考えられている。以下、3つの課題を詳細にみていく。

3.1 形式の課題

生命科学分野において、実験や測定を通して得られたデータはデータベースとして公開されることが多い。データベースに収められているデータは、カテゴリーから絞り込んで、あるいは検索キーワードから、断片的にアクセスするのが一般的である。反対にデータベースに収められているデータが丸ごとダウンロードできる形で置かれていることは稀である。そこで、NBDCでは既にデータベース化されたデータや種々の理由によりデータベースが閉鎖されたため、提供が終わってしまったデータを、丸ごとダウンロードできるようにしている（「生命科学系データベースアーカイブ」。以下、アーカイブ（<http://dbarchive.biosciencedbc.jp/>）（図1）。具体的には、人が理解しやすく、特定のアプリケーションやソフトウェアに依存しない形式でデータを提供するために、サービス提供の当初からデータをCSV形式でZIP

表1 公募要領にデータ提供の協力依頼が掲載されている事業や補助金の一覧

事業や補助金の名称	協力依頼がはじめて掲載された時期	備考
文部科学省委託事業	平成 20（2008）年度～	
JST 戦略的創造研究推進事業 （さきがけ、CREST）	平成 23（2011）年度～	
厚生労働省 科学研究費補助金	平成 23（2011）年度～	「人体に由来するデータを取り扱う研究は除く」とのただし書きあり
科学研究費助成事業 特別推進研究 基盤研究（S・A・B・C） 挑戦的萌芽研究 若手研究（A・B）	平成 24（2012）年度～	
JST 戦略的創造研究推進事業 先端的低炭素化技術開（ALCA）	平成 25（2013）年度～

多様なデータ形式

データをCSV, JSON, Turtle, RDF/XMLの形式で提供している

メタデータの付与

データベースやデータの説明情報(メタデータ)を付与している

ライセンスの明示

データの利用条件について、標準ライセンス(CC 表示-継承 2.1 日本)の付与を推奨している

図1 生命科学系データベースアーカイブの特徴

ファイルとしてダウンロードできるようにしている。また2013(平成25)年はじめ頃からは、機械がデータの属性を理解できるような形式に対応するため、データをJSONと2種類のRDF(TurtleとXML)形式でも提供している。

3.2 意味の課題

他人がデータを再利用するため、理解するためには、例えば、そのデータが取得された実験条件や機械の情報、測定単位などの説明情報(メタデータ)は欠かせない。言い換えると、データに文脈が与えられていなければ、公開されたデータは単なる数値の羅列でありデータを取得した本人以外は利用できない。このような点を踏まえて、アーカイブでは、データごとに2種類のメタデータを付与している。1つ目は、データの由来するデータベースについてのメタデータである。このメタデータには、データベースの名称、データベースの作成者の氏名や所属、連絡先、扱っている生物種、関連する文献の情報などの記述が含まれる。2つ目は、データそのものに対するメタ

データで、データ名、ファイルサイズ、データ取得方法、解析方法、データ件数などが含まれる。これら2種類のメタデータは、人がデータの意味を理解できるように、さらにはデータの再利用を促進するために付与している。人はメタデータの記載に表記揺れや別称が含まれていた場合でも、その意味するところを理解できる。同様に、概念間の関係性(同義なのか、上位下位の関係になるのかなど)も、それまでに蓄積された知識を頼りに理解することができる。しかし、機械に人と同じようにデータを理解させるためには、データの意味やデータとデータの関係性(例えば、あるデータでAと記述されている概念は別のデータでBと記述される関係と同じなのか、異なるのか、上位下位概念にあるのかなど)を理解できるようにするための表現、オントロジーが必須となる。NBDCでは生命科学分野におけるオントロジーの開発も支援しており、例えば、微生物の生息環境を記述するために開発されたMEO(Metagenome/Microbes Environmental Ontology)は、生命科学分野のオントロジーが世界的に集積されている

BioPortalからも提供されている (Bio Portal: <http://bioportal.bioontology.org/>, MEO: <http://bioportal.bioontology.org/ontologies/3009>)。

3.3 法上の課題

データを商用の目的をもって利用してよいのか、改変は許されるのか、加工後のデータを公表してもよいのかなど、どのような条件のもとでデータの利用が許されているかわからなければ、他人はデータを利用できない。そこで、アーカイブでは、クリエイティブ・コモンズ・ライセンス (CCライセンス) の表示-継承日本2.1を標準的なライセンスとして採用して、そのライセンスのもとでデータを提供している。

CCライセンスとは、インターネット上に公開した自分の作品を第三者がどのような条件で利用可能かをマークで表せるツールである。4つのマーク (表示, 継承, 非営利, 改変禁止) を組み合わせた6つのライセンスがあり、提供者は選択したライセンスに

該当するマークを作品に付与して公開し、利用条件を視覚的に示すことができる^{注7)}。利用者は表示されたマークを見れば、作品の利用規約がわかり、提供者へ問い合わせることなく利用できる (図2)。また、CCライセンスの特徴は、機械にライセンスの意味がわかるような様式で提供されている点にもある。アーカイブで推奨しているCC表示-継承は、(1) 作品の原作者の氏名、作品タイトルとURLを表示し、(2) 改変した作品を公開する際にCC表示-継承を付与することを条件に、例えば、作品を複製したり、改変したり、商用目的で利用することが許される。一方で、営利目的で利用することを禁止するライセンスや改変を禁止するライセンスの選択は推奨していない。なぜなら、公的な支援を受けて産出され公開されたデータは、その性質上、一定の団体やグループの利用に限ることなく、誰もが等しく利用できるべきからである。改変を禁止することについては、先人が築いた知識の上にさらに新しい知識を積み重ねることが科学の営みであり、特徴であることから、こ

図2 クリエイティブ・コモンズ・ライセンス (CCライセンス) の種類と利用のしやすさ

れに反すべきではないとの理由による。

3.4 その他の課題

前節までは、他人が取得したデータを誰もが自由に再利用できるようにするために最低限クリアすべき課題について取り上げた。一方で、生命科学分野で扱うデータには、ヒトに関するデータ、例えば個人のゲノム情報や脳画像、健康情報なども含まれることがある。これらのデータは、個人情報の保護や倫理的な観点から、データの特性に応じて適切なアクセス制限や情報セキュリティ対策を講じた上で共有され、再利用される仕組みが望ましい。NBDCでは、これらの点を踏まえて、多様なヒトのデータを共有するためのガイドラインを策定して公開している (<http://humandbs.biosciencedbc.jp/guidelines>)。本ガイドラインでは、データの提供と利用のルールと、データの取り扱いに必要な情報セキュリティ対策の基準を定めている。このガイドラインに基づいて、近くヒトに関するデータの提供も開始する。

4. 今後の課題

このようにNBDCでは、生命科学分野のデータを共有するために解決しなければならない課題に取り組み、研究者がよりデータを共有しやすい環境や仕組みを作ることを目指してきた。しかし、データを共有することは、研究者にとっていまだにハードルが高い。2008（平成20）年の終わり頃に行われた、欧州における研究のためのデジタル情報基盤の整備に向けたアンケート調査（欧州、米国など各国の研究者1,400名が参加。参加者は、生命科学分野の研究者に限らず、あらゆる学術分野の研究者が含まれた）^{注8)}によれば、研究者の大多数は、他の研究者が取得したデータを利用したいと答えている。一方で、自分で取得したデータを誰もが利活用できるように共有していると答えたのは回答者の25%と報告されている。このように研究者にとって他人のデータを再利用する

ことと自らのデータを提供することの意識には大きなギャップがある。このような意識の差は、先のアンケート調査から5年経った今でも変わらない。その大きな要因の1つにデータを共有することのインセンティブが高くないことが挙げられる。まず、制度面において欧米と比較した場合に、わが国の公募要領におけるデータ提供の協力依頼は、その強制力が弱い。わが国では研究費を申請する際にデータをどのように共有するか、計画を提出する義務もなければ、データを共有した実績が研究費の獲得に影響することもない。次に、研究者は発表した論文の数や質で評価され、共有したデータの数や質では評価されない点も理由の1つとして挙げられる。現在の科学では、学術雑誌に発表された論文が評価されるように、共有したデータについて評価するための指標もなければ、データを提供した者の功績を認めるための引用という習慣も確立されていない。このような問題を解決しようと、トムソン・ロイター社はData Citation Indexと呼ばれる仕組みを開発して、同社のWeb of Scienceにおいて2012年の後半頃から提供し始めた^{注9)}。このプラットフォームでは、データをあたかも論文のように扱っており、例えば、データごとにその作成者や関連論文、利用回数などの情報を得ることができる。さらに、データを利用した場合の引用方法を具体的に提示して、データの引用を促進しようと試みている。その他にも、共有されたデータを評価する指標として、Altmetricsの活用も期待される。Altmetricsとは、ソーシャルメディアを活用した新しい指標で、論文だけではなくデータなどの研究の成果物が、社会や研究コミュニティに与えた影響を定量的に測定しようとするものである^{4), 5)}。このような仕組みや指標が現れることによって、データを共有することの価値が見直され、その評価も高まることが期待される。一方で、今後も、データを提供することのインセンティブを高める仕組みや制度の設計に継続的に取り組む必要がある。

2つ目の課題は、データの利用規約に関するもので

ある。生命科学分野にかかわらず、データの利用規約としてCCライセンスを付与してデータを共有する傾向が世界的に見られる。一方で、CCライセンスはもともと著作物の利用を促進するために設計されており、データを対象としていない。加えて、データには数値、文字、文章、図、音声、動画などがあり、どこまでの範囲のものに著作権、そして著作権が認められるかがはっきりしない。わが国では事実には著作権は認められないため、CCライセンスのもと提供されているデータであっても実質的に、その多くが保護の対象外である懸念が残る。これらの問題を踏まえて、クリエイティブ・コモンズでは科学データはCC0（ゼロ）で提供し、科学界に特有の引用という慣習でデータを共有した者の功績を認めるべき、と提案している。CC0とは、著作権および著作権に関する諸権利の行使を、法律で認められる限り、放棄、または差し控えるというライセンスである。しかし、現状は、CC0よりもその他のCCライセンスを付して科学データを提供の方が一般的である。このため、今後もデータの共有に最適なライセンスの議論が注目される。

5. まとめ

本稿では、生命科学分野におけるデータの共有を実現するために、NBDCにおいてどのような取り組み

を行ってきたかを、欧米の状況を踏まえながら概説した。また、今日のように至るまでに5年以上の歳月がかかっており、関係者の努力の賜物であることをここに記したい。ここまで読んでくださった読者には生命科学分野におけるデータ共有は道半ばであること、多くの方の支援や協力なくして進まない営みであることを少しでもご理解いただければ幸いである。

研究者が産出したさまざまなデータが広く公開されて、利活用される世界では、今は想像もできない発見や新しい価値に富んでいる。そんなワクワクするような未来を描いて筆者は生命科学分野におけるデータ共有の進展に携わってきた。もし、本稿を読んで共感いただける方が1人でもいるようであれば、ぜひ、生命科学分野におけるデータ共有と一緒に取り組んでいただきたい。

謝辞

本稿の執筆にあたり、大学共同利用機関法人 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の川本祥子氏に貴重な意見やコメントをいただいた。ここに感謝の意を表する。また、図2はクリエイティブ・コモンズ・ジャパン (<http://creativecommons.jp/licenses/>) のWebサイトからCC-BYにより提供されている図を使用した。

本文の注

- 注1) 23andMe. <https://www.23andme.com/>, (accessed 2013-05-20).
- 注2) Human Genome Project. "Bermuda Sequence Policies". http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml, (accessed 2013-05-20).
- 注3) National Institutes of Health. "NIH Data Sharing Policy". http://grants.nih.gov/grants/policy/data_sharing/, National Science Foundation. "Dissemination and Sharing of Research Results". <http://www.nsf.gov/bfa/dias/policy/dmp.jsp>, (accessed 2013-05-20).
- 注4) National Science Foundation. "Grants.gov Application Guide". <http://www.nsf.gov/pubs/policydocs/grantsgovguide0113.pdf>, (accessed 2013-05-20).

- 注5) European Union. "Scientific data: open access to research results will boost Europe's innovation capacity". (Press Release). http://europa.eu/rapid/press-release_IP-12-790_en.htm, (accessed 2013-05-20).
- 注6) Wellcome Trust. "Data management and sharing". <http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Data-management-and-sharing/index.htm>, Biotechnology and Biological Sciences Research Council, "Data sharing policy". <http://www.bbsrc.ac.uk/organisation/policies/position/policy/data-sharing-policy.aspx>, (accessed 2013-05-20).
- 注7) クリエイティブ・コモンズ・ジャパン. "クリエイティブ・コモンズ・ライセンスとは". <http://creativecommons.jp/licenses/>, (accessed 2013-05-20).
- 注8) Permanent Access to the Records of Science in Europe. "Insight into digital preservation of research output of in Europe". http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf, (accessed 2013-05-20).
- 注9) Thomson Reuters, Web of Knowledge. "Data Citation Index". http://wokinfo.com/products_tools/multidisciplinary/dci/, (accessed 2013-05-20)

参考文献

- 1) 高木聡一郎. 欧州におけるオープンデータ政策の最新動向. 情報管理. 2013, vol. 55, no. 10, p. 746-753.
- 2) 白木澤佳子, 高木利久. ライフサイエンス分野のデータベース統合を目指して: バイオサイエンスデータベースセンター (NBDC) の発足. 情報管理. 2011, vol. 54, no. 3, p. 144-151.
- 3) 畠中秀樹. 生命科学系データベース統合に向けた取り組みの現状と課題. 情報管理. 2012, vol. 55, no. 5, p. 347-353.
- 4) 坂東慶太. Altmetricsの可能性: ソーシャルメディアを活用した研究評価指標. 情報管理. 2012, vol. 55, no. 9, p. 638-646.
- 5) 林和弘. "研究論文の影響度を測定する新しい動き: 論文単位で即時かつ多面的な測定を可能とする Altmetrics". 科学技術動向2013年 3・4月号. 科学技術政策研究所, 2013, レポート2, p.20-29, <http://www.nistep.go.jp/wp/wp-content/uploads/NISTEP-STT134J-2.pdf>, (accessed 2013-05-20).

Author Abstract

Data sharing has the benefits of reducing duplications in the grant awarding process, increasing the reproducibility of research results, speeding up research, and fostering new businesses. The Human Genome Project, an international research project that set out to determine the DNA sequence of the human genome, was a milestone in the history of data sharing in the life sciences. This article gives an overview of the data sharing activities that have been carried out by the National Bioscience Database Center, Japan Science and Technology Agency. It will highlight on Japan's data sharing policy in comparison with that of the United States and Europe, and on the three conditions that must be fulfilled for data to be truly reusable.

Key words

life science, data sharing, reuse, database, data sharing policy, application guideline, data format, metadata, license