第 2 回 SPARC Japan セミナー2017

「プレプリントとオープンアクセス」

生命科学分野におけるプレプリントの位置付けや経験について，統合 TV について

小野 浩雅

（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

生命科学分野では、2013 年にはコールド・スプリング・ハーバー研究所による「bioRxiv」が立ち上がったことを契機に、ここ数年でプレプリントの活用が進んでいる。査読前の論文をあらかじめオープンに共有することで、特に、日々刻々と変化する研究情勢にキャッチアップする必要のある研究者にとって必要不可欠な基盤となりつつあり、周囲の研究者の間では、プレプリントの話題になることが増えてきている。演者自身はプレプリント投稿の体験はまだ無いが、論文投稿の際にあらかじめオープンレポジトリに全てのデータを登録・公開する経験をし、今後このようなプロセスが標準となるのではないかと考えている。

小野 浩雅

日本大学大学院生物資源科学研究科博士後期課程単位取得退学。在籍中は哺乳類細胞における脱分化機構の網羅的解析をテーマに研究。2010年より情報・システム研究機構(ROIS)ライフサイエンス統合データベースセンター(DBCLS) に勤務。生命科学分野の有用なDBの使い方を動画で紹介する「統合TV」の編集や遺伝子発現等の大規模データ解析および可視化、活用支援を行う。2012年6月より現職（特任助教）。博士（生物資源科学）。

先ほどの生長さんが化学分野だったので、私は生命科学分野のプレプリントの現状等についてご紹介したいと思います。

1.自己紹介

今はデータベースセンターでコンピュータを相棒にして仕事をしていますが、学部・大学院時代は、細胞などを相手にして顕微鏡を見ながら実験をするというスタイルで研究を進めてきました（図 1)。それを、生物業界だけかもしれませんが、wet 研究といいます。一方、コンピュータをメインに使って生物学の謎を解いていくことを dry 研究といって、最近は wet と dry の融合という話が頻繁に出てきます。

そういう意味では、私はもともと wet で仕事をしてきたというバックグラウンドがあります。皆さんもたくさんお持ちだと思いますが、脂を貯める脂肪細胞を

（図 1）

他の細胞に変えるという研究をしていました。一番分かりやすい例は山中伸弥先生の iPS 細胞です。あれも結局は分化転換といって、全能性のある他の細胞に変えるという研究でノーベル賞を取られたわけです。そこの一分野と捉えていただければと思います。

この分化転換、脱分化という現象は不思議な現象なのですが、十数年前にそこを追い掛けるにはコンピュータの力を借りないといけないという状況に直面しました。縁があって、今日のモデレーターである坊農さんのところに、コンピュータで実験する手法を学びに行ったというのが私の最初の出会いです。博士後期課程のときに、コンピュータを使ってこういう生命現象を解いていくということに取り組みました。

そして、データベースを使って生物学を解いていくことを推進するライフサイエンス統合データベースセンターが、ちょうど私が博士の学位を取るぐらいのときに出来上がり、面白そうだなと思いました。周りには優秀な wet の研究者がたくさんいたのですが、こういう研究をする研究者はあまりいないのではないかと思って飛び込んだのがきっかけです。

私は学生のころからアルバイトとしてライフサイエンス統合データベースセンターに参画していたのですが、当時、ライフサイエンス研究の有用なデータベースやウェブツールが出はじめてきていて、私はそれを動画で紹介する試み「統合 TV」を始めました。自分で動画を編集して、ストーリーをつくって公開していきました。

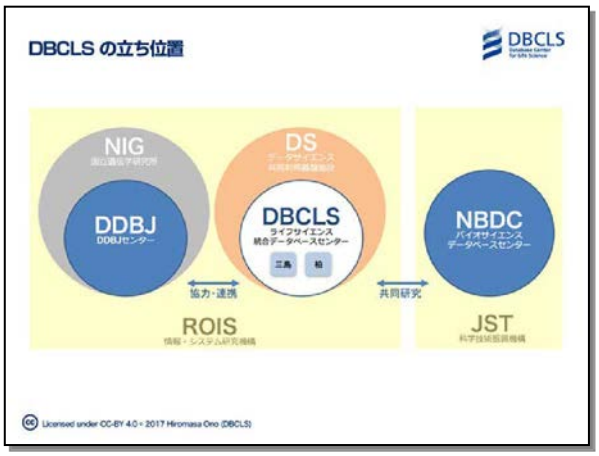
今年の 7 月か 8 月で開始から丸 10 年たち、動画が今 1,300 本に届こうというぐらい貯まっています。私がアルバイトで始めた当時は、1 日行って 1 本動画をつくるという形でしたが、今はいろいろなお手伝いをしてくださる方のご協力を頂いて、私自身はつくらず、編集者として、きちんとした内容になっているか、分かりやすくなっているかをチェックしたり、ウェブサイト自体を運営したりということをしています。最近言いはじめたのですが、これはまさに生命科学分野の「オープンエデュケーション」ツールではないかと自負しています。

もう一つ、生命科学の研究者が実際にどういうデータベースがあると日々の研究活動をより進められるようになるかということを考え、便利なデータベースをつくっていくことも、われわれのセンターの役目の一つです。われわれは実際に実験をしているわけではないのでデータを出すことはできないのですが、既に世界中に便利なオープンデータはたくさんあるので、それを再利用して、例えば、正常組織や細胞株の遺伝子発現データを簡単に検索できるデータベースなどをつくっています。クリックしていくだけで、例えばある遺伝子が心臓でどれぐらい出ているというのが簡単に分かります。このデータベースを RefEx と呼んでいます。

また、生命科学分野のデータベースをどうやって活用していくかというのを、われわれは基本的にはウェブで発信していくのですが、やはりフェース・ツー・フェースで会ってハンズオンにしないと、「これは便利だから使ってみよう」という話にはなかなかならないところがあるので、大学や研究所に出張して、データベースの講習会をよくやっています。そこで講師としてデータベース活用に関する教育もしています。

2.DBCLS の立ち位置

ライフサイエンス統合データベースセンターは、長いので DBCLS と呼ばれています（図 2)。国立情報学研究所と同様、情報・システム研究機構（ROIS）の



（図 2）

中にあります。似たような研究施設として、国立遺伝学研究所が静岡県三島市にあり、そこと兄弟研究所という形になっています。

DBCLS は三島と柏にラボがあります。DNA を研究者が解析したときは必ずどこかに登録しないといけないのですが、その日本における元締めが DDBJ という DNA のデータバンクで、遺伝学研究所の中にあります。DDBJ は毎日ものすごい量のデータが集まってくる施設です。われわれはデータベースセンターなので、これらの「ビッグデータ」を活用するために、三島にもラボを構えて DDBJ と密に連携しています。

もう一つ似たような名前のバイオサイエンスデータベースセンター（NBDC）というものがあります。NBDC は科学技術振興機構（JST）の傘下にあって、DBCLS はここと共同研究をしています。

謎の組織 DBCLS をどう紹介するのが適切かと考えたのですが、このセミナーは図書館系の方が今回多いと聞いたので、皆さん FAIR はご存じですよね。あの FAIR が実は、DBCLS の活動の中で出てきたと言ったら、どうですか。DBCLS は筋がいいなと思われた方がおられれば幸いです。見つけやすくて（Findable）、アクセスしやすくて（Accessible）、相互運用できて（Interoperable）、再利用できる（Reusable）、の頭文字を取った標語です。これについて書かれた論文が Scientific Data に出ています。

（https://www.nature.com/articles/sdata201618）

このファーストオーサーである Mark Wilkinson さんが DBCLS の片山俊明さんと仲が良く、二人の長年の構想で実現したのが、10 年前から行っている BioHackathon（バイオハッカソン）というイベントです。今年で 10 回目を迎えました。

Hackathon という言葉は最近 IT 系などでよく出てきます。Hack と marathon の造語で、1 週間ぐらい合宿で hacking し続けるのです。ハッキングとは悪い方のハッキングではなくて、コンピュータで仕事をすることを hack といいます。昼は議論しながら仕事をして、夜はお酒も飲みながら議論する、それを 1 週間ぶっ続けでやるイベントです。ここで FAIR という概念が大いに議論されたということがこの論文にも書いてあって、Acknowledgements に、NBCD/DBCLS BioHackathon 2015 のおかげでできたというようなことが書いてあります。そういう組織、集団だと思っていただければよいと思います。

3.統合 TV

3-1.概要

次に、統合 TV について紹介します。統合 TV は、生命科学分野の使えるデータベースやツールの使い方を動画で手取り足取り解説するウェブサイトで、図 3 のようなインターフェースになっています。生命科学研究を進める上で最低限知っておかなければいけないバイオ系のデータベースを網羅しています。

生命科学分野には、バイオインフォマティクスといって、コンピュータを使って生命科学現象を解き明かしていく分野があります。この分野は、先ほど wet と dry という言葉があると紹介したと思いますが、やりたくない人はやりたくない分野で、黒い画面に緑色の文字が流れてくると思考停止してしまう人がおられるのですが、そういう人に向けて、「実は怖くないのです、簡単にコピペをするだけで緑の画面を克服できるのです」ということも紹介しています。全ての動画に DOI を付けており、引用ができるようになっています。

この動画は YouTube に全て同じものが上がってお

（図 3）

り、おなじみのインターフェースで閲覧することができるようになっています。おかげさまで再生数も右肩上がりで、90 万弱まで来ています。

3-2.講習会の動画

動画の種類は、単純にスクリーンショットを撮って動画の使い方を紹介するというのが最初だったのですが、われわれはいろいろなところで講習会をしているので、その講習会の様子を動画で撮って、後から振り返れるようにアーカイブしています（図 4）。講習会では「今日は何をやります」という形で、説明書や手順書を用意されると思うのですが、それを実際にウェブサイトに載せておいて、動画を見ながら、手順書を使って、自分の手元のコンピュータで再現できる、再学習できるという用意をしています。このページには DOI が付いていて、再利用できます。

この分野には最新の解析手法があって、今、ヒトゲノムは非常に安い値段で解析できるようになっています。1,000 ドルゲノムといわれているぐらいで、昔は 1 個解析するのに何百億円とかかっていたのに、今はそれぐらいの値段で、しかも短時間で解析ができるようになりました。その一つのキーワードが次世代シーケンサー（NGS：next generation sequencer）です。これが開発されたのが大きなブレークスルーだったのですが、この解析手法が非常に難しいのです。

生物学はコンピュータを使わないと解析ができないレベルになってきています。その最近の解析手法の講習会の動画（図 5）は、コンピュータリテラシーとサーバ設計の話から始まっており、そういうレベルのデータを扱う必要が生命科学者の研究者の中にも出てきたということを意味しています。後の方になると生物学的な話になってきます。

この講習会は 3 年前に第 1 回があって、2014 年分は延べ 43 時間を超えています。43 時間も YouTube の動画をご覧になったことがありますか。1 本は 2 時間半や 3 時間で、これでも黙っている時間などをカットとしているのです。この講習会は、1 週間通しの集中講義形式で東大農学部で行いました。

NGS で職を得ることはできると思いますが、プロとしてやっていこうということであれば、番組を最初から最後まで見てやってみて、やっていけそうかどうかを試してみるといいと紹介しています。毎年、時間数が増えているのですが、今年もあって、絶賛編集中です。

3-3.生命科学分野の静止画素材

動画だけではなく、静止画の素材も統合 TV で提供しています。もともとは Togo Picture Gallery という名前の別のサービスでした。しかし、ウェブサイトが別だとなかなか普通の人はたどり着けないというご意見を頂き、それだったら有名な方の統合 TV にくっつけようということで、1 年ぐらい前に合体しました。

これもわれわれのセンターでコンテンツをどんどんつくっていくために、RA さんを呼んで手伝ってもら

（図 4）

（図 5）

いました。ある学生が、動画をつくるのはあまり上手ではなかったのですが、絵を描くのが非常に得意で、「絵でもいいから描いて」という軽い気持ちで始まったのですが、今や 500 本ぐらい素材がたまっています。これはクレジットを明記するだけで、どのように使っても、商売に使っても構いません。誰でも自由に、閲覧するだけではなく、使うことができる、ライフサイエンス分野の画像やイラストです。種類も生命科学分野に何かしら引っ掛かっていればよいというポリシーでやっているので、各種取りそろえていて、研究発表のスライドや資料などに使っていただけます。よく広告の挿絵に使いたいというお問い合わせも頂いています。

最近は細胞やマウスなどの 3D アニメーションも取り込んでいます。3D プリンターをお持ちの方はなかなかおられないと思うのですが、最近、街で 3D プリンターを貸し出しているところもありますので、使っていただくと楽しいと思います。

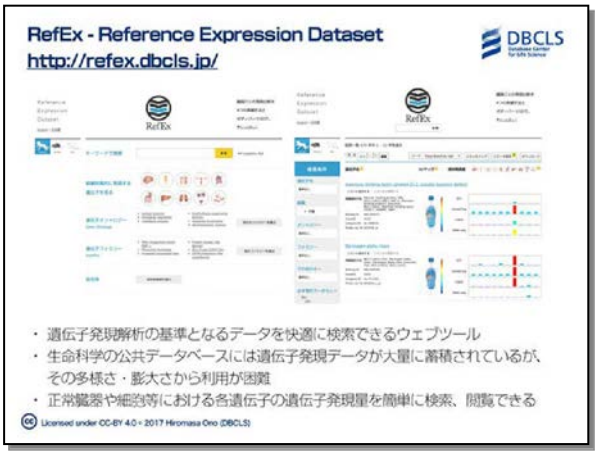
3-4.RefEx（Reference Expression Dataset）

冒頭で紹介した RefEx は、正常組織や細胞の遺伝子の発現量を簡単に検索できるサービスです（図 6）。遺伝子発現解析の基準となるデータを快適に検索しようというものです。キーワード検索もできますし、「組織特異的に発現する遺伝子を見る」という欄から臓器の絵をクリックして検索することもできます。

2017 年 8 月の終わりぐらいに、この RefEx に関する論文が出ました。奇しくも FAIR の論文が出たのと同じ Scientific Data というデータジャーナルにです（図 7）。これはまさにオープンアクセスです。私自身は実はプレプリントを投稿したことはまだないのですが、データジャーナルに投稿した経験を話したいと思います。

この Scientific Data 自体も生命科学分野で珍しく、普通のジャーナルとは違い、よりデータにフォーカスを当てて紹介するという、チャレンジングな試みです。もともと、研究者が出した、既存のデータベースに載っている一次データセットを他の人が再利用しようとするとき、適切に再利用するための情報が少ないという声がありました。それを一生懸命書いても一次データを出す人にとっては何の業績にもならないからです。これでは Win-Win ではないということで、データジャーナルが出てきたのだと考えています。まさに「一次データセットについて測定の対象、方法、品質を記述する」ために始まったジャーナルだったのです。

しかし、われわれは自分自身でデータを出しているわけではなく、それを再解析して使えるウェブツールをつくっているので、この規定とはそぐわないところがあったのですが、ちょうど運良くタイミングが合い、Scientific Data が再利用を決定的に促進するシステムや技術についての独自のレポートとしてアーティクルタイプの論文を新しく受け付けることになりました。ただ、アーティクルで出すときも、Data Descriptor と同じように、使ったデータについてしっかり記述しなけ



（図 6）

（図 7）

ればなりません。

再解析したデータを全て DOI 付きで、全ての人が参照できるところに出さなければいけないということで、われわれが使った全てのデータを figshare に公開し、43 個ぐらいのデータセット全てに DOI を付けました（図 8)。

また、どうやって再解析をしたかというプログラムも全てつまびらかにしなければいけないということで、プログラマーが使うリポジトリのサイト GitHub に、公開データの再解析に用いたプログラムや何をどう処理したかという文書をきちんと整理して置きました（図 9)。

4.生命科学分野のプレプリントサーバ

次はプレプリントサーバについてです。生命科学分野のプレプリントサーバは、bioRxiv が非常に有名で皆さんご存じと思います。一番初めのプレプリントサーバは arXiv.org でしたが、arXiv.org に生命科学分野のものが全て入っているかというとそうではなくて、quantitative biology といって、一部分の生物物理寄りのものだけ入っていました。生命科学全般に対応したようなものはあまりなかったのですが、今は図 10 のようなものがいろいろ出てきているという状況です。

プレプリントサーバを検索できるサービスも出てきています。図 11 の PrePubMed は、まさに PubMed に載る前のプレプリントを、各種プレプリントサーバ横断的に検索できるサービスです。PrePubMed に面白い統計が載っていたので引っ張ってきました（図 12)。今年 9 月の最新のプレプリントの統計情報で、一月で約 1,500 本のプレプリントが追加され、オーサーも約 750 人増えているということです。

このグラフで非常に面白いと思ったのは、緑色が

（図 8）

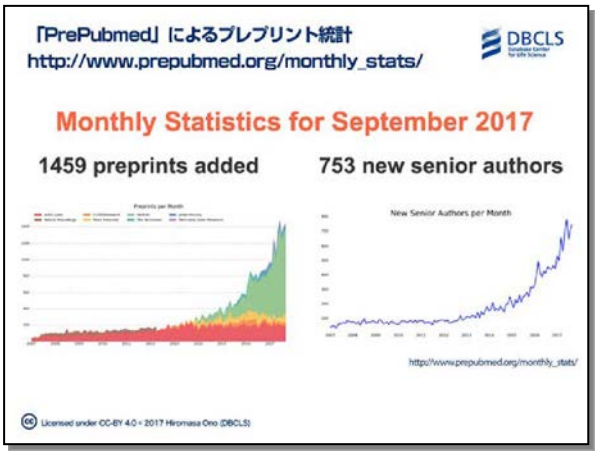
（図 10）

（図 9）

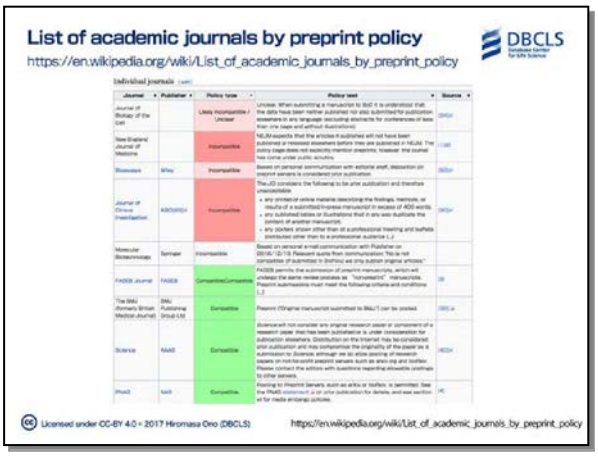
（図 11）

bioRxiv で、面積が急激に増えていることです。bioRxiv が出てきたのを契機にプレプリントがはやってきたということが分かると思います。オーサーも、ちょうど 2014 年ぐらいに増えてきたということが見受けられます。

bioRxiv は 2013 年 11 月にアメリカのコールド・スプリング・ハーバー研究所が開始しました。投稿自体は自由なのですが、論文の体を成していない、内容がふさわしくないというものは選別されているようです。これはどこのプレプリントでもそうだと思いますが、査読後に出版される前の研究成果の迅速な流通とフィードバックの活性化を目的としています。bioRxiv は現在までに 16,000 報のプレプリントが投稿されています。最近のアップデートで、有力な学術誌にそのまま転送して投稿できる機能も付いたそうです。e メールや RSS、Twitter でのアラートシステムが充実しているというのが特徴の一つです。



（図 12）



（図 13）

図 13 が生命科学分野のプレプリントのポリシーです。Wikipedia にこのようなエントリーがあって、随時更新されています。どのぐらいのジャーナルでプレプリントを認めるのか調べてみたら、プレプリントを認めないところが意外と非常に少なかったです。これらは独立系のジャーナルで、五つぐらいはまだ認めていないようですが、それ以外は compatible な状況のようです。

DBCLS の同僚の Tazro Ohta（大田達郎）さんは、図 14 のお花見メタゲノムについての論文を bioRxiv に投稿しています。お花見メタゲノムとは、お花見をして、桜の細菌叢をみんなで集めてそのゲノムを読むという面白いプロジェクトです。bioRxiv に載せているだけではなく、ジャーナルに投稿中で今リバイズをしている状況です。

bioRxiv への評価を Twitter などで見てみると、非常に好意的な方が多いです。好意的な意見を集めてきている面があるので話半分に聞いてほしいのですが、ある人は「bioRxiv にプレプリントを置いたら世界中から反響が来て、論文はまだアクセプトされていないけれど、別の国際共同研究が始まった」「有名どころが自分たちの材料を使って実験してくれていて楽しい」ということを言っています。

もう一つは、実感できる内容だと思うのですが、「読みたい雑誌の記事が有料だったから読めなかったけれど、タイトルで検索したら bioRxiv に PDF があっ

（図 14）

ていて読めた、ワーイ」というコメントです。このように bioRxiv を使っている方もいるという状況なのだと思います。

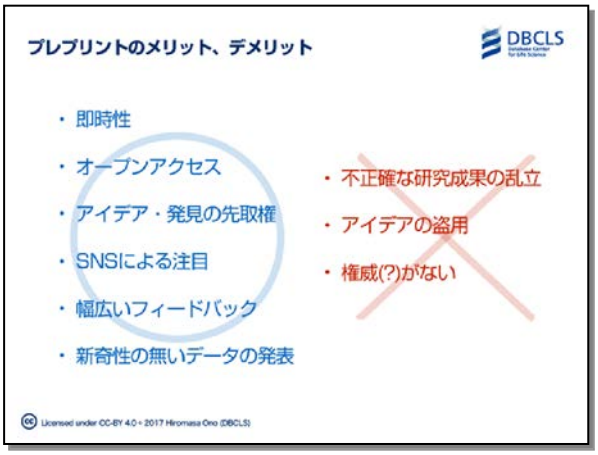
5.プレプリントのメリットとデメリット

最後に、生命科学分野に特化した話ではないと思いますが、プレプリントのメリットとデメリットを自分なりにまとめてみました（図 15）。

メリットは、まず、即時性があること、オープンアクセスであること、アイデア・発見の先取権が取れることが挙げられます。また、今はみんながスマートフォンを持ち歩いている時代で、気軽に SNS で情報が行ったり来たりしているので、注目されるというメリットが一つあると思います。さらに、それによって幅広いフィードバックが得られます。

もう一つは、新奇性のないデータを発表できるということだと思います。生命科学に限らず、オリジナリティのない論文はジャーナルベースで世にあまり出ません。再現性を確認しただけの実験や、やってみたのだけれどうまくいかなかった実験は、既存の媒体に載りづらい性格のものですが、既存のジャーナルのアーティクルの形式にのっとった形で公表できる媒体はメリットだと思います。

デメリットは、不正確な研究成果がたくさん出てきてしまう懸念があることです。また、コアのアイデアを盗まれて転用されてしまうことを心配する現場の研究者の声もあるように思います。そして、権威がないというか、今までの既存の評価軸に乗らないところがあるので使いたくないという意見もあるように聞いています。



（図 15）

●**フロア 1**　NICT の研究者です。生物はプレプリント文化を先進的に切り開いてこられたということで、学会内でどの程度コンセンサスがあるのか、若手は業績に対して安心感を持ってできているのか、研究者の日常の目線から見るとどんな感じかを共有いただけたらありがたいです。

●**小野**　私が見聞きしている範囲ですが、やはりまだ「プレプリント、ああ、何か聞いたことあるね」というレベルが平均的な生命科学系研究者の感じ方だと思います。今のところはアーリーアダプター、アンテナの感度が良い人が、先ほど紹介した Twitter もそうですが、たくさん発信もして、発信されたものも受け取っていて、「いいね」という雰囲気になっているのだという感じがしています。それが学会としてとか、普通のインフラとなっていくには 2〜3 年、あるいはそれ以上かかるかもしれません。

評価されるということに関しては、例えば科研費の申請書には、bioRxiv に挙がっているけれど、「(査読なし)」ときっと書くわけですよね。それを見た評価者が、「何だ、査読なしでは駄目だ」と評価するのか、「プレプリントを使っているとはなかなか筋があるな」と捉えるかどうかという、評価者個人がどう思っているかにまだ依存するところがあると思います。そこでどのようにコンセンサスを得ていくかは、生命科学に限らず、学会なり、国なりがどういう施策をしていくかということに依存するのではないかと思います。

●**フロア 2**　NBDC の職員です。小野さんがご存じかどうか分からないのですが、「Nature」の記事だったと思いますが、bioRxiv に載っている論文は、厳しめ

のライセンスを付けている人が多いということが載っていました。クリエイティブ・コモンズ（CC）を付けることもできるのですが、そもそも CC を付けずに普通の権利を付けていたり、CC でも特に厳しい改変禁止を付けていたりするものが非常に多いという話があります。

一方では、どうもその記事を読んだ印象では、bioRxiv のポリシーでは論文に関してはテキストマイニングが全てフリーで行えるということなのに、大半の著者はそのことを意識せずに投稿しているのではと思うのですが、その辺について聞いたことはありますか。

●**小野**　私もその話を聞いて、「へー」と思ったところです。やはりまだ bioRxiv についても、行き届いていないという批判的な記事も見つけました。プラットフォームとして今一番有名なのでみんなこれを使っていますが、まだ知らないでライセンスを付与しているところがあるので、その辺をきちんと意識してライセンスを付ける、付けないという意思表示がきちんとできるように、まだまだプラットフォームとしての改善の余地があるのではないかと思います。

●**フロア 3**　NII の北本と申します。データジャーナルへの投稿で解析済みデータを figshare に寄託したということについて、例えば DBCLS そのもので公開することはできないのかということと、日本のデータが figshare に流れてしまうことについて、こうしたらいいのではないかという意見をお願いします。

●**小野**　非常に重要なご意見ありがとうございます。われわれは公開リポジトリに出したのですが、Scientific Data が認める公開リポジトリであればどれでもいいという形になっています。NBDC のデータベースアーカイブというサービスも実は Scientific Data が認める公開リポジトリなのです。しかし、これは言っていいのか分からないですが、やはり figshare は楽なのです（笑）。これはインターフェース的な問題です。われわれも NBDC と共同研究をしているので、日本のところに置くべきだろうというご意見を頂くのはごもっともなのですが、やはりいざ論文を自分で出そうとすると便利な方に流れてしまったというところがあります。

ただ、「figshare に流れてしまう」という表現をされていましたが、彼らが囲ってしまうわけでもないですし、プラットフォームとして用意されているので、個人的にはそれでいいのかなという気持ちです。NBDC さんにはその辺を頑張っていただきたいなという思いです（笑）。

●**坊農**　補足します。この責任著者として言わせていただきますが、これを始めたときは、NBDC でそういう指定はなかったのです。だから、どこかに上げないといけなかったので、Dryad か、figshare に上げるかと調べたら、Dryad は有料、figshare は無料ということで、お金がないので figshare になりました。

●**フロア 4**　NISTEP の林（SPARC Japan 運営委員）です。GitHub の例を出されたように、オープンソース文化が特に dry なバイオサイエンス分野にあるので、私は何となく FAIR 原則はもっと文化として取り込まれているイメージを持ちがちなのですが、それでもプレプリント文化のようなものはまだ進まないのですか。情報系バイオサイエンティストと wet 系バイオサイエンティストの差のように単純に分けていいのか、もう少し複雑な研究者文化背景があるのか、教えていただけますか。

●**小野**　まさにご指摘のとおり、バイオインフォマティクスといわれている人たちは、こういう FAIR のようなものは意識していたか、していないかは別として、頭の中にはあった概念だと思います。

●**フロア 4**　GNU ライセンスなどはもとからそうで

すよね。

●**小野**　そうですね。もともとそうなので、むしろ何でオープンではないのか、というところです。私たちの周りでも、バイオインフォマティクスの人たちがかんでいるものが積極的に bioRxiv などにも投稿されています。当然、存在も知っているし、「こういうのに出していこうよ」「今、これが一番受けているんだよ」という感じでプッシュしているという印象があります。ただ純粋に、実験系のラボ、医学系のラボなどだと今のところ「何これ」という話になっていると思うのです。当然そういうことを教えてくれる人もいません。

●**フロア 4**　所属する組織の文化に引きずられるのですか。

●**小野**　そうだと思います。周りの付き合っている人の文化は多分にあります。おっしゃったように、IT 系というか、インフォマティクス系は全然これに対する抵抗はないのです。むしろ「何で出さないの」と言われるぐらいです。ただ、それが全部の分野に適用できて、公開前から出していいのかという話にも当然なるので、これが大正義というわけではないと思います。そこは純然たる深い溝がまだあると感じています。