第 3 回 SPARC Japan セミナー2013

「オープンアクセス時代の研究成果のインパクトを再定義する：
再利用と Altmetrics の現在」

生命科学分野の大規模データ利用技術開発の現状と今後の展開

坊農　秀雅
（ライフサイエンス統合データベースセンター）

講演要旨

DNA マイクロアレイや新型 DNA シーケンサ(Next Generation Sequencers)といった大規模解析による実験データの量は膨大でそのデータハンドリングは実験生物学者には困難であるが、論文発表に伴って公開されたデータを蓄積した公共データベースをフル活用する新しい研究スタイルが注目されてきている。DBCLS ではそれらを再利用する利用技術を開発し、実験生物学者の情報技術的な自立を促すための情報提供を行ってきた。本講演ではその現状を紹介し、今後について展望する。

坊農　秀雅

理化学研究所においてFANTOM(Functional annotation of mouse)プロジェクトの立ち上げに関わった後、埼玉医科大学ゲノム医学研究センターを経て、2007年7月よ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(DBCLS)にて統合データベースプロジェクト(統合DB)に従事。統合DBの広報・普及活動として統合TVや統合データベース講習会の立ち上げに関わり、現在は大規模データの利用技術開発を担当。京都大学博士(理学)。

ライフサイエンス統合データベースセンターは、国立情報学研究所と同じ、情報・システム研究機構のセンターの一つです。ただ、私どもはライフサイエンスのデータに特化していますので、そちらのお話だけになります。

NBDC のデータベースカタログ

これまでわれわれは、たくさんのデータを出してきました。ところが、それがうまく使えていないので、何とかそのデータをリサイクルし、簡単にリユースし、リテインできるようにしていこうとしています（図 1）。組織としては、JST に NBDC（National Bioscience Database Center）というファンディング機能も持っているところと、DDBJ とわれわれ DBCLS が連携しています。DBCLS は技術の開発を行い、DDBJ がデータのアーカイブを担い、それ以外の理研や東京大学を

（図 1）

はじめとするさまざまな国内の大学・研究機関がデータのオーガナイザーとして、ライフサイエンス分野のデータベースを使いやすくする、使えるようにするという取り組みです。

NBDC のウェブサイトでは、どのようなデータベースがあるかというカタログをリンクしています（図 2)。データベースに対する Google 検索的な横断検索を提供しているものと、データベースが面倒を見られなくなった場合に、その面倒を見るという形のサービスも提供しています。他のデータベースもいろいろまとめられていますが、そちらは専門的になるので、主にこの部分が皆さんの役に立つのではないかと思います。

カタログは毎週更新され、どちらかというと日本で作ったデータベースを集めて公開しており、どのようなデータベースが死んでいったか、死のうとしているかということも見ています（図 3)。毎日データベースをたたきに行って、反応がなくなったらそのような管理ポストへ入れています。自分のところがつながっているかどうか分からない、自分のインターネット環境が悪いのではないかと悩む生命科学者が多いので、そういうサービスでカタログが形成されています。

今日は DBCLS の取り組みについて、「Big data in lifescience 解釈のために」ということで、データベース統合化技術開発と信頼できるコンテンツ作成という話をご紹介したいと思います。

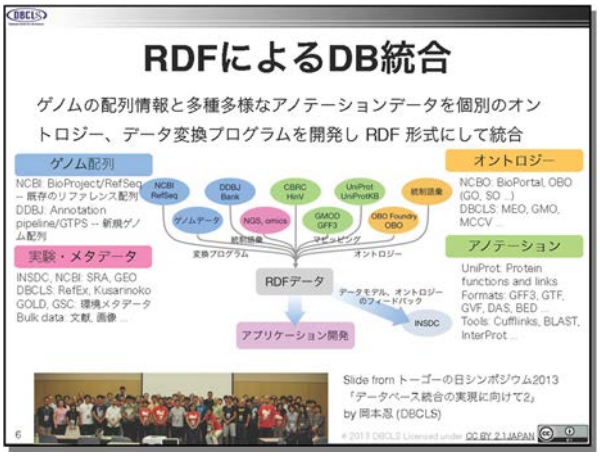
データベース統合化技術開発

技術開発の面では、RDF によるデータの統合をメインに取り組んでいます（図 4)。最近、ヨーロッパの Wellcome Trust にある EBI（European Bio-informatics Institute）が RDF データ形式でデータベースを提供するということをアナウンスしましたが、一体化した形式でやることで、アプリケーション開発がより便利になります。そのような技術開発のためにバイオハッカーソンと呼ばれる集まりを開き、みんなで一所に集まって開発会議などをしています。

さて、ビッグデータというと、ビジネスの分野ではツイッターのタイムラインにある大量のデータなどを想像されると思いますが、ライフサイエンス分野においてはシーケンスデータです（図 5)。今はこれが次世代シーケンサーという形で、すごい勢いで開発されて、ムーアの法則を上回る勢いでデータの生産が増え

（図 3）

（図 2）



（図 4）

ており、ハードディスクの容量の増加を超える勢いの塩基配列のデータが出てきて、どうするのかと騒がれていましたが、最近は落ち着いてきました。

出てくるデータは 1 回当たり 1 億行を超えていて、1 個のデータが 1GB 以上のサイズです。出てくるデータがツイッターのように誰でも見られるものならいいのですが、ライフサイエンスの分野では、そこに倫理的な問題が生じます。個人の塩基配列データを全世界に公開してしまったら、非常に問題です。そのような問題を含んだビッグデータが出てきています。

また、ツイッターならタイムラインで 140 字以内という分かりやすいデータですが、ライフサイエンスのデータは、シーケンサーを作った会社ごとにフォーマットが少しずつ違います。それをどう生かすかというのも、シーケンスのデータはただ見るだけではなく、新しい生物のシーケンシングに使うとか、実際にどのような遺伝子が動いているかという発現を見るアプリケーションに使うなど、いろいろあるのです。

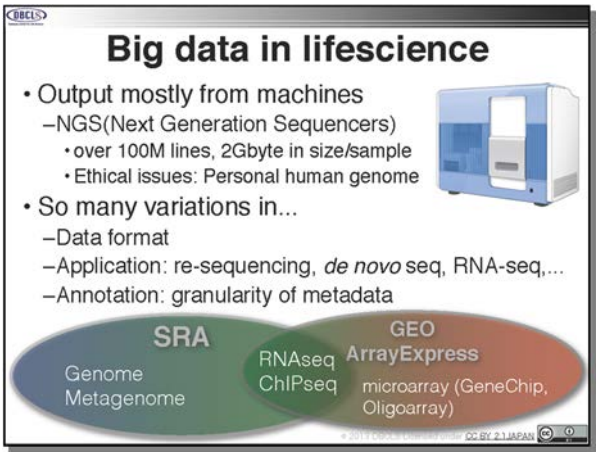
さらに、そのデータがどういう実験だったかというメタデータの使い方も、粒度が全く違います。ある研究者は熱心に書いてくれるのですが、書かない人もいます。そうなると、どうしても書かない方に合わせて統合し、解析しなければいけません。それがビッグデータの現状で、誰でも見られるデータベースとして SRA や GEO、ArrayExpress がありますが、最近では今言ったような問題があるので、研究者だけに見せるコントロールアクセスのデータベースなどもできつつあります。

その中で、次世代シーケンスのデータベースだけでも大きく、全員を把握するのが非常に困難な状況なので、それを見やすくしようと、DBCLS で開発したのが DBCLS SRA です（図 6）。これは簡単に言うと、データにアクセスするためのイエローページ（電話帳）です。例えば、皆さんが再利用するときにやりそうなデータを探して、ダウンロードしてクオリティーをチェックしてくれるというイメージです。

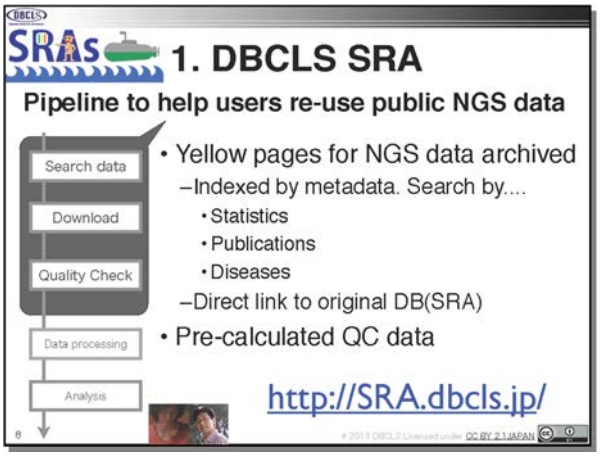
全容の把握には、定期的に自動で統計を取る仕組みをつくります。データは累積的なので、単調増加だと思われるかもしれませんが、実際のバイオのデータベースでは、データの見直しがあって突然減ったり、方針が変わってデータの納め方が変わるとレコード数が減ったりします（図 7）。

グラントの制約で、論文が出る前にたくさんのデータがエントリーされていくのですが、実際に皆さんが見たいのは論文が出たデータです。それを優先的に見れるように、PubMed の論文のエントリーと結び付いたテーブルとして実際のデータを提供しています（図 8）。

そんなのは簡単だろうとよく言われるのですが、実際にはデータが論文より先に出るので、このデータがどの論文に対応するかという情報が書かれていないことが多いのです。そこで、論文に対してテキストマイニング的なことをして、実際のどういうデータと関係があったかという対応関係をつくり、それをこういう



（図 5）



（図 6）

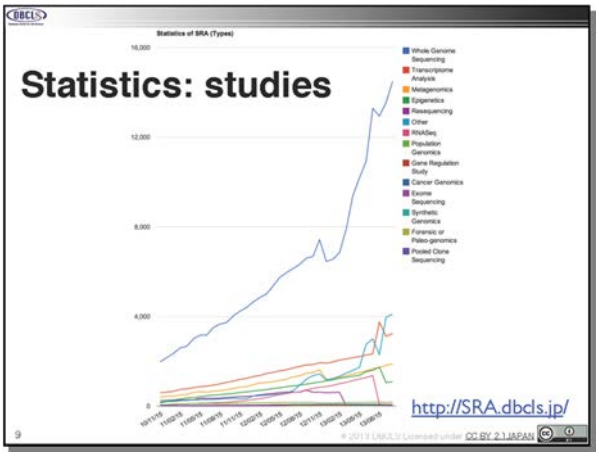
ものの提供していっているというサービスです。これは研究者には割と好評です。

また、病気ごとに次世代シーケンスによる研究結果のデータが出てきたかを、PubMed の中の MeSH term のツリーを使ってアクセスできるようにしています（図 9）。それに関して書いた論文が、今週 PLOS ONE に出ました。2 日前に出た論文なので、まだ 2 回しか更新されていませんが、メトリックとして、ビューとフェイスブックとツイッターのカウント数が出ています。

信頼できるコンテンツ作成

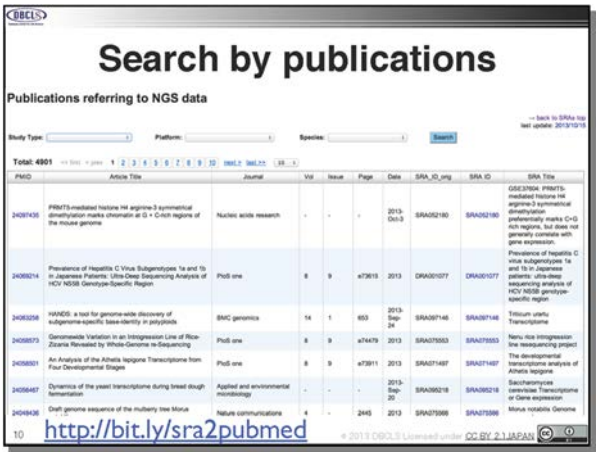
コンテンツ作成に関しては、まず、新着論文レビューというものをやっています（図 10）。これは、「Nature」や「Science」など、昔の基準で言うトップジャーナルに出た日本人の論文を日本語で書いてもらうのです。これは山中先生の iPS で、まだノーベル賞を取る前に書いてもらったお話ですが、自分の出た論文を日本語で書いてもらって、画面の図が再利用できるというようなことを、クリエイティブ・コモンズ表示ライセンスをメインに公開してもらいます。そうすることでこの辺の図を再利用できるということを 2～3 年前から始めています。

この新着論文レビューに加え、去年から新たな取り組みとして、領域融合レビューというものも始めています（図 11）。日本の学術学会の人たちにお願いして、その分野で活躍されている先生に、その分野の少し長めのレビューを書いてもらうのです。これはまだ一つ一つのフィギュアにまでは付いていないと思いますが、DOI が付いて公開という形になっています。われわれがウェブ上でやっているものも研究者の人たちが見直してくれるようになったのではないかと、個人的に

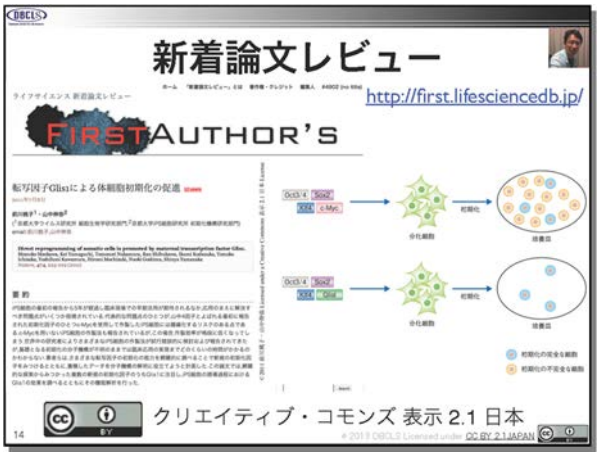


（図 7）

（図 9）



（図 8）



（図 10）

は思っています。これももちろんクリエイティブ・コモンズ、CC BY2.1 で公開しています。

それ以外に、私が主に進めてきたこととして、統合 TV というものがあります（図 12）。これは、ライフサイエンス分野のデータベース、ツールの使い方の動画チュートリアルです。現在 700 本ぐらいの動画があり、YouTube などにもロゴを付けてアップしています。

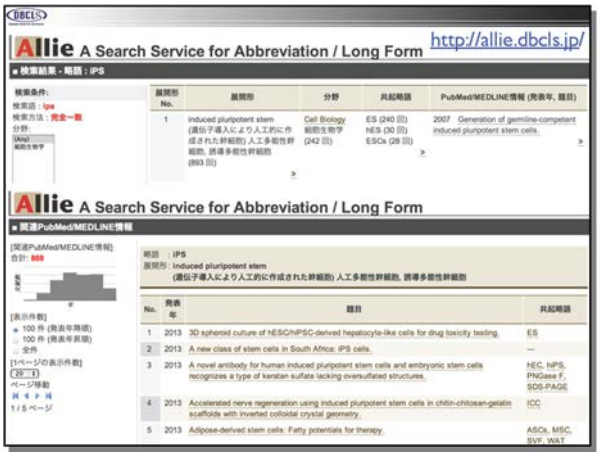
主に図書館の方が使っているのではないかと思うのが CiNii の使い方を紹介する動画です。これはライフサイエンス以外の文系の大学からも割とアクセスが多く、そういうところにも役立っているのではないかと思います。もちろんプロの方はこんなものを見る必要はないのですが、初学者向けにこのような動画で紹介するコンテンツがあるといいのではないかと思っています。

さらに、Allie という略語のアクロニウムのデータベースを PubMed から生成しています（図 13）。例えば iPS という言葉を入れると、「induced plu-ripotent stem」という正式名称が出てきて、加えてその単語が何年ごろから出てきたかということも分かります。また、それが毎年どのくらいの頻度で使われているかも示されていて、iPS であれば、2007 年から始まって一定数使われているということが分かるようになっています。

もう一つ、PubMed/MEDLINE の逐次表現検索（インクリメンタルサーチ）を行う inMeXes というものがあります（図 14）。例えば different と入れたい場合、d-i-f-f と、4 文字以上入力すると、逐次的に検索が始まります（図 15）。

検索結果をクリックすると、Life Science Dic-tionary という、京都大学で作られている生命科学の辞書のページにリンクして、実際のコロケーションを調べるこ

（図 11）



（図 13）

（図 12）

（図 14）

とができます。日本人研究者には英語を書くのが苦手な人が多いので、こういうものも用意しています。

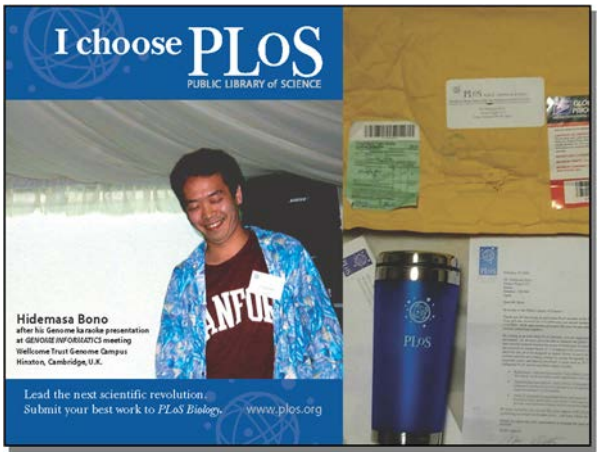
今後どうあるべきか

ライフサイエンス分野では、われわれ DBCLS などのセンターがデータベースの統合化に向けて、さまざまな取り組みを進めています。現状はまだ生命科学者に紹介して使ってもらうというフェーズで、これを組み合わせて次に行くというふうにはなってないのですが、その一方で測定機器は非常に進化して、データ量は大変な勢いで増えていっています。普通のパソコンで、ドラッグ・アンド・ドロップで使えるようなソフトでやれる勢いをはるかに超えて、本当ならば、UNIX のコマンドにターミナルを開いてやってもらわなければいけないような状況になってきています。

では、今後はどうあるべきなのかというと、まずはデータを出したがらない状況をぜひ変えてほしいと思っています。データの適切なサイテーションがなされるように、また、公的研究費から得たデータを売る行為を撲滅させる必要があります。それから、データを流通させることに利益があるということを普及させるため、トラッキング機能の充実と成功事例の充実を図る必要があります。今は私のような半分研究をやっている人間が、「どんどんデータを流通させると御利益がある」と言っても、みんなが「えっ？」と思っている段階なので、本当にそれで論文が出た、良い研究ができたという事例を積み上げていかなければならないと思っています。

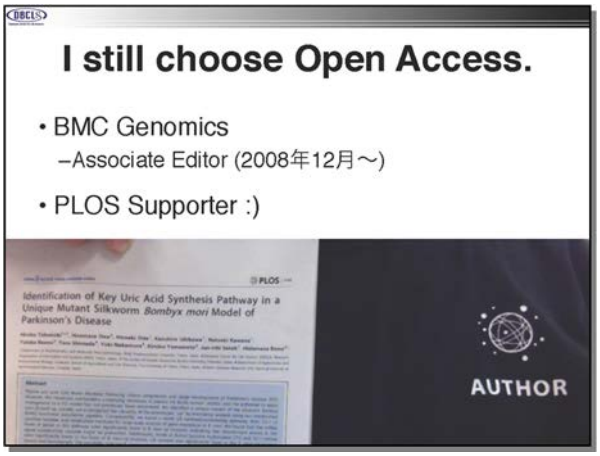
私自身はオープンアクセスを昔からやっていて、実は 10 年前から PLOS を信奉しています（図 16）。PLOS を始めた人は主にマイクロアレイを開発した人間たちなのですが、私はその人たちと昔から一緒に研究をしていたので仲が良く、これをやりましょうということになりました。Wellcome Trust のときに撮られた写真を使って“I choose PLOS”ということでやっています。

PLOS のサポートも昔からやっています（図 17）。現状でもオープンアクセスをもちろん選んでいますし、最近 PLOS の論文を出したと言いましたが、今、「BMC Genomics」の Associate Editor をしていて、ここでは古典的なものをやっていますが、今日の話を聞いて、新しい形のジャーナルなどにもっと力を入れていきたいと思っています。



（図 16）

（図 15）



（図 17）