第 2 回 SPARC Japan セミナー2016

「研究データオープン化推進に向けて：インセンティブとデータマネジメント」

医学生物学分野における データのオープン化とそのインセンティブ

仲里　猛留

（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）

CC BY

講演要旨

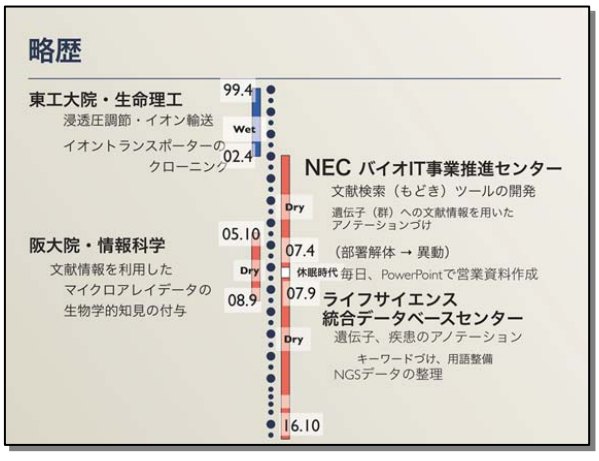
医学生物学分野では、文献情報が MEDLINE に、 塩基配列情報が GenBank にと 1970 年代からデータベースとして保存されており、インターネットの普及とともに誰でも広く利用可能となった。現在は米国 NCBI、欧州 EBI、日本の DDBJ によって、これらに加え遺伝子発現や化合物情報なども公共データベースとして蓄積・運用がされている。このように自然に生命科学データが集積されたのは、出版社が投稿規定内に実験データを公共データベースに登録することを義務化した背景がある。近年は、機器の発展やコストの低下により、診断などヒトを対象にした研究も広くなされるようになり、場合によってはデータをオープンにすることにより個人情報を侵害するのではないかとの懸念も持たれている。

仲里　猛留

東京工業大学大学院生命理工学研究科修士課程修了。在籍中はウナギの海水適応機構をテーマに分子生物学実験の日々を過ごす。修了後、NECに入社。バイオインフォマティクスの部署で文献情報を活用した大規模データの解析のためのソフト開発に従事。2007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。DDBJと連携して公共データベース中の次世代シーケンサー（NGS）データを検索するサイトの運用や、それら大規模データの生物学的な解釈を行うための基盤整備を行う。2011年より現職。大阪大学大学院情報科学研究科博士後期課程修了。博士（情報科学）。

私が所属するライフサイエンス統合データベースセンター（DBCLS）はあまり聞き慣れないと思いますが、親組織が情報・システム研究機構で、国立情報学研究所、国立遺伝学研究所、統計数理研究所、国立極地研究所と一緒です。

私の年表は図 1 のようにまとめられます。今日は研究者の立場からデータのオープン化についてお話ししたいと思います。大学院、マスター（修士課程）までは魚から遺伝子を抽出して研究していました。例えば、ウナギを買ってきて 30 匹ほど解剖し、そこから血圧調節に関係する遺伝子を薬品を調合して抽出しました。ウナギはサケやマスと一緒で、淡水と海水の両方に行ける魚です。淡水のときは周りが真水なので、自分のイオンが抜けて水が入ってきます。海に行くと逆にな



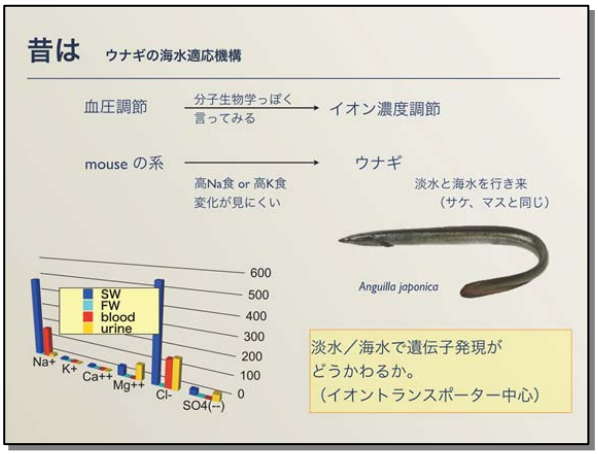
（図 1）

って、周りが塩辛いので、水が抜けてイオンが入ってきます。ちょうど逆になるので、淡水だけ出てくる遺伝子、海水だけ出てくる遺伝子を探していたのです（図 2）。

その後、私は実験に挫折し、IT バブルに乗って、NEC バイオ IT 事業推進センターに就職しました。そこも IT バブルがはじけて 5 年で部署がなくなり、その後、縁があって、DBCLS というアカデミックな機関に所属しています。

国立遺伝学研究所の中にある、日本 DNA データバンク（DDBJ）という機関は、DNA の A・C・G・T などの並び順をたくさん読める次世代シーケンサーから取ったデータを集めていますが、その検索サイトをつくる仕事が今の本業です（図 3）。

DBCLS の目的は、たくさん出てくる生命科学系のツールやデータをうまくリサイクルして研究者の人に使ってもらうことです。



（図 2）

（図 3）

生命科学分野におけるデータベース

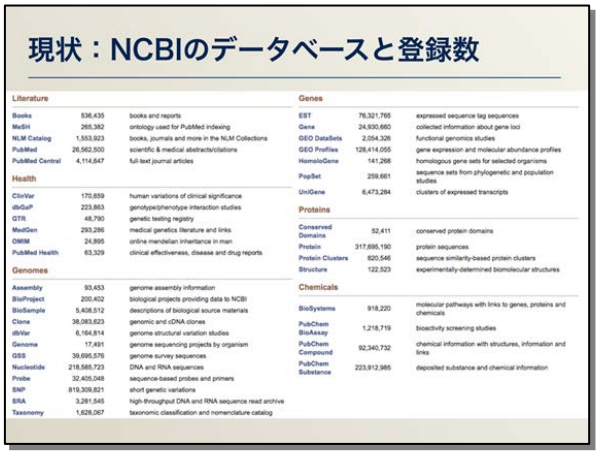
私は生命科学分野にいるので、そのデータベースの実情を最初にお話ししたいと思います。

バイオ系のデータは、アメリカ・ヨーロッパ・日本の三つでコラボレーションして集めています。私が学生のころは、アメリカが一番集めていて有名だったので、アメリカの国立生物工学情報センター（NCBI）のデータベースを使うように言われていました。NCBI のデータベースがどのようになっているかを図 4 に列挙しています。

文献関係、ヘルスケア関係、DNA 関係、化合物関係などいろいろなデータベースがあるのですが、左の下から 5 番目の「Nucleotide」にデータが約 2 億 1,800 万件入っています。文献関係では「PubMed」が上から 4 番目にあり、約 2,650 万件入っています。

私が研究室に入ったのは学部の 4 年生ですが、そのころから PubMed や Nucleotide などを見て、いいデータを探すような仕事をしていました。

図 5 は、コラボレーションしている国立遺伝学研究所のデータです。赤線が右の縦軸で Nucleotide の件数を表しています。約 2 億件です。水色のバーは左の縦軸でデータの量を表しています。bp（base pair）というのは塩基の数で、A や C などが何個入っているかだと思えばいいのです。単位は billion で 10 億ですから、2,000 億ぐらいとたくさん入っているのです。す



（図 4）

ごくたくさんのデータが入っていて、私が学部 4 年生だったのは 1999 年なので、ほぼ地を這うようなデータです。今は 2016 年なので、もう一つ棒が立つかどうかのところにいて、その何倍だろうかというぐらい急速にデータが増加しています。私が専門でやっている次世代シーケンサーは、ここ 5〜6 年ぐらいにデータがたくさん出はじめ、本格的に集められています。

データベースの変遷をまとめた図があるので、借りてきました（図 6）。NCBI で、先ほどお見せしたようにいろいろな種類のデータベースが集められているのですが、もともと NCBI で全部メンテナンスされていたかというと、そうではなく、あちこちから NCBI に集めたという経緯になっています。

例えば、Nucleotide は、ここでは GenBank と書かれていますが、最初は 1972 年、ロスアラモス国立研究所で、物理学のデータの一つとして DNA の並び順がストックされるようになり、その後、1990 年少し前ぐらいから NCBI に集められました。

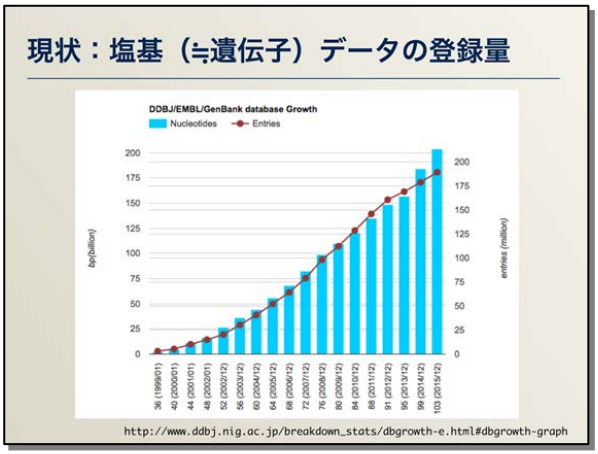
オレンジのバーは文献です。この後、PubMed が出てきますが、1960 年代からずっと集められています。疾患データベースの OMIM はジョンズ・ホプキンス大学で四十数年ぐらい、McKusick 博士が、「病気はこのような感じです」とカタログしていたのを、最近はインターネット時代なので全部デジタル化して、NCBI などから検索できるようにして、この中に一応入っていたのですが、最近は予算が切れて、またジョンズ・ホプキンス大学に戻りました。

このような感じであちこちで集めていたものを NCBI に集め、最近は無料でパブリックが自由に閲覧・検索できるような状況になっています。

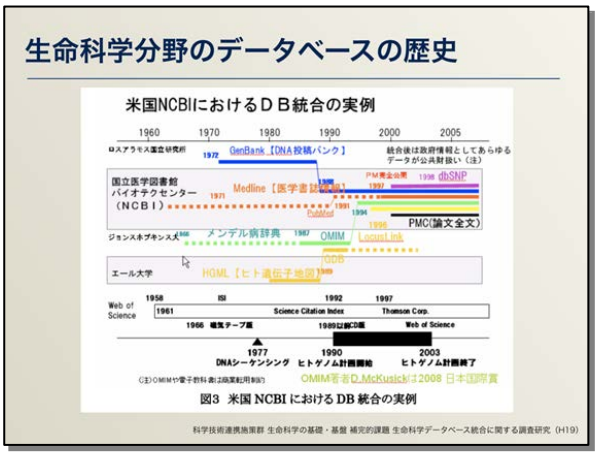
アメリカの NCBI、ヨーロッパの欧州バイオインフォマティクス研究所（EMBL-EBI）、日本の DDBJ は、国際塩基配列データベース（INSDC）というコンソーシアムのようなものを組んでいます。例えば、NCBI に入ったデータも大体は DDBJ から検索でき、逆もしかりで、DDBJ に集まったものが NCBI から検索できるという形になっています。

ちなみに、NCBI は、上位の組織がアメリカ国立衛生研究所（NIH）で、NIH の下にアメリカ国立医学図書館（NLM）という図書館の部門があり、この下に NCBI というデータベースをメンテナンスしている組織が入っているという構造になっています（図 7）。

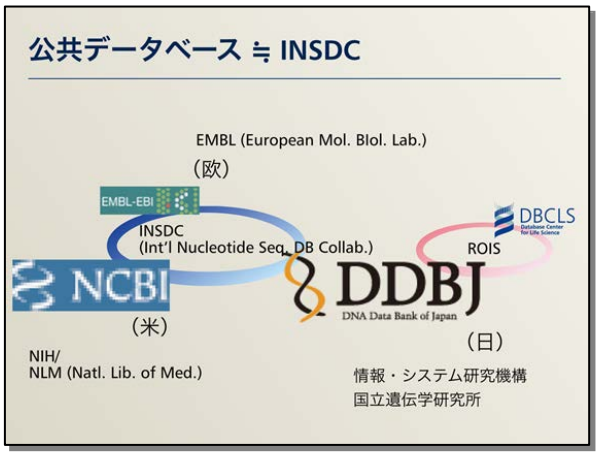
私は DBCLS にいるのですが、この 3 局で集めたデ



（図 5）



（図 6）



（図 7）

ータについては、特に DDBJ の場合は予算の兼ね合いもあって、集めるのにリソースを割いていました。私たちはデータをリサイクルさせるのが目的なので、DDBJ が集めたデータを検索できるようにします。これを同じ情報・システム研究機構の下位組織同士でやっているということです。

NCBI のデータベースはかなりたくさんあり、いろいろ集まっていると感じたかもしれませんが、データベースは NCBI だけではありません。「Nucleic Acids Research」という雑誌があり、これは年に一度、1 月に「Database Issue」というデータベース特集号、7 月に「Web Server Issue」というウェブツールの特集号を組んでいます。

図 8 は 2016 年のもので、23 回目の特集号です。一つのデータベースで一つのペーパー（論文）ではありませんが、ここには 178 のペーパーが入っています。そのうち 62 が新しいもの、95 がアップデートと書いてあります。これが毎年出ているので、ここに載らないようなデータベース、「Bioinformatics」「BMC Bioinformatics」など、いろいろな雑誌に載っているデータベースも合わせると、年間 1,000 ぐらいできているのではないかと言う人もいます。NCBI など大手のものもあれば、各研究者がつくるような小さいものもあるのですが、どんどんデータベースが生まれています。もう少し言うと、お金が切れると、マシンが壊れるとともに死んでいくということがあったりするのです。

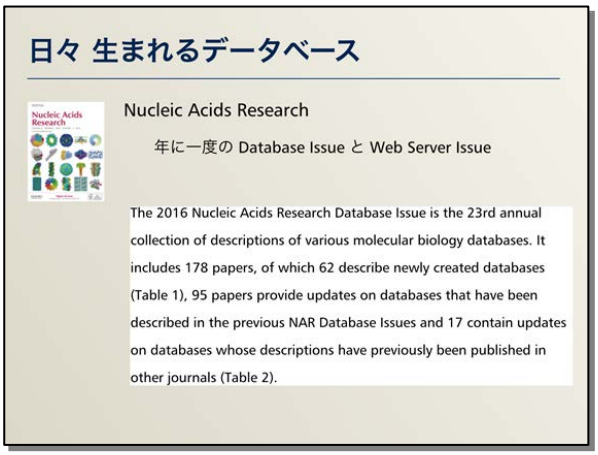
主要な生命科学データベース

特に注目すべき、私がまだ実験をしていた学生のころによく使っていたデータベース、ウェブツールが二つあります。一つは PubMed という文献のデータベース、もう一つは BLAST という類似した遺伝子を検索するツールです。

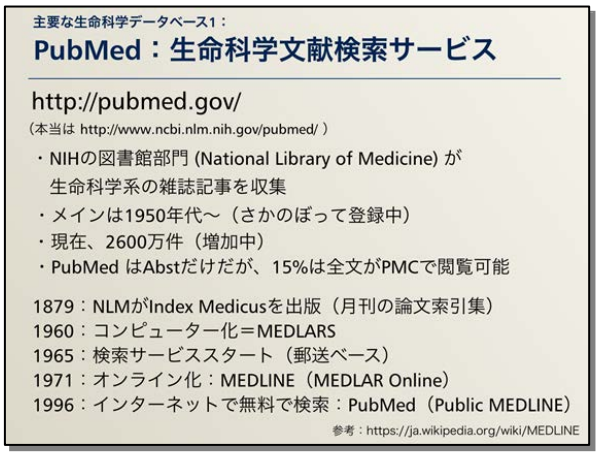
まず、PubMed について述べます。図 9 は NCBI の PubMed で、実際の検索結果の画面です。がん、糖尿病などのキーワードを入れると、関連する文献のリストが出てきます。詳細画面には、文献のタイトル、著者、掲載雑誌、発表日などの情報、Abstract（要約）が書かれています。バナーをクリックすると PDF を取れます。

ちなみに PubMed Central（PMC）という論文の全文を収載したものもあります（図 10）。PubMed にはデータが約 2,600 万件あるのですが、そのうちの約 15%

（図 9）



（図 8）



（図 10）

は PMC という別のデータベースにも収載されているのです。そこでは、論文の Abstract だけではなく、イントロ、メソッド、ディスカッションまで全部読めるようなものがあります。

物理で言うと、例えば Google Scholar、Citeseer などが昔やっていたように思いますが、生命科学分野では PubMed があって、今は全部読めるようにだんだんなってきています。

その歴史を見てみると、もともとは 1879 年に NCBI の上位組織 NLM が、「Index Medicus」という月刊の論文索引集を出しました。それが 1960 年代に MEDLARS としてコンピューター化され、その検索サービスが 5 年後ぐらいにスタートしました。それが MEDLARS Online としてオンライン化されたのは 1971 年です。電話をかけて取得したり、郵送でお願いしたりするのが、この時期のやり方でした。

私が研究を始める少し前の 1996 年に PubMed ができました。MEDLINE がパブリックになったので PubMed という名前が付いたと聞いています。インターネットに接続さえすれば、誰でも MEDLINE に入っている文献を検索して無料で見られるようになったという流れで、文献のデータが公開されてきました。

次はもう一つの BLAST についてです（図 11）。私は学生時代にウナギの血圧調節などに関係する遺伝子を探していました。当時は DNA シーケンサーを使うと、うまくいくと DNA の A・C・G・T の並びが 1,000 文字ぐらい取れました。それを BLAST に投げると、あなたの投げたものはここからここまであって、赤いものが似ているもの、緑がほんのり似ているもの、そしてだんだん似なくなってくるのですが、この辺がどのくらい似ているかというチャートが出てきます。具体的に A・C・G・T のどの辺りがどのぐらいマッチしたのかという詳細が出てくるというものになっています。

文献が MEDLINE に登録されると、PubMed でそれを検索できるようになります（図 12）。同じように、遺伝子の塩基配列が GenBank、EMBL、DDBJ などの DNA のデータベースに登録されると、BLAST で検索をかけられるようになります。BLAST は DNA の塩基配列情報がオープンでないと検索できないのです。結局、BLAST も PubMed も全てオープンデータの上に成り立っているツールだということです。

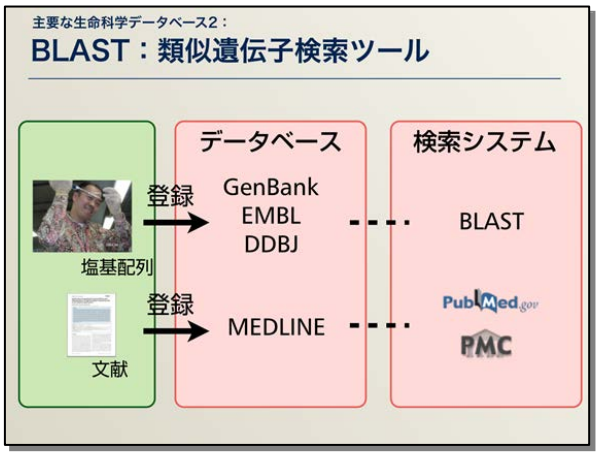
このようなものがないとわれわれは研究ができないと、学部 4 年のときにも言われていました。私が研究を始めたときには既にこのようなものがあったので、空気のようにオープンなデータを使っていたのだということを今あらためて感じています。

公共データベースにデータが集まる理由

それでは、なぜこのような公共データベースにデータが集まるのでしょうか。

理由の一つには、論文の投稿規定で要求されているということがあります。図 13 は「Nature」の投稿規定です。タンパク質配列、DNA 配列、遺伝子発現な

（図 11）



（図 12）

どのデータの種類、その横に GenBank、DDBJ、EMBL などのデータベースの名前が書いてあります。このデータが出てくる論文があれば、このデータベースに登録しなさいということです。登録してアクセッション番号を取らなければ、論文を掲載してくれないのです。

研究者は論文を出して何ぼで、成果は論文で測られます。生命科学分野では特に、データベースにきちんと登録しなければ論文が出せず、成果にならないのです。半分、強制、仕方なくというところもあるのですが、投稿規定で要求されているので、自然とデータが集まります。

もう一つの理由がファンディングエージェンシー（研究資金の配分機関）からの要求です（図 14）。例えば、NHI Data Sharing Policy には、1 年で 50 万ドル以上のグラントであれば、データをどのように公開するのかプランを考えなければいけないということが明記されています。

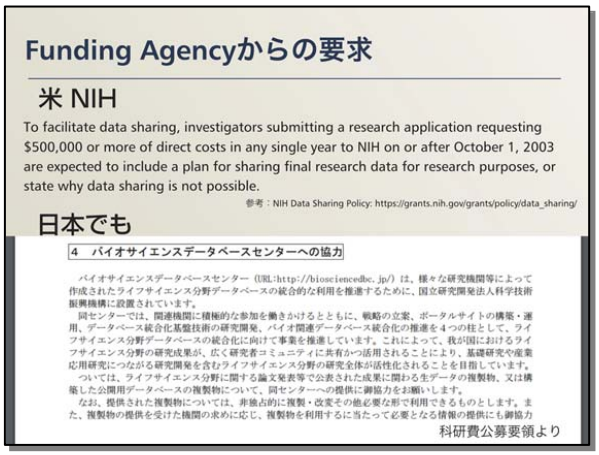
日本でも科研費の公募要領にも、「バイオサイエンスデータベースセンターへの協力」が掲載されています。バイオサイエンスデータベースセンターというのは科学技術振興機構（JST）の組織で、もともと私たちがしていたデータのリサイクル事業を今引き継いでもらっています。「論文発表等で公開された成果に関わる生データの複製物、又は構築した公開用データベースの複製物について、同センターへの提供に御協力をお願いします」と書いてあります。私たちも、データを出せ出せと言う圧力団体のような面があるので、このようにデータをオープンにするように公募要領に書かれるようになったのはそのような活動の成果です。

オープン化のインセンティブ

自分なりに、オープン化のインセンティブとは何かを考えてみました。研究者の視点から言うと、やはり究極の目的である論文が掲載されることでしょう。自分で論文を書くようになって分かったのですが、自分のデータが使ってもらえたり、論文が引用されたりするとうれしいのです。これは世代の差があるかもしれませんが、特にピペットや薬品を使うような生命科学の実験をする人は、もったいない精神にたけていて、少ないお金でデータを出し、それを自分だけで骨と皮しか残らないぐらいにしゃぶり尽くしたがる人が多いのです。

昔は、自分の物は自分の物、人の物は自分の物というジャイアニズムのように、データの囲い込みをする人たちが多かったです。しかし今は、オープンにした方がプレゼンスが上がります。例えば、論文をオープンアクセスのジャーナルに出すと、学会で「あなたの論文を読みました」と言われたり、読んだ人からコンタクトが来て、「私たちはこんなにいいデータを持っているから一緒にしませんか」と言われたりします。そうやって世の中を動かせるようになるので、研究者としても、実はオープンにした方がプレゼンスが上がるのではないかと思っています。そうなると、実際に

（図 13）



（図 14）

もらえる研究費も増えて、次の成果につながり、ハッピーなサイクルが回るのではないでしょうか。

これはお金のどろどろした話でしたが、別の観点で、付加価値の付与というインセンティブがあると考えています。私が本業で扱っている次世代シーケンサーのデータは大きいので、自分の手元のマシンで解析できない、そのスキルがないという人が結構いるのです。しかし、DDBJ にデータを登録すると、DDBJ の方で解析のパイプライン、ウェブツールを組んであるので、クリックするだけでそれにすぐに流し込めます。自分のところで解析せずに、国立遺伝学研究所のリソースを使って解析できるようになるのです。

また、自分が出したデータと、他のデータベースの他の人が出したデータが一緒になることで、自分のデータの付加価値が上がるということもあります。

別の観点では、オープンにすることで研究の再現性が上がるということがあります。図 15 は実は STAP 細胞の次世代シーケンサーのデータです。きちんと登録されているので、自分でデータを取ってきて、自分のコンピューターで追試できるのです。このデータについての論文がスピード感を持って取り下げられたのは、このようにデータが公開されていて研究の再現性が保たれていたからという理由もあります。

データのオープン化に求められること

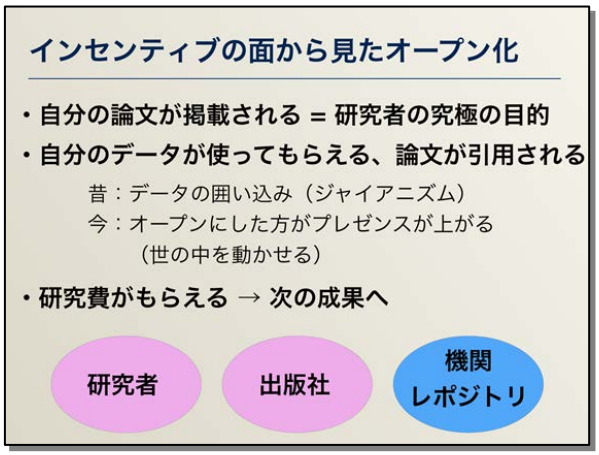
データをオープン化するときは、公共データベースに登録する、自分でデータベースを作成するという手段もありますが、例えば「Scientific Data」「GigaScience」などのデータジャーナルにサブミットする、または機関リポジトリを利用するという方法もあると思っています。

そのときに言いたいことが幾つかあります。まず、データを参照する仕組みです。遺伝子であれば、アクセッション番号や登録 ID などがありますが、そういうものできちんとポイントできるようにしてほしいのです。論文に番号を書いておけば、一意にこのデータだと分かるようにしてほしいと思っています。

また、実際にデータベースを維持しようとすると費用が掛かります。ミレニアム・プロジェクトなども、多額のお金が掛けられたのですが、お金が切れるとともにだんだん廃れていくので、私たちがレスキューしています。このような費用の問題は確かにあります。

そして、永続性です。例えば、researchmap の科学研究費助成事業データベースのリンクをクリックすると Not Found になることがあります。なぜかというと、科学研究費助成事業データベースの URL が変わったからです。名誉のために言っておくと、最近のものはきちんとリンク先のページが出ます。ただ、昔のものは Not Found になってしまっています。URL も一意のものとして機能するのですが、実際にはメンテナンスしている方で URL が変わってしまうとたどれなくなってしまうことがあるので、そのようなことも含めて永続性を強調したいと思います。私がインセンティブではないかと考えたことを図 16 にまとめました。

（図 15）

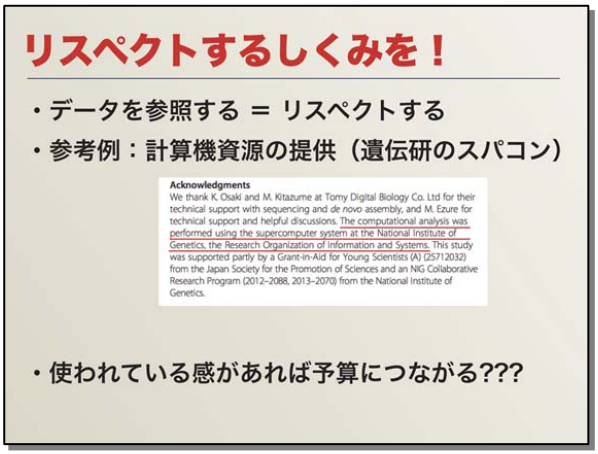


（図 16）

しかし、いろいろな立場があります。研究者の立場からすると、これは確かにプラスです。出版社も、例えば Scientific Data を維持している Nature などは、掲載したデータ論文を引用してもらえるのでプラスです。しかし、機関リポジトリにとって、これがインセンティブになっているかというと疑問もあるので、この後で議論できたらと思っています。

強調したいのは、データが水や空気のようにオープンになってしまっているということです。お金を掛けてデータを出しているので、それをリスペクトする仕組みがあると、インセンティブの一つになるのではないでしょうか（図 17）。

例えば、国立遺伝学研究所のスパコンを使った成果に関しては、「Acknowledgements」に「使いましたよ」ということが書かれています。このように、使われている感があれば予算につながるかもしれません。

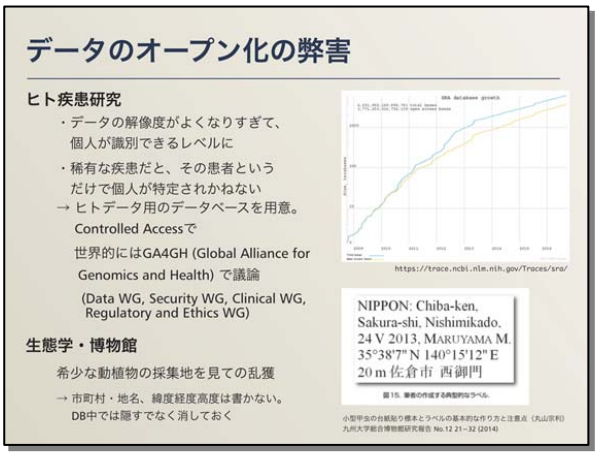


（図 17）

データのオープン化の弊害としては、最近、ヒトのデータがよく出てくるのですが、解像度が良過ぎて個人が特定できるのではないかという問題点があります。（図 18）

また、生態学や博物館では、データにラベルが付きます。詳細な地名、緯度・経度まで書いてあるので、貴重なランなどの植物は、それを見た花屋さんが採りに行くのです。それで絶滅がさらに加速されてしまうので、これらの情報は今全て消しているという状況になっています。

そして、オープン化はしていきたいのですが、実際には、図 19 のようにスライドガラスが山と詰まれて、倉庫に鍵が掛けられているので、カタツケていけたらと思っています。



（図 18）

（図 19）