第 3 回 SPARC Japan セミナー2016
「科学的知識創成の新たな標準基盤へ向けて：オープンサイエンス再考」

研究データ共有の理想と現実，そして実践可能性
～地球環境分野の研究基盤に関する意識調査から～

小野　雅史
（東京大学地球観測データ統融合連携研究機構）

講演要旨

「研究データの共有は科学の発展に貢献する良い考えだ」という理念については、研究者のみならず、政策立案者、資金提供者、市民等、多くのステークホルダーが基本的には同意するはずである。しかし、実際には、多くの研究データが、共有または公開される状況に至っていないのが現状である。このオープンサイエンスの理想と現実の間には、どういった課題が存在するのだろうか？これを明らかにするために、我々は地球環境情報分野の現場の研究者を対象として意識調査を実施した。そこで、本講演では、この意識調査の結果をもとに、我々がオープンサイエンスを前に進めるために実現可能な取り組みについて、考察した内容を紹介する。

小野　雅史

東京大学地球観測データ統融合連携研究機構（EDITORIA）の特任研究員。DIAS（データ統合・解析システム）という研究データ基盤を中心とする事業「地球環境情報プラットフォーム構築推進プログラム」に参加している。過去に、地理情報標準ISO/TC211の仕様検討委員、GEO（Group on Earth Observations）のオントロジータスクチーム、Belmont Forum E-Infrastructures and Data Management Collaborative Research ActionのData sharingグループのメンバーとして活動。

本日のセミナーのテーマは「オープンサイエンス再考」ですが、私の報告のタイトルにある「研究データ共有」は、オープンサイエンスの要素の一つです。研究データ共有の理想、広義にはオープンサイエンスの理想といったときに、皆さんは恐らくこういうものをイメージされると思います。「研究者がみんなデータを提供して、それをみんなでシェアして、それが科学の発展を促進し、ひいては社会への貢献に資する」。これは非常にいい考えです。みんないいことだと思っていると思います。でも、その一方で現実はどうなっているのでしょうか。

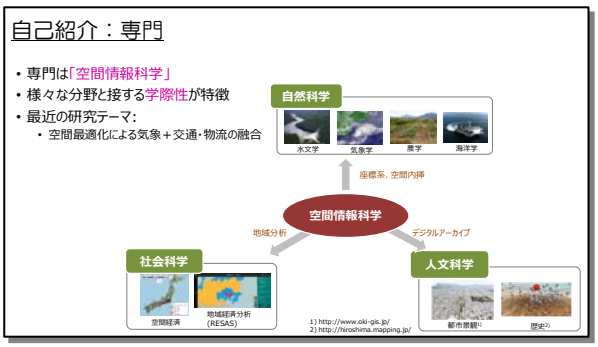
実はこれをとても簡単に知る方法があります。例えば、誰でもいいので研究者を捕まえて、こういう質問をしてみてください。「先生、あなたが持っている研究データのうち、何パーセントぐらいを他の人とシェアしていますか」。これは言うまでもなく、非常に歯切れの悪い答えが返ってくると思います。これはオープンサイエンス推進派といわれている先生でも、そんなに変わりはないと思います。つまり、こういうことなのです。みんないいことだと思っているのに現実はそうなっていない。ですから、この理想と現実の間にどういう空白があって、その隙間を埋めるために何をしたらいいのか、そのためのヒントになるような話を今日はできたらと思っています。小賀坂さんや小野寺

さんのお話とも非常に符合する部分があるので、思い出しながら聞いていただけたらと思います。

自己紹介

私の専門は空間情報科学です（図 1）。この分野の特徴は、自然科学だけでなく、社会科学、人文科学などいろいろな分野と接することが多い学際性です。ある分野とある分野を融合させるということも比較的やりやすくて、最近の私の研究テーマは気象と交通・物流の融合です。研究とは全く関係ないのですが、その他、色、調味料といった民間資格を持っています。私は料理が結構好きなので、家に調味料が 80 種類以上あって、それを組み合わせて毎回新しいカレー粉をつくる、そんなことをしていたりします。一見これは全然関係ないように聞こえるかもしれませんが、分野の融合のように、異質なものを組み合わせて何か新しいものをつくるのが好きな人間、それが私です。

こうした私自身の特性を生かしながら、これまでさまざまな異分野の融合関係のプロジェクトを実施してきました（図 2）。



（図 1）

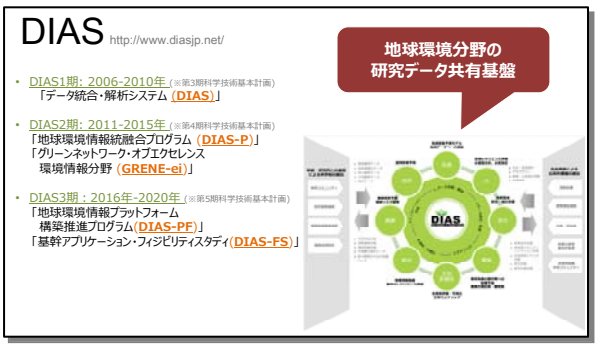
（図 2）

一例として、ISO/TC211 の geographic information standards に関する仕事、国内の地球環境データ統融合の Data Integration & Analysis System（DIAS：データ統合・解析システム）というプロジェクト、DIAS の国際版のような Group on Earth Observation（GEO：地球観測に関する政府間会合）のオントロジー等のタスクグループでの仕事、ベルモント・フォーラムのデータシェアリンググループのメンバーとしての活動などを行ってきました。

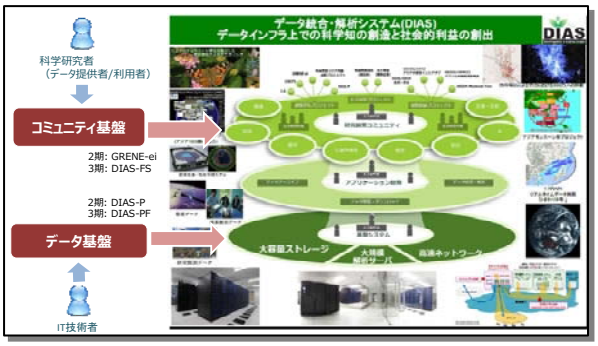
Data Integration & Analysis System（DIAS）

今でも私のメインのプロジェクトは DIAS です（図 3）。DIAS とは、地球環境分野の研究データ共有基盤のことです。DIAS という名前が使われだしたのは、第 3 期科学技術基本計画の 2006 年からで、ちょうど今年から DIAS 第 3 期に入りました。

この DIAS の特徴を一つ挙げると、過去の SPARC Japan のイベントで、北本先生からもお話があったと思いますが、コミュニティ基盤とデータ基盤の二つのレイヤーを持っていることです（図 4）。このデータ



（図 3）



（図 4）

基盤を支えるのは主に IT 技術者が中心で、コミュニティ基盤の中核となるのが、いろいろな分野のサイエンスの研究者です。このコミュニティ基盤の上で地球環境に関する課題を取り上げ、その課題であれば私はこういうデータを持っている、そのデータを使わせてくれれば私はこういう分析ができるというような議論をしながら、研究を進めているという枠組みになっています。

このような形で研究を進めていくことに力を入れているのは言うまでもなく、その他にも力を入れている領域があり、その一つが教育です（図 5）。東京大学での講義、サマースクールなどの企画、アジア地域でフィールドスタディなどを行っています。

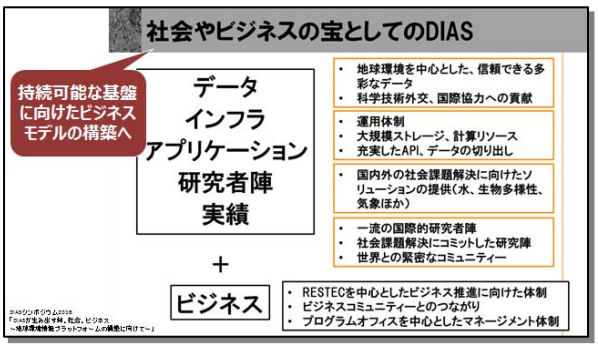
他に産学連携もしています。図 6 は河川のダム操作支援システムの事例です。例えば、東京大学がモデルや手法を考案して、電力会社は電力データを出して、日本工営などの建築コンサルはシステムのオペレーションをするという枠組みで共同作業を行います。

このように民間企業との仕事経験を蓄積していて、今期の DIAS の第 3 期で力を入れている領域がビジネスモデルの構築です（図 7）。なぜ研究基盤がビジネスなのかと思うかもしれませんが、キーワードは「サステナビリティ」です。国家財政もこれから緊縮していく中で、やはりデータ基盤は自分で独自に継続していくことを考えなければいけないのです。European Open Science Cloud でもサステナビリティの議論はいろいろ出ていて、コマーシャルサービスの利用も挙げられていると思うのですが、今、DIAS でもそちらの検討に力を入れています。非常にチャレンジングな課題ですが、いろいろ考えているところです。

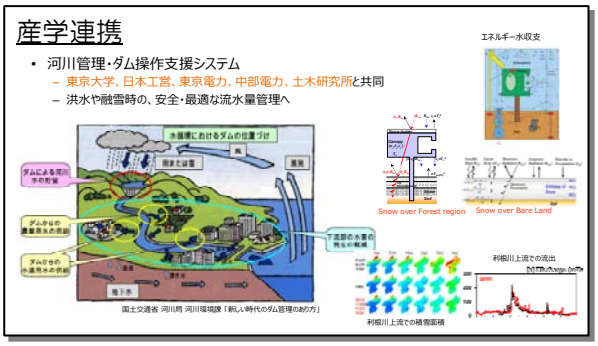
ただ、ビジネスモデルをつくるといっても、別に DIAS のメンバーが MBA を取りに行くというような話ではなく、ポイントになるのは人だと思っています（図 8）。つまり、DIAS の考えにどれだけの人が乗ってきてくれるか、それがポイントになると思うので、潜在ユーザーがどういうところにいるかを調査しながら、リーチする活動をしています。

このように、いろいろな活動を続けているのですが、当然、全てが順調にいくわけではありません。中にはデータ利用者側のニーズとデータ提供者側のシーズが

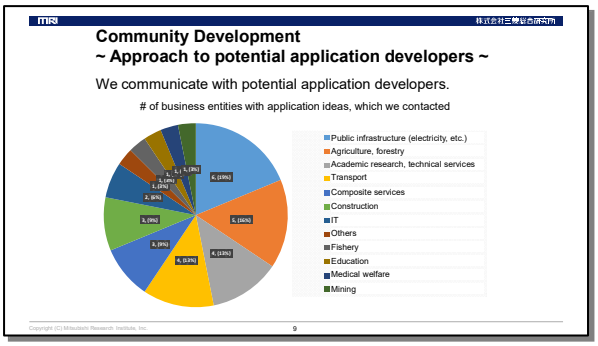
（図 5）



（図 7）



（図 6）



（図 8）

うまく折り合わないといったことが出てくるという問題があります。

データ共有とデータ公開

問題の詳細に入る前に、データ共有とデータ公開の二つのコンセプトについて説明します（図 9）。

データ共有（data sharing）の定義は、特定の個人・組織間でルールを決めてデータを提供することです。一方、データ公開（open data）とは、特別な制限なしで任意の他者にデータを提供することをいいます。

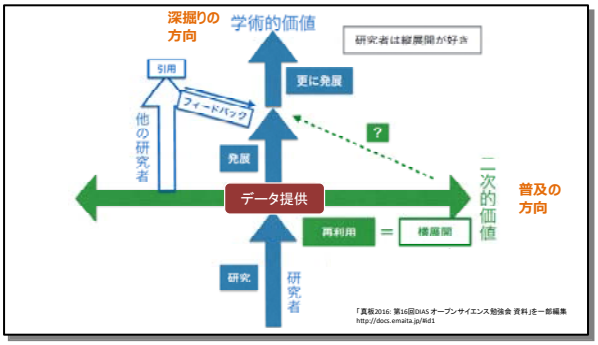
これはオープン・データ・イニシアティブや、ベルモントでも同じような意味で使っていたので、国際的にも通用する定義だと思います。オープンサイエンスというのは、広い意味でこの両者を包含する概念であると考えます。

こうして比較してみたときに、ひょっとすると皆さん、データ公開の方が制限がない分、いいことなのではないかと見えるかもしれませんが、一概にそうとは言えないのです。なぜなら、データ提供者の視点から見ると、利用者からのフィードバックを得やすい枠組みは、どちらかというとデータ共有の方だからです。

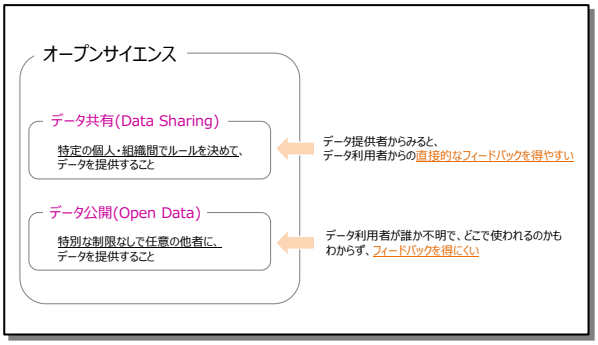
これは少し考えれば分かると思うのですが、特定の範囲でデータだけではなくルールもシェアしてデータを提供し合うという枠組みなので、相手の姿もイメージしやすいですし、コミュニケーションが取りやすいのです。一方で、データ公開の方は、任意の第三者なので要は相手が誰かも見えにくい、あるいはどこで使われるのかも分からない、そもそも連絡が取れるのかも分からないというような問題があって、フィードバックを得にくい枠組みです。

このフィードバックを得られるかどうかは、研究にとって非常に重要な要素です。図 10 は、DIAS の勉強会で国立環境研究所の真板英一さんがベースをつくられた図です。平たく言うと、研究には普及の方向と深堀りの方向の二つの軸があるということです。尖った研究者で、一つの研究テーマをどんどん深掘りしていくタイプは、縦軸の方向に進みます。でも、オープンサイエンスの基本戦略は、どちらかというと横軸の普及の方向性です。

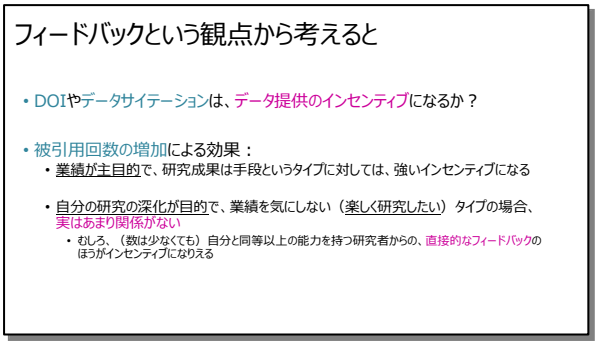
つまり、オープンサイエンスは、オープンにしてみれば誰かが使ってくれて、何か面白いことが起こるかもしれないというような割に楽観的な推論に基づいているのです。確かに誰かが面白い使い方をしてくれるかもしれないのですが、一つの研究テーマを深掘りするタイプにとっては、それは直接的にはあまり関係ないのです。ですから、そういうタイプの人にはどちらもフィードバックが非常に重要で、これが不可欠になります。こうした観点で見ると、DOI やデータサイテーションはインセンティブの本丸といわれていますが、



（図 10）



（図 9）



（図 11）

これがインセンティブとして効く相手と効かない相手がいるということも分かってきます（図 11）。

例えば、業績が主目的で、研究成果は手段というタイプに対しては、強いインセンティブになります。その一方で、自分の研究を深掘りするのが目的だったり、他の人がどう思っているかは関係ない、私は楽しい研究がしたいのだというタイプに関しては、実はあまり関係ありません。そういうタイプにとっては、自分の分野で、かつ、自分と同等以上の能力を持つ研究者からの一言という方が、強いインセンティブになるのです。このような議論をわれわれ DIAS でしてきたので、DIAS はデータ公開よりはデータ共有の方をアイデンティティとしてやってきたという経緯があります（図 12）。

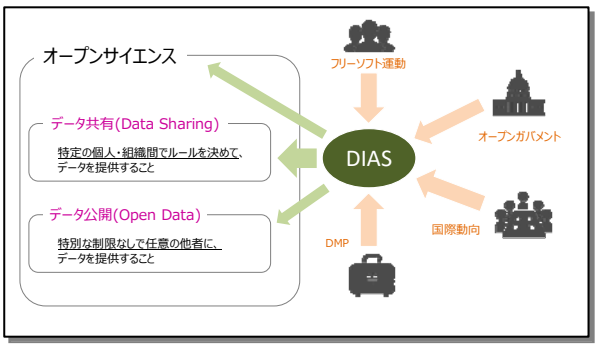
意識調査の実施

このような中、オープン化に対するプレッシャーがだんだん強くなってきて、その方向性も考えなければいけない、対応していかなければいけないという状況になったので、今後の指針を得るために調査を企画しました（図 13）。この詳細については「情報管理」11 月号に掲載されているので、そちらをご覧いただければと思います。今日はその調査の一部と、そこから分かったことを中心にご紹介します。

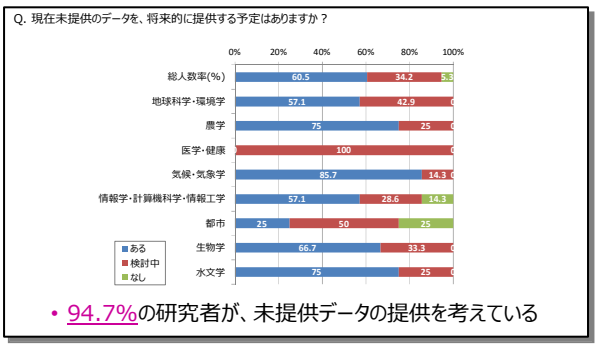
まず、研究者はそもそもデータの提供に反対しているのかという素朴な疑問があります。それに対する答えが図 14 で、要はそんなことはないという結果でした。実に 94.7%の研究者から、まだ出していないけれど、提供の予定がある、あるいは検討しているという回答が得られています。

次に、そんなに考えてくれているのなら、なぜそれを出してくれないのかという疑問が沸き起こると思います。

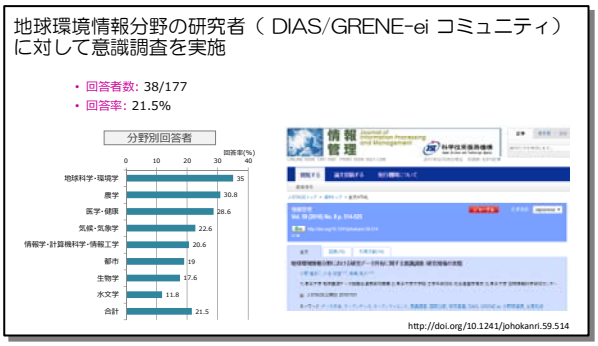
それに対する答えは、「時間がない」ということがやはり一番強い理由になるかと思います（図 15）。この理由が「インセンティブがない」という理由の倍近くを占めています。ここから、仮にインセンティブとして、DOI、データサイテーションが進んだとしても、何らかのサポートがないと思うとおりにデータの提供は進まない可能性があるということが見えてきます。



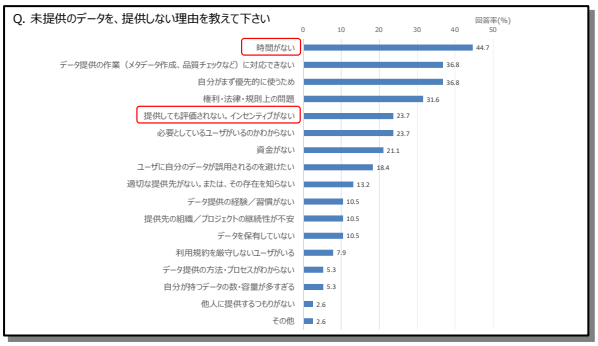
（図 12）



（図 14）



（図 13）



（図 15）

誰にだったら提供してもいいのかという疑問に対する結果は、図 16 です。青い方がメタデータに対する答えで、赤い方がデータに対する答えです。上の方に行くほど公開の範囲が広めで、下の方に行くほど狭めという整理の仕方をしています。やはりメタデータに関しては別に任意の人に見せてもいいけれども、データに関しては、「関係する分野だったらいいけれど、ちょっと公開は」という結果が見えてくると思います。

続いて、いざ提供するときに、データ共有あるいはデータ公開するときの条件はどういうものがふさわしいかという質問をしました。

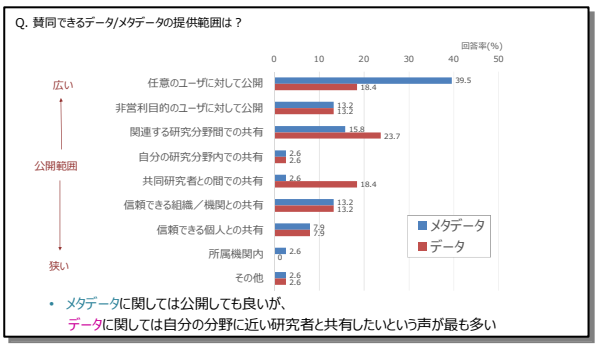
図 17 が、データ共有の条件の結果です。左側が私の調査の結果、要は国内の結果で、右側が私が調査の参考にした DataONE という国外のプロジェクトの結果です。青いバーが、データ提供側としてこういう条件を入れてほしいという結果です。赤いバーが、データ利用者として、その条件であればのんでもいいという結果です。

この結果から面白いことがいろいろ分かりました。三つだけポイントを言うと、一つ目に、国内の結果でも国外の結果でも同様に、データ共有の条件として、そのデータを使ったということを論文で引用すること、データ利用規約を守ることがトップの方に来ています。

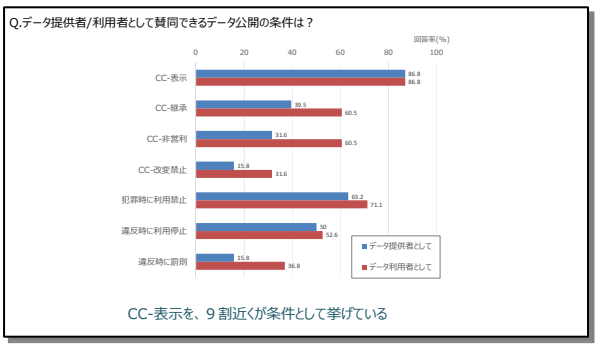
二つ目は、国内の結果では、データの提供者よりも利用者として受け入れ可能な条件が多い、つまり、データを使わせてくれるのなら、その条件は少々厳しくてものむという結果が出ているということです。

三つ目に、プロジェクトへの参加機会を提供するという条件で、国外の結果と国内の結果で顕著に差が出たことです。つまり、あなたのプロジェクトに参加させてくださいというデータ共有の条件を、海外では重視している、国内では重視していないという結果です。これはどういうことかというと、データ共有をコラボレーションのチャンスとして見る向きが海外では強く、国内では弱いという結果なのではないかと解釈しています。

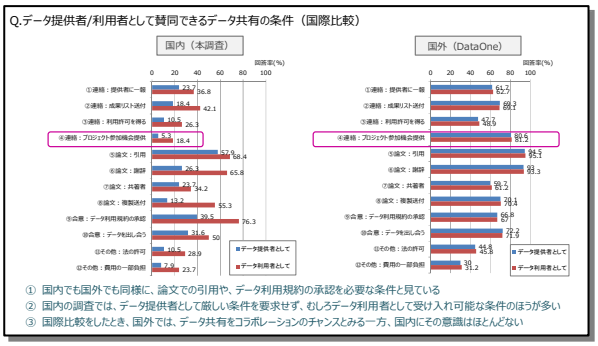
続いて、図 18 はデータ公開の条件の結果です。クリエイティブ・コモンズ・ライセンス（CC ライセンス）は表示がマストだという回答が 9 割近くあることがまず分かりました。



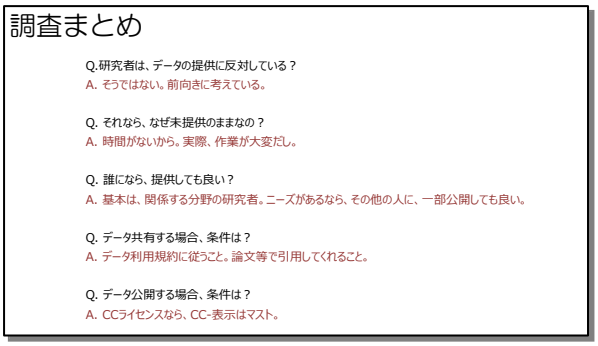
（図 16）



（図 18）



（図 17）



（図 19）

CC ライセンスは、それを守らなかったときも特にペナルティがなく、フリーライディングが起きやすいという問題が指摘されていました。ではペナルティとしてどういうものが適切なのか、私が個人的に興味を持ったので、そのような質問を入れてみました。それによると、罰則まではいかないけれど、利用停止処分はしたいという結果が見えました。

私の調査を FAQ 形式で簡単に整理すると、図 19 のような結果になりました。この調査をやってみて、出てきた結果の全てをまだ消化しきれていないところがあって、それに対する具体的なソリューションがすぐには出てこない状況です。しかし、これが研究者の現在の考えであるということはきちんと受け止めて、今後何をするかを考えていく必要があると思っています。

私たちはこれから何をすればいいのか

これから何をすればいいのか、考えたことを、簡単にご紹介したいと思います。

まず大前提として、焦って間違った方向に行ったり、強制力のあるルールをすぐ採用したりしない方がいいのではないかと思っています。最近、特に気になっているのは、オープンサイエンス系の議論だと、とにかく海外は進んでいて日本は遅れているから早くやらなければいけないという論法を使う人が結構いることで、これに私は違和感があります。

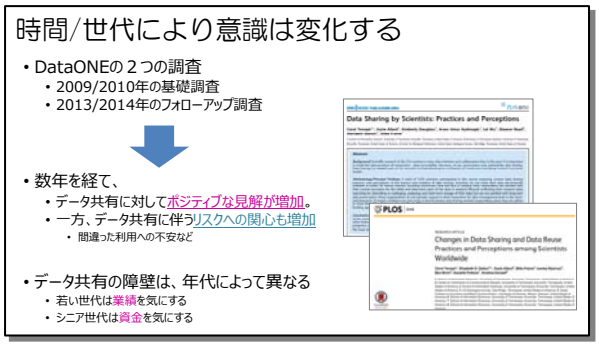
というのは、私の経験からも、確かに海外の方がいろいろなチャレンジ、トライアル、新しいコンセプトが出てはいるのですが、実際に全ての研究者の実態を調査すると、全分野に進んでいる人も遅れている人もいて、その統計結果を出すと、そんなに変わらないのではないかという印象があります。

それを裏付けるかのような調査結果も出ていて、例えばイギリスはオープンデータですごく進んでいるというイメージがあるかもしれませんが、実際には日本と、sharing と not sharing のレートはほとんど変わらないのです（図 20）。ですから恐らく、海外でも進んでいる人は進んでいるけれど、全体としてはそんなに変わらないのです。日本でも海外でも同様に試行錯誤の最中であるというのが実態に近くて、国別に進んでいる、進んでいないというよりは、ライフサイエンスのように進んでいる分野もあれば、全然まだ進んでいない分野もあると言った方が、私にとっては実感として腑に落ちる感じがします。

また、焦って何かをやらない方がいいということを裏付けるかのように、時間がたてば、やはり意識は変化するということが分かってきています。DataONE が基礎調査とフォローアップ調査を行い、数年たって研究者の意識がどう変化したのかを調べた有名な論文では、やはり研究データ共有に対してポジティブな見解がだんだん増えていったという結果が出ています（図 21）。ただ、その一方でリスクへの関心も増加しました。これは私なりに解釈すると、最初にいいことばかりを言い過ぎて、現実の方があまり追い付いていないということがあったので、この落差を見て、本当に大丈夫なのかという不安が増加したのではないかと思います。

ですから、何か結果を出そうと焦ってやるよりは、北本さんからもスピードの話がありましたが、（物事

（図 20）



（図 21）

を進めるときには）適切な速度というものがあるので、そこを見極めながら進めていく必要があるのではないかと思います。

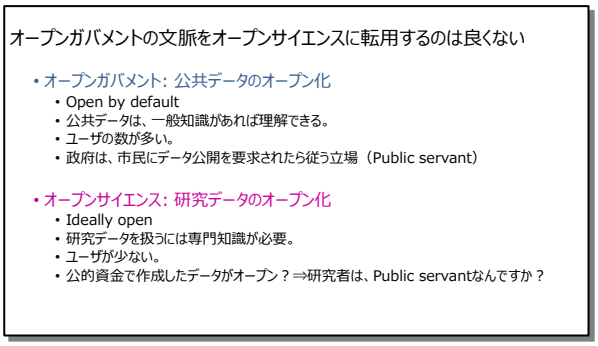
この論文は他にも面白いことが書いてあります。やはり年代によって考え方が違って、若い世代は業績を気にする、シニア世代は例えば教授だともう地位は安定しているので業績はあまり気にせずにプロジェクトで予算が取れるかどうかを気にするということで、インセンティブは年代によってもだいぶ違うのではないかということが分かりました。

ということで、焦ってやるよりは継続の方が重要です。DIAS もこの名前で活動しているのは 2006 年からですが、30 年以上かけて今のところまで来ています（図 22）。

オープンガバメントとオープンサイエンス

その他に、オープンガバメントの文脈をオープンサイエンスに持ち込んで、例えば政府標準利用規約に従って研究データをオープンにしろと言うポリシーメーカーの方がたまにいるのですが、この二つは全く違う概念だということを説明しておこうと思います（図 23）。

（図 22）



（図 23）

例えば、公共データの人口統計データを考えたときに、これは別に特別な知識がなくても使えます。その一方で、研究データは誰が使うのか。先ほど河川の例を出しましたが、水循環の分布型流出モデルのチューニング化された初期値で使うパラメーターセットを誰が使うのだということです。こういうものを第三者にオープンにすることは、あまり意味がないのではないか、その辺のバランスも考えながら進めなければいけないのではないかと思います。

もう一つ重要な論点として、立場の違いもあると思います。例えば、民主主義の基本原理から言って、市民がデータを公開しろと言えば、政府はパブリックサーバントなので出さなければいけないのです。でも、公的資金で作成したデータがオープンという論法は、国に言うのなら分かるのですが、研究者に言うのは適切なのか、私の方でも違和感が残っています。研究者はパブリックサーバントなのかということです。

これは突き詰めて考えると非常に深いテーマだと思います。やはり市民は間違えることがあるのです。そういうときに、こちらの道が正しいのだと誰が言うのか。その役割を負っているのが研究者ではないか、だから研究者はパブリックサーバントというよりは、ファクトに対するサーバントであるべきだと思っています。いくら国や市民が言ったとしても、それを全て鵜呑みにするのは、民主主義を正常に回すために良くないのではないかということを考えています。

データ・マネジメント・プラン（DMP）

私も海外で実際につくられているデータ・マネジメント・プラン（DMP）はどんなものか見てみたのですが、文章で説明して書いてあるものから、データの細かいスペシフィケーションを書いてあるものまで、まちまちでした（図 24）。これをつくることにどのくらい実効性があるのかという疑問が出てきて、きちんと検証した方がいいのではないかと思いました。しかし、小賀坂さんの報告を聞くと、その辺は慎重に丁寧にされているということなので、安心しました。逆に、

日本のアドバンテージは、いったんルールが決まると丁寧にやることだと思うので、国内で丁寧な作業、仕事をして、日本はうまくいっているという事例を見せられたらいいのではないかと考えています。

私たちはこれから何をすればいいのかをまとめると、すぐに結果を出そうと焦らない、やるべきことは継続する、やること（やったこと）は有効だったかどうか検証することではないかというのが私の考えです。

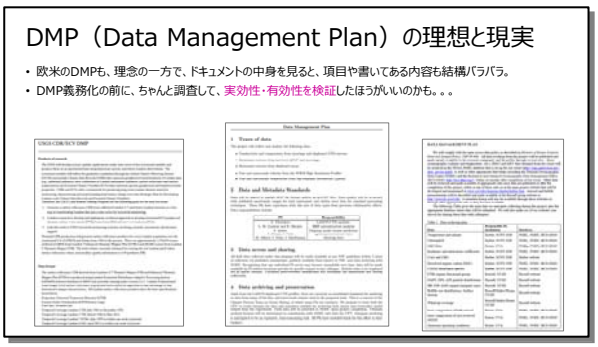
文献とデータの比較

最後に、SPARC Japan セミナーは図書館の参加者が多いイベントで、図書館は、文献管理で得たノウハウが研究データに転用できるのか悩んでいる方が多いのではないかと予想したので、文献の世界とデータの世界を少し比較してみました。

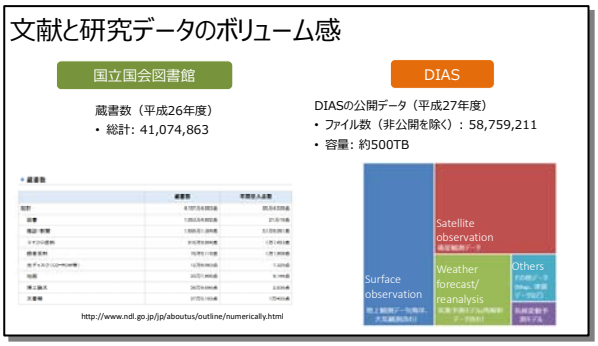
まず、文献と研究データのボリューム感は、国立国会図書館の今の蔵書数が約 4,100 万であるのに対して、DIAS の公開データが昨年度時点で約 5,800 万でした（図 25）。DIAS の本体は非公開領域なので、それを入れると数倍以上に膨らむかもしれませんが、このぐらいのボリューム感です。これだけのボリュームがあると、メタデータをどうやって付ければいいのかが悩ましいところで、適切なメタデータの粒度に悩んでいるところです。皆さんも、国会図書館のメタデータをもう一回ゼロから付け直すと言われたら、かなりうんざりで、それは無理だと思いますよね。そういうデータの世界で悩んでいるのがわれわれだということです。メタデータの粒度問題について本当に悩んでいるので、もしいいアイデアがあれば教えてほしいです。

図 26 は大きさの比較です。世界一大きい本は通称『オバママニア』という本で、厚さが約 34cm だそうです。世界一大きい画像データはアルプス山脈の図で、約 46TB あるらしいです。

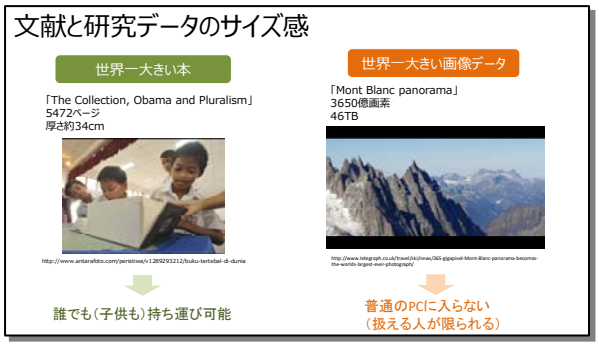
言いたいことは、データの方が大きいとかそういうことではなく、大きい本は子どもでも何とか持ち運べるサイズである、でもデータは移動させるだけでも特別な環境がないとできない、こういう違いがあるということです。この違いは結構大きいのではないかと思っています。



（図 24）



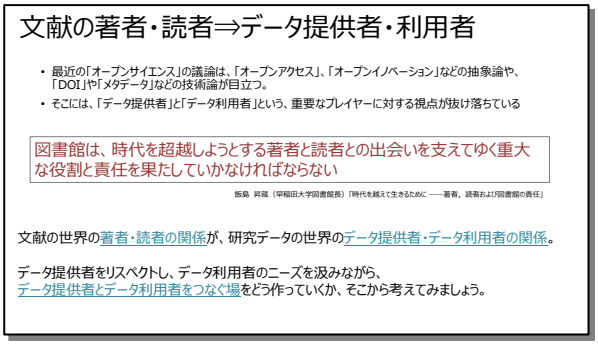
（図 25）



（図 26）

まとめ

最近のオープンサイエンスの議論は抽象論、技術論に引きずられている印象があります（図 27）。そこにはデータ提供者とデータ利用者という重要なプレーヤーに対する視点が抜け落ちていると感じています。データ提供者と利用者は、文献の世界で言う、著者と読者の関係です。その著者と読者を歴史を越えてつなげていくという役割が図書館にあるのではないかと思っています。ですから、データ提供者とデータ利用者をどうつなげていくか、そこのアーキテクチャをどうつくっていくかという視点で今後を考えていっていただければうれしいです。データキュレーター、データライブラリアンの役割もそこにどう関われるかがポイントになるのではないかと考えています。



（図 27）