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」

英国における研究データ管理支援の動向

池内　有為
（筑波大学大学院）

講演要旨

研究データの公開は、再利用の効果や研究の信頼性の向上を背景として自然科学分野から社会科学分野に至るまで拡大している。さらに、助成機関や学術雑誌による義務化も拡がり、研究データの公開は研究のライフサイクルに組み込まれつつある。しかし、公開には、資金、時間、技術といった物的・人的資源が不可欠であり、研究者にとって負担が大きい。そこで各国の大学図書館は研究データ管理の支援に乗り出している。本講演では、英国のデジタルキュレーションセンター、エディンバラ大学、グラスゴー大学の取り組みについて、訪問調査の結果を紹介する。

池内　有為

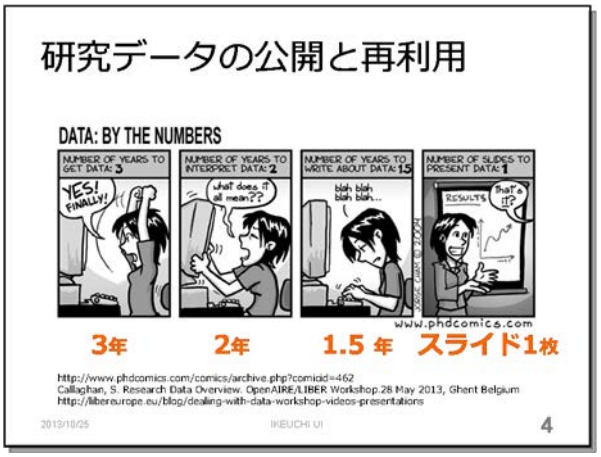
博士論文のテーマとして研究データ管理に取り組んでいる。1995年慶應義塾大学法学部政治学科卒業。1997年慶應義塾大学大学院図書館・情報学修士課程修了。1997-2005年フェリス女学院大学附属図書館勤務。筑波大学大学院図書館情報メディア研究科博士後期課程2年次在学中。

本日は、まず、研究データの公開と再利用の背景について概説した上で、8 月に訪問したエディンバラ大学、グラスゴー大学、そしてイギリスの大学の研究データ管理活動を強力にサポートしているデジタルキュレーションセンターでのインタビュー調査の結果から、イギリスの大学図書館における研究データ管理支援についてご報告させていただきます。そして最後に、今後の日本の研究データ管理支援に向けて、自分なりの考えを述べさせていただきます。

研究データの公開と再利用の背景

この漫画は博士課程の大学院生が、データの取得に 3 年、その解析に 2 年、記述に 1 年半、全部で 6 年半かかって、発表のときはスライドわずか 1 枚という、研究者として笑えないオチが付いています（図 1）。

もし彼女がこの研究データを公開したら、それを再利用する利用者は、最初の 3 年間をスキップして、わずか 3 年半でスライドが 1 枚作れることになります。



（図 1）

素晴らしいです。

これが研究データ公開のメリットで、ゲノムデータの共有による生化学分野の目覚ましい発展は、皆さんご存じのとおりです。そして、研究データの公開と共有は、天文学や海洋学などさまざまな分野に広がっています。

研究データ公開の別の側面として、研究結果の検証があります。つい最近も、偽論文をオープンアクセス・ジャーナルに 304 本投稿するという実験をしたところ、157 誌がアクセプトしてしまったという調査が「Science」に掲載され、話題を呼びました（図 2)。

査読や出版社の問題、そして論文捏造など、科学の信頼性に関わる事件は後を絶ちません。インパクトファクターが高い雑誌ほど捏造が多いという残念な報告もあります。データの公開によって研究結果の検証を可能にし、研究の信頼性や透明性を担保することもデータ公開の大きな推進要因とされています。

さらに、研究公開の義務化も大きな牽引力となっています（図 3)。アメリカ政府や G8、欧州協議会といった政府機関、各分野の学術雑誌、助成団体、大学などの学術機関によって、データ公開ポリシーが続々と発表されています。ここで雑誌の例として挙げた英国医学会の BMJ は、オープンデータキャンペーンを展開しています。一市民として、医学分野のデータ捏造は切にやめてほしいと思います。こうした潮流を受けて、2012 年には、50 カ国の助成機関の連携組織である Global Re-search Council や Research Data Alliance といった機関が相次いで発足しました。

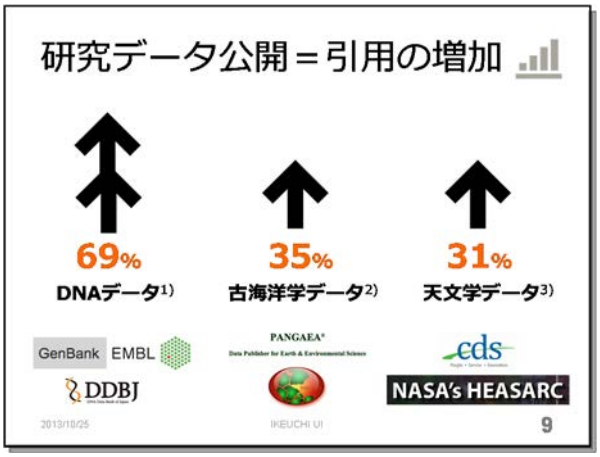
では、研究データを公開したらどうなるのか。研究者にとってのメリットの一つとして、データを公表することによって、論文の引用までもが増加することが明らかにされています（図 4)。例えば、2007 年には、DNA データを GenBank などのリポジトリに公開している論文は、そうでない論文に比べて、引用率が 69%も上がることが示されました。ちなみに、この研究は ImpactStory のコファウンダーでもある Heather Piwowar さんによるものです。

また、古海洋学や天文学分野についても、データ公開によって論文の引用が増えることが明らかにされています。今後は、引用の間接的な増加ではなく、データそのもののインパクトも直接測られようとしています。

しかし、データ公開の実現にはさまざまな障壁があ

（図 3）

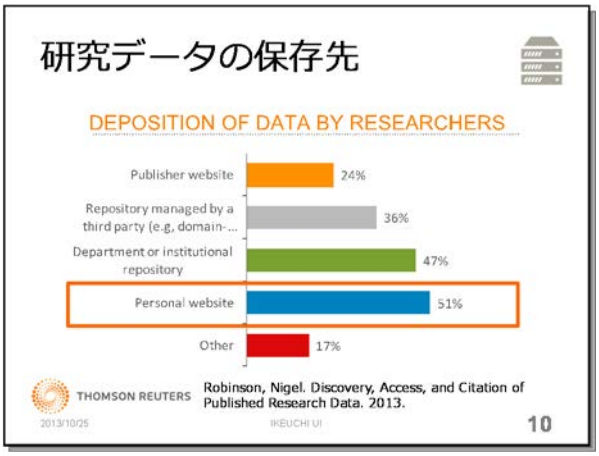
（図 2）



（図 4）

ります。例えば、データの保存先について、データの過半数は研究者個人のウェブサイト（Personal website）に保存されていることが報告されています（図 5）。私は現在、学術雑誌のポリシーを調査しているのですが、「個人サイトへの登録は、データの永続性が保証されないので認めない」というポリシーも見られます。では、研究者が投稿しようと思った雑誌が、「個人のウェブサイトは駄目、出版社の方（サイト）でもデータの掲載を認めない」という場合、研究者はどこに助けを求めればいいのでしょうか。

ここで白羽の矢が立ったのが、大学図書館です。データ公開と共有については科学雑誌でたびたび取り上げられていますが、今年 3 月 28 日に出た「Nature」には、「再起動する大学図書館」という記事が掲載され、大学図書館によるデータ公開支援に期待していると述べられています（図 6）。この記事では、大学図書館はもう何世紀にもわたって情報を整理して、保存して、提供してきたのだから、データも少し複雑だけれども、同じだろうと書かれています。



（図 5）

（図 6）

いやいや、同じではないでしょう。ちょっと待ってください。私はフェリス女学院という小さな大学図書館で働いていましたが、「池内さん、明日からデータリポジトリ担当、よろしくね」と言われても、「はい」とは言えないでしょう。私は割と怖いもの知らずで、何でもポジティブに引き受けてしまうのですが、「ちょっと考えさせてください」と言うと思います。

例えば大規模サーバーを扱ったり、勤務先に医学部があったとして、「医療分野のデータが来たから、メタデータを振ってね」と言われても、対応できる気がしません。でも、海外の大学図書館ではやっているらしい。

では、ライブラリアンはどうやってデータ管理支援の知識やスキルを身に付けているのか。費用は一体どうしているのか。問題点はどのようなことがあるのか。幸いにも筑波大学と日本図書館情報学会から助成金を頂くことができたので、イギリスに行って直接聞いてみることにしました。

英国の大学図書館における研究データ管理支援

訪問先として選んだのは、早い時期から研究データ管理支援を行ってきたエディンバラ大学とグラスゴー大学です（図 7）。イギリスの大学による研究管理支

（図 7）

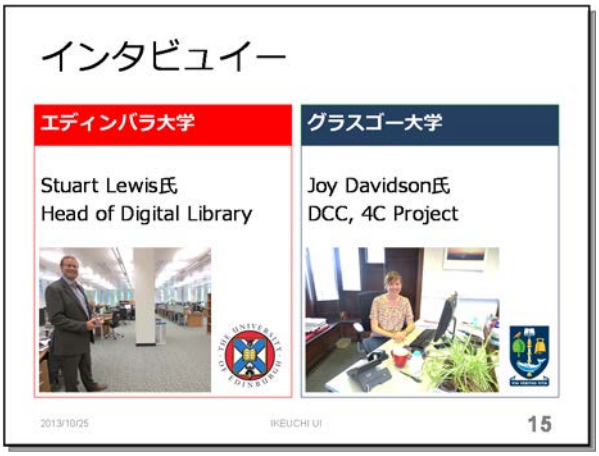
援をサポートしているデジタルキュレーションセンター（DCC）の本部があるのがエディンバラ大学です。ちなみにグラスゴー大学も DCC のメンバー館で、4C プロジェクトといって、データ管理に掛かる費用を計算しているチームの一員でもあります。

インタビューを引き受けてくださったのは、エディンバラ大学は Digital Library のヘッドの Stuart Lewis さん、グラスゴー大学は DCC や 4C プロジェクトのメンバーでもある Joy Davidson さんです（図 8）。

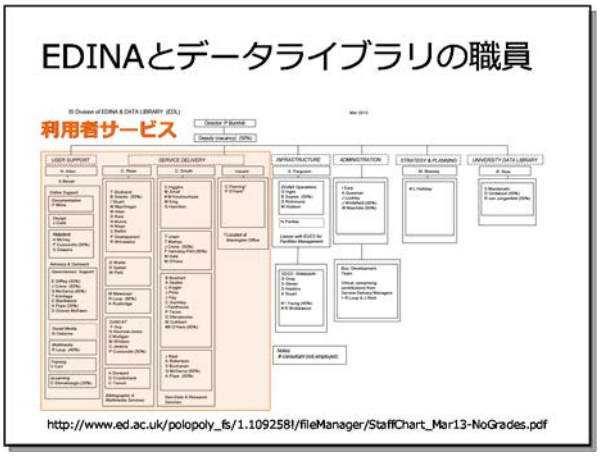
まず、先日ノーベル賞を受賞された Peter Higgs 博士を擁するエディンバラ大学についてお話しします。長い歴史と Charles Darwin や Graham Bell といったそうそうたる出身者がいる大学です。エディンバラ大学には、イギリスのデータセンターである EDINA と、図書館のデータライブラリーの二つがあります。職員についてもサイトに掲載されているのですが、スタッフが多過ぎて、とてもスライドには入りきらないくらいでした（図 9）。この他にも大勢いらっしゃいます。

では、職員構成図を見せようと思ったのですが、全体を見せると字が小さ過ぎて見えなくなってしまうくらいの職員構成です（図 10）。ただし、フルタイムだけではなくてパートタイムも入っていると Stuart は言っていました。6 部門 10 グループのうち、オレンジの網掛けをした部分が利用者サービス部門です。つまり、これだけのスタッフがエディンバラ大学に所属する研究者に対してパブリックサービスを行っているのですが、直接的な対面サービス以外にも、研究者が自分でデータ管理に当たれるようにウェブサイトも充実させています。

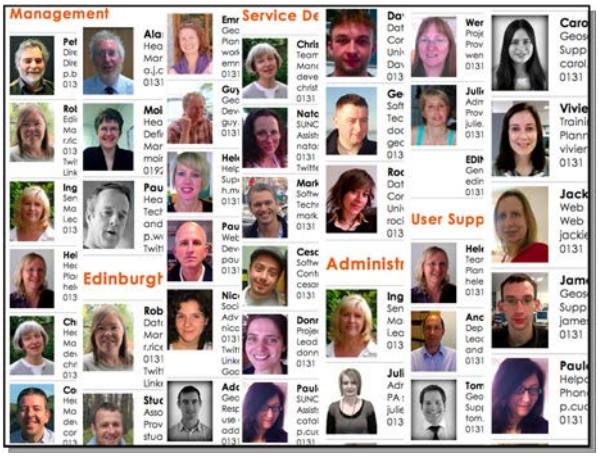
研究データ管理のガイダンスのページ（図 11）や、MANTRA という研究データ管理のトレーニングのオンラインコース（図 12）も用意されています。これ



（図 8）



（図 10）



（図 9）

（図 11）

は研究者だけでなく、サービス担当者にも向けて作られています。英語ではありますが、日本からもアクセスできるのでぜひ使ってほしいとおっしゃっていました。

まとめますと、エディンバラ大学は、大規模なリポジトリがあって、大勢のスタッフがいて、対面でもオンラインでも充実したトレーニングコースを提供している、いわば王道の研究データ管理支援サービスを実施しています。

続きまして、グラスゴー大学です。皆さんもご存知のとおり、1450 年代に設立されたイギリス最古の大学の一つで、Adam Smith や James Watt の母校でもあります。ノーベル賞受賞者も 7 名輩出しています。

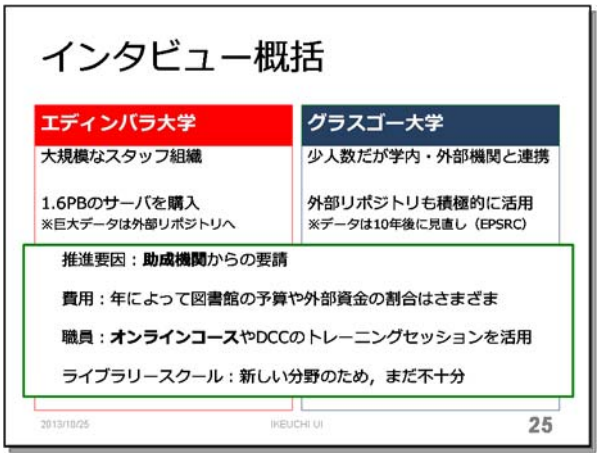
グラスゴー大学でも、エディンバラ大学と同様の研究データ管理支援サービスを提供しているのですが、こちらはごく少人数のチームでした。インタビューを受けてくれた Joy にスタッフについて尋ねたところ、5 人のスモールチームなのだと教えてくれました。その代わり、学内の部署、例えば助成申請の部署や IT サポート部門などと連携して支援を行っているのだそうです。また、DCC サービスも積極的に活用しているので、5 人のスモールチームだけれどもやっていけるということでした。これはなかなか心強い、少人数でもやっていけるのだという例だったと思います。

二つの大学の相違点と共通点をまとめてみます（図 13)。エディンバラ大学は、大きいスタッフ組織があり、データセンターも大きいのですが、先日、1.6PB のサーバーを購入したというので、幾らしたのか聞いてみました。「値段は非公開」だそうなので、具体的には言えないのですが、数十億円くらいです。それでもなお、Higgs 博士が扱っているようなビッグデータは CERN のような外部リポジトリに出して、自分たちはスモールデータだけを管理しているのだと言っていました。スモールデータといっても 1.6PB ですが・・・。

グラスゴー大学もデータリポジトリを持っているのですが、外部リポジトリも積極的に活用していこうという話をしていました。figshare はどうかと聞くと、経費の節約になるから、そういうものも積極的に取り入れていきたいということでした。

もう 1 点、面白かったのは、データを永久保存ではなくて定期的に見直すのだという話です。私は元図書館員なので、データを保存するといったら永久に保存しなければならないと考えていました。そうすると、エミュレーションやマイグレーションといった問題が出てくるし、その容量だけではなく、管理コストが莫大になると懸念していました。しかしグラスゴー大学では、10 年後にいったん見直して、あまり使われていないデータや不要なものがないかを見直すというのです。誰がそれを見直すのかという話をしたら、「研究者かな」と言っていました。見直すことも考えると、やはりデータにしっかりと DOI を振ることや、利用の動向を把握できるようにすることが、大事だと思います。

（図 12）



（図 13）

両者で共通していたのは、まず（研究データ公開の）推進要因です。ここ 1～2 年間で助成機関からの要請が一気に厳しくなりました。研究者は、助成金を取って研究をして、業績を積まなければならないので、助成機関の要請は受けざるを得ません。それで、データ公開に図書館が関与せざるを得ないのだと言っていました。

気になる費用について、図書館がお金を出しているのか、大学が出しているのか、助成金は出るのかと聞いたのですが、まだ始まったばかりということもあって、年によって図書館の予算を使ったり、大学がお金を出してくれたり、JISC などから外部資金をもらったり、割合はさまざまだということです。

職員はこのような新しい仕事にみんな対応できるのかと聞いたら、先ほどのようなオンラインコース、DCC のトレーニングセッションを活用して頑張っているということでした。では、未来のライブラリアンが通っているライブラリースクールの教育はどうかと聞くと、DCC でもグラスゴー大学でもエディンバラ大学でも、「まだ新しい分野なので、教える人もいないし、カリキュラムもうまくできていない。アメリカはデジタルキュレーションのコースがあるが、それは仕方がない。まだ不十分だけれども頑張る」ということでした。

一番厳しかったのはエディンバラ大学の Stuart で、「若い人たちにはもっと専門知識を付けてほしい。半年の研修ではなく、その分野の専門家、研究者になるくらいの知識を付けてほしい」とおっしゃっていました。

研究データ管理のオンラインコースの例をご紹介します（図 14）。こちらは JISC が助成した RDMRose というもので、八つのセッションがあって、大学の講義を受けるように研究データ管理について学ぶことができます。イギリスでもアメリカでも、ウェビナー（webinar）がよく開催されていて、私も 1 回参加してみたのですが、私以外全員が現職のライブラリアンで、研究データ管理の担当になったから受けに来たという感じでした。

デジタルキュレーションセンター

イギリスの大学図書館による研究データ管理支援をサポートしているのが、デジタルキュレーションセンター（DCC）です。インタビューを受けてくださったのは、Director の Kevin Ashley さん、Jonathan Rans さん、Angus Whyte 博士です（図 15）。

図 16 は、DCC が提供している研究データ管理サービスの構築ガイドです。まずは、ポリシーと計画を立てることから始めます。これは当たり前のように思えるのですが、皆さん口をそろえて「ここを頑張れよ」とおっしゃるのです。「そんなに大変なのですか」と聞いたら、とにかく大きくて古い大学が大変だ。教授がすごく強くて、コンセンサスを取るのが大変で仕方がない。逆に新しい大学の方が、図書館員がどんどん

（図 14）

（図 15）

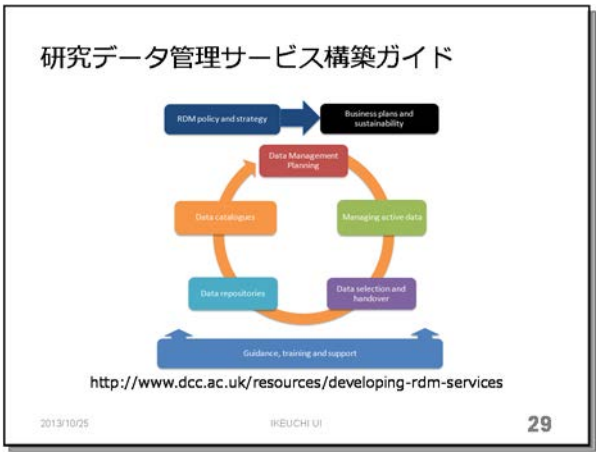
イニシアチブを取って決められるので簡単だとおっしゃっていました。グラスゴー大学では、意見を集約するだけで半年ほどかかったそうです。

DCC の支援活動として、国内のベストプラクティスを紹介したり、助成機関のポリシーを取りまとめて紹介したりしています。この表はイギリスの各助成機関のポリシーの一覧です（図 17)。機関によって違いがあるので、各大学の担当者はターゲットとする助成機関のポリシーを確認して、例えば、出したいところにリポジトリが提供されていない場合は、自前で用意することを考えるとか、適切な外部リポジトリを紹介するといった計画を立てていきます。

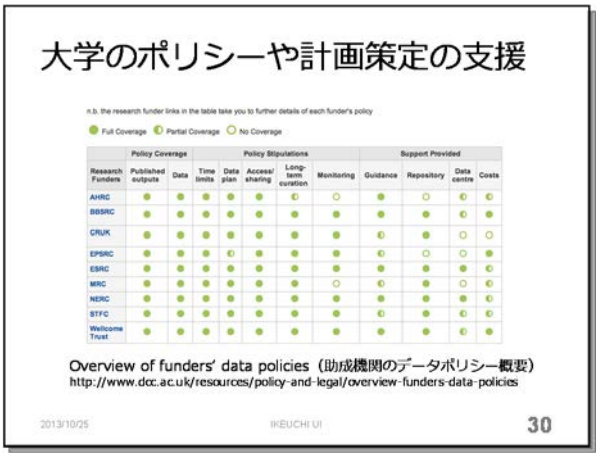
DCC が実施したトレーニングの開催大学の一覧がサイトに掲載されています（図 18)。ここで示しているのは一部だけで、ほとんど全てのイギリスの大学でやったとおっしゃっていました。グラスゴー大学の Joy が南アフリカでもやったと言うので、「日本でもやってくれる？」と聞いたら、大丈夫だろうと言っていました。サイトで確認したら本当に南アフリカでもやっていて、ただし、ウェビナーでした。多分、日本もお願いすればウェビナーの配信をしてくれると思います。

このリストを見ると、何度もグラスゴー大学が出てきていて、なるほど Joy はうまくこれを使って学内のトレーニングをやっているのだと思いました。対象としては、（ライブラリアンだけではなく）私のような大学院生や若手研究者向けのもの、サービス担当者向けのものもあります。

トレーニングに力を入れていることも分かりました。それでも、「やはり私に医療分野のメタデータを付けてと言われても、付けられないのですけれど」と、さらに食い下がってみました。そうしたら、ディレクターの Ashley が、いそいそとパソコンを壁のモニターにつなぎ始めて、「私たち DCC は分野別のメタデータの標準の例をまとめて、ここに掲載している。だから、この中から適切なものを選び出して使えばいい。自分で一から考えなくてもいいのだ」と言いました（図 19)。力強く励まされて、後進は要領よく先例をうまく使ってやっていけるのかなと、かなりポジティブな気持ちになりました。



（図 16）



（図 17）

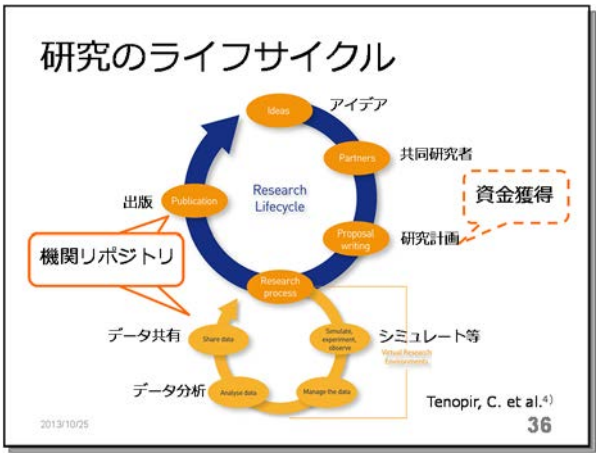
（図 18）

日本の研究データ管理支援に向けて

図 20 は、研究のライフサイクルを図式化したものです。今、大学図書館が関わっているのは出版の部分、機関リポジトリです。もし、大学図書館が研究データ管理も行うことになったら、データ共有の部分、場合によっては助成金申請の資金獲得の部分にもコミットするかもしれません。一方で、さまざまな外部の商用無料のデータリポジトリ、サービスも提供されています。

こうしたサービスをうまく使って、自分の大学の研究者のニーズに照らし合わせて管理計画を立てることが大切なのだと思いました。（インタビュイーの）皆さんが繰り返し、計画が大事だと言ったのはここの部分なのだなと思いました。「ライブラリースクールの学生に必要な能力は何か？」とそれぞれに聞いたのですが、グラスゴー大学の Joy は「とにかく専門知識より、コミュニケーション能力だ」と言っていました。「コミュニケーションか・・・」と思ったのですが、やはりこういうことを考えていくのに当たって、コミュニケーション能力、つまり研究者のニーズを的確に把握して、支援すべき部分を明確に理解して、提供することが大事なのだと思いました。

（図 19）



（図 20）

研究データは 2009 年に 0.7ZB だったのが、2020 年には 35ZB（35 兆 GB）になると言われています。イギリスは 2020 年までに統合検索システムを開発する予定だそうです。

アメリカでは、NSF によって、今年 1 月から、研究業績として、論文や特許だけでなく、データセットやプログラムコードも評価対象になる、すなわち業績として書けるようになることが決まりました。やはり研究者の識別子、名寄せがますます重要になると思います。

また、個人的に私が申し上げたいのは、論文へのリンクを徹底してほしいということです。メタデータが付いていても、やはりデータだけを見てもよく分からないものが結構あるのです。でも、そこに論文がくっついていれば、格段に理解が深まります。さらにわがままを言うならば、再利用した論文もリンクしていただけるとありがたいと思います。データから逆引きするというか、データから論文を見られれば、研究の参考にもなり、評価にもつながっていくのではないかと思います。

もし日本の大学図書館が研究データ管理支援を始めるならば、まずは先行事例や周辺知識について海外のトレーニングプログラムを活用することが大切です。もちろん英語のではありますが、何とか頑張って、先行事例を使って進めていければと思います。

そして、図書館が支援するべき部分を明らかにして、先人の言うことをよく聞いて、しっかりポリシーと計画を立てることです。

最後に、データリポジトリなのですが、全ての大学がエディンバラ大学のように 1.6PB のサーバーを買うことは不可能なので、構築できない場合は、適切な外

部のリポジトリを紹介することも必要ではないかと思いました。

メタデータについては、DCC の影響を受けて、まずは標準化されたものを使っていこうと今は思っています。そして、くどいようですが、大事なことなのでもう一度言わせていただくと、論文のリンクをぜひぜひお願いします。データから見ると、論文は最強のメタデータだというのが、今、私が研究していて最も強く感じているところです。

●Q1　核融合科学研究所に勤めています。私は図書館員ではなく研究者なのですが、データリポジトリやデータ収集については非常に興味深く聞かせていただきました。

一つ質問させていただきたいのは、ここでターゲットとするデータはどのレベルのものを想定されるのかということです。例えば、実験のときに計測器から上がってきた 1 次データなのか、それを機械処理した 2 次データなのか、あるいは研究者がそれに対して評価した 3 次データなのか、どのレベルでの収集をターゲットにされているか、もし分かれば教えてください。

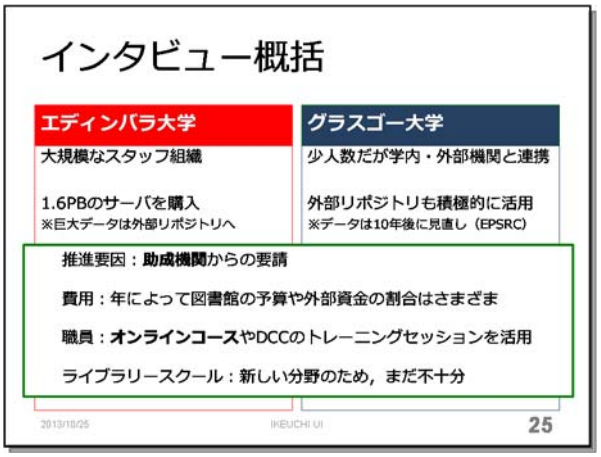
●池内　具体例として、エディンバラ大学やグラスゴー大学がどこまでやっているかは直接聞いてこなかったのですが、各研究データのリポジトリの例を一つ一つ見ていきますと、本当に生データを上げてしまっているものもあれば、ある程度 2 次加工をして分かるようになっているものも確かにあります。コンピュータープログラムであれば Readme ファイルや解説書が付いているものもあります。ですから、統一されたフォーマットなどはないと思いますが、データを格納するときには、そのリポジトリや雑誌ごとに、どこまで上げなければいけないかは決まっているようです。

●Q2　遺伝研の研究者です。こういうデータはもちろんいろいろな人に有効で、リポジトリすればすぐにでも使う人がいるかもしれないのですが、例えば 10 年前のデータ、あるいはすごく古いデータを見つけて、全く新しいノーベル賞級の仕事をすることもあり得ると思うのです。そういうことを考えると、グラスゴー大学の 10 年後に見直しをするというポリシーは、どういう論理で 10 年間なのでしょうか。

●池内　すみません、説明が不足していました。図 21 に小さく「EPSRC」と書いてあるのですが、これは非常に有力な、お金をたくさん出してくれる助成機関です。こちらのポリシーとして 10 年でいったん見直せという指示があるそうです。10 年間で 1 回も使われなかったデータは、その時点で保存しなくてもいいことになり、代わりに、その間に 1 度でも使われたものは、そこからまた 10 年間保存するというルールが、EPSRC に関してはあるそうです。

ちなみに私はデータ公開に関する学術雑誌のポリシーを調べているのですが、分野によっては、データは 5 年間、コードプログラムは 2 年間、必ずアクセスするようにしなさいという形で期限を区切っているものも存在します。ただ、保存期間については、統一された世界規格があるというわけではありません。

●Q2　もちろん、ローデータをすぐに使いたい人がいることは分かりますし、10 年間 1 回も使われなかったデータは要らないのではないかという論理は分かります。一方で、どこかに隠れているデータを 10 年



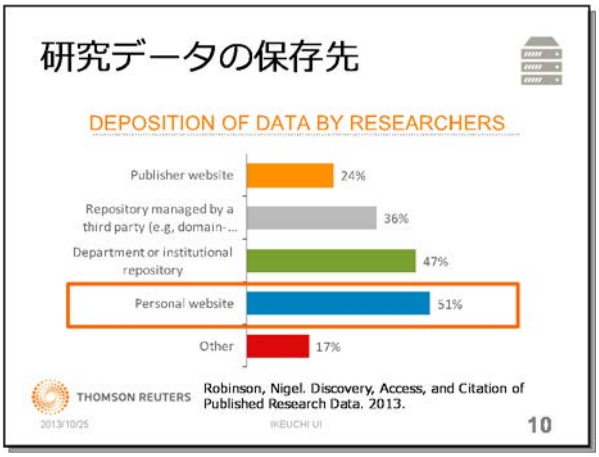
（図 21）

後に発見することもあるだろうし、そもそも学問の転換を考えると、10 年ではなく 50 年くらいのタイムフレームで動いているのではないかと思うので、10 年というのはお金の制約ではないかと思うのですが。

●**池内**　やはり管理と容量の問題の両方があると思います。本当は、できることなら全てのデータを永久保存するのが究極の理想だと私も思います。

●**Q3**　ピッツバーグ大学の職員です。図 22 にあったデータ保存先のパーセンテージをトータルすると、175%になってしまいます。そうすると、Personal web-site も 30%以下になってしまうのですが、このパーセンテージはどこから出てきたのでしょうか。

●**池内**　これはトムソン・ロイターの調査なのですが、恐らく複数回答の項目で、研究者が「あなたのデータはどこに保存していますか」と聞かれたときに、「Publisher website にも Personal website にもある」という回答をしたのではないかと予想します。うかつにも気付かず使ってしまったので、後ほど原典に当たり、お返事差し上げます。



（図 22）