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」

研究データマネジメントと日本の大学

青木 学聡

（京都大学情報環境機構）

CC BY NC

講演要旨

オープンサイエンスへの貢献を含め、大学、研究機関において、研究データを適切に管理することは、新たな知の創生、研究成果の社会への還元、公正な研究活動の維持等、多くの理由からその整備が強く求められている。研究データの入手、保管、公開のルール、すなわち研究データマネジメントの確立は、欧米においては近年、必須の事項とされたが、日本ではまだ初期段階にあり、各大学は手探り状態にある。 本講演では、欧米における研究データマネジメントの位置づけ、大学での取組みを紹介し、日本の大学における研究データマネジメントの今後の展開を議論する。

青木 学聡

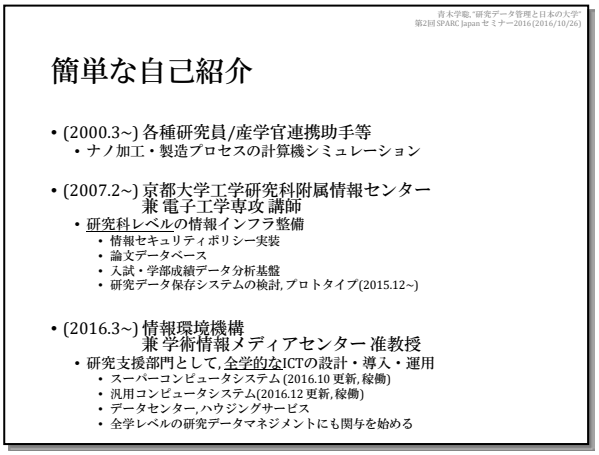
博士（工学）。ナノスケール加工・計測技術のシミュレーションを中心テーマに、2000年より各種研究プロジェクト研究員等として活動。2007年2月より工学研究科講師。教育研究活動と並行し、同附属情報センターにて研究科内の情報セキュリティ、情報インフラ整備、データ分析業務を担当。2016年3月より現職。 大学全体にわたる研究者、研究プロジェクト支援のためのICTシステムの計画、設計、運用に携わる。

1.研究データマネジメントへの関与

私は 2000 年にドクターを取得し、ナノ加工・製造プロセスの計算機シミュレーションを専門に仕事をしていました（図 1）。

2007 年から京都大学工学研究科附属情報センターという非常にマイナーな組織に属して、一方で健啖として電子工学専攻の講師をしていました。そこでは、研究科レベルの情報インフラ整備を行っていました。特に、必要だけれどインセンティブが起きにくい情報セキュリティポリシーの実装をしました。また、京都大学は論文リポジトリを持っていますが、工学研究科でも論文データベースを持っています。ともすれば形骸化しそうなものを、実体を持ったものにするために、それをどうやって研究評価に生かすかも検討しました。入試や学部成績データ分析の相談も受けました。本セミナーのテーマに関係する話では、2015 年 12 月から、研究データ保存システムの検討、プロトタイプの作製にたずさわりました。

このような仕事はあまり外に出ない仕事で、ついつい研究がおろそかになってしまって、どうしようかと



（図 1）

考えていたときに、全学機構である情報環境機構へ異動させていただくことになりました。現在は研究支援部門として、全学的な ICT の設計・導入・運用を行っています。例えばスーパーコンピューターシステムの運用支援、あるいは汎用コンピューターといい、京都大学は幾つか仮想サーバーサービスを提供しているので、それの基盤となるシステムの運用支援をしています。場合によっては、先生たちが持っているサーバー機器を空調の効いたデータセンターでお預かりすることもしています。

最近、全学レベルでの研究データマネジメントをどうすればいいのかという話が出てきて、こういうものにも関与を始めている状況です。

2.研究データ管理の多義性

私は研究データ管理に関わりはじめて半年です。半年間で人からいろいろと話を聞き、このように整理するのがいいのではないかと思った話をします。

研究データ管理には、三つの意味合いがあるのではないかと思います。「研究公正とコンプライアンスのため」「オープン（データ）サイエンス促進のため」「学術領域の発展と社会貢献のため」です。これは最初から順番に、研究者にとってみればつまらないものから面白いもの、といえるかと思います。

2-1.学術領域の発展と社会貢献のため

まずは、一番アトラクティブですが、あまり手を出せない、「学術領域の発展と社会貢献のため」についてからお話しします。

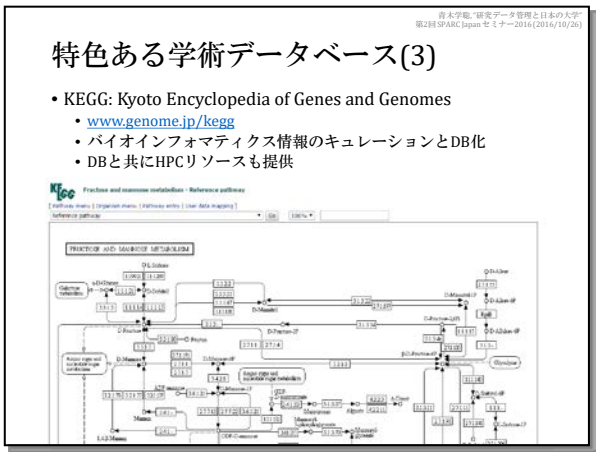
京都大学はやはり大きい大学で、しかも各部局の独立性が高いです。いろいろな小さい研究ユニットやセンターを持っていて、そもそもデータを集めて公開することが存在意義だというセンターもたくさんあるので、非常にたくさんの学術データベースがあります。

例えば、京都大学が保有している貴重資料画像を高解像度でお見せする図書館のデータベースがありますし（図 2）、Kyoto Encyclopedia of Genes and Genomes（KEGG）というバイオインフォマティクスのデータベースなどもあります（図 3）。このあたりのデータのアクティビティや位置付けについては下田先生や仲里先生に伺うのが一番いいと思いますが、このようなデータベースを各研究者が独自に運営し、世界的に有名になっているという状況です。

ただし、こういった研究・学術データは、主にコミュニティレベルで、部局やセンター、研究ユニットのような小さいユニットで、多大な労力と時間をかけて醸成されています。先ほどの KEGG は 20 年かけて育ててこられたという実績を伺っています。また、自発的な取り組みなので、研究者とその研究コミュニティが使いやすいように、独自に改良が重ねられてきました。

そういう意味で使いやすいデータベースになってい

（図 2）



（図 3）

ることは確かです。その一方で、持続的に発展させる、維持させることが大きな課題になっているということが、私が学術領域の発展と社会貢献の視点から見たこれらのデータベースの特徴です。

ただし、大学側は、一体どれぐらい学術データベースがあるのか、全容を全く把握できていない状況です。今後、偉い人が「大学としてのコンテンツ発信力を評価するぞ、もっと頑張れ」という新たな評価軸を設ければ、恐らく皆さんはせっせと調べるのでしょうけれども、まだそういうことは出てきません。

また、現行の運営側は、自分たちが頑張って運営してきたもので、お金がないから何とかサポートしてほしいという本音と、自分たちが好き勝手やるのだから、大学の上に余計な手出しをしてほしくないという相反する思いがあります。このあたりについては今、皆さんが自発的に、しかも先頭を走っているデータベースについて、研究データ管理をどうしなさいということを提言できる状態ではないというのが私の考えです。

2-2.研究公正とコンプライアンスのため

次は、研究公正とコンプライアンスのためのデータ管理についてです。研究現場でどんなインパクトがあったかという話をします。恐らくグッドプラクティスではなくバッドプラクティスになるのではないかと思います。

2014 年 8 月、文部科学省が「研究活動における不正行為への対応等に関するガイドライン」という文書を出し、各国公立大学・私立大学は対応しなくてはいけなくなりました。2015 年 2 月に京都大学は、「京都大学における公正な研究活動の推進等に関する規程」を出しました。これは公開されている内容なので、皆さんも読むことが可能です。ここには、教職員・部局（研究科や学部）・全学の役割、責任の所在、研究データ一定期間保存義務が明記されました。

続いて 2015 年 7 月に、京都大学は、年度内に研究データの管理方法、保存方法を部局ごとに制定するように求めました。現在、2016 年度は、部局ごとに決定した管理方針ないしデータ保存計画を基に行動しているはずです。

私は部局ごとの制定が求められたときに工学研究科にいて、先生たちに研究データを 10 年間保存するように言っても、絶対にこれが一律に守られるわけがないということがあったので、慌てて研究データ保存システムを立ち上げることとなりました。工学研究科や情報環境機構は ICT のインフラを整備する底力がありました（図 4）。そのシステムを使って、研究科ないし情報環境機構という組織でデータを預かり、バックアップを取る活動を始めています。これはまだ試行段階です。

ファイルを zip ファイルにしてアップロードし、論文のテーマ、著者、発表日、発表メディア、書けるのであれば DOI を書きます。研究成果の正当性を保証する研究データがどこに保存されているかをトラックできるようにするための非常に簡単なシステムです。

それで始めたのですが、やはりやっつけでつくったので、いささか運用・安定性に不安があります。一方で、何もデータをきちんと管理するということは、研究データに限りません。事務の文書でも何にしても、データをきちんと保存したいという要求はあるはずなので、次に更新する計算機システムの中に、大容量の光ディスクを使ったアーカイブシステムを導入することにしました（図 5）。

これと連携させて、この上で、エンタープライズ・コンテンツ・マネジメントシステム、文書管理システ

（図 4）

ムを導入し、ユーザーは研究データとメタデータを文書管理システム上に登録する。ダークアーカイブをしなければいけないときは、コンテンツを光ディスクにコピーして、そちら側は誰も触れない状態でデータの正当性を保証するという仕組みを今検討して立ち上げようとしています。

ただし、研究データ管理の観点からすると、これはあくまでも組織あるいは研究者を保護する保険であり、オープンデータではありません。これはある種のバッドプラクティスの一つではないかと思います。とにかくデータを長期保存しなくてはいけないということで、2 年前にいきなり上から降ってきて、慌てていろいろなルールを決めて、すごく駆け足でやってしまったので、ルール（policy）が先行して、現実的な実施手順があまり検討されないままずっと走ってしまっているという現状があります。ルール（policy）を実施手順（procedure）と実施（deployment）が後追いしているのです。

私のいる情報環境機構は全学に対してシステムを提供することをミッションとしているのですが、この仕組みを使うと、各部局が定めた研究データ保存のプロシージャーポリシーに一致するかどうか、というレビューはまだ受けていない状況です。それはこれから擦り合わせが必要です。いびつな構造で、システムとポリシーと実施手順がかみ合っていない状況で、とにかく走らせなければいけない、という状態にあるということです。

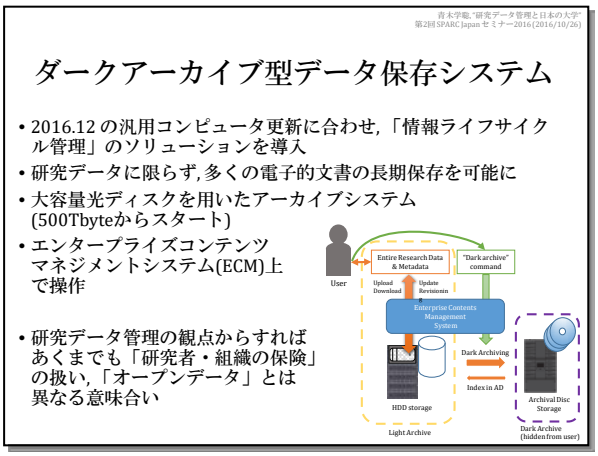
また、困ったことに、ほとんどのデータが死蔵されるものです。ここに収められたデータは何か研究不正が疑われたときに、ここに研究データがあると出すためのものだからです。だから、実はそれほど力を入れて構築したくないシステムで、できるだけ安く、それでなお皆さんに信頼されるシステムにする難しさも抱えています。研究公正とコンプライアンスだけのために IT システムを構築をしなさいと言われると、非常にやりにくい ICT システムになってしまうという事例になってしまっているのではないかと考えます。

2-3.オープン（データ）サイエンス促進のため

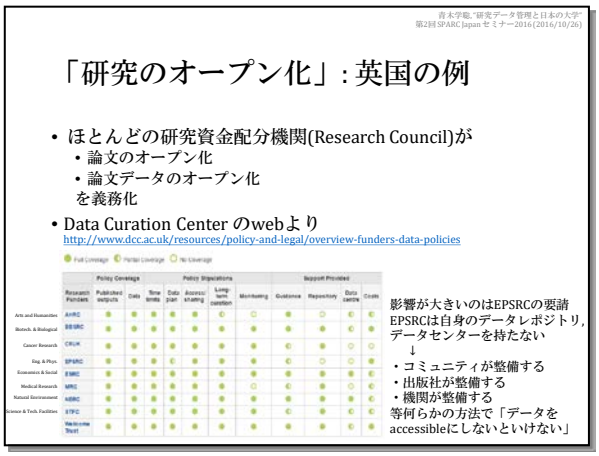
オープン（データ）サイエンスとは、今まで一部の研究者が自発的に実施してきたオープンデータ的な活動から広がり、研究者がデータを集めて、保存して、公開することを当たり前の作業にしていく運動だと私は認識しています。これに対して世界はどうなっているのか、あるいは日本国内はどうするべきかという話をします。

2-3-1.英国の事例

先日、私は大学のプロジェクトの資金を頂き、英国を訪問し、英国での研究データ管理がどうなっているのかという話を伺う機会がありました。英国ではほとんどの研究資金配分機関（リサーチカウンシル）が論文をオープンにすること、さらに踏み込んで、論文に載っているデータ、研究の成果として公開していいデ



（図 5）



（図 6）

ータは、オープンにして電子的にアクセスできるようにすることを既に義務化しています。具体的に星取表を英国の Digital Curation Centre が提示しています（図 6）。この機関の資金の場合は、このデータをオープンにしなければいけないということが書いてあります。これは非常に有名な絵なので、いろいろなところでご覧になった方がいらっしゃると思います。

また、英国に限らず、アメリカ、ヨーロッパの例は、科学技術振興機構（JST）が 2015 年 5 月に研究データの共有ポリシーに関するレポートを書いているので、それを読んでいただくと一番分かりやすいのではないかと思っています。

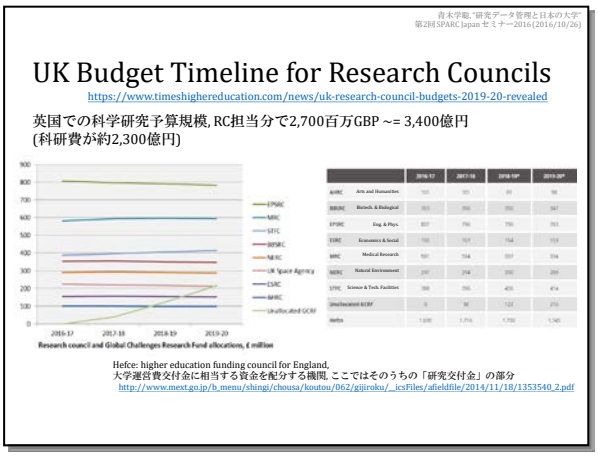
英国で一番インパクトが大きかったのは恐らく、英国工学・物理科学研究会議（EPSRC）から、工学と理学の分野はデータオープンアクセスにするように要請があったことでしょう。EPSRC はデータをオープンにするように言っていますが、データを置く場所は準備しておらず、自分たちで準備するか、コミュニティが準備しているリポジトリを使うようにと言っています。だから結局、工学・理学系の先生たちが研究をして、論文を出版したときに、そのデータにアクセスできるようにするためには、自分たちが所属している研究コミュニティがリポジトリを整備するか、出版社のデータジャーナルに投稿するか、自分自身でやるしかないのです。

英国の研究資金配分機関の予算規模は、約 2,700 万英国ポンド（約 3,400 億円）です（図 7）。科研費の規模が約 2,300 億円なので、それの 1.5 倍に相当する研究に対してオープン化を求めたということです。それ以外に、日本で言う厚生労働省系なども同様に、データをオープンにしなさいという宣言を次々出していて、ほとんど全てのファンドがオープンデータにするという方向に向かっているという現状です。

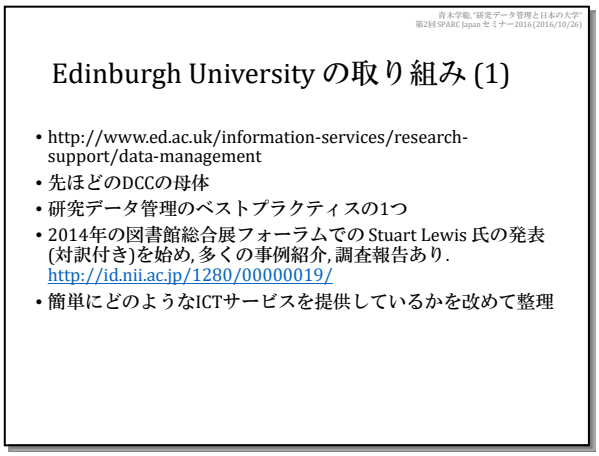
それに対して、英国の各大学はどうしているのでしょうか。一番先進的な取り組みをしているといわれているエジンバラ大学は、研究データマネジメントについて集中的に研究している Digital Curation Centre を生み出した母体になっている大学であり、研究データ管理を具体的に実践している大学の一つです（図 8）。

非常に有名な大学なので、何回も調査が行われ、調査レポートも出ています。一つ参考になるのは 2014 年の図書館総合展フォーラムで、ここのヘッドであった Stuart Lewis 氏が発表された内容です。英語と日本語の対訳のスライドが出ています。これを見ていただくと、歴史と今やろうとしていることが簡単にまとめられていて非常に参考になると思います。

私はインフラを提供する立場なので、ここで一体どんな ICT サービスを提供しているかに注目してお話しします。エジンバラ大学は、簡単なブローシャー「A guide to the Research Data Service」を出版しています（図 9）。研究を始める前、実際に研究を行っている間、研究が終わって成果を公開したいときという研究の各ステージ、そしてこれらの研究活動サイクルでどのようにトレーニングとサポートをするか、こうい



（図 7）



（図 8）

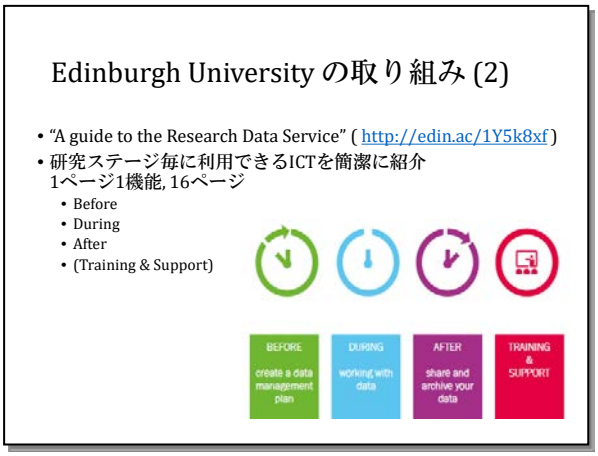
う四つの視点でエジンバラ大学では何ができるか、何を提供しているかをまとめています。1 ページ 1 機能で 16 ページ、ぱらぱらと読んで、なるほどと思える内容です。

最近、英国ないし米国では、研究申請調書を書く際に、こういうデータを集めて、どこで管理して、最後にどのように公開するかというデータ管理プランを 2 ページほど書くように、と言われます。エジンバラ大学では、出すファンドに合った Q&A がウィザード形式で表示され、それに答えていけばデータ管理プランを仕上げることができるツールを提供しています（図 10）。

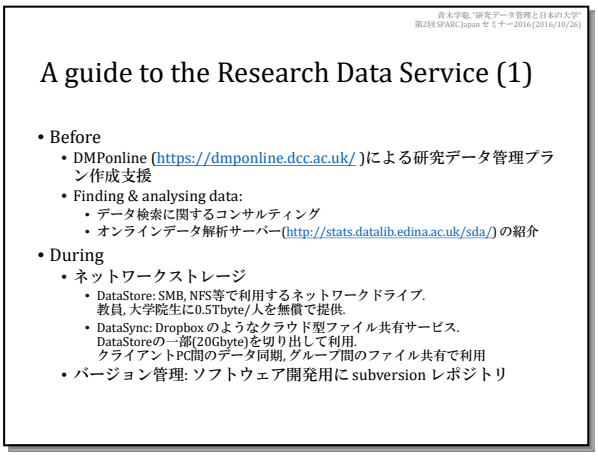
エジンバラ大学の経済・社会学系では、オンラインデータをその場で解析する仕組みも提供しています。このあたりはそれほどインパクトはないのですが、一番研究者にとって魅力的に移るのは、研究をしている最中にデータをどう扱うか、どう保存するかということです。そこでエジンバラ大学では、教員・大学院生に対して、0.5Tbyte、500Gbyte ぐらいのネットワークファイルストレージを使っていいという話をしています。当然、全員が使っているわけではないですが、最大これぐらいまでは自由に使っていいということ、だと推測しますが、少なくとも、自分たちの普段の生活で出てくるデータは全て大学が提供するネットワークドライブ上で扱ってよいというぐらいのキャパシティになります。また、共同研究者とデータがシェアできるように、Dropbox のようなクラウド型のファイル共有サービスを提供しています。

研究が終わった場合、論文を出版するということがありますが、PURE というシステムを持っていて、これは研究者総覧と機関リポジトリを合わせたような仕組みです（図 11）。内容はかなりいろいろなものが入っています。大学にどのような人がいるのか、どのようなプロジェクトが実施されたのか、研究成果（主にオープンアクセスの論文）、その他の研究活動（受賞等）などを、利用者の選択でアップロードして公開できるシステムです。これは恐らく日本の大学でだいぶ整備が進んでいるものだと思います。

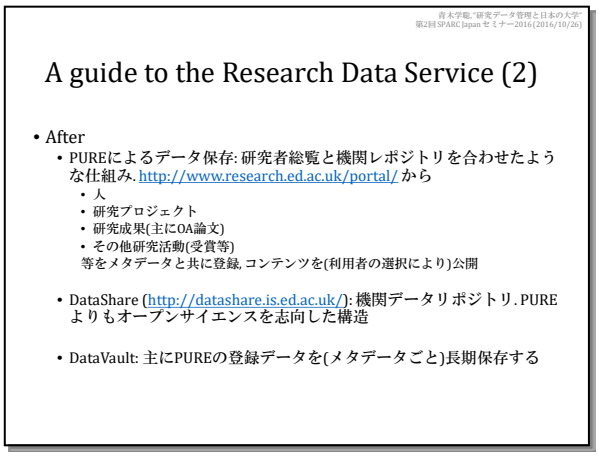
これとは別に、もう少しディテールにこだわった研究のデータセットを公開する機関データリポジトリも持っています。



（図 9）



（図 10）



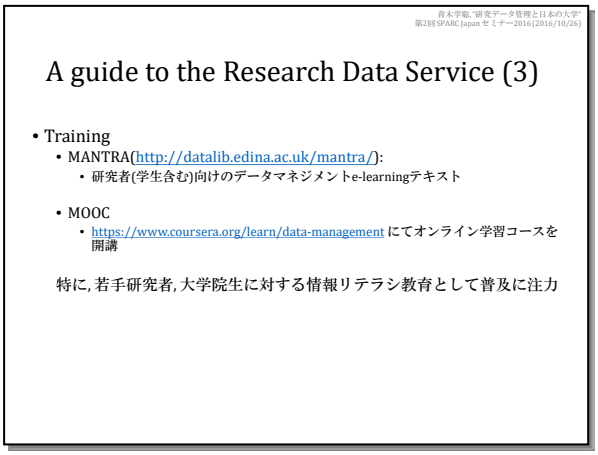
（図 11）

そしてトレーニングです（図 12）。向こうの方は、お年を召した研究者にいろいろ教えても駄目で、若い研究者、大学院生に対してしっかりと、「これからの時代はこうやってデータを管理して自分たちの研究をより良くするのだ」と教え込むことが重要だということを強調していました。今、研究データ管理は恐らくオープンサイエンスを促進したいということがバックグラウンドにあって、ファンドの要請に従って、データ管理プランを作成して実施しています。それに堪えられるだけの十分な研究用のインフラの提供（大容量のストレージ、機関リポジトリ）は、機関の体力勝負になるのではないかと思います。

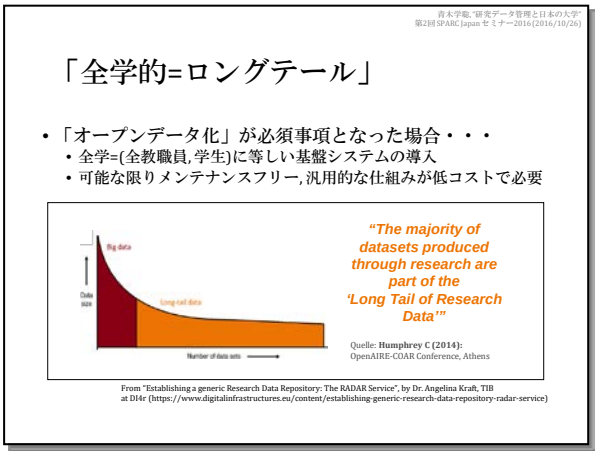
全員がやらなくてはいけないとなった場合は、非常にロングテールなものをサポートしなければいけません（図 13）。よく頑張っている、非常に頻繁にアップデートするデータリポジトリだけでなく、その後ろの方にある、何に使われるか分からないけれども残しておきたいデータまできちんとカバーしなくてはいけません。そうすると、可能な限りメンテナンスフリーで、汎用的な仕組みをしっかり入れなくてはいけません。ということは、できるだけコストを下げるという視点がどうしても必要になってきます。こういうものは果たしてあるのか、ないのかということを、これから各機関で整備しなさいと言われたら、どうしても考えなくてはいけないことになってしまいます。

2-3-2. ドイツの事例

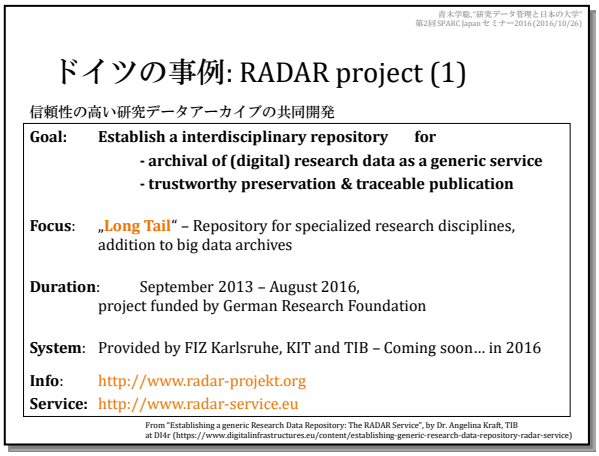
それに対応する形として、ドイツでは、RADAR project という、研究データアーカイブのシステムを共同開発して、それを各機関へ提供するというプロジェクトがあります（図 14）。ステージとしては、研究を始めていて、共同作業が終わり、研究が終わったとき



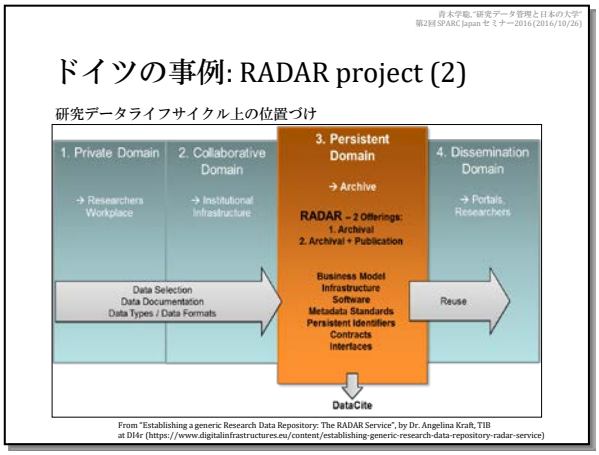
（図 12）



（図 13）



（図 14）



（図 15）

に研究の内容物を一気にアーカイブします（図 15）。目的は、一つはアーカイブするだけ、もう一つは研究者の選択によりその中の一部をパブリッシュすることです。

ビジネスモデルとしては、ベーシックサービスは何も言わずに黙って保管するだけです（図 16）。その中で、これとこれは公開したいと言われたときは DOI を付けて公開できるようにしようとしています。

各機関あるいは研究グループに提供するときの値段も決めています。2016 年 8 月にできる予定とされていますが現時点ではまだ公開には至っていない模様です。もうすぐ出てくると期待しています。

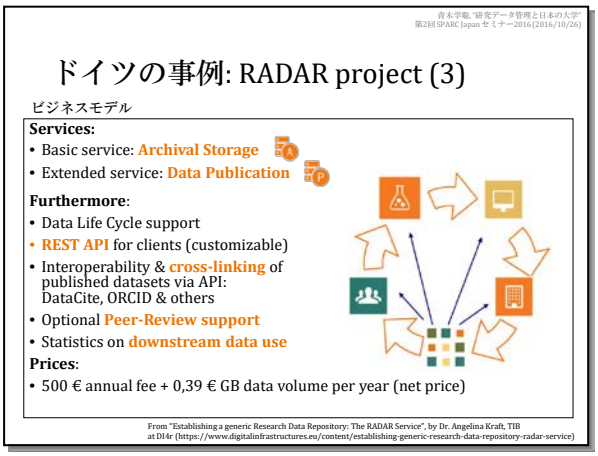
3.京都大学は何をするか

それでは、京都大学では何をしようということです。当初は、どこから手を付けたらいいのか分からない状況でうちひしがれていたのですが、図 17 のようなことをまとめています。本当にできるかどうかは分かりませんが、エジンバラ大学プラスアルファのものを理想高く目指すのであれば、これぐらいのものを準備すると完璧だ、という内容。まず、研究者・大学院生が自由に利用できる大容量のクラウドストレージは絶対に要るでしょう。そしてこれを、フラットな環境でできることです。アクティブな研究活動を行い、学内および学外でシームレスにデータ共有をし、きちんと協働できる環境は研究機関としては絶対に持つか、あるいはどこかで整備しなければいけないと思います。

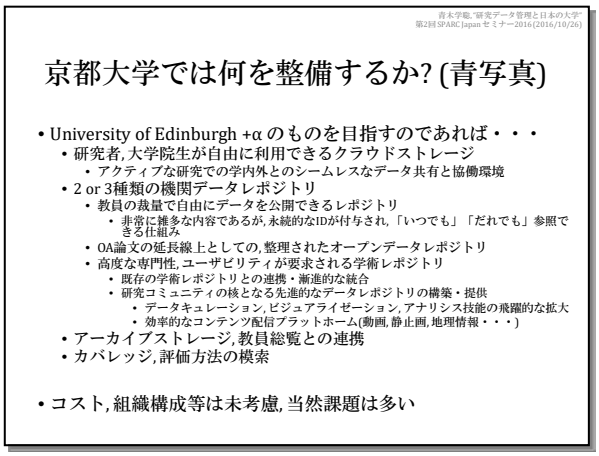
二つ目は、2～3 種類の性格の異なる機関データリポジトリです。データをアーカイブする場所、ためておく場所として必要になります。1 種類目は、教員の裁量で自由にデータを公開できるリポジトリです。先生たちが何かの機会で、このデータは面白いと思ったとき、あるいは、どこかの学会で発表して簡単なペーパーにしたので見てほしいというときに、それに永続的な ID、DOI かまたはそれに類するものを自発的に振って公開できる仕組みです。そこにはいわゆるキュレーションなどの作業はありません。つまり研究者が自発的に公開するドキュメントに対して永続的な ID を付けられる仕組みを持ったリポジトリです。

2 種類目は、今、各日本の大学はオープンアクセス論文を対象としてリポジトリを整備していますが、その延長線上として整理されたオープンデータリポジトリ、あるいはある種のキュレーションやある程度のフィルターを通ったリポジトリです。

3 種類目は、高度な専門性、ユーザビリティが要求される学術リポジトリです。これは、今業界の最先端を走っていて充実したリポジトリに迫る何かということです。現在アドバンストな、あるいはソフィスティケートされた学術リポジトリと連携・漸進的な統合ができるような基盤、さらに何か新しい学問領域を生み出すときには、新しいデータベース、新しい考え方、新しいメタデータを生み出せるような基盤になります。きちんとしたデータキュレーション、ないしはこのデータはこのように見るものだというデータビジュアラ



（図 16）



（図 17）

イゼーション、このように使うべきだというアナリシス技能を飛躍的に拡大させることができないことには、高度な専門性やユーザビリティにはなかなかたどり着けません。

ただし、これだけオープンデータが出てきたということは、今まで 20 年などすごい時間をかけていたものを、少しは早めることができる可能性があります。今まで 20 年かかっていたものが 10 年になる、もしかしたら 3 年ぐらいで立ち上がる基盤になるのであれば、ここを頑張る意味はあるのではないかと思っています。

研究公正のためのデータや、人には見せたくないけれども永年保存したいデータとうまく連携するような仕組みはやはり必要だと思います。

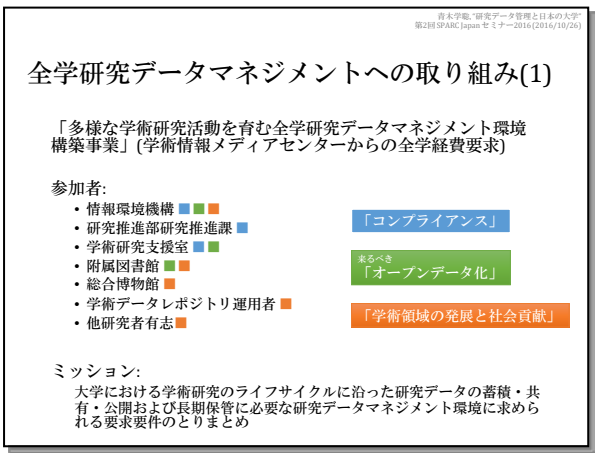
いろいろなデータベース、リポジトリが出現、整備された際には、結局、うまくいったのか、いかなかったのかということは絶対に言われると思います。パフォーマンスをどう評価するかは考えておかなければいけませんが、全くアイデアがない状況です。幾つかの研究でオープンアクセスは経済的価値を生んだというレポートは出ていますが、各機関がオープンアクセスにしてどれくらいの価値があったかということについては、何とも言い難い、評価のしようがない部分があるかもしれません。ここは非常に難しいところだと思います。

当然ながら、全部やるとなるとコストはうなぎ上りですし、幾つか選んで実施するか、NII が今準備しているようなフレームワークにうまく乗る、そういうところでお互いに情報連携を、パイロット的な事業に組み合わせて、どのように大学の中でうまくローンチさせるかを考えていきたいと思っています。

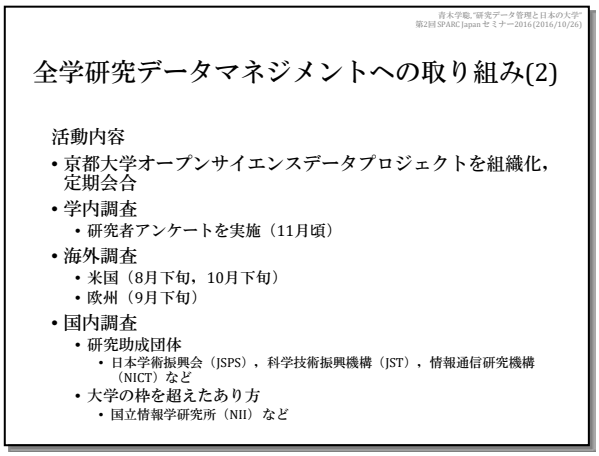
4.全学研究データマネジメントへの取り組み

こういうことを考える機会として、京都大学では今年度、どうやって研究データ管理の環境を構築するかというタスクフォースのプロジェクトを立ち上げました（図 18）。私が入っている情報環境機構の他、コンプライアン関係の研究推進部、そして学術研究支援室というファンディングのサポートを行う部署も入っています。実際にリポジトリで精力的に活動されている附属図書館、博物館、能勢先生をはじめとする、実際に学術データのリポジトリの運用をしていらっしゃる方、研究者有志、これらの方が集まって、大学における学術研究のライフサイクルに沿った研究データの蓄積・共有・公開、長期保管に必要な研究データマネジメント環境に求められる要求要件を取りまとめて、次年度以降に京大はどうしたらいいのかという提言ができれば、ということで、このような活動をしています。

具体的には、いろいろなミーティングを行い、関係各所からヒアリングをし、研究調査を行いました。私もエジンバラ大学をはじめ、幾つかの大学を回りました（図 19）。



（図 18）



（図 19）

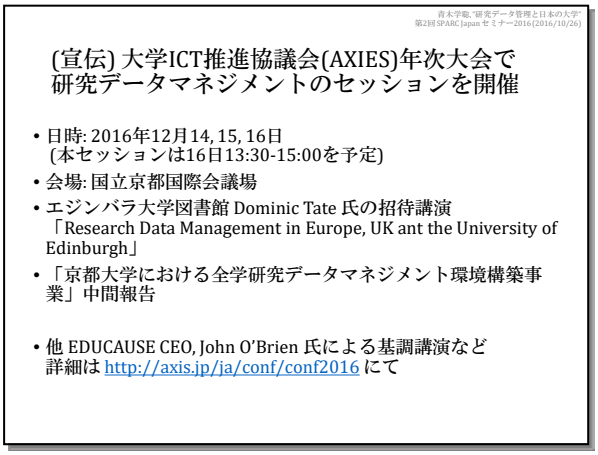
5.まとめに代えて

私からは、京都大学における研究データ管理の現状を、「コンプライアンス」「オープンデータ化」「学術領域の発展と社会貢献」に分類し、それぞれに一体何が求められているかという話をさせていただきました。

ただし、この「コンプライアンス」と「学術領域の発展と社会貢献」は、いずれはオープンデータという大きな概念として、ポリシー・手順・システムなどが集約されていくと考えて、準備を進めていくことが重要です。

オープンデータを義務化することに関しては、恐らく大学あるいは機関全体の ICT インフラの底上げに直結する内容になるので、各組織が十分なサポート体制を、義務化のタイミングを見計らいながら構築していくことが大切だと考えています。

最後に宣伝です（図 20）。大学 ICT 推進協議会年次大会で、研究データマネジメントのセッションを持ちます。2016 年 12 月 14～16 日で、本セッションは 16 日 13 時半～15 時を予定しています。エジンバラ大学図書館の Dominic Tate さんもお呼びし、私たちの活動内容を報告させていただきます。



（図 20）