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」

研究データ利活用に関する国内活動及び国際動向について

武田　英明
（研究データ利活用協議会/国立情報学研究所）

講演要旨

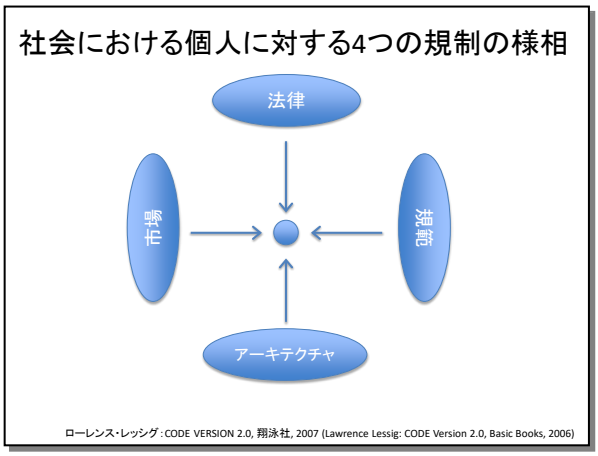
研究データに関する利活用に関する関心が近年、国内外で高まっている。本講演では、オープンサイエンスの流れの理解とその上での研究データ利活用の枠組みについて概要を説明する。その上で、国内及び国際的な動向について概観する。国内では DOI の RA(登録機関)であるジャパンリンクセンターが 2014 年に行ったデータ DOI 実験プロジェクトを契機に分野横断的なつながりができ、それが研究データ利活用協議会の発足につながった。国際的には RDA(Research Data Alliance)が 4 年前より活動を始めており、funder、研究機関、出版社等を巻き込んで、横断的なつながりを形成している。その活動を一部紹介する。

武田　英明
http://www.nii.ac.jp/faculty/informatics/takeda_hideaki/

現在、オープンサイエンスということが盛んにいわれています。しかしなぜ今、オープンサイエンスなのでしょうか。

なぜ今、オープンサイエンスなのか

一つだけ理解してほしいことがあります。それは、サイエンスをオープンにしたかったからオープンになったわけではないということです。それを理解するための一つの鍵はローレンス・レッシグの批判です。彼は、われわれ個人は四つの方向から規制を受けていると主張しています。それは市場（お金）、法律、規範、アーキテクチャです（図 1）。レッシグは法律家ですから、規制とは法律であると言いそうですが、そうではなく、市場（お金）からも受けているし、規範からも受けているとしています。最後のアーキテクチャが一番分かりづらいところです。これは社会の仕組みそ



（図 1）

のものから規制を受けているということです。かつてインターネットがなかった時代は、われわれは例えば鉄道網などの社会の規範から規制を受けていたけれど、現代はインターネットの仕組みで規制を受けているということが彼の主張です。

インターネットの世界、特にウェブの世界においてはオープンでフラットな仕組みが導入されました。それがわれわれのアーキテクチャになったということです。そのオープンやフラットということが、市場・法律・規範にも逆に影響を与えているのが現在なのです。

それと同じことが今、サイエンスにおいても起きているということがポイントです（図 2）。われわれはそのさなかにいるのでぴんとこないかもしれませんが、例えば研究者の規範として、成果はみんなで共有すべきといわれるけれど、インターネットがなかった昔には、みんなで共有するには出版するしかありませんでした。だから、論文出版が良かったのです。でも今は、インターネットだったらオープンで、もっとたくさんの人に見てもらえます。だからオープンデータにするのです。

われわれの規範は既に変わってしまっています。それはアーキテクチャが変わったからです。もちろん市場も変わりました。そうなったために、既存の出版社はそのオープン性を自分たちの商売へ入れ込んで、article processing charge（APC）を取るようになったのです。法律についても、各助成団体や国が制約を設けるようになりました。

オープンサイエンスの系譜

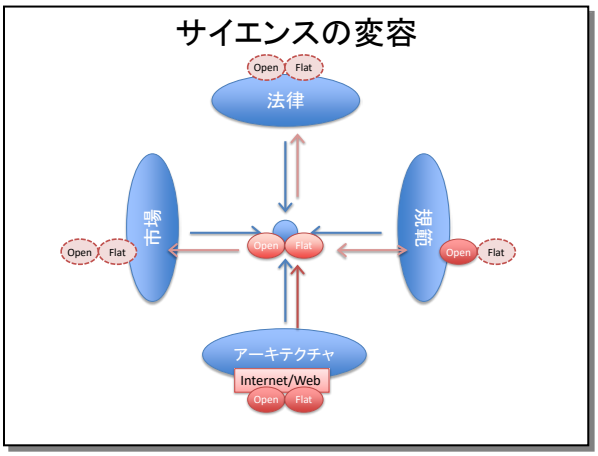
ネットワーク、コンピューターのアーキテクチャがどんどん変わりつつあり、それが直接・間接にわれわれに影響を与えているのが今のオープンサイエンスの世界です。図 3 の一番下が、ソフトウエア、コンピューター、ネットワークの世界で何が起きているかです。その世界から生まれてきたのがクリエイティブ・コモンズ、Web2.0 などです。

仲里先生の報告では、生命科学では 1960 年代から MEDLINE があったという話が出ました。でも、PubMed になったのは 1990 年代です。これは、もともと自分たちのコミュニティでデータを共有したいという考えがあったのですが、インターネットが発展してやり方が変わってきたという例です。

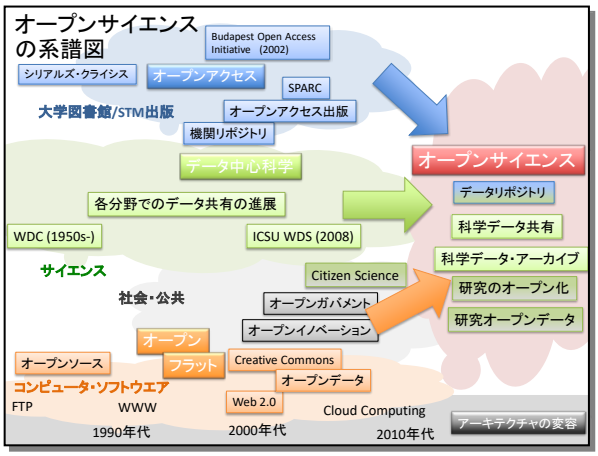
出版物も今までは紙で出版するのが良かったのですが、インターネットの発達によってウェブで出版することで、オープンとのつながりができてきたというのが一番上の青色です。

真ん中の部分は、むしろ社会・公共が変わって、オープンガバメント、オープンイノベーションが入ってきているということです。

オープンサイエンスは、この四つの絵から影響を受けて今があるということが理解を難しくしています。例えば、データリポジトリは図書館系とサイエンス系の中間辺りにあります。研究のオープン化、研究オープンデータは、どちらかというと政府・公共のオープン系と研究のオープン系の中間ぐらいにあります。そ



（図 2）



（図 3）

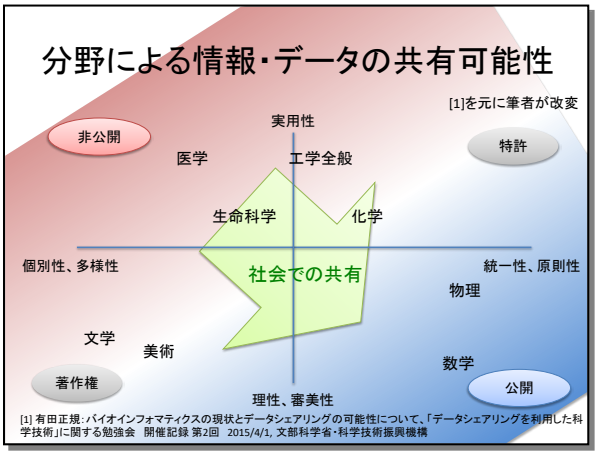
ういう状況で、まだら模様がオープンサイエンスの今だということです。

オープンサイエンスと研究データ共有

研究コミュニケーションの世界では、昔はファクスでやりとりをしたり、「あなたの論文を下さい」とはがきを書いたりすることがありました。それがオンライン化すると、雑誌購読料が高騰し、大学図書館に雑誌を購読するお金が足りなくなって購読雑誌が減少するというシリアルズクライシスが起きました。その対策として、SPARC やオープンアクセスが出てきました。

科学では、生命科学や天文学など、各分野でベストなデータ共有の方法を探してきました。ここで少し注意が必要なのは、分野ごとにデータ共有の特性が違っており、共有の度合いやデータ量、分散か集中なのかも違うということです。

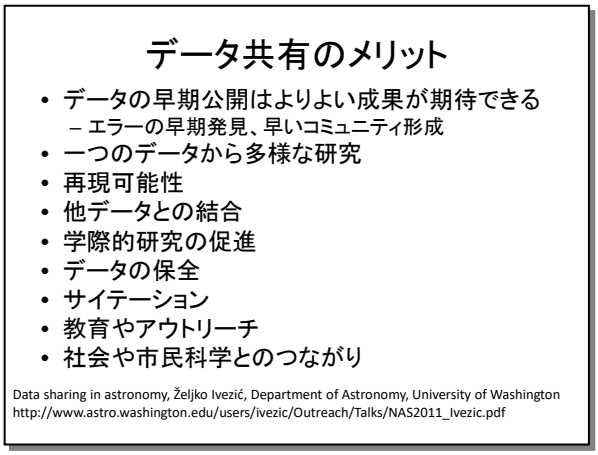
それを無理やり図にすると図 4 のようになります。二つのラインはかなり主観的に分けています。個別性・多様性を探求する学問なのか、統一性・原則性を探求する学問なのか、実用性を探求する学問なのか、理性・審美性を探求する学問なのかをマップしようとする試みです。例えば、理性・審美性を下に置いて、実用性を上に置いたときに各学問がどこにあるか。数学や物理は、より統一性や理性を探求するようなものなので、右下でしょうか。逆に工学は比較的上の方です。文学・美術は個別性・多様性を尊びつつ、ある種の理性・審美性を求めるので左下としています。



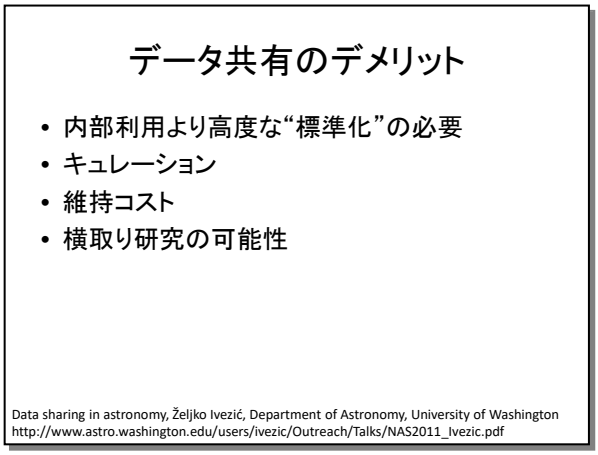
（図 4）

このとき、右下が公開、左上は非公開、その中間の右上と左下は、右上が特許で左下が著作権という世界が大ざっぱな理解で見えてくると思います。共有できるかどうかという観点では、右下に行くと比較的この問題はやりやすく、左上に行くと慎重にならざるを得ません。自分の学問分野によって立ち位置が変わります。この図で、自分はここにいるからこうなのではないかと理解していただけるといいと思います。そういう問題がある上で、各分野ではそれぞれの特性に合わせたことをやってきました。

図 5 は、天文学者が挙げたデータ共有のメリットです。天文学は最初のころ、300 年ぐらい前はデータを隠していたのです。ガリレオなども隠していました。下手をすると死ぬまでデータを出さない。それを天文学は早くに克服して、このようなメリットがあると言っています。



（図 5）



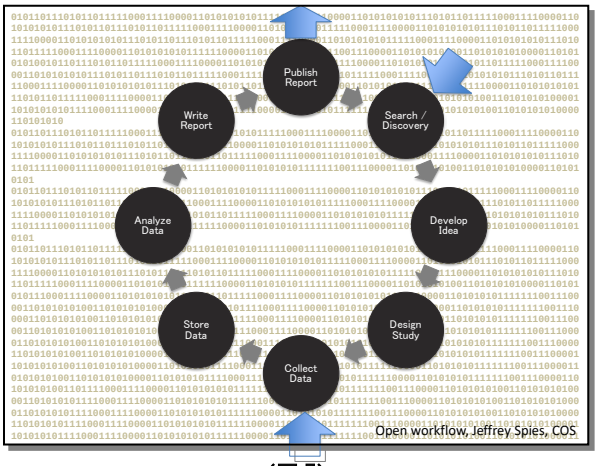
（図 6）

デメリットももちろん挙げています（図 6）。やはり人に見せるためにはコストが掛かります。キュレーションも必要です。維持するにもコストが掛かります。横取り研究の可能性もあります。ここまでが各学問分野での研究データ共有、メリット・デメリットの話でした。

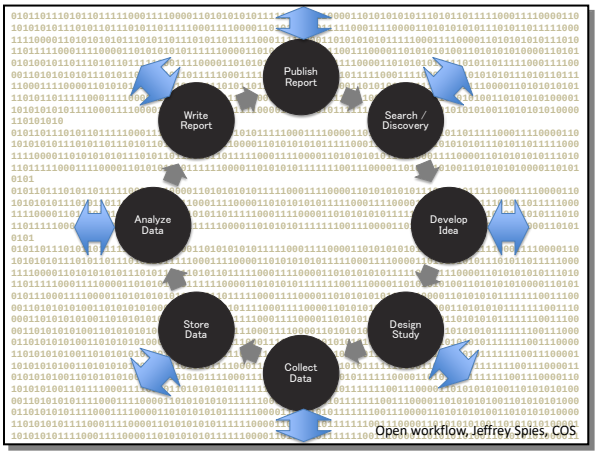
研究データ共有の枠組み

では一体、オープンサイエンスなりオープンデータは何をすることにならざるを得ないのかということです。デジタル化以前の研究者は論文を書いて、データも自分でつくっていましたが、だんだんデータをもらうようになり、論文もデジタルになり、データも論文もほとんど区別がつかなくなったのが現状です。

図 7 は、アメリカの Center for Open Science の Jeffrey Spies 氏がつくった Open workflow の図です。先ほどは



（図 7）



（図 8）

in と out しか描いていませんでしたが、もっと細かく描かれた、最初に探して、アイデアをつくり、実際に研究を計画し、関連するデータを集めて、分析し、レポートを書いてパブリッシュするというストーリーが回るのが研究のライフサイクルだというものです。誰が描いても大体このような図になると思います。

先ほどの私の図では、図 7 の青い矢印のところにしか入り口と出口はありませんでしたが、彼らは全体がオープンになるということを知っています（図 8）。研究プロセス自体がオープン化する。あるいは、最初は共有して、プライバシーやセキュリティが入っているものは永遠に公開されないかもしれませんが、そうでないものは時間が経つと公開されるという意味で、プロセス全体が公開されるという前提がわれわれの研究の未来像だと思います。これは今すぐにできるものではありませんが、5 年、10 年経ってみるとこれが当たり前の研究スタイルになっているだろうということは私も同意します。

ただ、研究プロセス全体というよりも、データ共有が当面の課題だと思います。データ共有のポイントはデータのライフサイクルです。研究者は研究途上のデータ、研究発表に使ったデータ、保存用データと切り分けますが、問題は担当が違うことです。研究中であれば研究者で、研究が終わるとそこから先は研究機関に任されます。データの作成から保存まで、研究データのライフサイクルを通してどうサポートするかということが課題です。

FAIR 原則はもともと FORCE11 がつくったものです（図 9）。FAIR というのは、Findable（見つけられる）、Accessible（アクセスできる）、Interoperable（相互運用可能）、Re-usable（再利用できる）の頭文字を取ったもので、研究データがどうあるべきかという原則です。今この FAIR 原則がコンセンサスになりつつあります。研究オープンデータがどうあるべきかという方向はかなり見えてきているように感じます。

Research Data Alliance（RDA）

ただ、それを実際にどう実施するかが問題です。今、大きく違う世界の人が関わってオープンサイエンスというコンセプトができているので、ステークホルダーが非常に多いのです。研究データ共有に関する国際活動として、Research Data Alliance（RDA）が 2013 年から、最初は 5 年という形で始まりました（図 10）。今まで研究に関するコンソーシアムは、研究者や研究機関が集まり、せいぜい拡張しても政府関係者、ファンディング、ファウンダーでしたが、RDA には研究者、研究機関、出版社、政府関係に加えて、社会の IT ベンダーや企業など、非常に多様なステークホルダーが入っていることが特徴です。

RDA は年に 2 回プレナリーミーティングというものを開いていて、先月デンバーで行われました。今度は 4 月にバルセロナで行われます。

（図 9）

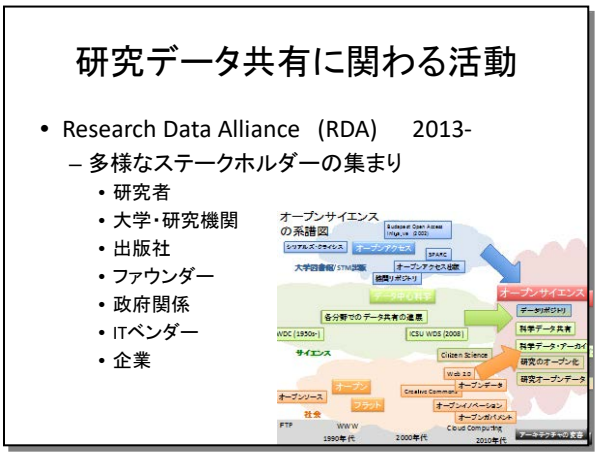
RDA は、研究データ共有と交換の障害を減らすためのインフラストラクチャーとコミュニティ活動の発展と世界的なデータドリブンイノベーションの加速に焦点を当てた国際的な活動です。皆が共有できるインフラをつくりたい、それを支える人の活動をつくりたい、プラス、データドリブンイノベーション、社会インパクトを与えることなど、かなりスコープを広く取っています。

RDA は、研究者とイノベーターが技術・ディシプリン・国境を越えてオープンにデータを共有するというビジョンを達成するために、データのオープンな共有を可能とする社会的・技術的架け橋をつくります。面白いことは、単に技術で応えるのではなく、社会的な仕組みでも応えるというところです。

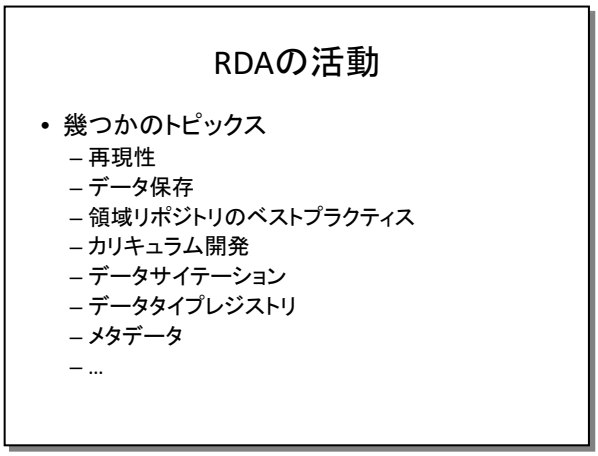
ここはトップダウンの組織ではなく提案ベースで、人が集まってインタレストグループとワーキンググループをつくり、リコメンデーションなどのアウトプットを出します。ワーキンググループでは 18 カ月で何かアウトプットを出すというプロセスだけが決まっていて、あとは自分たちで提案し、人が集まれば承認されるという制度です。

RDA で取り上げられるトピックスとしては、図 11 にあるものが挙げられますが、もっとたくさんあります。今日出てきたような話は大抵どこかのグループのテーマになっています。ポリシーをどうするかも取り上げられます。

RDA 自体は、組織はほとんどないも同然で、かな



（図 10）



（図 11）

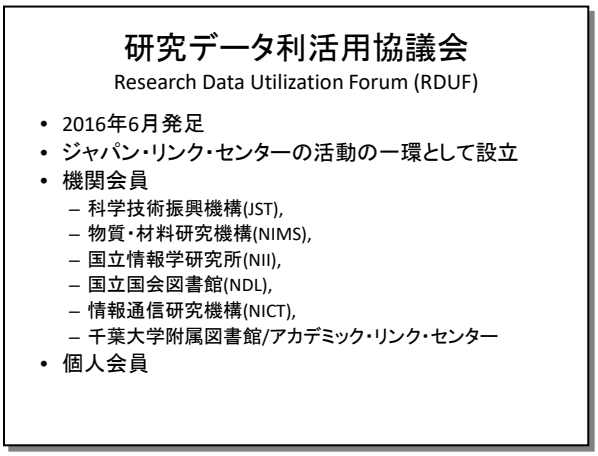
り小さな事務局があるだけです。ワーキンググループ・インタレストグループが勝手に集まって、勝手にオンラインで議論して、プレナリーミーティングのときにサマリーを報告する仕組みになっています。

ヨーロッパは RDA Europe、アメリカは RDA US があり、資金をもらって活動しています。それにオーストラリアを加えた 3 局が発足に尽力したと聞いています。でも、今影響力があるのはヨーロッパとアメリカの 2 局になっています。2016 年 3 月に、アジアで初の RDA のプレナリーミーティングが、本日の会場と同じビルの一橋講堂で、科学技術振興機構（JST）主催で開かれました。そこで、日本はどうしているのかとたくさん聞かれました。

研究データ利活用協議会

2016 年 6 月、研究データ利活用協議会が発足しました（図 12）。前身はジャパンリンクセンターの研究データへの DOI 登録実験プロジェクトというものです。DOI は文献に付けることが今までの多くの習慣でした。しかし、DOI の仕組みは文献に限らず使えるということは古くから気付かれていて、ヨーロッパでは 10 年前からデータサイテーションの活動があります。それが 5 年前から、DataCite という組織として活動しています。

その中で、研究データに永続的な識別子を付けるべきだという議論があり、日本はジャパンリンクセンターが主体となって、どうやってデータを見つけるべきかを一緒に考えようという実験プロジェクトを行いました。このときに初めて、データを扱う研究機関が一堂に会しました。図に挙げている以外にも、理化学研究所や海洋研究開発機構なども入っていただいて、一緒になって、データに DOI を付けることの意味は何か、どういう単位で付けるといいのか、どういう手順にしたらいいかを考えました。



（図 12）

それは 1 年で終わりましたが、初めて分野を超えて、まさにディシプリンを超えて、実務者レベルで顔を合わせて研究データについて議論することができた、非常に良い機会でした。それを母体にもう少し活動を継続したい、実験プロジェクトは終わったので別の名前を付けて広がりを持たせようということで「研究データ利活用協議会」と名乗ったのです。ジャパンリンクセンターのアウトリーチという位置付けで今は活動しています。

会員には機関会員と個人会員の二つがあります。機関会員は図に挙げている六つの機関です。上の四つはジャパンリンクセンターの共同運営機関です。それに情報通信研究機構（NICT）と千葉大学のアカデミック・リンク・センターが入っています。

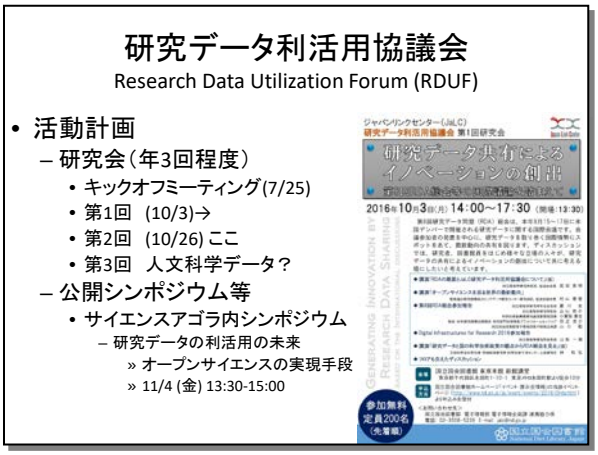
目的は、研究データに関する多様なセクター、特に実務者を集めることです。あまりお偉い方が顔を合わせる場所ではなく、実際にデータを取り扱う人たちが集まる場をつくりたいのです。集まって、研究データの共有と公開に関する課題を共有します。分野を超えると全く問題が違う、でも実は同じ問題もあるかもしれないということを共有したいのです。それを挙げるだけでも十分に価値があります。他には、うちではこういう技術を使っているということを共有できるとうれしいです。

その上で、研究データの共有と公開について、技術的・社会的な問題に関する議論を行いたいのです。自分たちの問題を解決するだけではなく、もっと広がりを持たせるためにどうしたらいいかということです。そして、RDA のような海外の関連組織とうまく情報共有やコラボレーションを図ることをミッションとし

ています。

活動計画としては、研究会を年 3 回程度開きたいと考えています（図 13)。キックオフミーティングを 2016 年 7 月に開き、第 1 回の研究会は 10 月 3 日に国立国会図書館で「研究データ共有によるイノベーションの創出」という題目で行いました。ちなみに、研究会は、毎回担当が代われば興味が変わってくるので、機関会員の持ち回り担当制で行おうと思っていて、この回は国立国会図書館に担当になって企画していただきました。第 2 回は今回のセミナーと合同開催です。第 3 回はまだ決まっていませんが、人文科学データについてできればよいと考えています。もうじき公開できると思います。

11 月 4 日に、サイエンスアゴラ内で 1 時間半の一般向けシンポジウムを行います（図 14)。これは本当に一般向けで、研究データそのものに興味を持ってもらうことが目的です。今年は水の話題が多かったので水の専門家をお呼びして、実社会と研究データの話をしたいと思っています。



（図 13）

（図 14）

まとめ

オープンサイエンスは、ウェブの発展とともに変わりつつあります。オープンサイエンスの重要なステップとしての研究データ共有が、今のわれわれの焦点です。データ公開の原則（FAIR）があります。そして、横断的な対話が始まっているということも重要です。研究者だけで閉じるような議論では今や駄目ですし、大学や研究機関を飛び越えて、社会ともつながっているのが現在ではないかと考えています。