

第 2 回 SPARC Japan セミナー2015

「科学的研究プロセスと研究環境の新たなパラダイムに向けて
- e-サイエンス, 研究データ共有, そして研究データ基盤 -」

発展途上のオープンデータ

Mark Parsons

(研究データ同盟 (Research Data Alliance) 事務総長)

講演要旨

近年、政府と研究機関は、オープンサイエンスの基礎的な構成要素として、オープンデータの必要性を強調してきました。私たちは、データそれ自身よりも多くのもの、データが再利用可能で有用であることを必要としています。記述的で機械可読なメタデータが必要であるのはもちろんのこと、データを完全に理解するためのソフトウェアやアルゴリズムも必要です。自分で選んだツールで簡単にデータを読み込み、解析できるよう標準とプロトコルが必要です。データの発生源と派生について信頼できるようでなければなりません。すなわち、相互運用可能なデータ基盤が必要であり、しかしそれは多様な文化と規模、技術を超えて動作する柔軟な基盤でなければならないのです。本発表では、基盤の概念を、データを取り巻いて構成される人と組織、機械が一体となったものとして、紹介する予定です。新しい組織である、Research Data Alliance が基本的なデータの共有と再利用を可能にするために、どのようにそれらの関係を構築し機能しているのかわかるでしょう。



Mark Parsons

<https://rd-alliance.org/about/organization/key-profiles/mark-parsons.html>

Research Data Alliance に関して、それを取り巻く文脈を踏まえてお話しします。私たちは、オープンサイエンスにつながるオープンデータに関心を抱いています。私はオープンデータを強く支持しますが、それはまだ不十分です。データをオープンで利用可能なものすることは、第一歩に過ぎません。

分かりやすく説明するため、最初に私自身の経験を紹介します。私は以前、米国雪氷データセンター

(National Snow and Ice Data Center) に勤務していました。主な成果物の一つは、南極・北極周辺の海氷（凍結した海水）の量に関するデータでした。特に北極は

興味深く、2012 年 9 月に北極の海氷は史上最小を記録しました（図 1）。紫色の線は、1979 年以降の 30 年間の中央値です。この動画は、1979 年から現在まで、毎年の海氷の最小値を示したものです。ここには、史上 4 番目に少なかった今年のデータは含まれません。2012 年は過去と比べ大幅に減少しましたが、そのほんの少し前の 2007 年にも最低記録があったのです。このことでデータセンターに大きな注目が集まりました。『ニューヨーク・タイムズ』紙の一面に記事が掲載され、アル・ゴア元副大統領が詳しい情報を求めて当センターを訪れました。私たちに関心が集まった一

大事件だったのです。

この背後にあるデータは、1980年代から公開されていました。基本となるデータは、地球から自然に放出される受動型マイクロ波の衛星リモートセンシングにより収集します。受動型マイクロ波の検知は、北極圏の研究に役立ちます。北極圏が闇に包まれる冬も、何が起きているか把握できるからです。北極圏を覆いがちな雲にも邪魔されません。受動型マイクロ波センサーはもともと、防衛省が軍事的な気象予測を目的に開発したものです。海水など誰の頭にもありませんでしたが、研究者が異なる帯域の受動型マイクロ波放射を判別できるアルゴリズムを開発し、これを氷や雪の検出に使い始めました。データの用途が、軍事的な気象予測から、海水をめぐる長期的な気象記録作成へと変化したというわけです。

実際、研究界では競合する複数のアルゴリズムが開発されました。どれにも長所短所があります。データセンターでは常に、どのアルゴリズムが一番優れているかと質問されました。むろん答えは「状況によります。どの分野に使うんですか？」です。これはあまり満足のいく回答ではありません。誰もが一番良いものを知りたいがります。研究者間でアルゴリズムをめぐる競争もありました。どのアルゴリズムも一長一短で、向いている分野と不向きな分野がありますが、アルゴリズム競合をめぐる科学界の駆け引きが明らかになりました。データセンターが誠実な媒介役として、各アルゴリズムの具体的な用途を提案できるようになるま



(図1)

で、しばらく時間がかかりました。私たちは勝者・敗者を選ぶのではなく、誠実な媒介役を務めている者として、研究界に信頼を確立するまでに時間を要しました。この最終成果物を生み出すまでに投じた努力には、そうした一面もあります。

異なるリモートセンシング・プラットフォーム上のさまざまなセンサーについて、時系列データも作成しました。これらのデータを一貫性ある時系列に統合し、長期的推移を把握できるようにしました。これ自体が、異なるデータセンサーや衛星プラットフォームを集約し、一貫した時系列データを作成するという研究でした。データをグラフ上にどう配置するか、どの予測を使うか等の問題も扱いました。

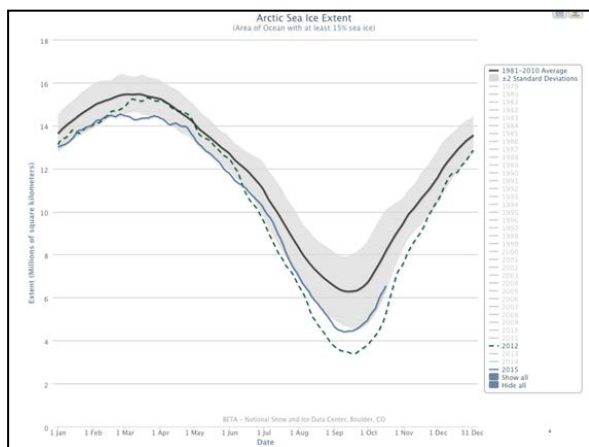
データセンターは数年前、全データセットを包括的に再処理し「これが、最善のアルゴリズムを使った一貫性ある時系列データだ」というベンチマークを作ろうとしました。ところが面白いことに、データを完全に再現することはできませんでした。その過程には、文書化されずアルゴリズムや時系列データで把握されていない、人間による判断があったのです。データ自体に問題はなかったけれど、検討すべき一定の不確実性があるのだと思います。

最後は文脈です。このデータをどう理解するかということです。これは単なる画像に過ぎず、氷縁がここにあるのは明白ですが、「氷縁」の定義は研究コミュニティによって違います。この氷縁の定義に用いたアルゴリズムは、気象学的な観点に立つもので、長期的な時系列や傾向分析に適しています。北極圏への船の航行に取り組む船舶会社なら、氷縁を全く違う形で定義し、このデータは必ずしも役に立たないでしょう。クジラやアザラシを捕獲する基盤として氷を利用する漁師なら、また別の形で氷縁を定義します。データの使われ方を理解することが重要です。アルゴリズム、時系列、予測、グリッド、文脈のうち、文脈とはその問題を特定のオーディエンスとどう関連づけるかです。最初に2007年、次いで2012年とこの問題で私たちは市民の関心を集めました。マスコミにはデータへの

理解を深める手段が必要でした。

データ発表の手段として、もう一つ作成したものがこれです。操作してさまざまな年を表示できる、オンラインチャートです（図2）。これは1～12月の年間サイクルの長期的な平均値です。海氷は毎年増減し、一般に9月末に一番少なくなります。灰色の部分は、標準偏差±2の範囲です。2012年は劇的に少なかったことが分かります。長期的な平均値を、標準偏差で3～4下回っているかもしれません。この長期平均に関し留意すべきは、この長期的な平均値は定常期でなく、減少期の値である点です。2012年に大きく減少したことが強調されていますが、この双方向的なグラフは、異なるオーディエンスに異なる文脈でデータを提示する一つの方法に過ぎません。青い線は今年、2015年の海氷の量を示します。2日前に私がウェブから集めたデータです。図3が実際のデータで、綺麗なグラフで見るよりはるかに複雑です。ここから、使用した多様なアルゴリズムの複雑性を感じ取れるでしょう。ほぼリアルタイムで処理されたデータもあれば、ユーザーへの詳しい説明が必要な他の成果物と統合されたデータもあります。

北極圏の変化を理解しようとするなら、海氷は非常に重要な変数です。複雑な変数なので、さまざまな問題に対処するため多数の分野のデータを統合する必要があります。複雑性の一例をあげると、茶色い四角はデータソースです。緑の五角形は、データソースから作成したデータ成果物です。青い楕円形は、センサー



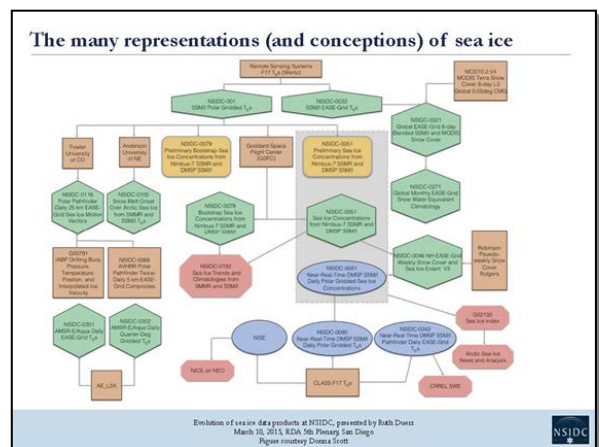
(図2)

での検知から24時間以内に作成されるほぼリアルタイムの成果物です。赤い八角形は解釈に基づく成果物で、気象学やグラフなどを含みます。黄色の角が丸い長方形は、最終的に消去される予備的な成果物です。これら全てを総合して、先ほど見せた図が作られます。この図には、灰色の枠で囲った項目が表示されています。

一つの成果物を紹介するには随分長い導入ですが、要は、データセットを利用可能・解釈可能・再利用可能なものにするために必要な、成果物同士、アルゴリズム同士、ひいては科学界の駆け引きの相互関係を伝えなかったのです。

データの生成価値

私たちは、データの生成価値を高めようとしています。データの生成価値とは、ジョン・ウィルバンクスがジョナサン・ジットレインの研究に基づき提唱した非常に興味深い概念です。ジットレインは当初、オープンネットワーク（特にインターネット）の生成価値を語りました。ジョン・ウィルバンクスはこの概念をデータに敷衍（ふえん）し、データの生成価値を「多様で広範なオーディエンスからの検閲のない貢献を通じて、予想外の変化を生み出す力」と定義しています。すなわちデータは、人々が操作しその結果何が起きるかを見るために開示されるべきです。順応性やアクセス可能性が高く、簡単に理解でき、相互接続性が高まり活用されるほど、データは一層生成的になるとウィ



(図3)

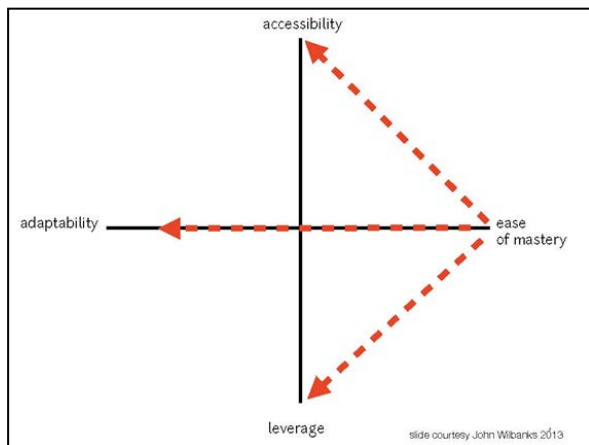
ルバンクスは主張します。データは、さまざまな環境や適用分野に順応できるものでなければなりません。簡単に理解できるというのは、データの使い方や各自の文脈に応じた利用の仕方がどれほど簡単に分かるかです。また、データは簡単にアクセスでき、それが影響を与える他のデータと相互接続できる必要があります。これは把握しにくい概念ですが、ここで考えるのは古典的経済理論でいう正味現在価値ではなく、正味将来価値であり、そのデータを長期的にどう活用できるかだといいます。生成価値を高める試みは、現在のデータ慣行と正反対に位置するものです。

ウィルバンクスは、これを四つの次元から説明します（図4）。私たちはこの四つを最大限に高めようとしています。これがその4次元です。順応性は、さまざまな用途にどれほど順応できるかです。理解の容易性は、どれほど簡単に使えるかです。アクセスしやすさも入っています。それに、ウィルバンクスが「活用」と呼ぶ要素、すなわちデータがどれほど相互に接続し影響力を持つかです。各次元は互いに正反対に思えます。海氷の画像は簡単に理解でき、専門家でなくても分かりますが順応性はありません。単なる綺麗な画像に過ぎず、それを使って詳しい分析はできません。時にはキュレーションで、データを公開前に完璧にしようとすることもあります。しかし、手を加えず開示し、どんな予想外の変化が起こるかを見た方がよいのです。

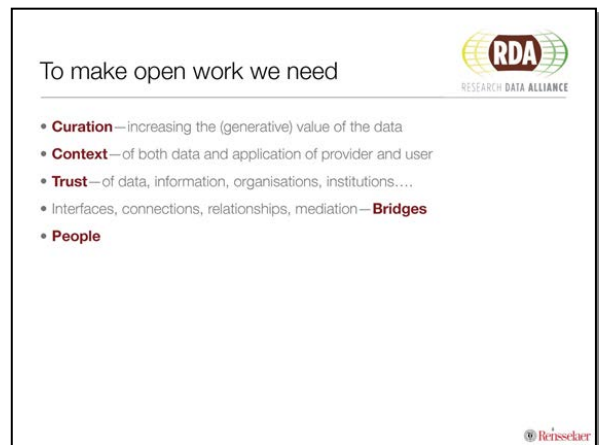
海氷を例にとると、ホッキョクグマの研究者は何年も前からデータセンターを訪れ「海氷データが欲しい」

と言っていました。海氷の後退はホッキョクグマの生息地や狩猟パターン、生活などに影響します。ホッキョクグマ絶滅の危険もあります。海氷の専門家は「このデータは不確実すぎる。氷縁のデータは、そちらの分野で使うのに適さないから使うな」と言いました。私は先日、アメリカ地球物理学連合の会合でこのデータを使ったホッキョクグマ研究者のポスター発表を4件見ました。データを作った人間が、そのデータの再利用法に関して助言する上で最適の人材とは限りません。これが想定外の変化に関する側面です。海氷研究者が不確実性を強調したこと自体は正しいけれど、その不確実性をどう解釈するか、自分の分野に活用するかどうかを決めるのは、ホッキョクグマ研究者です。その分野の専門家は彼らだからです。直感的には理解できないかもしれませんが、データを作成した専門家に、そのデータの最適な再利用法を尋ねるのがいいとは限らないのです。

オープンデータを機能させるには、幾つか必要なことがあります。幾つか要件を挙げますが、この一覧が全てではありません（図5）。データをより良いものにし、生成価値を高めるためキュレーションが必要です。データを使いたい無数の人に役立つような文脈に配置する必要もあります。信頼の問題には後で触れますが、データの信頼性、その背後にある組織への信頼性、誠実な媒介役としてのデータセンターへの信頼性を含め、信頼は大きな問題です。これら全ての要件の間に相互的なつながり、接点、結びつき、相関関係が



(図4)



(図5)

あります。仲介的な面があり、データセンターは原データ作成者と多数のエンドユーザーの媒介役を務めます。これが橋渡しで、この橋渡しという重要な例えについては後述します。もちろん人的側面もあります。これはデータ科学の研修、データ利用に関する研修、人々のニーズに対する理解なども指します。

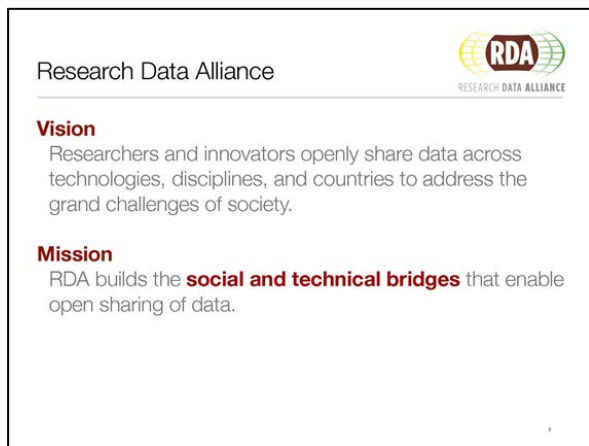
ある分野の一つのデータセット、一つの変数だけでこのように長い説明になります。これらが合わさって、幅広いオーディエンスに有用なデータになります。Research Data Alliance は、あらゆる分野の全データにこれと同じことをやろうとしています。簡単に言うと RDA のビジョンは「研究者とイノベーターが、社会に壮大な課題に対処するため技術や分野、国を越えて公にデータを共有する」ことです（図 6）。そのために、データのオープンシェアリングを可能にする社会的・技術的な橋を築きます。社会的・技術的な橋とい

うのは、重要なメタファーです。この例えが重要である一つの理由は、相互接続や橋渡しのためにはインフラを考える必要があるからです。私たちは、基本的にはデータインフラ構築を目指しています。しかしここから、ではインフラとは何かという疑問が生じます。それを理解するのは困難です。

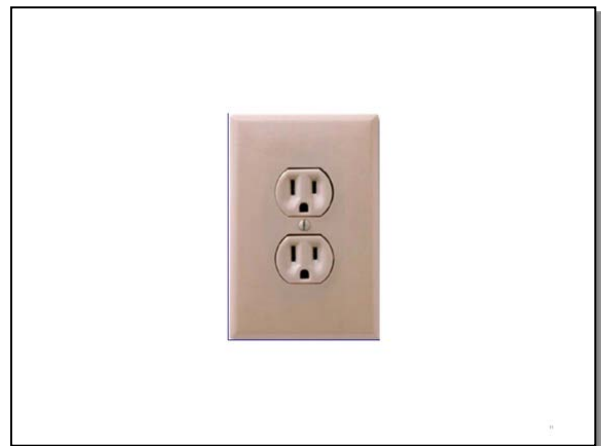
インフラの力学

現状は図 7 のような感じですが、米国向け（図 8）、欧州向け（図 9）のコンセントに当たるものを作りたいと考えています。だから、図 10 のアダプターを使うのです。単なる気まぐれではなく、実際にインフラはこうした接続を通じ機能しています。

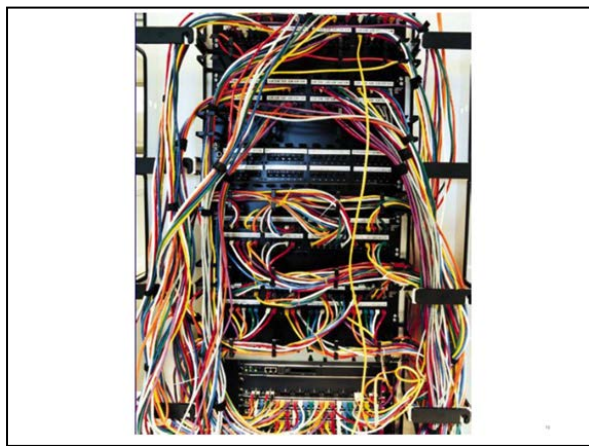
「インフラ研究」と呼ばれる学問分野があります。過去にインフラが構築された過程を研究する学問です。インフラ研究の報告書を読んでみたいという方がいれ



(図 6)



(図 8)



(図 7)



(図 9)

ば、エドワーズらが書いた『Understanding Infrastructure: Dynamics, Tensions, and Design』を強く薦めます。簡潔な要約がついていて、鉄道や銀行制度、インターネットまで過去の全インフラに関する調査が含まれます。その最初の考察は当たり前かもしれませんが、インフラは成熟するに従い普遍的でアクセス可能な、信頼性、透明性あるものになると理解することが大切です。透明性とは、故障するまでその存在に気づかないという意味です。例えば停電するまで、電気はあって当たり前だと思います。優れたインフラは日常生活の一部と化すため、目に見えなくなります。

しかし、透明なものは研究できないため、インフラ研究は難しいのです。過去のインフラの分析を通じて、報告書の著者らは、マスタープランに基づきトップダウンで作られるインフラはないことを発見しました。最初はマスタープランがあっても、計画どおりには決して行きません。むしろインフラは進化的な過程をたどり、最初のシステム構築段階では、特定の文脈の特定の問題を解決する個別技術を開発します。次にネットワーク間または技術移転の段階に入り、開発したシステムやツールがさまざまな場所で異なる文脈に使われます。相対立するシステムが現れることもあります。電気分野の代表例は、AC（直流）対 DC（交流）、テスラ対エジソンです。最後は統合の段階で、ネットワーク間に橋が築かれインターネットワークが生まれます。この統合段階は、異なるシステムの結びつきを可能にする通路や橋によって特徴づけられます。この

電気コンセントのアダプターは、私たちが当たり前とみなすものの例ですが、おかげで世界中どこのホテルでもノートパソコンを使えます。これがインフラの基本的な部分です。

本当はインフラとは「何か」ではなく、「どの時期にあるか」が問題です。果たしてどの時期に、橋や通路を特徴とする統合の段階にあるのか。データ界は今、その時期を迎えていると言えます。今がデータの統合段階です。けれど単に「インフラがどの時期にあるか」でなく「何がインフラか」という問題も生じます。

エドワーズらは、この通路や橋を、技術的問題に対処するための技術的な解決策、あるいは社会的問題に対処するための社会的な解決策とみなしてはならないと指摘します。実際には両者の組み合わせで、どちらか一方では決してありません。社会的選択を伴う技術的な解決策である場合も、多々あります。規格を開発すると、今度はその規格を一貫した形で使うという選択が必要になります。たとえその規格が機能するよう望んでも、実際の使用法を指定することはできません。ツールは提供できますが、人々に、そのツールを各自の文脈で適用するという選択をしてもらう必要があります。技術的か社会的かというのは誤った二分論で、両者の橋渡しが必要です。

私はインフラを、人間、技術、制度、組織の間の相関関係、相互作用、結びつきと定義しています。これら全てにより、データセットがさまざまな文脈で再利用可能な有益なものになります。RDA は、こうした



(図 10)



(図 11)

社会的、技術的な橋や通路、相互の結びつきを築くことを重視しています。

「Create, Adopt, Use（作成、採用、使用）」

「作成、採用、使用」が私たちのモットーです。このスライドは、フラン・バーマンが作成しました（図11）。RDAでは12～18カ月間の短期的なワーキンググループを設置し、最終的に実際に何らかの活動を実施します。単に提言や白書を作成するだけでなく、実行に移すことが必要です。政策の実践、システムの相互運用性、共通の類型・規格・メタデータ、持続可能な経済、地域社会での慣行採用、研修・教育・労働力などです。

持続可能な経済に関しては、地域社会の支持を得てシステムを長期的に維持する方法はあまり考えていません。これを示す興味深いメタファーとして、公共交通が挙げられます。地域社会での慣行採用については、私は道路交通規則をよく引き合いに出します。世界各地を旅すると、右側通行か左側通行かを除けば、交通規則はどこでも似ています。しかし、パリ、東京、北京、ニューヨークで車を運転するのは、全く違う体験です。道路規則を実際にどう適用するかをめぐって、地元の人なら誰もが知っている地域社会独自の規範があるからです。地域社会で慣行を採用すると、こうした規範的アプローチに行き着くのです。最後に、研修や人材教育を実施しています。RDAはこうした活動を行っています。

トップダウン式の計画はありません。ただコミュニティに「一緒に短期的なワーキンググループを作り、架け橋を築きましょう」と呼びかけるだけなので、一体感はあまりありません。ただし、一定の共通原則に従うよう求めます。私たちは、それぞれ違う成果物を生み出す1,000本の花のようなものですが、全員が共通原則に同意します。最初の原則はオープンネスです。これはオープンデータに限りません。プロセスを開示し、幅広い参加を受け入れるといったことです。それに、コンセンサスに基づく意思決定、特定の技術に偏

向しない均衡のとれたアプローチ。また調和に向けたアプローチで、さまざまな要素を結びつけるよう努めます。加えて地域社会主導、非営利性などです。

私たちはどんな架け橋を築いてきたか。データエンジンを例に取りましょう。これもフランが考えた例えです。彼女は、データエンジンの技術的側面に触れています。データの記述法を示す参照モデル、データ型のレジストリがあります。これはMIMEタイプのよなもの、と考えて構いません。MIMEタイプを見れば、メールに添付されたものがWordファイルかそれ以外の形式か分かり、どのプログラムでそのファイルを開けばよいか分かります。研究データも同じようにタイプを作成し、ツールを使ってデータ型を判定できたら素晴らしいでしょう。例えば、特定の単位で流動速度を示した水路図のデータ型があるかもしれません。小麦データの相互運用性に関する枠組みは、より特定の分野に限られます。この枠組みでは、小麦に関連する情報の値（収穫量、タンパク質含有量など）を意味論的に関連づけます。

前述した道路規則の場合、どのように取り組んでいるか。法的な相互運用性は、しばしば話題にされる大きな課題です。RDAではワーキンググループでなく、長期的なインタレストグループとして、法域を越えて機能する共通原則を作成しています。データにまつわるライセンス取得や著作権の問題は、国により法的アプローチが違うせいで必要以上に複雑化しています。このグループは、そうした問題の調和を目指しています。

研究者それぞれがデータエンジンを持ち、それを運用しなければなりません。私たちは特に途上国で、データ科学カリキュラムを開発するためCODATAと共にサマースクールプログラムを開催しています。

積極的なデータ管理計画は、私が気に入っている活動の一つです。多くの国が、データ管理計画を提案し、義務づけ始めています。米国国立科学財団は、数年前から義務づけています。他の国もこれに追随しようとしています。しかしデータ管理計画は、決して計画通

り進みません。研究自体が計画通り行かないからです。「計画は役に立たないが、計画作成のプロセスは欠かせない」というドワイト・アイゼンハワーの言葉を引用したいと思います。データ管理計画は持続的な動的プロセスでなくてはならない、そこが面白い点です。

初期の成果物、新たな成果物、今後の成果物

RDA の具体的な成果物を少し紹介します。恐らく多少難解ですが、2 年前に生まれた団体なのでまだ実質的な成果物はありません。最初の四つの成果物は、コンセンサスを経て正式に承認されました。コミュニティで合意が得られれば、RDA の協議会が「この成果物は承認済みである」と認めます。「継続的な異議がない」などコンセンサス形成の定義が非常に厳しい ISO のような規格認証機関と違い、RDA では、複数の組織に採用され使用されているかをもとに判断します。完全な採用でなくとも、誰かが使用していれば大丈夫です。

簡単な基本的用語から始めました。これはいわば非常に基本的なデータモデルです。「デジタルオブジェクト」「永続識別子」などの簡潔な用語が含まれます。一貫した形で議論を進めるために、これを承認しました。専門用語は、今後も論争が続くテーマでしょう。

次の成果物はデータ型のレジストリ、MIME タイプです。これはさまざまな場所で話題にされているようです。全てのデータ型に単一のレジストリを用意するのでなく、連合的な制度です。各研究者が、自分の適用領域、分野用にレジストリを作成しますが、全員が同じモデルに従って作成するため、レジストリ間に一定の相互運用性が生まれると期待できます。

永続識別子型のレジストリも同様に、特定されたデータの種別を永続識別子で定義します。論文なのかデータセットなのか、人物なのかといったことです。

それに「実際のポリシー」と呼ばれる基本的な機械稼働用ルール (machine actionable rules) があります。ポリシーといっても政治的な意味ではなく、機械稼働を可能にするためのルールなので、固定性チェック、

コピー、バックアップなど基本的なものです。ここでも信頼が問題になります。リポジトリがこの一貫した機械稼働用ルールを使用していれば、一定の信頼性が示唆されます。どれも些細で難解な技術的な事柄ですが、将来的にはより広範な架け橋作りを目指しています。

先月の RDA 総会で、幾つか新しい提言を発表しました。現在、正式な承認手続きを進めています。規格のメリットは、沢山の規格から選べる点です。メタデータ規格のディレクトリを使えば、数多くのメタデータ規格から選びやすくなります。例えば研究データ管理を引き受けようとしている図書館が、さまざまな分野の研究データを預かり、メタデータに関し選択を迫られた際に役立つでしょう。

ダイナミックデータの引用手法に関しては、確かに私たちは収集段階ではデータ引用法を知っています。必ずしも引用するとは限りませんが、引用の仕方は知っています。しかし、絶えず情報が付加され何度もバージョンが更新される、動的に変化するデータセットの一部を引用するのは非常に難しいことです。例えば特定の時点の海水データの一部に言及する場合、どうすればよいか。これは、そのためのアプローチです。

小麦を記述する用語については、嬉しいことに国連食糧農業機関に採用され、世界で広く使用されています。地球全体の 70 億人への食糧供給に関連する、基本的に重要な問題です。

データ記述レジストリの相互運用は、面白い問題です。これは、レジストリ同士を相互に結びつけるサービスです。多くのデータレジストリが存在しますが、共通キー、永続識別子、ID 付与、著者名などを通じて異なるレジストリを結びつけることができます。

今後の成果物のうち最初に紹介するのは、統一リポジトリ認証制度です。パブリッシャーや資金提供者らは、信頼できるリポジトリを把握する手段を求めています。リポジトリ認証には多くの規格があるため、World Data System や Data Seal of Approval を使ってさまざまな規格を調和させ、一貫性ある一つの規格への統

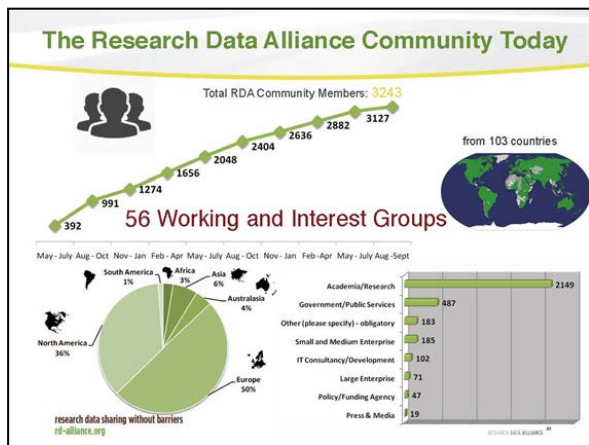
合を目指します。

次に、データ出版関連サービスにも関与しています。データ出版は最近流行りの言葉で、論文を入手でき、その論文とデータの密接な関連づけを行えるような形でデータを提供することを指します。しかし、そのインパクトをどのように測定するか。「文献計量学」というより、data-io-metrics という言葉の方が適切でしょう。インパクトファクター自体がそもそも欠陥のある測定法であることを踏まえて、データのインパクトファクターは何なのか。どうすれば、データを論文と関連づける形でワークフローを管理できるのか。これは、前述したレジストリや、レジストリの相互接続のシステムにも関連する問題です。

ここまで、RDA の成果物を高い次元から俯瞰しました。難解なのは分かっているので、詳細を知りたい方は私に質問するか、RDA ウェブサイトをご覧ください。設立後まだ2年半という点を強調しておきます。こうした初期的な成果物はありますが、さらに多くの取り組みが進行中で、大きな関心を集めているようです。

現在の Research Data Alliance コミュニティ

このスライドは既に古いですが、RDA コミュニティは急速に成長し、現在の会員数は約 3,500 人です（図 12）。100 カ国以上が参加していますが、今は北大西洋地域が中心です。これを変えたいという思いもあって、今回日本を訪れました。これは欧州・北米の



(図 12)

みの取り組みではないのです。会員の大部分は研究者ですが、民間部門、政府部門も増えています。民間部門に力を入れたいと考えていますが、数字的には増えており関心も高まっています。56 のワーキンググループ、インタレストグループが結成されています。

個人会員に加え、団体会員もいます。団体会員と違い、個人は無料で参加できるので皆さんに参加をお薦めします。ある程度のリスクを負って主体的に参加してもらうため、団体会員はわずかな会費を支払います。私たちは実施と採用を重視しているため、実際、団体会員に RDA 内でより多く発言してもらいたいのです。物事を採用するのは、個人ではなく団体であるべきです。研究者個人が自分の研究室で何か採用するのも結構ですが、真の影響を持つには団体が採用することが必要です。そうやって、成長していくのです。そのため団体の視点や彼らのニーズ、何を実施する予定か等に耳を傾けたいと考えています。

これは会員・関連団体のスライドですが、やはり情報が古くて、この後に 4~5 団体が加盟しています

(図 13)。関連団体には、出版社やスーパーコンピュータセンター、図書館、技術コンソーシアム、研究所など多様な団体が名を連ねていますが、民間部門は多くありません。けれど最近、数名の民間会員を獲得しました。民間部門との連携を拡大する必要があります。

会員数 3,500 人、56 のグループでさまざまな事に取り組んでいるので、多少混乱しています。全てを把握するのは難しいですが、多様性や相違点を越えて幾つ



(図 13)

かのテーマが浮上しつつあります。一つ目は「永続識別子」というシンプルな概念で、特定の情報を長期にわたり明確に一貫性ある形で参照できることです。これにはデータセットや論文だけでなく、人、制度、プロセス、手段も含まれます。この要素は、56 グループ全てに共通して見られるものです。どのグループも、相互運用性あるデータインフラを構築するには、永続的な参照が必要だという前提に立っています。

同僚のピーター・ウィッテンバーグは、インターネット誕生の経緯を踏まえ図 14 のスライドを作成しました。インターネットプロトコルは、全てのネットワーク技術と全ての付加価値サービスをつなぐ一種の腰のくびれのようにになっているという概念を表した図です。単純なインターネットプロトコルに、多くのものがかかっています。これはデータネットワークに似ていると言えます。データネットワークでは、永続識別子が、データソース、デジタルオブジェクト、データベース等と、アプリケーションや分析プラットフォームといった付加価値サービスをつなぐいわばくびれになります。永続識別子は、重要な概念です。

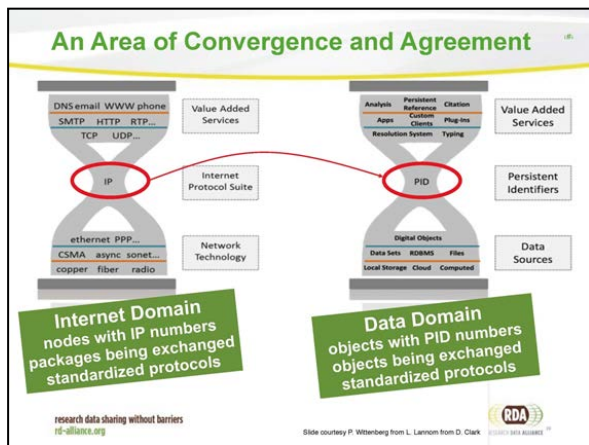
二つ目のテーマは、信頼性の認証という考え方です。どうやって何かを信頼できると言えるのか。信頼できるリポジトリを示す認証制度はありますが、常に信頼性を証明する必要はあるのでしょうか。集团的知性に頼ることはできるのか。データ出版という発想からデータのピアレビューが示唆されますが、これは全くの誤りです。データはレビューを受けないからです。デ

ータは、測定値または「事実」であり、論文の主張と同じようにレビューはできません。アプローチの仕方が違うのです。では、そのデータをどうやって信頼するのでしょうか。ホッキョクグマ研究者は、専門家が使うなと注意してもデータを信用しました。信頼性の証明というのは難しい問題です。

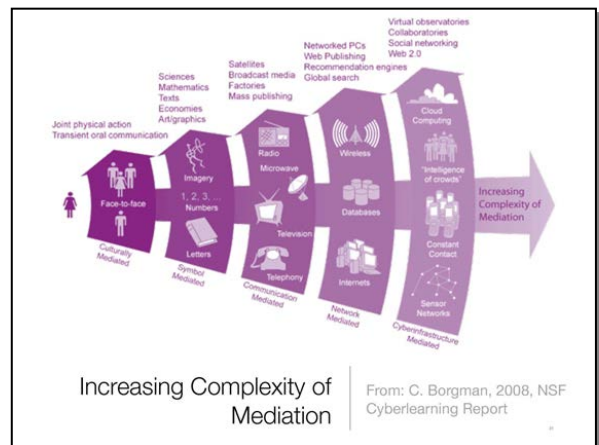
対話、関係性、相互接続、媒介の重要性については、既に繰り返し話しました。媒介は、私たちの取り組みの中心を成すものです。クリスティン・ボグマンは、時とともに媒介が複雑化する様子を描いています（図 15）。昔は文化が媒介役を務め、コミュニケーションは対面式でした。テレビや電話を媒介としたコミュニケーションを経て、私たちは現在、インターネットやデータベースによる媒介に慣れています。現代世界では複雑性がさらに増しつつあります。集团的知性だけでなく、キュレーターや媒介役自身の知性も関与できます。こうした仲立ちを務める人の役割があります。この媒介的接続が、RDA の取り組みの焦点です。

RDA は組織として、状況の変化を認め柔軟に対応することを目指しているため、媒介という考え方が関わってくるのです。

ある同僚が、アジャイルソフトウェア開発宣言を RDA 用書き換えてくれました。ソフトウェア開発の世界では昨今、アジャイル開発が大流行です。アジャイル開発用の宣言があり、それを同僚が、組織のあり方を踏まえて書き換えたのです。私たちは、プロセスやツールよりも個人と交流、技術よりも人間に価値



(図 14)



(図 15)

が個別に取り組んで得られた成果です。相互の関連性に気づいた彼らは、個々のピースを継ぎ合わせるデータファブリック・グループを新たに設置しました。このグループは実際、機能しています。これら小さな橋が互いを結びつけ始めています。

ピーターは、各グループの下にぶら下がる要素も示しています。これはデータ管理計画、リポジトリ認証等を示します。私には自信がないので詳しい説明は省きますが、この図はデータをめぐる取り組み全体を捉える一つの方法です。

一歩離れて、RDA を機能させている要因は何か考えたいと思います。なぜ短期間に 3,500 人もが RDA に集結したのか。それを支える熱意は何か。データシェアリングとは何を意味するのか。それは、保存、アクセス、再利用、文書化、プロトコルなど幅広い内容を意味し、この極めてシンプルな図に触発されたものです。私は基本的な形でデータ管理について考えてきました。RDA に来る前から、長年データ管理に関わってきました。この仕事の素晴らしい点の一つは、他のさまざまな専門分野が各自のデータをどう管理しているか、データをどう捉えているか分かります。データの捉え方は分野により違います。あるグループにとってのデータが、別のグループから見るとデータではないことがあります。その逆も起こり得ます。興味をそえられるところです。私はデータの機能の仕方を考えてきましたが、幾つか基本的な原則や共通点があります。目新しいものではありませんが、ここで繰

り返す価値はあります。この原則は、インフラ開発の参考になります。

ステップ 1 はデータの保存です。これは 10 年ほど前に実施した、空中センサー、モデル、現地実測を用いて、50 人が山に入り雪を計測した大規模な野外調査の写真です（図 18）。スキーで転倒しても、この男性はフィールドノートのデータを守りました。つまり収集の時点では、データ保存の意識が非常に強いのです。研究者は、データを保存する重要性を理解しています。

別の例をあげると、ある人が野外調査で小型のデータ記録装置を使って雪の深さを測っていると、測定途中で装置が止まりました。彼は翌朝早起きして出かけ、同じコースをもう一度計測しました。データが大切だったからです。現場では保存意識が強いのに、最初の使用が終わると意識が薄れます。どうやってこの意識を保つのかというのは、重要な課題です。保存は難しいですが、何をすべきか誰もが知っています。問題は、それをどうやって維持するかです。ある意味で私たちはこの問題を解決しましたが、ビジネスモデルはまだ見つかっていません。データ保存は難しく、その背後には強い研究者精神があります。

ステップ 2 は、データシェアです。私はこの写真の女の子の表情が大好きです（図 19）。彼女は少し不安げで、多分嫌がっています。私が強調したいのは、私たちに共有する文化がないということです。分野により大きく違いますが、保存とアクセスは別々ながら



(図 18)



(図 19)

互いに依存し合う問題であるという点に留意する必要があります。同じコインの表と裏なのです。アクセスがないのに保存するのは無意味だし、保存しなければアクセスは不可能です。

しかし共有の方法が分からないので、RDA ではとにかくこれに取り組もうとしています。オープンネス、調和、バランス、コンセンサスという前述の基本原則に従っていますが、共有の仕方が分かりません。

1,000 本の花、ボトムアップ型アプローチという RDA のやり方を疑問視する声もあります。計画が必要だということです。しかし私は、このやり方を信じていますし、上手く機能していると思います。小さな架け橋が幾つも生まれ、互いに協力し始めています。

信頼性

先ほど、多様性を越えて共通するテーマがあると言いましたが、それら全てを含む更に大きなテーマがあります。信頼性です。信頼の認証ではなく、もっと大きな意味での信頼性です。マーク・パイドは、デジタルオブジェクト識別子である PI（永続識別子）の誕生 20 周年にこう述べています。「アイデンティティ管理の普遍性を阻む障害が主に技術的なものでなく、同一の識別子システムを等しく適用して、学術論文とハリウッド映画を識別できるとすれば、障害は果たしてどこにあるのか？ 規格に疎い外部の人はしばしば、規格とは単に技術の話だと思込んでいる。最高の技術的ソリューションが、必ず勝利を収めると。実際には、規格作成に関わる人間は罪深い秘密を知っている。その一つは、規格実施で何より大切なのは、人々が進んで従う許容できる社会的規範を確立することにあるという点だ。どうすればそれが実現するか？ 答えは、一つの単純な単語の中にある——信頼だ」。

パイドはこう続けます。「・・・規格界では『中立性』が話題になるが、規格作成にはもう一つの隠された、あまり認識されていない秘密がある。完全に中立的な規格などあり得ないのだ。全ての規格には、暗黙の『観点』、規格全体の仕様の背景となる一連の前提

が含まれる」。しかし、「その仕様は、規格を実施する人々の間でコンセンサスが得られた『観点』を表現するものである点も重要だ」。規格や識別子に関するこの発言と同様、データシェアリングでも信頼性が重要になります。データシェアリングは、摩擦や違いが発生しやすい場所だからです。そんな場所では信頼が不可欠であり、信頼があつてこそ、実施に際してコンセンサスが得られた観点を維持できるのです。

信頼構築について、素人なりに考えてみました。信頼性をいつ証明するのか、本当に証明する必要があるのか。どうやって信頼性を証明するか。集团的知性に頼るのか、専門家の知恵を借りるのか、単に再利用で判断するのか。こうした疑問はまだ解決されていません。あれこれ手を加える余地を残す必要があります。公表したものを人々がいじれるようにする必要があるという点は、オープンソース・ムーブメントに似ています。何か発表し、それを他人がいじり操作するのを認めるのは信頼性の問題です。これは難しい問題です。

分散化を取り入れる必要があります。全てを一括管理する中央システムを作らず、集権化する場合はコミュニティが統制する必要があるという意味では、これもオープンソースモデルに似ています。政府や何らかの当局が行うのではなく、コミュニティ自体が分散化を支持することで、信頼が築かれます。

信頼は、経験共有を通じて築かれます。私たちは互いから学ぶため、RDA は経験共有の一つの手段です。半年ごとに総会を開き一堂に会します。そこで学びが生まれるのです。全ての規格に観点が含まれるため、RDA は関与の場、意見が対立する相手と関わる建設的な論争の場です。信頼性を示すもう一つの方法は、データやツールの再利用です。RDA は、使用されているかどうかを重視しています。使用されていれば、一部の人から一定の信頼を得ている証なので、それで十分なのです。

持続可能性に関しては、継続的に使われているかどうか、長期的に機能しているかです。RDA は、人々を結集させる国際的な取り組みとして、広範囲の連携

を目指しています。信頼性の概念に関する考察はこれで終わりですが、こうしたコミュニティの努力と関係構築が、信頼を築くのに役立つと思います。RDA は、国際的支援に基づく広範囲の連携も確立しようとしています。

地域 RDA と初期的なインパクト

私たちは国際組織ですが、地域 RDA という構想も持っています。データシェアリングという課題に、グローバルとローカル両方のレベルで取り組む必要があります。私は「グローバル (glocal)」という言葉が好きで、これは単に「グローバルに考えローカルに行動する」という意味ではありません。同時並行的に両方の場で活動することです。RDA ジャパンを例にとると、日本はオープンサイエンス・ムーブメントをめぐり独自のポリシーや優先順位を構築しています。従って国内で機能する形をとりますが、これを国際レベルに反映させてから再び国内に戻すことも可能です。私たちは、幾つかの国でそうした例を見てきました。これは一つには、資金調達の仕組みの影響です。団体・資金提供機関は、自国で資金が使われるのを望みますが、実際にはそれが有益な運営方法であると分かっています。

結論として、RDA は実際にインパクトを与えています。私たちには影響力があります。まず「ビッグデータ」が流行語になりデータに注目が集まっています。私はこの流行語が好きではありませんが、世間に広まったことは喜ばしいことです。私たちは数十年前からデータについて考える必要性を訴えてきて、ようやく今その意見が聞き入れられています。RDA は、こうした声の成果であるとともに、それをさらに増幅する取り組みです。私たちのメッセージを、さらに大きく訴えています。

コミュニティでは何度も、連携という言葉を目にします。RDA からもたらされる成果の多くは、連携の価値が生んだものです。例えば連携により調和が加速します。私たちはデータの引用法を知っていますが、

最近のことです。データ引用法の進化を見れば、実際には競合する複数のアプローチがありました。一定の対立があったと思います。さまざまなグループがこの問題を最初に解決しようと考えましたが、第 1 回 RDA 総会で全グループが集まり、協力の必要性を認識しました。そのため彼らは、データ引用の共通原則を開発しました。これは、データ引用原則に関するさまざまな団体の共同宣言です。

これは原則で、誰もが賛成できるレベルの高次の宣言に過ぎませんが、これで全員が力を合わせ、研究者にどうやってデータを引用させるかといった一層難しい課題に取り組むことができます。データの引用法は知っていますが、どうやって習慣化するか。ダイナミックな引用はどうするのか。単にコミュニティが協力することで、この大きな前進が得られました。

私たちは共通テーマを見つけています。永続識別子はどこにでもあります。誰も想定しませんでした。あらゆる分野・観点から永続識別子が生まれています。人々は、これは自分たちが把握できる共通テーマだと気づきました。偶然ですが、私の友人の大気科学者がデータ型のレジストリを発表し、ある海洋学者が船旅の話をしていました。大気科学者は、自分たちも計測用「プラットフォーム」を使用すると話しました。この 2 人は、共に抱える課題や共通の解決策を編み出す方法について、素晴らしい議論をすることになりました。研究者の間には、思いがけない接点があります。

古いパラダイムの再検討を通じて、新たな洞察も得ています。組織構造を持たない RDA という概念は画期的で、それがおおむね受け入れられたこと自体、大きな前進です。

もう一つのインパクトは、2 年間で実際の成果物を生み出した点です。単なる白書や提言でなく、早い段階での失敗から私たちは急速に学んでいます。18 カ月という制限があるため、みんな成果を出し、何かを成し遂げることに集中します。上手くいけば素晴らしいですが、上手いかわくても失うものはさほどありません。もう一度挑戦し、実際に何を実現できるか示

せます。

橋と橋の間に、更にそれらをつなぐ新たな橋があらわれ始めています。相互接続性が有機的に成長しています。

最後に、RDA は資金の節約にもなります。多くの人が、RDA で築いた人脈を使って新たな助成金獲得のチャンスを手にしたと言います。ある研究者は、RDA で得た人脈だけで3件の助成金を得ました。うち1件は欧州の Horizon 2020 プロジェクトから得たもので、明確に RDA に言及しています。他の2件では言及されていません。彼女はこの成功を、RDA で築いた人脈のおかげと考えています。

あと一つ、ジェイミー・シアーズから聞いた例を紹介します。彼は、大型ハドロン衝突型加速器のデータ全ての保存を任されました。データは現段階で100ペタバイトあり、更に増えています。彼はこの問題への自分の対応を、カエサルにならい「来た、見た、勝った」と表現します。彼は RDA 総会や、さまざまなワーキンググループ、インタレストグループの会合に参加しました。他の分野がこの問題にどう対処したかを学び、プロセスを発展させ洗練させました。現在彼は、このデータの保存に関するコストモデルを含む戦略を実行しています。これにより、2年の期間と約20FTEの人員数を節減できたと推定されます。浮いた人員は解雇せず、別の職務に割り振れるでしょう。これは目に見える経済的な効果です。彼の見積もりに1桁間違いがあっても、それでも大きな効果です。ここでも、RDA はインパクトを与えています。

今後もこのインフラパイプラインを発展させ、小さな架け橋づくりを推進したいと考えています。コミュニティ全体の連携を高め、新たな領域に手を広げたいです。純然たる研究者で RDA に参加する人は多くないので、間に立つ RDA の人間が研究者に働きかけ、彼らのコミュニティを巻きこむことが重要です。これが、アウトリーチやスタッフの教育、資金提供団体の巻き込み等に関係する活動です。将来はそうした方向に向かいます。

RDA は半年に一度、世界のさまざまな場所で総会を開催しています。次の総会は3月初旬に、ここ東京で開かれます。皆さん奮ってご参加ください。東京の次は、International Data Week の一環として米国で開催されます。ご意見、ご質問お待ちしております。ありがとうございました。

●フロア1 AIST でウェブデータ管理を担当している Julian Leblay と申します。RDA は興味深い取り組みですが、色々な点でセマンティック・ウェブを連想します。似通った要素が幾つかあり、活動の仕方も似ています。第一に、RDA の取り組みはセマンティック・ウェブと比べてどうですか。これに関連して、セマンティック・ウェブで生じたさまざまな問題に、RDA はどう対処していますか。例えば固有識別子に関しては、草の根的な活動なので、似通った目的に対して人々が別々の識別子を使い始めています。データ統合の問題もあります。これは多くの場合、技術的な課題になります。

●Parsons 第一に、セマンティック・ウェブは技術であり、ティム・バーナーズ・リーが考案した基本的ルールに基づく、リンクト・オープンデータという技術的アプローチです。それは相互運用性に向けた一つの手法です。RDA にもそうした要素はあります。例えば小麦のデータは、おおむねリンクト・オープンデータのアプローチに沿っています。語彙を具体的に検討する新たなグループもありますが、ご質問は、この包括的アプローチを追加できるかという点だと思えます。しかし世の中には多くのアプローチがあり、それらは時に相入れません。どうやってそれらを縫い合わせるのか？これは有機的アプローチで、いわば適者生存です。だから、私たち自身が解決策を持っているだけでなく、コミュニティに解決策を求めています。二つ、三つ、いえ四つの対立する解決策が生まれても全

く構いません。有機的な進化の中で、ダーウィンの意味で変化に一番適応したものが生き残ります。そういう方向に進んでほしいと願っています。だから組織的な採用を重視しています。組織に採用されれば、ある程度受け入れられた証拠になります。

私たちは、「競合する二つのシステムがあるから、こちらを使いなさい」あるいは「両者を関連づける方法を考えなさい」とは言いません。そういうやり方はせず、有機的アプローチを採用しています。実際、インフラはそうした形で生みだされるからです。インフラはトップダウンで構築されません。そのためセマンティック・ウェブは、十分に機能しないでしょう。万能な唯一の解決策などないからです。これは、ある種哲学的なアプローチだと思います。上手くいくかまだ判定できませんが、私たちは正しい方向に進んでいると思います。ただし、このアプローチには確実にリスクが存在します。

●フロア2 RDAの紹介があり、グループ結成から12~18カ月後に成果を発表するとのことでした。成果物が期待に沿うものか、評価する手段はありますか。

●Parsons それは今苦労している点で、前の質問とも関連します。生み出された成果物をどうやって維持するか。評価面では、技術的な承認には慎重です。一方が他方より技術的に優れているとは言いません。ただ、ある規格が実際に使用されているため、一部の人には有用であると言います。そこから支持が生まれます。

18カ月が過ぎた後、成果物をどうやって長期的に維持するかというと、多くの場合、公表されるのはパイロット版、ベータ版です。どうやって更新するかというと、成果物のバージョン2を作るため、新たなワーキンググループを結成する場合があります。一種のメンテナンス組織として、活動期間が長く広い権限を持つインタレストグループを結成する場合があります。未解決の根本的な課題の一つは、「not invented here

(専用に開発されていない)」という問題です。これに関して、皆さんの意見を喜んで伺いたいと思っています。グリッドコンピューティング、セマンティック・ウェブ、GISシステムいずれであれ、誰もが私たちの文脈では上手く機能しないと感じます。個々の文脈に応じた特別なツールが必要だと感じ、独自のツールを使うことを選びます。完璧なサイロ型ツールより、不完全なツールを共有する方が良いと、全世界を説得する必要があります。

●フロア2 RDAは、法律上の(de jure)規格ではなく、事実上(de facto)の規格を作ろうとしている。この理解で合っていますか。

●Parsons 私たちは規格を作成する団体ではありません。あえていえば、その両方です。RDAは規格実施団体です。規格はあるかもしれませんが、詳細な適用の方法となると必ず厄介な問題が生じます。これは、ワーキンググループが取り組む問題かもしれません。リンクド・オープンデータの規格を、どうやって小麦に適用するか。

規格作成前の手続きもあります。誰かが立ち上がり、良いアイデアがあるからIEEEやISO、他の正式な規格認証団体に持ち込もうと提案しなければなりません。RDAは両方の立場にあり、実施を重視しています。持続可能性の質問に戻ると、実施され採用された規格は、維持されやすくなります。関係者が規格存続に関心を持つからです。これが実際のどの程度効果的かは、また答えが出ていません。

●武田 今のお答えからすると、RDAは非公式な団体なので、成果物の評価は、フィードバックが基本となるのですか。

●Parsons 失礼、その点を明確にすべきでした。成果物が公表されると、意見募集の期間を設けるよう求められます。ワーキンググループ結成時も、意見募集

期間が要求されます。今 RDA のウェブサイトを見れば、新規に提案されたワーキンググループが、コミュニティの意見を募集しています。成果物を公表する際にも、意見募集期間があります。技術諮問委員会も、成果物を正式に評価します。団体諮問委員会もあり、参加団体が集まり、成果物が各団体に有用かどうか意見を出して議論に加わります。これが、コンセンサス形成プロセスの一部です。

評議会は RDA の上級機関で、ビジョンや戦略に責任を負います。評議会は彼らの観点に基づき、コンセンサスを評価しようとします。宣言を出すのではなく、コミュニティ内にコンセンサスがあるかどうか尋ねます。評議会は、集まった意見を検討して問題がないか、それとも致命的な欠陥があるか判断します。評議会は、各団体が実際に選択し実際に使用しているかを見ます。ここから採用されたかどうか判断します。コンセンサスが得られれば、RDA の承認印をもらえます。

規格の長期的な維持は、ものによりさまざまでしょう。小麦の場合は、国連食糧農業機関（FAO）が採択したので、その時点でもはや RDA の課題ではありません。

●武田 つまり 18 カ月議論しても、それで議論が終わるわけではないですね。

●Parsons そうです。私たちも、それに気付きました。最初の四つのワーキンググループは、おおむね活動を終了しました。まだ活動を引きずっていますが、成果物を出す期限を決めたのは有益だったと分かりました。RDA を始めた当初、私は懐疑的でした。国際機関が 18 カ月で何か産みだせると思わなかったのですが、今は考えを改めています。18 カ月でやることは可能です。期限があることで集中でき、意欲が高まる一方、分かりやすい成果物が生まれます。RDA はボランティア組織なので、ボランティアも、1 年半の間だけ参加しやり遂げればよい、果てしなく活動に関わるわけではないと感じられます。ただし、社

会的志向が強い成果物の場合、18 カ月は少し短いかもしれません。18 カ月の研究というルールを変えるつもりはありませんが、RDA には順応性があります。新たなアプローチを採用する可能性もあり、常に提案を受け入れています。

●フロア 3 今日は、RDA が実施を目的とする団体だと聞いて嬉しく思います。RDA に参加していない人は、RDA が何か理解していません。RDA は規格統一団体ではなく、優れた慣行や最善の形での実施を目指しているということです。

ダイナミックなデータ引用を例にとると、一定の成果物がありますが、それが最終地点ではありません。研究データを整理するための出発点に過ぎないのです。メタデータレジストリ、リポジトリ認証など、多くの要因が複雑に関係し合っています。ワーキンググループの成果物について、この種の重層的な議論を行い、混成的なシステムに新たな層を作り出すといったことはしていますか。

●Parsons それに対し、直接的なはっきりした答えはありません。これら三つ、四つの問題は、全て関係し合っています。さまざまなグループが、さまざまな形で結成されています。データファブリック・グループは、四つのワーキンググループから生まれたインタレストグループです。彼らは、四つの成果物がどのようにに存続し、相関関係を持つ可能性があるか探ろうと考えました。こうした議論の場が、今も続いています。

ほぼ有機的に結成されたもう一つのグループは、World Data System との連携です。書誌計量学、ワークフロー、論文・データの相互接続から成る、データ出版に関する成果物が幾つかあります。World Data System もインタレストグループで、今後 6~12 カ月間の採用数の増大に取り組みたいと考えています。彼らは、「not invented here」の問題に苦勞しています。彼らとしては、全ての出版社に、これが良いアプローチだと納得させるという点を重視することになるでしょう。

さまざまなグループが結成され、独自の活動クラスターを形成しています。

RDA は、こうした関連性を強めようとしています。何回かの総会では、グループ間の協力や合同セッション開催を求めました。それが上手くいく場合もあれば、いかない場合もあります。どのグループとどのグループが協力するように、といった指示は出しません。一度試みましたが、一回以上セッションを開催しなければ、別のグループと協力するしかないのです。先程のクラスター化を通じて、どのグループ同士が協力できるか把握できます。RDA のガバナンスレベルでもさまざまな努力がなされ、技術諮問委員会と共にグループ間の相互連携を促していますが、草の根レベルで相互連携に取り組んでいる方が多いです。

例えばレジストリ相互接続とデータ出版社サービスは、全く別のコミュニティでお互いのことを全く知りませんでしたが、RDA を通じて出会いました。両者が協力したアプローチを用いて、欧州の Open Air が開発を進めているシステムが、少なくとも一つ存在します。私たちは、連携するよう命じるのではなく、単に連携を促し、そのつながりが堅固なものになるよう願っているだけです。彼らが互いにつながり合う機会を提供しているのです。

●フロア3 現在、複数のワーキンググループを結びつける試みをしていますか？

●Parsons 多少試みています。今の時点でお見せできるものではありません。しかし若手研究者が RDA に参加できるよう、早期キャリアプログラムを実施しています。彼らは互いに結びつくための起点としてマッピング・エクササイズを始めました。これは技術諮問委員会が検討していることですが、提案に過ぎません。協力しろと言われても、人は協力しません。実際に一緒に活動させれば、協力していくでしょう。

●フロア4 AIST の Jason Haga です。二つ質問があ

ります。一つはデータ出版に関連します。どの出版社が実際にデータを発行しているか、あるいはデータセット受入と発行を検討しているか、教えてください。

二つ目の質問はクラスター化に関連します。バイオサイエンスなど特定領域の人の場合、利用できる規格を見ても、どれが自分たちに適用できるか分からないのではないのでしょうか。RDA はサービス提供者になりたくないという意図は理解できますが、特定領域の研究者を支援し、どれを活用できるかということや、採用方法を伝える取り組みはありますか。

●Parsons 最初の質問については、データ出版と、データに関連する従来型出版を区別したいと思います。個人的意見ですが、データ出版は、データに学術的評価を与える手段です。それは持続可能なアプローチではありません。出版は固定的である一方、データは非常に動的です。私は領域別リポジトリの出身です。領域別リポジトリは、データの利用可能性という点で出版社より優れた仕事をします。これは私の個人的意見で RDA の見解ではありませんが、データ出版グループは、この種のデータ出版について議論しているわけではありません。ドイツの出版社 Copernicus が持っている、地球システム科学のデータがあります。データジャーナルに関心があるなら、「Nature」には新しいデータジャーナルが幾つかあります。私は関心ありませんが。データと論文を相互に関連づけるのは、全く別の問題です。出版社は、データを保有したくないのです。彼らは、データを保管する信頼できるリポジトリを正確に参照し、主張が十分な証拠があること、その証拠とリンクできることを好みます。しかし、証拠を維持するのは彼らの責任ではありません。これを行っているジャーナルは、Elsevier、Ubiquity Press、Faculty of 1000、PLOS などです。このうち数社は、データ出版ワーキンググループで非常に積極的です。

領域・分野に関する二番目の質問は、非常に良い質問です。今後 RDA が力を入れて行きたい点です。これに関し幾つか取り組みを実施しています。第一に、

アンバサダープログラムを作成しました。多くの純粋な専門分野の研究者が、RDAに参加するとは思えません。データ志向の研究者は参加するでしょうが、大半はデータ相互運用性の詳細より、自分の研究上の疑問に関心があります。従って、応用コンピュータ科学者や図書館員など、データ実務に近いRDA内の人間が、研究者に働きかけることが大切です。

私が米国雪氷データセンター (NSIDC) のデータ管理者だった当時、研究コミュニティとの関係構築が重要でした。そのためにフィールドワークをしたのです。そうすることで、信頼性が得られました。アンバサダーを育て、各領域に働きかけることが大切です。しかし、領域特異的な問題か、一般的な問題かを見極める努力をすべきです。

私たちは、村山さんがおっしゃった連携を促そうとしています。総会と総会の間に、ワーキンググループ、インタレストグループの長を集めた会議を開いています。次回は12月にワシントンで開かれ、テーマは分野特異的グループです。どんな場合に、分野特異的になるのが妥当か、どんな場合は一般的であるべきか、その二つをどう区別するか。これは答えが出ていない問題で、明確な線引きがあると思えません。私は答えを持っていませんが、少なくとも議論が進んでおり、面白いものになるでしょう。

●フロア5 私はRDAに詳しくないので、永続識別子システムの現状を知りたいです。PIDの仕様が既に発表したのですか、それとも仕様修正を協議している段階ですか。PIDの目的はデータを表現することですが、一種のIPアドレスといえます。RDAは、データを表現し、IPアドレスに似たシステムを活用しようとしているのですか。

●Parsons 私の技術的知識の限界に達しています。これは最初の質問に関係するかもしれませんが。世の中には幾つもの永続識別子システムがあり、多くは同じ種類の事柄を表現するのに使用できます。正直なところ、

識別子システム同士の宗教戦争のようなものです。

これまでRDAは中立を貫き、特定の側につかないようにしてきました。DOIがいい、いやHandleがいい、あるいはLife Science ID、UUIDが好きだという話はしたくなかったのです。ある程度の調和が見られた分野は、永続識別子型のレジストリでした。永続識別子をメタデータと共に登録する方法のことです。恐らく、識別子を解決する方法、特定する方法といった技術的システムより、識別子をめぐるメタデータの標準化に取り組んでいるのでしょう。識別子の解決法や識別子が指す内容について、多少の情報を求めているのです。この辺りが私の技術的知識の限界ですが、これまでRDAは、完璧な永続識別子の包括的仕様を定義してはいません。そういう活動の仕方はしません。一貫性ある永続識別子が誕生すれば素晴らしいですが、近いうちにそうなるとは思えません。

●フロア6 今日はどうもありがとうございます。質問があるのですが、一つは、RDAという組織はできて新しいので、まだ活動的にいろいろなアイデアがどんどん出てきていると思うのですが、組織というのは長くやっていると、だんだん硬直化していったり、次の新しいフォーラム等でこれまでの過去の資産をどう生かしていくかというものが大事になってくると思うのです。そういう形で、このRDAの組織をこれからどうサステナブルにして、さらに同じような成果を出していくかということに対して、どのようなポリシーを持たれているか、教えていただけますか。

●Parsons 質問の意味を確認させてください。今の勢いをどう維持するのか、そのためのポリシーはあるか、さまざまな成果物をどう管理し把握するかという質問だと理解していいですか。

●武田 これまで2年半やってきて、5年、6年、10年後には多くの成果物が生まれるでしょう。RDAの方向性は何ですか。

●Parsons 第一に RDA は、10 年計画に基づき活動しています。10 年後も存続すると想定していません。存続しているかもしれないし、全く別の組織になっているかもしれません。問題を解決し役目を終えているかもしれません。変化が急速なため、遠大な未来は描いていません。ここ最近、今後の方向性を計画立案していますが、たかだか3年後までです。組織運営をどう改善するか、より多くのコミュニティとどう関わるかなど、戦術的な意思決定に重点が置かれています。しかし、あらゆる場所で持続可能性の問題に突き当たります。実に良い指摘だと思います。成果を維持し組織を維持することが問題です。

私たちは、組織を維持するため政府に頼るビジネスモデルを作成しました。政府は大量の研究データを生みだし、社会は研究データ再利用の恩恵を受けている。従って政府は、社会の代表としてデータ保守に関心を持つべきだという論拠です。それが、組織の持続可能性に向けた私たちの理念、ビジネスモデルです。アウトプットを維持できるかは、アウトプット自体とそれを利用するコミュニティに左右されます。RDA とワーキンググループ・プロセスを通じて維持したいというなら、それも構いません。別の団体が成果物を採用し、その団体が保守を受け持ってもよいのです。

機動性を維持し硬直的にならないという理念を、どうやって維持するかは、自分たちの姿勢を見守るしかありません。「これが昔からのやり方だから」という罫に陥るのは簡単です。個人的には、私が今の仕事を続けるのは2〜3年です。一つには、自分がこの仕事をずっと続けられると思わないため、事務局長になるつもりはありません。それが組織にプラスになるとも思いません。刷新が必要です。同様に、コミュニティの熱意を保つ必要があります。ボランティアをサポートしなくてはなりません。全員が、自分のアイデアは重要だと感じられ、RDA にアイデアを持ち込めるようにする必要があります。協力して何かをやり遂げられるグループ（特に国際的なグループ）があれば、私たちは支援します。

私が組織に植え付けようとしているポリシー、いやポリシーというより理念は、機動性（アジリティ）です。プロセス全体を通じ、ボランティアを重視し適応力を維持しようと努めています。また設立後間もないので、RDA が成功するかどうか分かりません。

●フロア7 RDA に参加したいと思ったときに、18 カ月のコミットメントが必要となると結構敷居が高いと思うのですが、何か一番簡単に参加する方法はどのようなものなのでしょうか。例えば普通の国際会議だと、論文を発表してそれで終わり、そういう参加の仕方ができますが、RDA の場合は、グループをつくって、18 カ月コミットメントして、レポートを書いて初めて参加できるという、何か結構難しいと思うのですが、もう少し参加しやすい方法は。せっかく東京でやりまし、参加を考える人もいると思うのですが、そのときにいきなりそれだと大変だと思うので、質問しました。

●Parsons 分かります。参加にはさまざまな形があります。RDA の実態や仕組みを理解する一番の方法として、総会に参加するようお勧めします。数日の参加ですみ、状況を実際に見て探ることもできます。ご自身が適切と思うレベルで、ワーキンググループやインタレストグループに参加できます。見守るだけでもよいし、メーリングリストに参加して、興味を持てる面白い結果が出るか見ているだけでも構いません。特に魅力的なものがあり、2〜3 カ月本格的に関わりたい、使用事例を発表したいというなら、それも可能です。アウトプットを試し、実際に自分に役立つか確認することもできます。18 カ月参加する必要はありません。そうするのは、問題を解決し、ワーキンググループで取り組みたいという熱意がある場合だけで構いません。

どんな組織もそうですが、RDA 会員 3,500 人のうち、積極的に活動しているメンバーは 10〜20%のみです。見るだけの人もいますが、大部分は来ては去って行き

ます。関心がある六つほどのグループの活動を見守り、特に興味深いものが出てくると対話に参加します。自分が一番関心があるのは何か見つけるのが、最も難しい部分だと思います。後で個別に相談に乗り、お手伝いすることもできます。5 カ月後またここに集まり、総会に参加してください。それが、RDA の仕組みを理解する一番の方法です。