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－」

ディスカッション　「研究データ共有は今後どうあるべきか？」

武田　英明　（国立情報学研究所）
Mark Parsons　（研究データ同盟（Research Data Alliance）事務総長）
北本　朝展　（国立情報学研究所）
池田　大輔　（九州大学システム情報科学研究院）
能勢　正仁　（京都大学大学院理学研究科）

●**武田**　それではパネルディスカッションに入りたいと思います。発表された 3 人の方に何か質問等がありましたら最初にお願いします。

●**フロア 1**　二つ質問があります。一つ目の質問は、池田先生のご講演に関係します。DIAS が格納しているデータで、約 8 割と一番メモリを消費しているのは、シミュレーションしたデータや、観測データを基に、地球全体で点数値をグリッド上に配分した、同化データ（assimilation data）といわれるものです。それはいずれもかなり容量が大きいので、簡単に転送できるものではありませんが、そのようなものは第 3 の柱と第 4 の柱のどちらに入るのでしょうか。データセットの数は少ないかもしれませんが、無視するには容量が大きいというのが私の質問です。

二つ目の質問です。北本先生の遺跡の話で、分野が違う中でのコミュニケーションは難しいという話がありました。せっかくネーティブスピーカーの Mark さんがいらっしゃるので伺いたいのですが、日本語で「分野」と言う場合、“discipline”と“domain”という言葉が乱立していて、区別がありません。そこが議論の混乱の原因になりそうなので、そこを簡単に説明していただけますか。

●**池田**　昔から転送できないほど大きいデータはあり、郵送するということもよく聞きます。基本的にはデータから何か発見するものがあれば、第 4 の柱にも入ってくると思います。例えば機関リポジトリ上の論文なども、それをマイニングするとなると、第 4 の柱に入ると思いますし、第 3 の方法から生成されたものも、第 4 の柱に入ることは十分にあると思います。

●**武田**　能勢さんが observational physics と experimental physics では、データシェアリングのモチベーション s が違うとおっしゃっていましたが、実は同じことをお二人の方がおっしゃっていたような気がします。同じことを言っていたつもりなのか、それとも違うのか。あるいは、今、言ったようなケース、観測データを組み合わせてつくったデータでするとしたらどうなるのかということをお聞かせください。

●**池田**　私は同じつもりでは全くなかったのですが、先ほどの能勢先生の話に合わせて言うと、シミュレーションから生成されたデータは、実験の physics と同じではないかと思います。モチベーションとしては低いというのも、多分そうではないかと思います。シミュレーションの方をものすごく詳しく知っているわけではないので、まず推測が入っているのを了解していただきたいのですが、スパコンなどは、プラットフォームがそれほどコモディティー化していないというか、

一般的なものではないですし、研究室で代々引き継がれているようなソフトウエア資産でつくっていると思うので、使い回しが利かないというか、シェアされていない分野だと思います。ですから、そのデータがシェアされるかと言われると、確かに違うかなという気がします。今、指摘されてはじめて気付きましたが、能勢先生の対比と同じようなことが言えるのかもしれません。

●**能勢**　私も指摘されるまでそのような視点はなかったのですが、今のご質問にあったようなデータの量ではなく、シミュレーションのデータという意義ですか。

●**フロア１**　観測結果を基にデータ同化をする、そのようなデータです。データ同化をした結果のデータです。高層大気だと、あまりデータ同化はしないのでしょうか。

●**能勢**　やっているグループもあります。やはりシミュレーションやデータアシミレーションのような、研究者のアイデアが入ってくるようなところは experimental になるのではないかと思います。シミュレーションの計算する前のコードも研究者は外へ出したがりません。それによって自分たちが研究をして結果を出したいからです。データをシェアというのはよく聞きますが、シミュレーションのコード、特に最新のコードをシェアするというのはあまり聞かないので、先ほどの対比では、experimental データという形に対応するのではないかと感じました。

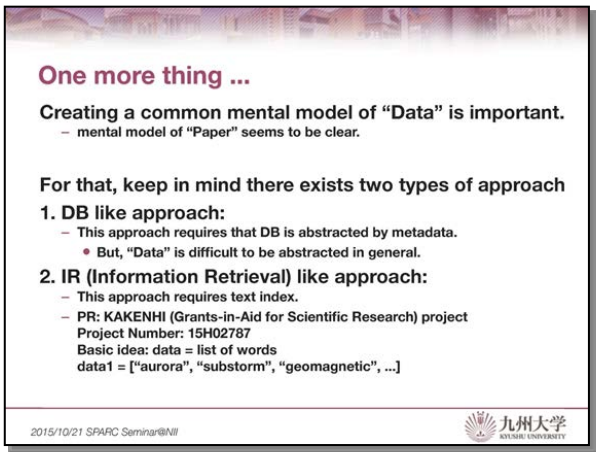
●**武田**　では２番目は Mark さんに質問なのですが、少し難しいですが、“discipline”と“domain”は違うのかということです。

●**Parsons**　domain の方が大きいと思います。正直なところ、私自身も最近ようやくこの区別を思いついたばかりです。domain は地球科学など非常に大きな分野を指し、海洋学などは discipline というのが私の解釈です。discipline は domain の中に含まれます。domain は非常に幅広いものです。

●**フロア２**　池田先生にご質問です。図１の“One more thing…”のところで、リポジトリのつくり方には、データベース的なものと Google 検索的なものの二つあると言われたのですが、最後は少し時間が短かったので、もう少しご説明を伺いたいです。

●**池田**　ここは図書館のコミュニティーの人が多いですが、図書館は先に分類がきちんとあって、どう分けるかという枠組みが定まっています。一方で、必ずしもいつもそれで満足できるという人はいなくて、例えば数学などは、少なくとも九州大学では独自分類をしています。ですから、その分類でいいかと言われると、それはユニバーサルではないということです。データ分野ごとに非常に違うものだと思うので、それをデータベース形式にして、先にスキーマを定めてしまうと、メタデータとしてはあまり意味がないというか、取得日やファイルサイズ程度のものになってしまって、探すのにもあまり良くないことになるのではと言いたかったのです。

探すためだけであれば、インデックスするように、データと単語をひも付けることができれば足りるだろうというのが、情報検索というつもりです。ただし、言っておかないといけないのは、本当にそのデータを



（図１）

使うのであれば、結局その分野のことを知らないといけないということです。北本先生などもおっしゃっていたと思います。探すときは Google 的に探せたとしても、結局はその分野に入っていく必要があるので、あまり楽ではないと思います。

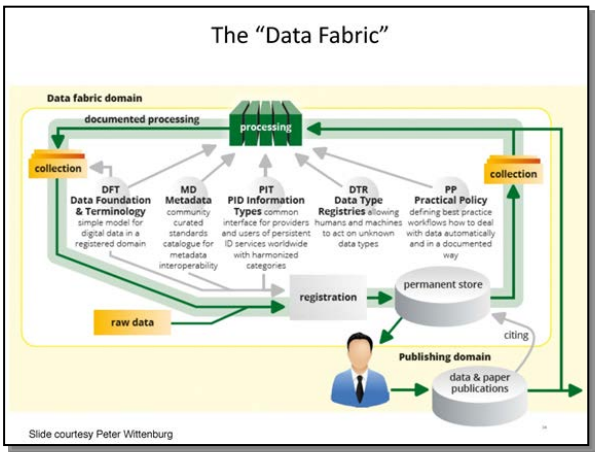
●**フロア 2**　今のはメタデータの探し方のお話だったと理解したのですが。

●**池田**　データを探すときに、データをメタデータで抽象化しようとすると、データは分野ごとに非常に違うので一般的なメタデータにしかならず、探すのには役に立たないのではないかということです。

●**武田**　ここにいらっしゃる 4 人の方々の間で、ご発表について、お互いに質問等はございますか。

●**池田**　Mark さんに質問です。図 2“Data Fabric”のところで、raw データを登録するのですが、例えばセンサーネットワークを考えたときに、データは常にずっと連続的に生成されますよね。どこを一つの単位として登録するというのは、非常にサブジェクティブというか、主観的な気がするのですが、そのあたりはどうお考えでしょう。

●**Parsons**　良い質問ですが、私にも明確な答えはありません。有用性の単位をどうするか、データのどの



（図 2）

要素に独自の識別子を付与するかは、文脈に左右されます。私が以前勤務していたデータセンターでは、データの大部分がファイルベースで、一部はデータベース化されていました。私たちは、各ファイルに、UUID や ARK などの識別子を付けようとしていました。それを時間・空間・手法など何らかのレベルで集計し、より大きな識別子を付与して、収集データにメタデータを付ける取り組みも進めていました。それが収集データであり、収集データの個々の構成要素です。

ご指摘のように、データとはどんなものかという一般的モデルはありません。率直なところ、それは果てしない哲学的議論の一つかもしれません。データというのは、用途があって初めて存在するものです。私が考えるデータは、あなたが考えるデータと同じではありません。あなたにとってのテキストが、私にはデータかもしれない。ある意味、データはその使用法によって規定されるのです。

●**武田**　それでは、少し共通の議論をしたいと思うのですが、その前に一つだけ私からの追加コメントです。先ほど、実験物理の場合は、実験データは独立であまりシェアするモチベーションがない、観測物理の場合は、観測データはシェアするモチベーションがあるという話がありました。例えばバイオサイエンスだと、どちらかというと結果の方をシェアします。遺伝子の構造の発見、相互作用の発見などは、発見そのものがデータで、それを共有するというケースもあります。この場合はいろいろな実験をして、それをさらに計算して、何らかの発見がシェアされます。ですから、これもまた違うデータのタイプで、モチベーションも違います。この場合は、むしろそれを公開しないと、成果として認められないという文化があります。

私からの質問はここから先です。この場合、研究あるいはデータのクレジットを誰に与えるのか。研究者あるいは関与した人に、どのような立場があって、現状でそうである、あるいはこうなるべきだろうということに関して、皆さまのご意見が聞きたいと思います。

これは難しくて、論文ですら、著者に誰を並べるのかという段階で、分野ごとに問題が起きています。その辺も広げて考えていないと、逆に言えば、単にデータというだけでは答えられないかもしれませんが、誰が関与して、誰の名前を載せるべきか。特にデータを公開するにはどうしたらいいのかということについて、皆さまの立場からコメントいただきたいと思います。

●Parsons　それは非常に大きな問題です。幾つかの視点から意見を言わせて下さい。私もデータ引用に触れましたが、他の講演者のお話にも何度か登場しました。私は 90 年代からデータ引用という概念に関わってきました。私がデータ引用に携わった本来の動機は、データ共有のインセンティブの付与、知的労力を注いで優れたデータセットを作成した人へのクレジット付与でした。その人物は研究者、現場の技術者、コンピュータプログラマー、それら全てかもしれません。

データ引用の動機が、時とともに大きく変化しました。今は、論拠の証明や再現性のため、実際に使用したデータを正確に参照するという意味で引用が注目を集めています。ここから、データ引用は研究共有のインセンティブなのかという疑問が生まれます。現在まで私は、データ引用が共有のインセンティブだという確かな証拠に出会っていません。そのため、クレジットとは何か、何にインセンティブを与えたいかという概念全体を、一歩離れて見直す必要があります。何に見返りを与えたいのか。同様に、さまざまな物事に対して誰が説明責任を負うのか。

研究界では、教授や教授を目指す人にとって、定評あるハイ・インパクト・ジャーナルでの引用には価値があります。それは重要な財産です。他方で現場の技術者で、専門が実験デザインプロトコルや道具の設計である場合、発表実績で評価されることはありません。自分のデザインがどれくらい成功したか、開発した道具がどのくらい堅牢かによって評価され、引用など気になりません。仕事の内容に応じて、違う形でのクレジットを望むかもしれません。論文もこれと同じで、各著者が果たした異なる役割を記載するよう求めるジャーナルが増えています。研究全体に含まれる多様な役割を区別し、役割毎にクレジット表示することを目指す、Project Credit のようなプロジェクトも登場しています。でも少し先走りましたね。誰のために、どんな文脈で何をクレジットするのか。

データセットの著作者という概念は、一部の研究者にとって、一部の文脈で重要ですが、それ以外の人には、全く無関係です。彼らがクレジット表記を欲しがるのは、上司に良い仕事をしたと認めてもらえるから、専門学会の人名録に名前が載るからです。何のために誰にクレジットを与えるか考える前に、まず何にインセンティブを付与するか考える必要があります。質問に質問でお答えしました。

●**北本**　今の Mark さんの、データサイテーションの目的が変質してきたという話を興味深く聞きました。私はサイテーションはクレジットが目的というのが理想の姿だと思っているのですが、結局それがなぜ reproducibility の方に行ってしまったかというと、今のジャーナルから見て何が便利かという発想でいくと、どうしてもそちらの方にねじ曲がっていく傾向が出てくるからではないかと思います。そちらから見ると、サイテーションは再現可能の方に行った方が便利で、サイテーションによるクレジットはもう論文で満たされているので、わざわざデータでやる必要がないと考えると、どうしても再現可能性の方に行ってしまう。やはりジャーナルが支配する世界における力学として、そちらに行くのだろうと思いました。

ですから、データに対してクレジットをどうやって持っていくかは大きな問題で、この前のプレゼンで私が提案したのは、いろいろな学会で「巨人の肩賞」をつくってはどうかということです。「巨人の肩」は英語で“shoulders of giants”というのですが、「研究とは巨人の肩をつくることである」とニュートンが言ったといわれるように、重要なキーワードです。要は研究の価値は巨人の肩をつくっているかどうかで測られる

べきであって、論文は一つの基準でしかなく、データのインフラをつくることなども、どれだけそれが巨人の肩なのかというところで評価して、全ての学会がそれで評価すればいいのではないかということです。サイテーションは一つの基準にすぎませんが、コミュニティーの人はみんなそれを誰がつくったかを知っているということはよくあります。明文化されていないけれど、「あの人のデータだったら信用する」というのがあります。要はクレジットであってアウォードなので、そのようなところで、いろいろなところの基準をつくっていくのがいいのではないかと思っています。

●**池田**　午前中、Dynamic Citation Working Group の話が何回か出ました。昨日は Mark さんを含めて登壇者で少し懇親の場があったのですが、そこでも何人かから dynamic citation という言葉が出てきました。私は 3 月にサンディエゴで開かれた RDA Fifth Plenary Meeting に出たときに、初めて dynamic-data citation ということを聞きました。それは情報系のウィーン工科大学の Rauber 先生が主導的にされているのですが、データセットの中の任意の部分をサイテーションできるようにするというものです。全体ではなくて、ここの部分だけを引用することができる。言い方を換えると、そこの人にクレジットを与えることができるのだと思います。

昨日も能勢先生と話していたのですが、そんなことをする必要はないのではないかという意見もあると思います。この前私は能勢先生のワールドデータセンターのデータを使いましたが、「ワールドデータセンター」と一言しか書かないと、実際にデータを取った方が誰かというのは直接出てきません。しかし、もっと細かい単位であれば、この機関だと〇〇先生と〇〇先生というクレジットを与えることができます。何年も前のものだと、もうご存命ではない方にもきちんとクレジットを与えることができるかもしれません。そうなるべきかはどうかは分かりませんが、技術的にはそのようなことができるのではないかと思います。

●**武田**　サイテーションまで話が広がりましたが、私はクレジットに誰の名前を記すべきかという、もう少し小さい問題で考えていたのです。

●**能勢**　私はデータセンターで仕事をしていて、研究も自分の職務ですが、データセンターでデータを管理して、データを配布することも業務です。われわれの分野ではデータのシェアリングは一般的なカルチャーで、当然のように行われていて、普通はデータセンターのデータを使っても acknowledgement に名前が載る程度で、あまりそのデータを誰が取ったかなどを評価することは今までなかったのです。ですから、言葉は悪いですが、ただ働きという感じがありました。やはりデータを取った人、データを管理した人、データを提供した人はクレジットを与えられるべきです。先ほど Parsons さんが「観測をしている人はあまりそこまでは考えない」とおっしゃいましたが、やはりデータを取った人のクレジットは与えられるべきで、メトリクスのような形で業績として評価できるようなものが理想的だと考えています。

●**Parsons**　私の方が武田先生の質問をよく理解していると思うので、フォローさせて下さい。クレジットに誰の名前を書くべきかというと、多くの人がさまざまな形でクレジットを与えられるべきです。データ引用でのデータセットの作者に関しては、そのデータセット作成に知的労力を注いだ人です。アルゴリズムを設計した、実験プロトコルをデザインしたなど、物理的にデータ収集していなくても、彼らは知的労力を注いでいます。

物理的にデータを集めた人も、違う形でクレジットを受けるに値します。データを編集しメタデータを付与し、公表したデータセンターも、クレジットを受けるに値しますが、これも違う形でのクレジットになります。恐らく著者でなく、編集者や発行者に近い形です。こうしたさまざまな役割を考える必要があります。

北本さんの「巨人の肩」という考え方が私は大変気

に入りました。現在は、重要なデータセットよりも、論文でありさえすればそれに大きなクレジットが与えられます。データは、学術界を越えて使われている点を理解する必要があります。多くは地球・宇宙科学関係のものですが、こうしたデータは農業予測、災害対応、土地利用計画などあらゆる用途に使用されます。こうした利用は、一部の学術的用途と同じくらい、いやそれ以上に重要です。これが引用に反映されることは決してないでしょう。そのため再利用に目を向け、データの広範な再利用はより大きなクレジットに値すること、クレジットを受けるに相応しいさまざまな役割があることを考える必要があります。

●**武田**　今の話だと、観測系だと従来はあまりクレジットを書いていませんでしたが、これからデータがもっと広い範囲でシェアあるいは活用されることを考えると、将来クレジットはもっと明示していった方がいいだろうという方向でしょうか。

●**能勢**　個人的な興味からの話になるのですが、私がDOIやデータに興味を持って参加しているのは、データを提供した人やデータを取った人にクレジットを与えたり、acknowledgeを計量化してメトリクスに載るような形にしたりするために、DOI、データサイテーションを進めていきたいと考えているからです。Parsonsさんが“deserved to”とおっしゃいましたが、データセンターのサイテーションを決める側が、誰をサイテーションしてもらうべきかを決めることができるので、データサイテーションは分割を進めていくと、きちんとしたクレジットが反映できるのではないかと考えます。

●**武田**　昔はデータは非常に固定的でしたが、先ほどdynamic-data citationの例もあったように、今はいろいろな方法が可能になったので、いろいろな人をたくさん並べても問題ないし、サイテーションしたものをどんどんトラックすることもできるようになるので、それが可能だからこそ、これからはそのようなことをもっとやっていくというのは、確かにデータシェアリングの活動の一つの方向としてはいいのではないかと思います。

ありがとうございます。一つの話題が続いたので、もう少し別の話をしたいと思います。データをつくるところは、今までも研究者や研究所がミッションとして、あるいは自分のモチベーションから行ってきましたが、いざ実際にデータをシェアしたいと思ったときには手間が掛かります。先ほどMarkさんが、“Saving data is hard, sharing is harder.”とおっしゃいまいした。“harder”の部分は、研究者自身だけでやるのではなく、いろいろな助けがあった方がいいわけですが、それは誰が、どこが、どのように助けてくれたらいいと思いますかという質問です。つまり、つくったデータをシェアするコストを誰が負担すればいいかということです。

能勢さんの場合は、もともとセンターがそのようなミッションで、今もそうだから、あまり変わらないということでしょうか。

●**能勢**　データはある程度中身が分からないと、他の人に提供することもなかなかできないので、やはりドメインのサイエンティストが関わることは必要だと思います。今回は太陽地球物理学という分野の中だけの話をしましたが、将来的に、もっと広く地球科学でデータシェアをしていこうとなった場合に、データをキュレートできる背景を持っている方の力は必要だと思います。丸投げするわけにはいかないと思うのですが、広く大きなデータになった場合に、研究者プラス、そのような知識を持った方の助けは必要になってくると思います。

●**武田**　データキュレーションをやる人が別にいるといいということですか。

●**能勢**　私もデータキュレーションという言葉を正し

く捉えているのか分からないですが、先ほど Parsons さんが、キュレーションはデータの価値を高めるとおっしゃったと思います。太陽地球物理学で言うと、一つの観測データがあってもあまり意味はないのですが、もう一つ、もう二つとデータが集まってくると、ある現象がどう伝わっていったかが分かるので、「1＋1」が 2 ではなく、もっと広がるのです。ですから、いろいろなデータの在りかを知っていて、このデータとデータを組み合わせることができるという広い視野でデータ全体を押さえられるような専門の方がいると、うまくシェアできて、さらに意味があるものになると考えます。

●**武田**　例えば美術館・博物館等のキュレーターは、その館が持っている美術品等の価値を見極める能力を持っているし、企画展のようなもので実際に集めてくるという仕事もしたりします。そのような意味では同じなのかもしれません。それは同じ能力があってできることではないかと思いました。

●**Parsons**　能勢さんにおおむね賛成です。データを収集した専門家の主観的な知識は絶対に必要です。研究者が、自分のデータを再利用可能にするため一定の役割を果たすのは筋が通ったことです。それは専門的なスキル、責任であり、そうしたスキルを教えるべきです。方法論の授業をとらずに、大学院を修了できません。データの授業を受けず、大学院を修了できてはならないと思います。

とはいえ、それが研究者の最大の責務だとも思いません。主観的な知識は必要ですが、研究データは主に公共財とみなされるべきで、データが有用で再利用可能なものとなるよう保証する公的な責任があります。そのデータを極力有用なものにするため、専門家にお金を払う必要があります。キュレーターやデータ管理者、データ科学者などです。彼らは主観的なレベルの専門家と密接に協力し、解釈や再解釈、さまざまな文脈を提供してデータを幅広く利用可能にする必要があります。従って、それは協業的な責任なのです。とはいえ研究者が、自分も一定の役割を担っていること、同時に全てが自分の責任ではないことを理解することが大切です。全ての負担を研究者に負わせるのは不当です。実際、研究者はデータシェアリングの価値を実感すべきです。太陽地球物理学などの分野では、データセンターを通じてデータを利用できるため、研究者も価値を分かっています。他のデータと照らし合わせると、さらに価値が高まります。その他の分野では価値が分からないかもしれませんが、専門的なデータ管理者は、データを再利用する人間だけでなく、データを共有する人にもシェアリングが価値をもたらすよう取り組む必要があると思います。

●**北本**　キュレーターが重要だということには私も全く同意なのですが、キュレーターだけで十分かというと、それだけでは駄目な気もします。Mark さんからインフラストラクチャーだという話がありましたが、インフラストラクチャーなのであれば、やはりアーキテクトやシビルエンジニアが要ると思います。かなり建設事業だと思うのです。建設だとすれば非常に幅広い人が必要で、アーキテクトとオーナーがいて、相談しながらつくり上げていくようなプロセスになると思うので、そのくらいのスケールで考えるとすれば、もう少しスケールアップして、もちろんそのためには、情報学の研究者が入らざるを得ないと思います。ですから、情報世界のアーキテクトやシビルエンジニアのような人が出てきて、共同作業として建設していくようなイメージなのではないかと思っています。

●**池田**　私は機関リポジトリの研究にも携わっていますが、論文をどうやってシェアするかという分野では、各ディシプリン、ドメインなどでデータベースがシェアされてきました。例えば arXiv（アーカイブ）や PubMed などがあって、一方で institutional repository もあります。ただし、ドメインごとの分野リポジトリは、幾つか超巨大なものがありますが、後発の小さいとこ

ろはかなり苦戦していると思います。3 年ほど前に、disciplinary repository のワークショップに出たときに、小さい後発のところ、例えば哲学の分野でどうやってシェアしようかということで、彼らは PubMed や arXiv など、分野リポジトリの巨人を見て、いかにしてビジネスモデルをつくって大きなユーザーをシェアするかということを考えていました。

けれども、先発のところはうまくいっているのでいいかもしれませんが、それはあまり研究者のコミュニティーの仕事ではないように思えたのです。僕らはそこで、永久に保存する役割は機関リポジトリが担って、例えば、分野別にキュレーションするところは、その上で仮想的なリポジトリとして実現すればよいのではないかと言って、数学分野の分野リポジトリをこうして構築したという発表をしてきました。

データに関しても同じようなことが言えるのではないかと思っています。インフラとして、例えば大学や研究機関などの機関では、データ登録を義務付ける方向に行くと思います。実際、九州大学や京都大学は、研究データや研究のラボノートは、パブリケーション後 10 年間は保存しておきなさいという方針を決めました。多分、他の大学もこの方向に行くと思います。そうするとその大学は、今度は保存するインフラを整備すると思います。そこにはキュレーション的な知識はあまりなくて、単に raw データに近いものがどんどん載っていくと思います。

その上で、保存に関してはあまり考えなくてよい各分野のところで、それぞれのキュレーターが各分野の知識を与える。あるいはもっと素人がやってもいいのです。天文分野などは素人がかなり強い分野だと聞くので、そのような分野ではその人たちが集めたような仮想アーカイブがあってもいいと思います。そのような 2 段階になるといいのではないかと思っています。

●**安達**　今の池田先生の意見は極めてデリケートな問題で、Mark Parsons さんにお尋ねしたいのですが、日本では昨年、研究不正で非常に大きな社会問題が起こって、研究不正を防ぐために、glacier store、要するに氷漬けにしてデータを保存しておくようなことが現実的に議論されるようになって、それとオープンデータの話がごっちゃになって議論されるということが起こってきました。先ほど池田先生が言われたように、大学で実験データやノートを全部そのまま永久保存しておくために、ストレージが要る、システムが要る、そのためのコストが掛かるという話と、今日の話題である、主として科学を前に進めるためのデータのオープンという話が一緒になっています。

特に大学のアドミニストレーションの人たちと話をするときには、研究倫理の問題の方がより比重を占めるという、僕の個人的な意見としては、全く別のことをごっちゃに扱うという状況になっています。欧米の RDA が関係するような議論では、このような問題がきちんと整理されて議論されているのでしょうか。

もう一つ似たような問題で、日本で必ずいわれるのは個人情報です。医療に関する情報のときには必ずそのようなものが引っ掛かり、まずオープンにすることに対してネガティブな障壁が出てきて、全てのその分野のイノベーションが止まるような事態になっています。そのようなことが欧米のオープンデータの動きの中でどのようになっているのか、ぜひ教えていただきたいと思います。

●**Parsons**　欧米でもこの問題は未解決です。懸念点を分けて考えるべきだというご意見には、賛成です。あまりに色々な問題が絡み合っています。クレジットは、アカウンタビリィとは表裏一体の関係にあります。不正データや再現不可能性は、日本に限られた問題ではありません。研究の背後にあるデータが誤っている、あるいはデータを入手できない、検証できないことが発覚した研究が、数多く存在します。そのせいで引用が、クレジットの手段でなくアカウンタビリティの手段という方向に進んでいます。

広い倫理的議論を切り離すことが大切ですが、そうするとさまざまな問題に突き当たります。データシェ

アリングの倫理を研究者にどう教えていくかも、考える必要があります。そもそも私は、自分のデータへのアクセスを拒むのは倫理に反すると思います。他方でデータを公開する場合、倫理的な形でアクセス可能でなければならない。アクセスがプライバシー権を侵したり、保護された種を危険にさらしたり、先住民との関係を損なってはなりません。数多くの倫理的な制約がありますが、それはデータシェアリングを阻む所有権的な制約でなく、倫理的な制約であるべきです。

私もこの問題が解決済みとは思っていないので、ご質問への答えにはなっていませんが。全てをひっくるめて考える傾向がありますが、さまざまな懸念を分けて考える必要があります。データ引用という個別の課題に関しては、懸念を切り分けて考えるべきです。クレジットの問題は再現性とは別で、再現性と来歴も別の問題です。引用は、これら全ての問題に役立つかもしれませんが、万能の解決策ではありません。さまざまな問題に万能的な解決策を提案する姿勢には、極めて慎重になる必要があります。

●**池田**　私は逆に機関の方に、つまり、わざとタングルさせてもいいのではないかと思います。特に個人情報について言うと、オープンにさせてはいけないデータはもちろんあるのですが、オープンにしてよいものだけをオープンにするのは情報の力で何とかなると思っています。一方で、例えばオープンに科学を進めるということだけだと、例えばオープンアクセスも同じような理念がありましたが、ビジネスモデルとしては全く成り立たないわけですよね。機関リポジトリが少なくとも件数として非常に多くなったのは、機関に任せたというところが大きいと思います。その機関が、例えば学生を呼ぶのに効果的だと思うかもしれないし、それはある意味どうでもいいのではないかとも思えるのではないかということです。

●**安達**　僕自身の今までのデータベース、特に数値データベースと昔言っていたものとの付き合いの経験から言うと、いみじくも北本先生が言われたように、名誉教授になった先生が死ぬまでにデータを公開したいなどというのが一番たちの悪いデータで、データベースはずっとメンテナンスしないと死んでしまうのです。死んだデータベースを預かってくれというのは、研究者へのかなり不遜な要求のような気がします。それを例えばデータキュレーターなどに任せれば生き永らえるのかというと、そうではないと思います。

先ほど来、成功しているデータセットは、やはり全部組織的につくっているのですね。データセンターなどがあって、個々人の属人的な努力を超えた形で、コミュニティーは維持していかなければいけないというのが何とか残っていて、オープンデータなどの話になっているのです。

逆に言うと、スモールサイエンスで、自分でデータを取っているような分野は、このような動きから取り残されています。そこが一番重要なところだと思っています。わが国には共同利用施設などの制度があるにもかかわらず、コミュニティーのためにデータベースを維持してつくっていこうという志向性を見せているところは非常に少ないです。そのようなところが自分の研究コミュニティーのためにデータベースをつくって維持していく、公開していくことが重要な仕事であると意識してやるのが、日本の環境の場合だと、最初の取っ掛かりなのではないのかと思います。個々の研究者に悪い環境の中で努力を求めるのは、失敗する可能性の高いアプローチのように思います。

そのような意味で、分野ごとに差が出てくるのは、当面の場合は仕方ないし、池田先生などがやっている分野はそのような苦労を既に十分してきたので、そうなっているという気がします。

●**武田**　まさに私の最初の質問の意図がそれです。データシェアリングは、研究者に全てを任せるのではないというのは恐らく明らかなので、それは誰なのか、あるいはどういう仕組みなのかというのがここでの共通認識です。もちろん答えは一つではありません。ま

ず、キュレーションのような仕事があることを認識しましょうということや、コンピューターサイエンスの力で何とかなるところもあるのだから、そのような力は積極的に借りるべきだということ、コミュニティーやディシプリンごとにそのような活動を明示的につくっていくということが、皆さんが言われたことかと思います。どれも別々ですが、一緒にならないと多分うまく回らないというあたりでしょうか。

また、オープンデータは、オープンライセンスといった、基本的に再利用が何でも可というものが多いですが、リサーチデータはそれが全てとは限らないと思います。ライセンスに関して何か考え方があれば、皆さんに伺いたいです。この場合のライセンスというのは、データを誰がどう使っていいかを規定した規約のことですが、それに関して、そのようなものは CC0 でいいのだという考えや、ある程度必要だろうという考えなど、考え方に違いがあると思いますが、それに関してショートコメントを何か頂ければと思います。

Mark さんはデータライセンスをどう思いますか。

●Parsons　公的な資金を使って集められたデータは、公共財とみなすべきです。従って、そのデータを維持する公的な責任、研究界全体の責任があります。それは単に研究者個人の責任ではありません。公共財として、倫理的な制約の範囲内で公的に利用可能・アクセス可能であるべきです。私自身は、CC0 Public Domain Mark が好きです。データをパブリックドメインに保管する限りですが。対照的にクリエイティブコモンズ・ライセンスでは、「ライセンス」という言葉自体が本来的に制約を課します。CC0 の良い点は、データをパブリックドメインに置きつつ、コミュニティーの規範を参照する機会も得られるところです。この問題には法的な対処でなく、学術的な行動規範を通じた対応をすべきだと思います。

もとの情報源の引用を義務づける法律はありません。引用がなければ単純に出版できないでしょう。それが学術界の仕組みだからです。データシェアリングを優れた倫理的慣行とみなし、データを出来る限りオープンにしパブリックドメインに保管するという、規範に基づく姿勢を育てる必要があります。以上が私の意見です。

●**北本**　私の考えは、オープンからクローズドにスペクトルのように連続的に広がっているので、いろいろなチョイスがあり得て、いわゆるオープンデータというのは極限にすぎないと思っています。ですから、いろいろなバリエーションを用意したいのですが、無限に用意すると機械可読にならないので、やはり何種類かに整理しなければいけないとは思っています。

実は DIAS プロジェクトもいろいろなライセンスがあって、整理しきれていないのが現状です。それを整理したいとは思っているのですが、一つ一つ違うので、そう簡単ではありません。ですから、あくまでいわゆる CC0 のようなオープンデータは極限という理解です。ビジネスの方の完全クローズドまでのいろいろなバリエーションがあった方がいいと思っています。

●**池田**　僕は研究データとライセンスについてはあまり考えたことはなく、オープンにしてしまえば自由に使えるものかなと何となく思っていました。一方で、アクセス制御の方が大事なのではないかと思っていました。そのようなものは楽に使えるように、情報の方でインフラとして簡単に使えるような仕組みを用意すべきだと思っています。

●**能勢**　自然科学データに限ったことになると思うのですが、最終的にはオープンでシェアが当然だと思います。よく何人かの研究者からは話を聞くのですが、結局、研究者は税金で研究をしているわけです。個人のポケットマネーで行った研究は、その人が勝手にクローズドにすればいいのですが、どこのどういうサポートでデータを取ったかということを究極的に考えると、もちろんデータを取るまでの努力は大きいので、先に論文を書くなど、先取権は認めていいと思うので

すが、やはり最終的にはオープンにしていくのが普通ではないかと思います。といっても、そのデータをどう取ったかなどは、その観測機器をつくった人でないと分からないので、クレジットを示した上で、もしくはその人にコンタクトを取った上でのオープンが最終的には目指していくところではないかと考えます。

●**武田**　いろいろ立場が違うところもあって、逆にそれがあるということが今日認識できてよかったのではないかと思います。

以上で第 2 部のパネルセッションを終わりたいと思います。皆さま、ご参加いただいてどうもありがとうございました。