第 3 回 SPARC Japan セミナー2016
「科学的知識創成の新たな標準基盤へ向けて：オープンサイエンス再考」

ディープラーニングとオープンサイエンス
～研究の爆速化が引き起こす摩擦なき情報流通へのシフト～

北本　朝展
（国立情報学研究所）

講演要旨

講演者はこれまでオープンサイエンスの意義を利便性、透明性、参加という 3 軸から考えてきたが、これだけでは科学研究にオープン化がどう貢献するかを理解しづらいという問題があった。そこで「スピード」という新たな軸を導入することで、オープンサイエンスの意義を再考してみたい。その背景にある仮説は、研究のスピードが極限まで高速化すると、情報流通もそれに追従して高速化せねばならないため、情報流通の妨げとなる「摩擦」を取り除く方向に進化して、結果的に科学研究がオープン化するというものである。こうした仮説を検証するのに最適なフィールドが、人工知能および機械学習の一分野である「ディープラーニング（深層学習）」である。この分野ではいったい何が起こっているのか？この例外的な研究分野から得られる教訓がどれほど一般化できるかは未知数であるが、これまでの動向を分析することで、オープンサイエンスの一つの可能性を先取りしてみたい。

北本　朝展

1997年東京大学工学系研究科電子工学専攻修了。博士（工学）。現在、国立情報学研究所コンテンツ科学研究系准教授、総合研究大学院大学情報学専攻准教授、情報・システム研究機構人文学オープンデータ共同利用センター準備室長。画像データの分析を中心に、データ駆動型サイエンスを人文科学や地球科学、防災等の分野で幅広く展開する。文化庁メディア芸術祭アート部門審査委員会推薦作品、山下記念研究賞などを受賞。オープンサイエンスの展開に向けた、オープン化や超学際的研究コラボレーションにも興味を持つ。

私の研究分野は情報学で、画像処理や画像データベースの研究をしていたのですが、その後、データ駆動型サイエンスに発展し、気象情報、地球環境情報、人文科学情報などの分野で研究を行うようになりました。この流れで、最近はオープンサイエンスの問題にも関わっています。

1.オープンサイエンスの背景

1-1.オープンサイエンスへの収束

オープンサイエンスとは、「オープン」という言葉をてこにして、サイエンス（研究）の方向を変えるものだと考えています。共通する部分は、今よりもよりオープンにしようという方向性です。いろいろな方向に、いろいろな運動があるのですが、それを「よりオープンに」という一語で束ねるとどういう世界が見えるかという話になると思います。

ただ、個々の活動で「オープンサイエンス」の意味は異なるので、単一の定義は困難です。よく私が言っているのは、オープンサイエンスは同床異夢であるということです。オープンにしたいという思いは共通しているのだけれど、どんなオープンを夢見ているかという認識を共有しないと、実は違う夢を見ていたとい

うことがよくあります。私は、図 1 にある、オープンアクセス、オープンデータ、データマネジメント、シチズンサイエンス（市民科学）、研究の再現性など、透明性・参加・協働・共有の視点で動いている多様な活動は、全て「オープンサイエンス」というキーワードでまとめられると考えています。

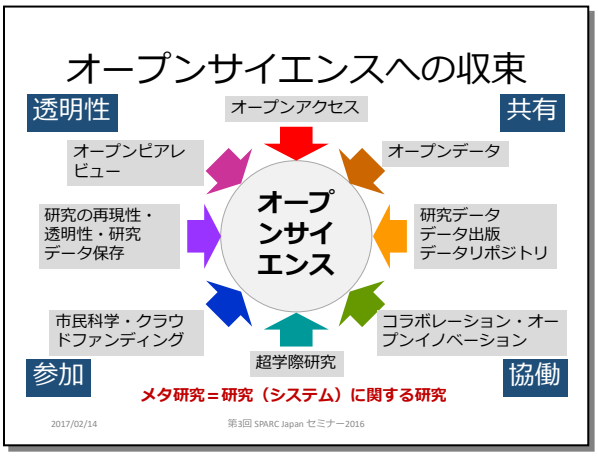
「オープン」には三つの側面があると考えています。一つ目は、他者が使えるということです（再利用）。外部の人が研究成果を自分の目的に再利用できます。これにはオープンデータやオープンアクセスなどが当てはまります。

二つ目は、他者が検証できるということです（透明性）。外部の人がエビデンスを検証し、正当性を判断できます。政府の政策を検証するオープンガバメントも同じような目的を持っていますが、他者の行った研究を再現・検証することも非常に重要なテーマになってきています。

三つ目は、他者を受け入れることです（参加）。研究のプロセスに他者を巻き込むことで、今まで得られなかったような成果を生みます。これにはオープンイノベーションや市民科学が当てはまります。

1-2.制度を分析する四つの視点

オープンサイエンスを分析するときには、法、規範、市場、アーキテクチャの四つの視点があると考えています。この四つの視点は、Lawrence Lessig が書いた“Code: And Other Laws of Cyberspace”（邦題『CODE ―インターネットの合法・違法・プライバシー』）という本で挙げられていて、Lessig は制度をどうやって変えていくかをこの四つの視点から考えることができると主張しています。



（図 1）

日本語で言うと、法は「しなければならない」、規範は「すべきである」、市場は「した方が利益がある」、アーキテクチャは「せざるを得ない」と言い換えられます。このような道具を使って制度を変えていく方法は、オープンサイエンスでも有効ではないかと考えます。

例えば、法によるオープン化は、資金提供機関が決めるルールを変えることによってオープン化を進めていくというものです。データ管理計画なども基本的にはそうで、ルール変更が重要な役割を果たします。例えば、資金提供機関が、どれだけオープン化したかで研究評価するというルールになれば、研究のやり方が変わっていきます。ただ、これは上からの押し付けで副作用も大きいので、本当にそれを適用していいのかという見極めは非常に重要です。ゆっくり慎重に進めていくことが求められます。

規範によるオープン化は、そもそもサイエンスはオープンであるべきであるという考え方です。これは分野によっては一般的な考え方です。例えば、地球科学では地球全体のデータがないと研究できないので、シェアしなければ研究が進まないという状況があります。こういう分野は、データ共有が不可欠なので、オープンな文化が発達したと言えます。

それだけではなく、世代の差もあります。上の世代より若い世代の方がもともとオープンな文化に生まれているので、共有に抵抗がありません。

ただ、こういう規範によるオープン化は、文化圏が違うと同じ文化を共有できないので、全分野で共有することはなかなか難しいです。

市場によるオープン化は、「した方が得だ」というものです。オープン化によって報酬が得られるのであれば、みんなオープン化するだろうということです。研究成果をオープン化すると例えば引用が増加するな

らば、オープン化した方が得だからオープン化しようという考えを持つ人も出てくるでしょう。ただ、これは報酬と裏腹に損失への不安があります。他者に成果を横取りされるのではないか、報酬は労力に見合うのか、こういう損失への不安があると、よほど報酬が確実でないとなかなかオープン化できなくなります。

アーキテクチャによるオープン化は、Lessig の本でも情報社会特有の方法であると述べられているのですが、何らかのプラットフォームの上で作業すると、必然的にオープンに誘導されてしまうというものです。例えば、何かツールを使っていて、ワンクリックでオープン化できる機能があれば、オープン化へと誘導されてしまうかもしれません。プラットフォーム上の見えないルールで、いつの間にか誘導されるというのがこの方法の特徴です。

また、苦痛の軽減という視点も重要です。オープン化は大変だけれど、有償のサービスを使えば簡単にできるとなると、そういうプラットフォームの上に乗ってオープン化を進める状況も当然起こり得ます。こういうものは良くも悪くも企業にとってはベンダーロックインのチャンスであり、そうした利益がオープン化を進めていくという未来もありえます。

このように四つのオープン化は方向が全く異なりますので、オープン化の話題ではどのオープン化に焦点を合わせているのかを明確にすることが有益ではないかと思っています。

1-3.研究データのオープン化

研究データの問題についても触れたいと思います。私は研究データを三つの種類に分けています。研究資源データ、論文付属データ、研究過程データです。

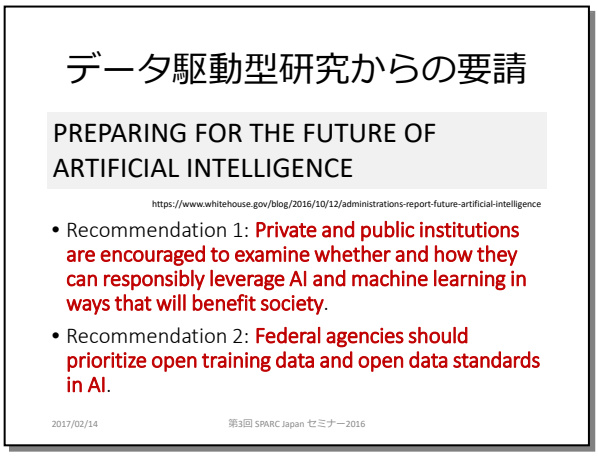
研究資源データは、研究の入力となるデータです。観測データや、評価用データセットなど、これを使って研究するというものです。これは再利用が目的ですから、再利用しやすいオープンデータになっているかという点が評価基準になります。

論文付属データは、研究の出力となるデータです。典型的なのは論文のエビデンスとなるデータです。これは、再利用も可能ではありますが、それよりも透明性、すなわち研究が再現できるかの方が重要ですので、再利用とは違った視点でのオープン化が求められます。

研究過程データは、研究の入力と出力の間で生み出されるデータです。例えば、日々の研究活動のエビデンスとなるようなデータです。このデータは透明性のために重要で、研究不正防止のためのデータ保存の義務化などもこれに関連しますが、基本的にオープン化が目的ではなく、クローズドに長期保存し、本当に必要があれば限られた人に対してオープンにするものです。

私が関心を抱いているのは、研究資源データをどうやって再利用のためにオープン化するかということです。このニーズは、データ駆動型研究が広まるにつれてますます高まっています。図 2 は、「Preparing for the Future of Artificial Intelligence」という、2016 年 10 月にホワイトハウスから出たものです。ホワイトハウスのウェブサイトは最近大幅に変わりましたので、今アクセスできるかは分かりません。ここでは、マシンラーニングや AI を発達させるには、データをオープンにすることで研究を活発化させることが重要であると強調されています。この AI とマシンラーニングが今後のオープンデータの一つの焦点になります。

データについて、考えるべきはライフサイクルです。ライフサイクルとは、データが誕生してからたどるプロセスを段階の遷移として表すモデルです。研究者は



（図 2）

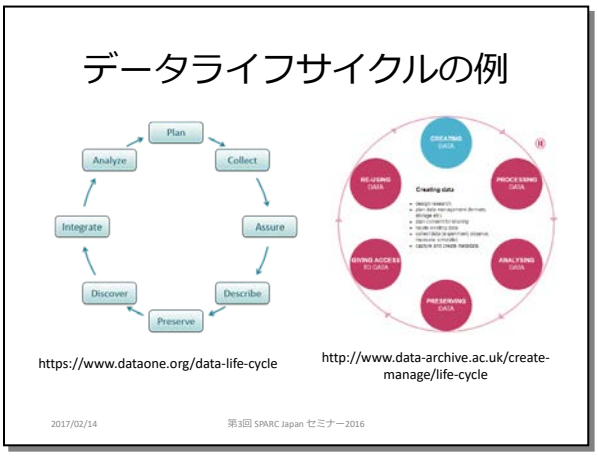
全体に関わりますが、そのうちの一部にはライブラリアンも関わります。例えば、データ管理計画などはライブラリアンが関わることが期待される段階です。

例として、Plan、Collect、Assure、Describe、Preserve、Discover、Integrate、Analyze という段階があって、これがぐるぐる回るのがデータライフサイクルです（図3)。個々の段階で何をするかということが重要になるのですが、左のサイクルと右のサイクルをご覧になって分かるように、データライフサイクルはそれぞれ違います。ここには示していませんがもっとたくさんの図がありますので、唯一の正解があるというよりは、プロジェクトに適したライフサイクルは何かを考え、その中で誰がどこの部分を担当するかを考えることが重要だと思います。

1-4.ライブラリアンの役割変化

こういう状況の中で、ライブラリアンも役割が変化していくことが考えられます。従来のライブラリアンは本を扱っていました。本はこれ以上加工されない最終の生成物であり、これを利用者（読者）のために整理していればよかったのです。しかし、データ管理計画など、データを扱わなければいけないとなると、データライブラリアンが必要だという話になります。

データは、本のような最終生成物ではなく、変わりうるものです。また、データにどんなメタデータを付けるのかという話になったとき、メタデータに決まったフォーマットがあればいいですが、ない場合は作者のところに行って、このデータをどのような目的でつくったのかインタビューしながらフォーマットを決めるプロセスが入ります。本で言えば、作者に作品の意図を聞いて、それを記述するというような、編集者や出版者に近い役割を果たさなければなりません。ですから、従来の最終生成物だけ扱っていればよかった時代と、役割もかなり変わってくるのです。



（図 3）

ライブラリアンは、データに標準的な価値を与えて使えるようにすることが目的ですが、一方でキュレーターは、データに付加価値を付けることが仕事です。データライブラリアンが整理した段階で、標準的なメタデータが付いて誰でも使えるようにはなりますが、そこにさらにデータキュレーターが関わり、このデータにはどんな価値があって、どういう目的に使えるかを売り込む。そんな人がいると、データの利活用も進むでしょう。このようにライブラリアンの役割変化が今後進んでいくと考えています。

現段階でデータライブラリアンに求められるスキルの一つ目は、データライフサイクルに関する知識と、DOI などの識別子を与えてデータをアクセス可能にすることです。

二つ目は、メタデータを付けないとデータにアクセスできないので、メタデータ規格や語彙に関する国際標準との相互運用性などに関する知識です。

三つ目は、データは置いておくだけでは使われないので、リポジトリに入れるなど、データの共有や利活用を促進する方法に関する基本的な動向の把握です。

四つ目に、上級編として、研究者コミュニティとの議論やエスノグラフィ的観察に基づき、データ管理システムを設計し構築する能力です。これができるデータライブラリアンは大変ありがたいと思います。

ただ、10 年後を考えると、こういう仕事のかなりの部分は AI に代替される可能性があります。今からデータライブラリアンになろうという方は、10 年後のキャリアを目指しますから、そのときまでに代替される知識を学んでしまったら、人生戦略上、非常にまずいです。ですから、10 年後どうなるかを予想しな

がらデータライブラリアンに必要なスキルを考えなければいけません。

メタデータの付与についても、例えば、アクセスが目的であれば、項目や品質の考え方も変わりますし、AI 向けメタデータであれば、人間が付与するかどうかも分かりません。ですから、AI との関わり方は必ず考慮すべきポイントになります。この場合、データライブラリアンには二つの道があると思います。一つは、AI を使いこなす側に立って、データライブラリアン兼データサイエンティストというスーパーマンになる道です。もう一つは、AI の下で働く道です。現実世界のダーティデータを AI が使いやすいクリーンデータに整理して提供する。スーパーマンになるのもいいですが、下で働くというのも必ずしも悪いことではありません。今の世の中も実際にはそうなっていて、Google のためにデータをきれいにしてあげるというのは、みんながやっていることです。ですから、データライブラリアンになるのであれば、10 年後も維持できる価値を考えなければいけないというのが前半部分のお話です。

2.オープンサイエンス再考

AI が出てきましたが、後半は AI の研究の世界で起きていることの話です。

2-1.オープン化の第四の軸「スピード」

これまでオープンサイエンスの意義は、「再利用できる」「透明性がある」「参加できる」といった三つを考えてきました。ただ、これだけだと研究者にはあまり実感がなく、これだけではオープン化が進まないという感じがありました。そこで、オープン化の意義として、「スピード」があるのではないかという点を今日はお話したいと思っています。研究のスピードが極限まで高速化すると、情報流通もそれに追従して高速化しなければいけないので、情報流通の妨げとなるような摩擦を取り除く方向に情報流通プラットフォームが進化し、結果的に科学研究がオープン化するという

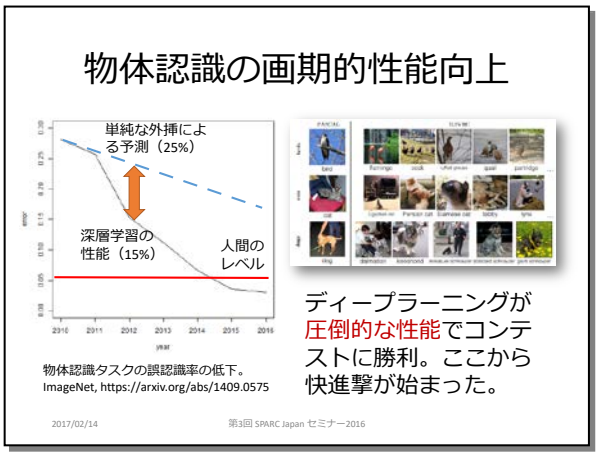
ような仮説です。その例をディープラーニング（深層学習）という分野の現状から検証したいと思います。

2-2.ディープラーニング登場

皆さん、ディープラーニングはお聞きになったことありますか。AI という言葉はニュースで日々お聞きになっていると思います。今、第 3 次人工知能ブームと人工知能学会会長も言っていますが、第 3 次人工知能ブームの中心的存在がこのディープラーニングで、技術的には「ニューラルネットワーク」と呼ばれるものです。ニューラルネットワークの中で特に層が多いものが「ディープ」と呼ばれます。原理は昔から知られていたのですが、ビッグデータとアルゴリズムの改良で画期的な性能向上があり、それが衝撃を与えて人工知能ブームにつながっています。

画期的な性能向上は、まず画像認識の分野から始まりました（図 4）。画像だけを与えて、それが何かを認識する物体認識タスクのコンペティション「ImageNet Large Scale Visual Recognition Challenge（ILSVRC）」が 2010 年から始まりました。緩やかなスピードで誤認識率が改善していくと多くの人が思っていたところ、2012 年には誤認識率がいきなり 25%から 15%まで下がり、ディープラーニングが圧倒的な性能で優勝しました。みんながその結果に驚き、その後は一斉にディープラーニングに走り、今は人間よりも性能が高くなっています。

ここからいろいろな分野にディープラーニングの適



（図 4）

用が始まったのですが、私が一番好きなのは囲碁のソフトウエアの AlphaGo です（図 5）。人間のチャンピオンも AlphaGo に負けました。過去のデータはもちろん学んでいますが、自己対戦によって戦略を深化させ、人間とは異なる戦略を用いて勝利したことが非常に衝撃的でした。

私も感動して、AlphaGo 観戦ツイートをまとめて Togetter ページをつくりました。私は囲碁というゲームがプレイできますので、この勝利がどう衝撃的かということも具体的に分かりましたし、この勝利は今までとはちょっと違うと感じました。この AlphaGo をつくっている会社が DeepMind です。この会社、今は Google の傘下にありますが、何のプロダクトも公表されていない時点で数百億円相当の金額で買収されたとして話題になりました。この会社が今、ディープラーニングでは重要なキープレイヤーになっています。

まずディープラーニングで面白いのは、オープンソース競争が始まっているという点です。ディープラーニングの最先端のライブラリを各社が競ってオープンソース化しています。Google の TensorFlow、Microsoft の CNTK、Facebook の Torch、他には日本でも Preferred Networks という会社が Chainer というライブラリを公開して頑張っています。

本当に最先端のライブラリは知的財産の塊なのに、それがどんどんオープンソース化されているのはなぜでしょうか。やはりオープンソースに人がおびき寄せられることを狙っているのでしょう。オープンソースのコミュニティができ、協力者が増えれば、そこから創出される価値も増えます。例えば、それを使ったライブラリや可視化ソフトなどいろいろなものが出てくれば、全部自社で開発しなくてもよくなります。まさにオープンイノベーションですが、そこでは競争すべき領域と協調すべき領域があります。差別化できる部分は守りつつ、外部の力を使えるところは使うというのが今の状況です。

（図 5）

では、Google がオープンソース化したからといって、みんなが Google と同じことができるかというと、それはありません。ソフトウエアは確かにオープンですが、例えば AlphaGo を実行するのに、数十億円のコンピューター資源が必要になるなら、誰でもできることではありません。また画像認識でも、数億枚の画像データを使わなければいけないとなると、そんなデータを持っているところは限られます。だから実際は同じことはできないのですが、少なくともソフトウエアについてはオープン化がどんどん進んでいます。

このようにオープンソース化が急速に進み、論文の実験コードなども再利用や再現性の観点からオープンソース化され、誰でも試せるようになりました。共通基盤データも、例えば先ほどの物体認識タスクの実験データがオープン化されているので、誰でもソフトウエアをつくれば性能が比較できます。さらに応用分野がどんどん広がっているので、他の分野の研究者や一般の人々も大挙参入しています。ですから、一刻も早く成果を公表しないと、自分の成果が古くなってしまうのではないか、という恐怖を皆さん持っています。スピードへの強烈な圧力がかかるという、他の分野にはない独自の状況が生まれているのです。

2-3.arXiv の利用

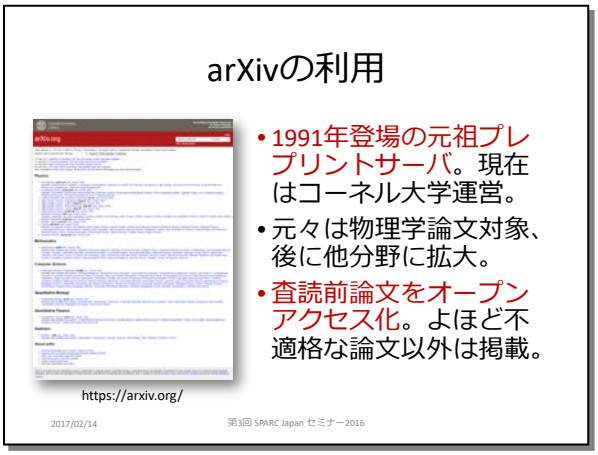
そこで登場するのが arXiv です（図 6)。なぜ今さら arXiv なのだと疑問を持たれる方もいらっしゃると思います。これは 1991 年登場の元祖プレプリントサーバーで、登場してからもう 25 年以上たっているからです。現在はコーネル大学が運営しています。実は今、この arXiv が大きく変わっているのです。

arXiv への投稿数はずっと上昇しています（図 7)。2010 年ぐらいから加速しているように見えます。分野別の投稿状況を見ると、もともと high energy physics が多かったのですが、割合としては減ってきました。math も多いですが、cs（コンピューターサイエンス）が近年になって割合を増やしてきています。これはディープラーニングの盛り上がりと同期した変化です。

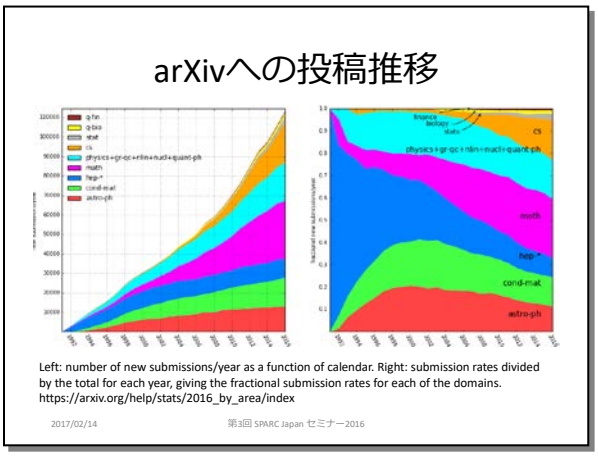
ディープラーニングの分野では、この arXiv が主戦場になってきています。AlphaGo をつくった DeepMind のサイトには Publications というページがあり、134 件（雑誌 10 件、arXiv 系 34 件、著名国際会議 90 件）の出版がリストになっています。雑誌は「Nature」などです。著名国際会議のリストもすごいのですが、そのような国際会議や「Nature」に混ざって、arXiv が同格でリストに並んでいます。別に彼らがそれらを区別している感じはありません。実際、DeepMind が出している論文は、arXiv で終わってしまうものもあります。つまりレビューは受けていない。arXiv にまず最新の成果をどんどん流す文化があり、その中で必要なものだけ査読を受けているような印象を受けます。

その arXiv がどう主戦場かを見てみましょう。例えば図 8 は 2016 年 10 月 10 日に arXiv に投稿された、東京大学の松尾豊先生らの論文です。この論文を、10 月 11 日に arXiv に投稿された論文が引用しています。この論文を書いたのが DeepMind の人なのです。DeepMind の人は arXiv を毎日毎日チェックし、1 日前の論文でも引用して投稿しています。まさに引用の爆速化が起こっています。ネットの世界から見れば、Twitter なら 1 分前でも引用できますので、1 日前の情報を引用すること自体は別に珍しくないと思います。ただ、従来の学術情報出版の基準から見ると、考えられないようなスピードで引用が進んでいると言えます。

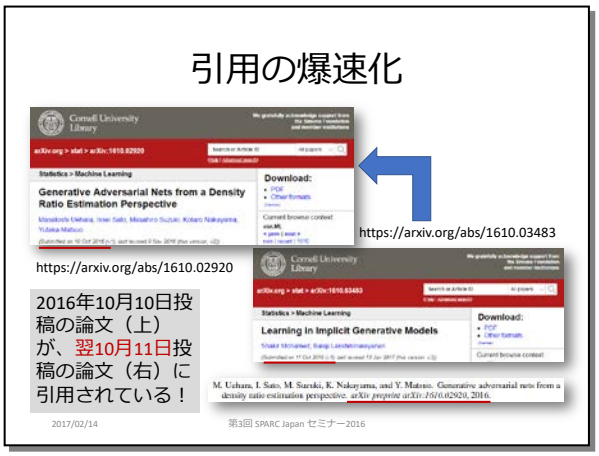
さらに、最近進展している分野なので、2016 年や 2015 年の論文が必然的に多くなります。そうなると、論文のほとんどが arXiv にあるので、arXiv を読んでい



（図 6）



（図 7）



（図 8）

れば基本的な動向は追えることになります。よほど古いものでない限り arXiv を見ていればいいということになれば、過去論文データベースの重要性が低下してきます。

皆さんにとって心配なのは、arXiv に投稿することが二重投稿にならないかという点でしょう（図 9）。二重投稿と arXiv がどう整理されているか。ディープラーニングの中心的な国際会議である Neural Information Processing Systems（NIPS）は、“Prior submissions on arXiv.org are permitted.”と明言しています。他の会議では、International Conference on Machine Learning（ICML）も“e.g., in arXiv”と書いています。一方、Computer Vision and Pattern Recognition（CVPR）という会議は、arXiv.org への投稿は出版ではないからオーケーと言っています。この「出版ではない」というのは微妙な線引きですが、レビューされてないから出版ではないという解釈を取っています。ただ NIPS では数語で済むところ、CVPR は数パラグラフを使って arXiv がなぜ出版でないかを説明しており、この長さの違いが文化の違いを反映していると考えられます。

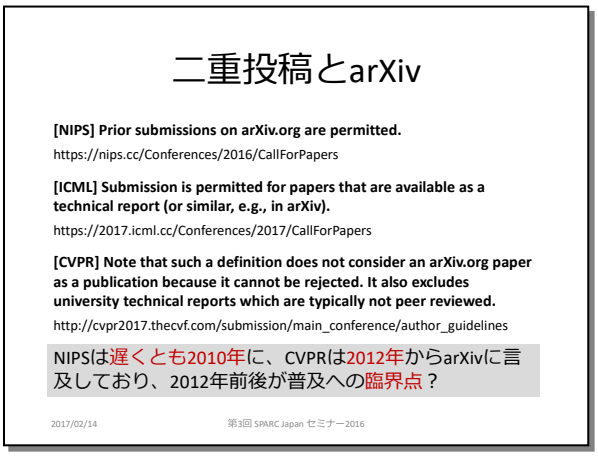
NIPS は遅くとも 2010 年にはこう書いているので、昔からそうであったと言えます。一方 CVPR は 2012 年からこのような注意書きが入りました。ですから、2012 年前後が普及への臨界点だったのではないかと思います。

さらに、図 10 は 2016 年 12 月 9 日に出たプレスリリースで、「なお、本研究成果は 2016 年 12 月 5 日に『Computing Research Repository』に公開されました」と書いてあります。この「Computing Research Repository」というのは arXiv です。つまり、arXiv に投稿したのでプレスリリースしますというものです。

このプレスリリースを行った大阪府立大学の黄瀬浩一教授および岩村雅一准教授は知り合いだったため、個人的にメールを出してプレスリリースの背景について幅広く意見交換させて頂きました。特に、通常のプレスリリースは査読済み論文が出版されるタイミングが多いのに対し、査読がない arXiv への投稿の段階でプレスリリースを行うことへの考え方をお聞きしました。すると、「慎重にすべきという意見もあったが、速報性を重視してこのタイミングでリリースすべきと個人的に判断した」とのことでした。

このような成果の迅速な公開は、実は Ingelfinger Rule と呼ばれるルールに反します。これまで査読論文に関連する内容の公表を遅らせてきたのは、プレプリントやプレスリリースを論文の公開前に行ってはいけないというルールのためでした。しかし速報性を重視する圧力が高まっており、そうも言っていられない状況になってきています。

ディープラーニングの会議で、International Conference on Learning Representations（ICLR）というものがあります。ここは面白くて、著者が arXiv に論文を投稿し、そのリンクをプログラム委員会に連絡することで、査読が始まるというプロセスを採用しています。つまり論文を最初に arXiv に投稿しなくてはいけない



（図 9）

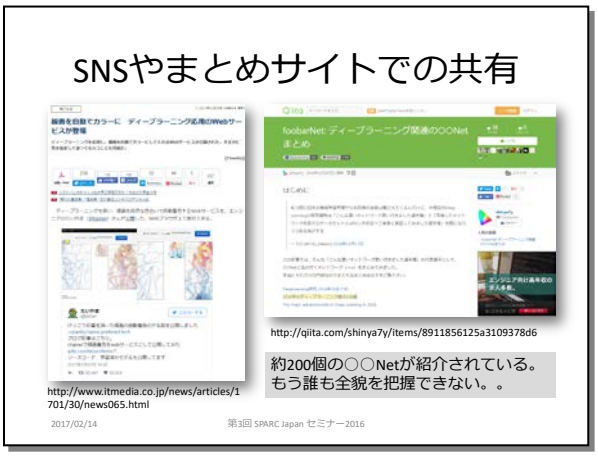
（図 10）

のです（図 11）。

最近はさらにオープンピアレビューに移行しています。OpenReview というシステムを使い、投稿した瞬間に論文が公開されるだけでなく、その後のレビューも公開されることになります。ICLR のサイトには、“The friction in our publication system is slowing the progress of our field.”、つまり friction（摩擦）があるようなシステムは問題であり、私たちは新しいやり方に変えるのだという宣言が書いてあります。私は本報告のテーマ「摩擦なき」に決めたとき、実はこの言葉を知らなかったのですが、やはり摩擦というアナロジーに行き着くのだなと感慨を覚えました。

このような急速な進歩は、SNS やまとめサイトでどんどん共有され、どんどん広まっていきます（図 12）。この「ディープラーニング関連の○○Net まとめ」には、○○Net というものが約 200 個紹介されていて、

（図 11）



（図 12）

もう誰も全貌を把握できないという状況です。このようなソーシャルな方法で、情報共有が加速度的に進んでいます。

これはある意味、市場によるオープン化であって、オープンソース、オープンデータにする方がお得だという動機からオープン化が進んでいきます。オープンアクセスリポジトリも、最初に「出版」して、その後に査読に行く世界になってきています。ソーシャル化で研究成果は SNS で迅速に共有され、公開サービスもいろいろな人が立ち上げ、日々どんどん進化している状況です。さらに、その公開サービスは、多くのユーザーがすぐに試すので、迅速性や頑健性も日々ネット上でチェックされます。そういう時代になってきました。

Michael Nielsen は、邦題『オープンサイエンス革命』という本を書いた人です（図 13）。実はこの英語のタイトルは“Reinventing Discovery: The New Era of Networked Science”です。「オープンサイエンス」とはタイトルでは言っていないのですが、オープンサイエンスに関することも論じています。

この本で描かれているオープンサイエンスの世界が、ディープラーニングでかなり実現してしまっているのではないかというのが私の感想です。いろいろな形でオープン化が進んで、知識が共有されています。面白いのは、Nielsen が 3 冊目に書いた本がディープラーニングの本なのです。両方ともネットワークという点で、2 冊の本には多くの共通する面があるようです。

（図 13）

法によるオープン化も着々と進んでいます（図 14）。ゲイツ財団は、財団の助成を受けた論文の、財団の OA 方針に対応していないトップジャーナルへの掲載は認めないこと、ウェルカム財団はプレプリントを助成金申請に記載してもよいことを表明しています。

3.摩擦なき学術情報流通へのシフト

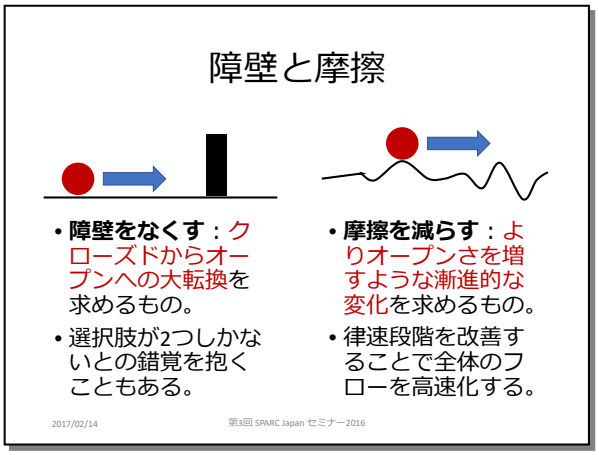
このような動きの中で起こっていることは、摩擦なき学術情報流通へのシフトだと思います。そこで考えなければいけないのは、我々が本当に欲しいものは何かということです。「人々が欲しいのは、ドリルではなく穴である」という有名な言葉があります。ドリルを買いに来た人がいて、ドリルがないときに、別のドリルを勧めるのではなく、そもそもなぜドリルが必要なのかを聞き、穴が必要なのだったら穴を開ける別の機械をお勧めすればよい。目的と手段は何かと考えることを忘れてはいけないという言葉です。

私は昨年の SPARC Japan セミナーで障壁について話しました。障壁をなくすのは壁を取るという考え方で、時としてクローズドからオープンへの大転換を求めることにもなりますが、摩擦を減らすというのは動きをスムーズにするということで、そうした漸進的な変化も必要ではないかと考えています（図 15）。

摩擦の代表例として査読があります。arXiv への投稿も、要は査読の問題をどうするかに集約されます。査読があることで公表が遅れる、不当な査読がある、査読者を見つけるのが大変で持続的でないといった問題があります。

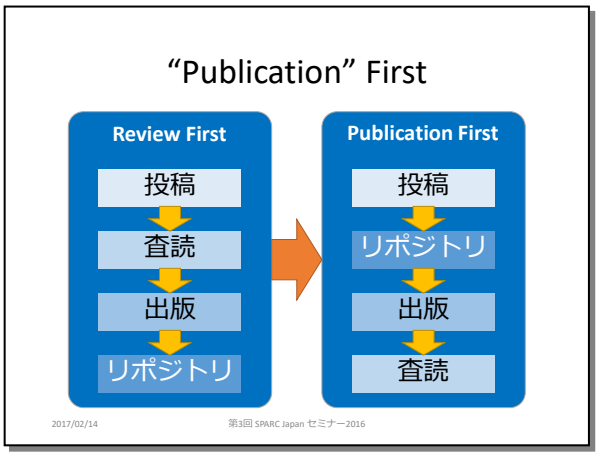
これまでの査読はゲートキーパー、つまり優れた論文だけを選別する門番の役割を果たしていました。これは確かにスペースが有限であれば有効な方法です。紙面や時間枠の制約があるので、門番があなたは通す、あなたは通さないと選ぶ方法が有効でした。しかし今はもうネットとなってスペースは無限なので、むしろ利用者の注意力が有限であるという点が制約になってきます。利用者にとって重要なものを、既に出ているものからどうやって拾ってくるか、ということが大きな課題となります。

そのとき、一つの予測として、選ぶという行為が付加価値サービスになる可能性があります。誰にとっても、自分が読むべき論文を選んでくれるとありがたい。これはある意味では、出版物のオープンアクセスが、著者支払いモデルから購読モデルに戻るとも言えます。



（図 15）

（図 14）



（図 16）

つまり、あなたのために論文を厳選してくれる付加価値サービスを購読するのです。これを私はSubscription model 2.0 と呼びたいと思います。ただこのような状況では、どのリポジトリにデポジットするかが流量勝負となるかもしれません。そうなるとarXiv のようなリポジトリに比べ、機関リポジトリなどは難しくなるかもしれません。いずれにしろPublication が最初、後で査読というモデルに移行する可能性が高いと思います（図 16）。

出版は人類の知の共有が目的で、評価は人類の知の継承が目的だと思います。ですから、オープンな出版と、多様な評価を組み合わせることで、出版と評価を分離できればいいのではないか。最終的に欲しいのは門番ではなく、この人の言うことは信用できるというブランドではないかと考えます。

4.まとめ

オープンサイエンスの重要な駆動力としてスピードがあるというのが本報告の主張です。ディープラーニングはスピードへの圧力が強烈な特殊事例かもしれませんが、決して孤立事例ではないと考えています。物事は臨界点を過ぎると急速な変化が起きるものであり、そういう時期を既に迎えた分野としてディープラーニングを捉えることもできます。本セミナーは図書館の参加者が多いと思いますが、この未来に図書館が果たせる役割は何か、次のアクションをぜひ考えていきましょう。