



■ 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 -

2015年 10月 21日(水) 国立情報学研究所 12F 会議室 参加者:100名

“研究データ”は、大学などの研究機関や学術情報流通のコミュニティにおいて、昨今取り上げられている特に重要なテーマの一つとなっている。これまで研究データは、研究過程において生成され、研究者自身やその所属する研究コミュニティが所有してきた。90年代以降の情報通信技術の革新は、とりわけインターネットとウェブ、その後の無線通信とモバイルにおける技術基盤のコモディティ化によって、様々な研究分野において大規模で多種多様な研究データを生み出し、共有と処理を容易にした。このような研究データをとりまく環境の変化は、科学的研究プロセスの変革を予測した“第4のパラダイム”の指摘につながっていく。これは、e-サイエンス、データ駆動科学、データ中心科学の意味するものと同じといえる。並行して、学術界において、特に論文を対象としたオープンアクセス政策が広範に普及、定着した。オープンという概念は研究データの文脈にも影響を与えることとなり、ここ数年、オープンデータやオープンサイエンスの名の下、データ共有が謳われるようになってきている。今、我々はこの新しい科学のパラダイムに向かって研究データ共有という新しい問題に直面しているのである。

本セミナーは、研究データに関するステークホルダー、すなわち、研究者、研究マネージャー、エンジニア、出版者、および政府関係組織、そして、研究支援に責任のある図書館員、さらにこの新しい科学のパラダイムに関心のあるすべての人が、このトピックの本質を理解し、研究支援環境として何が必要であるかを議論するきっかけを提供することを目的に開催された。

現各登壇者の講演要旨、配布資料等を含め詳細は、SPARC Japan のWEB サイトをご覧ください。

(<http://www.nii.ac.jp/sparc/event/2015/20151021.html>)

第1部 基調講演

“Open Data is not Enough”

Mark Parsons (Secretary General, RDA)



第2部 サイエンスと研究データ

講演

- ①「研究のバリア」を打破する研究基盤デザインと研究データ利活用
北本 朝展 (国立情報学研究所)
- ②オープンなプラットフォームが研究に与える影響を帰納的に考える
池田 大輔 (九州大学システム情報科学研究院)
- ③超高層物理学における研究データ共有
能勢 正仁 (京都大学大学院理学研究科)

パネルディスカッション

「研究データ共有は今後どうあるべきか？」

モデレーター: 武田 英明 (国立情報学研究所)

パネリスト: 第1部 第2部登壇者

武田氏: 研究データのクレジットを誰に与えるか？ またどうあるべきか？

Parsons 氏: データサイテーションは共有の動機になると考える。誰が責任を負うかという事が重要になる。RDA のプロジェクトにおいても色々な役割を与えている。データサイテーションの目的が変化してきている。

北本氏: 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う。しかし、再現可能性の方に意識が向かってしまうのは、現状のジャーナルの仕組みから発想しているからだろう。論文のサイテーションでクレジットは満たされていると考えると、データのサイテーションに意識が向かわない。とはいえ、データへのクレジットをどうやって行うかはやはり大きな問題である。それに関連して、各学会において「巨人の肩」賞を作ってはどうかという提案を以前に行ったことがある。論文だけではなく、いかにデータ共有やデータインフラを進めたかも含め、どれだけ「巨人の肩」を作ったかという基準から貢献度を評価できるのではないかな。

池田氏: マークさんの話でダイナミックデータサイテーション WG の話があった。3月のRDAで初めてダイナミックサイテーションの話聞いた。データセットの任意の部分をサイテーションできるようにすることで、データセット全体ではなく部分的に引用することができるので、その部分の対象者にクレジットを与えることができる。賛否はあると思うが、テクニカルな面からはそういう事が可能になると思う。

武田氏: サイテーションまで話が広がると少し大きくなる気がする。単純に誰をクレジットすべきか？

能勢氏: 私は、研究もデータの管理も職務としている。我々の研究分野ではデータシェアリングは当然のことになっており、いままでデータにクレジットをとすることは考えなかったが、データを取った人、データを管理している人、データを提供した人にはクレジットを与え、マトリックスのような形で業績を評価できるのが理想的ではないかな。

武田氏: これからデータがもっと広い範囲で活用されることを考えると、クレジットは明示したほうが良いのではないかな。

能勢氏: 私が DOI やデータなどに関わっているのは、データ取得者やデータ提供者へのクレジットや評価を計量化できるようにしていきたいから。データサイテーションの文化を進めていくことで、きちんとしたクレジットに反映することが出来るのでは。

武田氏: 今の技術ではクレジットの量が増えても問題なく表示できるので、クレジットをしっかりと行う事は、今後のデータシェアリングの方向性として良いのではないかなと思う。

武田氏: データをシェアするときに誰がどのように手助けしたらよいか？



能勢氏: ある程度、ドメインの研究者が関わらざるを得ない。内容がわからないと提供できない。研究者と大きなデータを扱えるデータに広い視野を持っている人(データキュレーター)が必要。

Parsons 氏: 能勢さんに同意。専門知識なしには行えない。

北本氏: キュレーターが重要なのは同意。ただし、インフラという建築・土木のアナロジーで考えれば、アーキテクトやシビルエンジニアのような多様な専門家も必要。そのような人たちと情報学の専門家との共同作業になるのでは。

池田氏: 論文をどうやってシェアするかという事は、arXiv.orgのようにドメインごとにデータベースを構築しているが、後発のところは苦戦している。本来それは研究者の仕事ではないのではないかと考える。データの保管は義務として大学が担って、キュレーションは専門家が行うような二段階になるとよいのでは。

安達氏(フロア): 日本では研究不正の話とオープンデータの話がごっちゃになっているが、欧米ではどうか？ また、個人情報の扱いについて伺いたい。

Parsons 氏: 二つの問題は分けるべきというのは同意。しかし、この問題は教育とか倫理的な境界線があるが深く関わっている。データサイテーションは全てを解決するわけではない。データは組織的に作成され集中化されるべき。

池田氏: 私は、機関に任せてもよいのではと思う。個人情報の方で合わせて言うと、オープンにするデータとさせてはいけなデータを選択は情報の力で何とかなる。一方で、オープンに科学を進めるというだけだとビジネスモデルとして成り立たないのでは？ 機関リポジトリの数がこれだけ増えたのは、機関に任せたからではないかな。

安達氏: データベースは継続的にメンテナンスしないと陳腐化してしまう。データキュレーターなどに任せれば良いというものでもない。成功しているデータセットは組織的に作られ、集中化され管理している。個々の研究者にデータメンテナンスの努力を求めるべきではない。日本ではそれを支える事ができる基盤があるのでは。

武田氏: データのシェアについては研究者だけに任せるべきではないというのは共通の認識になったのではないかな。答えは一つではないと思う。キュレーションのような仕事があるという認識と、コンピュータサイエンス力を積極的に借

りべきだということ、またコミュニティや分野ごとに明示的に活動を作っていくと言うようなことが一緒にならないとうまく回らないということかと思う。

武田氏: データを使うときのライセンスについて

Parsons 氏: 公的な資金で集めたデータは公共財であり、ある一定の倫理の下で CC0 (クリエイティブコモンズ 0) でよいのではないかとパブリックドメインとするべき。

北本氏: いろいろなチョイスがありうるが、無限にチョイスがあると扱いはづらひるので、数種類のバリエーションに整理していけないか。CC0 というオープンデータは、選択肢の分布の中でオープン化を極限まで進めたものと捉えられる。

池田氏: ライセンスよりもアクセス制御の仕組みが必要。

能勢氏: 自然科学データに限られるかもしれないが、公的資金で得たデータはオープンであるべき。ただし先取権という考え方は必要ではないか。

第 3 部 日本の研究データ基盤

講演

① 日本の研究データ基盤

加藤 斉史 (科学技術振興機構)

② 超高層大気研究のためのデータベース

～IUGONET プロジェクトの活動～

田中 良昌 (国立極地研究所)

③ 研究資源としてのデータセットの共同利用について

大山 敬三 (国立情報学研究所)

④ 事務職員のためのオープンデータ入門

星子 奈美 (九州大学附属図書館)

パネルディスカッション

「研究者の期待と研究データ共有への関わり」

モデレーター: 蔵川 圭 (国立情報学研究所)

パネリスト: 第 3 部登壇者

蔵川氏: 研究データ基盤をどう担うか? 歴史的な観点やニーズの変化などそれぞれの立場からお聞かせ願いたい。

加藤氏: 研究データの DOI 登録の中で明確になってきたことは、DOI を付与することによって最低限識別できるようにし、メタデータについてもシンプルで、ドメインによっては物足りないかもしれないというのが現状。一番近い将来の目標としては、ドメインごとの活用についてはそのドメインにお任せするが、論文と研究データのような異なる種別のもや別のドメイン同士を結びつけるという役割を果たしたいと考えている。

蔵川氏: JST の JaLC はビックインフラだと考えるが、ユーザーからのニーズに変化があって論文からデータにも DOI をというようになったと理解するがそのような声はあ

たか?

加藤氏: 一つはビックインフラストラクチャというお話があったが、日本から出て行くものを逃がさないで集めよう思想で JaLC の一つの基盤として位置づけている。もう一つは、論文から研究データへという中で、識別してそれを評価に使いたい、何処で引用されているかきちんと把握したいというようなニーズが届いているので、DOI 付与によってそのあたりを定量的に評価できるように取り組んでいきたい。

田中氏: 昔は自分でとったデータを解析すれば論文を書けたが、現在ではとにかく一つの現象に対して色々なデータを使って検証をした方がより成果が出るというスタイルになっている。そういった意味で、IUGONET のニーズは大きくなってきている。

大山氏: 一つはデータベースの規模と精度の変化。以前は限られた条件下での小規模データを丁寧に作る事を意識していた。現在は統計的な手法が発達したこともあり、大規模な生データを切り口を変えて分析する事で何かを生み出す手法が主流となった。もう一つは、従来は狭い分野の技術の研究で閉じていたが、最近はメディアをまたがる研究や人や社会との関わりに関する研究テーマが大きくなってきた。情報学では、人間や社会が研究対象になる。星子氏: 図書館へのデータに関する相談は、URA や事務部門から受けたことはあるが、研究者のニーズについては現時点では把握できていない。

蔵川氏: 研究データのオープンを進める上で実務上何かクリアにしなければならない事や注意点等はないか?

加藤氏: 一つはユースケースが少ないのでメタデータのデータセットが適正なのかの検証が進んでいないということ、もう一つは全体の流れが確立しつつあるので、スムーズな展開を心がけていきたい。

田中氏: IUGONET では、ライセンスやクレジットについては気にしている。IUGONET 自体でデータポリシーを制定しているわけではなく、各データのポリシーは各 PI が決めている。現在は利用者のモラルに依存する部分が多く定量的に把握することが出来ていないので、そのあたりはプレッシャーになっており、解決したい。

大山氏: データを提供してもらう時に条件を慎重に詰めるようにしている。



星子氏: ディスカバリーサービスの導入により、貴重書等の画像データへのニーズが高まっており、データ管理の

重要性を実感している状況である。

-----参加者から-----

(大学/図書館関係)

- ・研究分野によって温度差があることが実際どういうデータを扱っているかを聞いてよくわかった。
- ・全てを理解できたというよりも、現在の傾向や状況を知ることができました。また研究者の方の世界におけるデータについての考え方を知ることができたことがよかったです。

(大学/大学・教育関係)

- ・データ作成者のクレジット、費用負担など、重要な問題の存在を再認識した。

(大学/研究者)

- ・来年の RDA プレナリーに向けての情報が得られました。

(その他/図書館関係)

- ・分野によってコンテキストが異なり、議論の方向性を理解するのに努力が必要でした。
- ・オープンデータ、オープンサイエンスについて研究者の視点からの話を聞きたかったので、大変ありがたかったです。

(その他/大学・研究関係)

- ・オープンデータの実態、実例について現状を把握することができた。

(その他/その他)

- ・RDA についてよくわかった。構成、他テーマもよかったです。

-----企画後記-----

😊 機関リポジトリのムーブメントが定常状態に入って、ここ3年ぐらいの間で学術情報環境に関わって何か取り上げるべきことと言ったら研究データぐらいだろうと思っていたところ、SPARC Japan の企画ワーキングの依頼を頂いた。著者の名寄せを中心とした電子図書館を研究テーマとしていた身としては、研究データというのは今でも遠い存在である。サイエンスを対象とした研究データは、研究で取り扱うモデルの観測変数や潜在変数の数値列で、そのモデルを理解しなければ意味を理解することは不可能である。少しは対象データに関するモデルを理解しようと、分野の教科書を漁ってみるが、それを理解するために今度は数学の復習へとさらに深みにはまる。文献のメタデータはあくまで本や記事のモデルについての記述であって、研究データのメタデータは分野のモデルについての記述である。図書館員が文献リポジトリのメタデータを難なく取り扱うことができるのは、それが本や文献に関するものであり、図書館員はその専門家だからである。研究データも分野の研究者でない別の専門家が扱えるようにと、たとえばパッケージのように扱われる慣習ははたして今後訪れるだろうか。

蔵川 圭 (国立情報学研究所)

😊 本セミナーに企画および発表者として携わったことは、図書館の日常業務で具体的な実務へ落とし込む段階には至っていない、研究データ管理という大きな課題を考える機会となりました。Mark Parsons 氏の基調講演では、RDA の活動内容や今後の方向性について興味深いお話をいただきました。また、研究者の方々のご発表では、多様なデータとその管理・共有の実例を分かりやすく教えていただき、研究データをより身近なものとして捉えることができるようになりました。多くの皆様にセミナーへのご協力、ご参加をいただきましたこと、御礼申し上げます。

星子 奈美 (九州大学附属図書館)

😊 研究者の方々と共同で企画するセミナーということで、とても刺激になりました。また、研究者から見たオープンデータ、オープンサイエンスに関する図書館職員への期待は現状あまり高くないと感じました。しかし、図書館職員はこのテーマに積極的に参画し、研究支援に関わる図書館の立ち位置を確立して行かなければならないのではないのでしょうか。そんなきっかけとなるセミナーになったと思います。

梶原 茂寿 (北海道大学附属図書館)

OPEN  International
ACCESS WEEK

October 19-25, 2015

openaccessweek.org