

第2回 SPARC Japan セミナー2017

「プレプリントとオープンアクセス」

全体議論



- 坊農 秀雅** (情報・システム研究機構 ライフサイエンス統合データベースセンター)
引原 隆士 (京都大学図書館機構長/arXiv.org 会員コンソーシアム代表)
生長幸之助 (東京大学大学院薬学系研究科/化学ポータルサイト Chem-Station 副代表)
小野 浩雅 (情報・システム研究機構 ライフサイエンス統合データベースセンター)

●**坊農** 全体議論を始めたいと思います。質問し足りなかった方はぜひお話ししてください。多岐にわたってお話しいただいたので、個別の先生へのお話などもあると思います。

●**フロア 1** 産業衛生学会のもので、私はジャーナルの編集をしているのですが、海外の医学系ジャーナルでのプレプリントサーバについて、どれくらい広まっているのかなど、状況をどなたかご存じですか。以前担当していた医学系のジャーナルで、中国からの投稿にクロスチェックをかけたところ全部一緒になって、結局プレプリントサーバに入っていたということがありました。そのジャーナルはプレプリントサーバへの投稿を認めないとしていたのですが、小野先生の資料を見ると BMJ は認めているし、以前に Elsevier の方から Elsevier ではほとんどが認めていると聞いたことがあります。海外では実際にはどんな感じかお分かりになりますか。

●**小野** 私も医学系ではないので残念ながら具体的な話は分からないのですが、Wikipedia のリストを見ると、incompatible なものは割と医学系、独立系の雑誌が多かったです。医学系が悪いわけではないのですが、保守的な感じがあるのかなど。先ほどバイオインフォマティクスの方が先に進んでいるという話がありましたが、そういう意味では、中身のセンシティブさに時

代が追い付いてきていないところがあるのではないかとこの印象を持っています。

●**フロア 2** 慶應義塾大学の市古 (SPARC Japan 運営委員) と申します。アメリカ国立衛生研究所 (NIH) がプレプリントに対する方針を今年出したことによって、PrePubMed という動きになってきたのだと私は理解しています。今、保守的だという話があったので少し付け加えますが、確かに慶應義塾大学の中でどの程度 bioRxiv に投稿されているかを見たところ、臨床医学の積極的な投稿は目立たなくて、理工と医学系に関わる方、それから SFC で情報科学に関わる方と医学研究を一緒にやっている方の投稿が見受けられました。

●**坊農** 私自身も少し医学系をやっているのと、BMC Genomics でアソシエイトエディターをしていたこともあるので、それで見ていると、プレプリントに出したものは、リファレンスとして挙げられることはよくあって、普通の論文のように扱えるという反面、基本的には査読を経っていないので信用できないという扱いになっていたと思います。

ただ、小野さんの出してくれた Wikipedia の表も動的に変わって行って、incompatible だったジャーナル誌が急に compatible になっていくという傾向があります。ざっくり探すときの参考情報としては Wikipedia はいいと思いますが、情報のソースが最新ではないの

で、実際のジャーナルの判断をそれぞれのところへ見に行き、正しいかどうかをよく見た方がいいと思います。ただ、劇的に変わっていつている現状があると思います。

医学系で全く同じものを出してくるというのは、論文の本数を稼ごうとしている人がやりがちで、そういうものは情報技術で対抗して取り締まるのが大事だと、私自身やっています。それをやり過ぎで逆に、前の論文と同じ手法を使った論文がはじかれたということが最近ありました。自分が前に出したものとほとんど同じということでプログラムのチェックではじかれてしまって、他のジャーナルに出さざるを得ないという事態になったこともあります。

プレプリントで出したものと全く同じなど、おかしいものはおかしいとチェックできるようになってきたという点では、良い時代になってきたのだと思います。

●フロア 3 国立極地研究所の図書館員です。引原先生にお伺いします。コーネル大学で、arXiv.org のエフォート管理をしながら、ライブラリの中で実際に回しているということですが、arXiv.org は今後、拡大路線で行こうとしていて受け取りました。そのときに、ライブラリで今後どのような運用体制を取っていくのでしょうか。また、その視点から、SSRN では商業ベースに移行したという話がありましたが、そちらの可能性についてはどう考えられているのでしょうか。

●引原 それは非常に微妙なところで、Gordon さんの講演で言及のあった、SSRN が移行した理由を Twitter で拝見したのですが、確かに経営的な問題というか、人のお金の問題は大きいです。arXiv.org の場合はタスクが増えていくのに合わせて、エフォートがあるときは増やしたり、あるときは減らしたりしています。今、従来のシステムから新しいシステムに移るときに増員しています。ファンディングに対して増員を要求してお金を付けてもらうという形を取っていて、それは経常的ではありません。いずれそれは減るので、

その分が今度はチェックする人たちの雇用の方に回るのであるというのが、経営状態を見ていても分かりません。

かなりの部分はオートマチックになっていますが、iThenticate に引っ掛かったものなどのケアは丁寧にやっているのですが、それが増えてきたときにどうなるかというのは確かに問題かもしれませんが、現状は回っていると思います。

●フロア 4 NISTEP の林 (SPARC Japan 運営委員) です。先ほど引原先生にできなかった質問を投げたいと思います。リソースシェアという意味で、機関リポジトリがせつかく 744、今はもっと増えているはずですが、あるので、アーカイブの受け先を日本の場合は機関リポジトリにしようというアイデアはどうなのでしょうか。つまり、会費を払うだけではなく、ストレージの置き場やナビゲーションのところのコントリビューションで組む、商業主義的なものへの対抗として考えられるアイデアとしてそういうものが思い浮かんだのですが、個人的な感触で構わないのですが、そういう方向性をまず示さなければいけないのではないかと、ということが 1 点と、機関リポジトリをそもそも具体的に使うことに関してはいかがでしょうか。

●引原 今年の arXiv.org の Member Advisory Board (MAB) で、われわれ日本から、機関リポジトリについてディスカッションしてほしいと要求を出しました。会議の日の昼休み (13~14 時半) に昼食を取りながらディスカッションしたところ、リポジトリ自体についての考え方がヨーロッパと日本とアメリカでは違うということで、日本の場合はリポジトリがプレプリントサーバ的なところがあるのでその可能性はかなりあるのですが、やはりなじまないという議論がありました。

それは機関がどこまで全体のデータベースのサポートにタスクを出せるかという問題があると思うのです。今はコントロールできるミラーサイトと Amazon のク

クラウドでやりたいというのが正直なところで、京都大学の基礎物理学研究所にあったサーバがなくなったのもそれが一つの理由です。今後はクラウドになっていったときに、日本でクラウドをどこかが管理するということは当然あると思います。けれども、現状のシステム開発をしている段階では、やはり全体をマネジメントしたいという方が強いようです。

●フロア 2 少し下世話な話なのですが、arXiv.org は経営的には問題ないのでしょうか。慶應義塾大学の例なのですが、arXiv.org に協力するかどうかというお金の分担の話が来て、誰がどう払うかとなったときに、図書館員が価値を示していくのがなかなか難しい状況にあります。私としては arXiv.org に協力することはいいことだろうとは考えているのですが、お金の面が大丈夫であれば学内でもう少しいろいろ議論できます。将来的にこれが成り立たなくなることがあるとすれば、図書館としては何をどうすることができるのかということをお話いただければと思います。

●引原 今日はお金の議論は下世話になると思って外していたのです。ここ 2~3 年は開発の経費のこともあって増えていますが、それはいずれ落ち着きます。ただ、2018~2022 年のモデルとして、どれだけのユーザー負担を考えるかを議論した結果、利用率に応じた負担を頼むということで今まで来ています。

現状では日本に負担が来ていると思えるのですが、かなり考慮を要求しました。日本の場合はファンドを直接アメリカに送ることは、内閣府や文科省に聞いてみてもやはり難しいということがあって、国立情報学研究所 (NII) などが取りまとめてコンソーシアムで払うという形態しかないだろうというのが向こうとの認識にありました。

それは日本だけではなくて、アジアの多くの国が多分そうだろうということを言っています。もし仮に日本のファンドを、S が付いている会社などに出していただいたら、これは大きな違いがあるかもしれません。

私の認識としては、経費を出すということと同時に、今まで arXiv.org の単なるユーザーであった日本が発言権を持って、いろいろな提案をしてきています。昨年からはかなりして、それで存在感を示すことができています。arXiv.org を中心として、先々プレプリントサーバが統合していく動きになると思うのです。そのときに発言権をかなりキープできるのではないかと思っているのも、ここは単にユーザーとしてのお金ということだけではなくて、次のものを得るための投資として考えた方がいいのではないかと考えています。

ですから余計に、例えば「ビューワーをもっとこうしてください」と発言したっていいですし、「メタデータもこういうのはどうですか」と提案してもいいわけですが。今までは単なるユーザーだったからできなかったもので、やはりそういう方向に転換した方がいいのではないかと私は思っています。

●フロア 5 先端医療振興財団のものです。私は医療系に携わっていて、各学会との連携もあるのですが、学会同士のデータシェアリングなどの課題があります。ただ、それぞれの学会でいろいろな考え方があるので難しいという状況だと思います。それで、日本で独自にそういうデータベースを作成されるお考えがあるかどうか、それをハーモナイズして海外に展開する構想はあるか、もしないとすればどういうところが難しいか、どの先生でも結構ですので教えていただけたらと思います。

●坊農 私たちライフサイエンス統合データベースセンターは、予算規模的に基本的にはライフサイエンス分野で多くの方に使ってもらえるデータベースの維持・管理ぐらいしかできません。ある分野のデータベースをつくることに関しては基本、関与できないのが現状です。なぜかという、お金というよりも、まず人が足りないからです。そういうことをやったらいいのにと皆さんから伺い、確かにそのとおりだと思うのですが、分野オリエンテッドなデータベースはその分

野でやりたい方に始めてもらって頑張ってもらって、困ったことがあれば、われわれが技術的にお手伝いするというのが現状だと思います。

●フロア 5 それをもし商業ベースでどこか出版社にお願いするとしたら、非常に高額になって、実現が不可能になってきます。今ある海外のものを使うということに関して、それぞれの学会の先生の考え方があって、ここは使うがここは使わないということで、結局ばらばらのデータベースに入ってしまった、それらが結び付かないことが今後問題になってくると思うのですが、そのあたりはいかがでしょうか。

●引原 データベースに関して、海外の出版社の狙いはまずデータを取得することで、次はクロス的に AI などを使ってマイニングしていくことなので、日本の学会がこぞって海外のデータベースにデータを置くことは非常に危険です。ジャーナルで負けたところを、最後データで維持しているのですが、それがぎりぎりの状態にあるということを学会の先生方に理解していただかないといけないのかもしれないかもしれません。

それをどうしたらいいかというときに、クラウドでも、日本の NII を中心とした形でもいいですから、機関がデータサーバをパブリックにサポートするシステムを立ち上げなければいけないのは明らかだと思います。それをやらずに、学会が学会の維持のためにやるということは危険ではないかと思われま。アーカイブが商業出版に吸収されてしまったという例がありますが、それはアーカイブを維持するためという非常に矮小な議論の結果だと思います。そこは学会が、データのユーザーとしての研究者や企業などと協議して決めていかなければいけないのではないかと考えていただきます。

●坊農 NBDC から、公開データベースではないものの話はないですか。紹介していただけないですか。

●フロア 6 NBDC で提供しているものとしては、私が担当している生命科学系データベースアーカイブというのがあります。これは基本的には全部公開していて、あまり突っ込んだ加工はできないのですが、ある程度再利用できるような形に持っていくことはできると思います。それとは別に、エンバーゴのようなものですが、将来公開することを前提として、プロジェクト期間内はプロジェクトの中だけで共有するというサービスも展開しているので、個別に相談していただければお力になれるかと思います。ただ、基本的には、データをこういうふうにしたいというのはそれぞれ研究コミュニティの方に主導していただきたいと思いません。

●フロア 7 これはデータの保存ということであれば、また別の面があります。2017 年 11 月 6 日 (月)、国会図書館と研究データ利活用協議会 (RDUF) で、データサイエンスを巡る世界の動向におけるデータのアーカイブといった問題について議論するので、そこで続きの議論ができるかと思います。

(<https://japanlinkcenter.org/rduf/events/index.html#s001>)

活用するデータにはあまり適切ではありませんが、データを保存するだけであれば国会図書館も使えるかもしれません。要するに活用するデータはそういうのはあまり適切ではないと思うのです。データとしてとにかく保存したいというのであれば、またアーカイブという視点では別の策があるのではないかと思います。

●フロア 4 今日しなければいけない大事な議論は、どうやってプレプリントが各研究者コミュニティに受け入れられるかということです。なぜ最初に物理分野でうまくいったのか、そのうまくいった要素は他に展開できるのか。私の分野の化学で、生長さんの話を聞いて一つ作業仮説を立てるならば、そのコミュニティのトップジャーナルのエディターが受容するかどうかということがあり得る気がします。そのあたりはパネルの皆さんの領域でどうですか。トープジャーナル

のエディターが「プレプリント、オーケー」と言うと、傘下のジャーナルみんな横並びになりそうな気がして、それが物理で起きていて、化学ではまだ起きていないということだと思うのですが、いかがでしょうか。

●**生長** 全くおっしゃるとおりだと私も思います。端的に言うと、日本化学会がプレプリント投稿を受け付けて、化学会全体でこの流れを進めようと言って、評価基準をみんなて修正していこうという流れをつくれれば、時間がかかるかもしれませんが、徐々にそうなっていくのではないかと思います。その過程で、プレプリントは有益なものであるということがコミュニティに受け入れられていけばそれで問題はないかと思えます。でも、誰がそれをやるのだというのが難しいところだとみえています。

●**フロア 4** お隣に物理学会の関係者もいらっしやいますよ。

●**フロア 8** 物理の中でも、先頭を走ったのはやはり高エネルギー物理なのです。ビッグサイエンスで、データは共有物という意識もあったと思いますし、何とんでもコミュニティとしては小さいので、そこでは合意が得やすかったということがあります。次第に物性物理にも広がっていったのですが、私の見るところ、その契機は、高温超伝導ではなかったかと思うのです。あのときに研究のスピードが爆発的で、査読付きの論文なんか待ってられないと、みんなプレプリントで勝負したのです。それで一気に広がったというのが私の個人的な印象です。

●**フロア 4** これは NII の北本さんが以前ディープラーニングで紹介した例とまさに同じです。研究が早過ぎて待ってられない。

●**坊農** 北本さん、何かコメントはありますか。

●**フロア 9** NII の北本です。昨年の第 3 回 SPARC Japan セミナーで話したのは、ディープラーニングで研究の速度が爆発的に速くなって、論文の査読を待ってられないからみんなプレプリントに流れたということです。高温超伝導と全く同じというのは今日初めて知ったのですが、それは非常に大きな要因かと思えます。

●**坊農** どうもありがとうございます。昨年出ていない人にもお分かりいただけたかと思えます。

今の、化学ではこれから起こるとか、そういう話を聞いていて、われわれ生命科学の分野はそういう意味では不思議な立場だと思いました。ビッグラボなどが何かやってというところが日本ではあまり起きていなくて、海外で分野の最先端を行っているところが最新の機器を使ったデータなどをプレプリントで出しているということは、たまに聞きます。日本では生命科学関係ではそういうことがあまり起きていない感じがします。化学などは、まだ学会がきちんとオーガナイズする力を持っているのではないかと思えました。やはり日本化学会に頑張ってもらって、プレプリントをやりましょうと言うのが一番いいのでしょうか。

●**生長** プレプリントの普及の順番として、先に物理があって、次に生物が来て、化学に行くという順番は分からなくもないです。実験に時間がかかり過ぎて、査読の速度とキャリアプランに齟齬が起きやすいのが生物学の世界だと思っています。生物学の研究を少ししているので、最近何となく感覚が分かるようになってきました。先ほど言われていた高温超伝導の話は、研究のスピードの方があまりに速過ぎてということですが、評価やキャリアパスなどの人間的なところとサイエンスのスピードのギャップが出てきた瞬間に、プレプリントが普及していくのではないかと思います。

化学の世界では、今までは評価やキャリアパスなどのサイクルと研究サイクルがちょうど良くかみ合っていたので、プレプリントの必要性がひょっとしたらな

かったのかもしれませんが。けれども、化学の世界でも生物との融合的な研究をかなりたくさんの方がやるようになった結果として、プレプリントが必要という話になったという気がしています。それをトップの方が理解してそれをやろうとするかなとは思っています。

●フロア 7 北本さんの補足ですが、コンピュータサイエンスのプレプリントが普及したのには、もともとジャーナルに権威がないという事情もあります。カンファレンスがベースで、カンファレンスにも権威があるもの、ないものがあって雑多な世界です。しかも昔からテクニカルレポートも大学ごとに出すという文化もあって、そういうことに関してもともと無頓着なところがあったのです。今までも、例えばページランクの詳細な論文はスタンフォードのテクニカルレポートしかないということがあったぐらいで、それがたまたまアーカイブに載るようになったのが割と最近です。コンピュータサイエンスの場合は逆にジャーナルにさほど権威がなかったのでそういう方向に流れてもそんなに摩擦はなかったと私は理解しています。

●坊農 確かに考えてみると、私が生命科学でプレプリントでと例に出したのも次世代シーケンサーで、より長く早く安く読もうという競争があって、早く論文を出したがついている分野です。小野さんに出していただいた例も、シングルセル、1細胞の技術という、今競争の非常に激しいところです。生命科学の bioRxiv を見ていると、そういうところしか出ていないような印象を受けています。厳しい競争がプレプリントを促すというまとめでよろしいのでしょうか。そういう競争が化学でもいろいろありそうですが、化学だと次は何がいきそうですか。

●生長 私が属している有機合成は研究サイクルが短い分野なので、タッチの差は当然あるのですが、プレプリントで評価される、されないというほどの差ではない気がしています。どちらかというと化学はもっと

研究スパンの長いもので評価が決まってくるようなものが増えてきたと思います。

●坊農 それでは時間となりましたので、全体議論を終了いたします。先生方、どうもありがとうございました。