第 2 回 SPARC Japan セミナー2018

「オープンサイエンス時代のクオリティコントロールを見通す」

学術コミュニケーションのエコシステムの今後 ～arXiv の現状から考える～

武田 英明
（国立情報学研究所）

講演要旨

インターネットの利用とオープンアクセス活動の浸透により、学術コミュニケーションのエコシステムが大きく変わってきている。伝統的なジャーナル投稿掲載という方法以外にも様々な方法が生まれている。現在の学術コミュニケーションのエコシステムを概観し、その中での preprint server の立ち位置を確認する。特に代表的 preprint server である arXiv を取り上げ、現状と課題を確認する。

武田 英明
SPARC Japan運営委員会委員長。
http://www.nii.ac.jp/faculty/informatics/takeda_hideaki/

本日はいつもの話と少し違って arXiv、プレプリントを中心に話をしたいと思います。なぜ私がプレプリントの話をするかというと、10 月 2 日に arXiv の Member Advisory Board（MAB）というミーティングに行ってきたからです（図 1）。arXiv というのはメンバー制度になっていて、大学等がメンバーです。日本の場合は日本コンソーシアムをつくっていて、幾つかの大学が入っています。そのコンソーシアムの代表が毎回 MAB に出ています。コンソーシアムの代表は、本当は京都大学の引原隆士先生なのですが、代表が代わる時期で、コンソーシアム代表代理という形で私が MAB に参加してきました。

（図 1）

MAB は 1 年に 1 回行われます。Member Advisory Board と、もう一つ Scientific Advisory Board（SAB）というボードがあり、この二つで arXiv はマネジメントされています。MAB の参加者の多くは図書館で、一部は Jisc などの研究助成機関です。研究者代表の方もいるのですが、今回は欠席でした。場所はコーネル大

学です。私は初めてコーネル大学に行きましたが、きれいで広大なキャンパスでした。

arXiv の現状と課題

現在、arXiv には約 140 万論文があり、1 日に約 600 件、新しい投稿が来ます。1 秒に 7 論文がダウンロードされます。モデレーターという、論文のチェックをする人が 162 人います。プレプリントは論文査読はしません。ただし、分野ごとにモデレーターのグループがあって、そのモデレーターが、その分野の論文として適切かどうかを評価します。それに通れば 1～2 日で掲載されるようになっています。

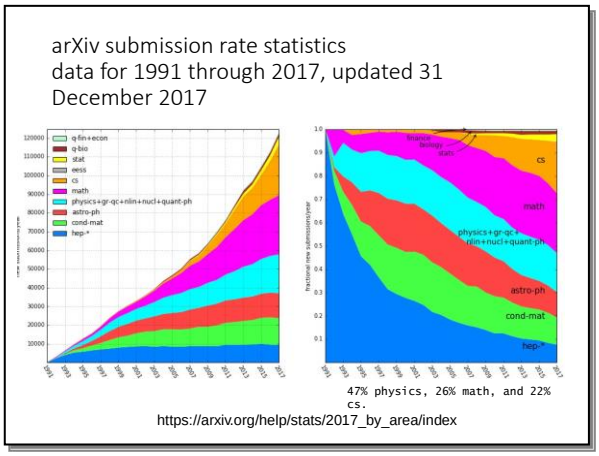
最近の傾向を見ると、2012 年のころはサブミッションが年間 84,000 件でしたが、2017 年は約 120,000 件で、5 割増ぐらいのペースで増えていることが分かります（図 2)。ポイントは、この増えた量がどこから来ているかというところです。

arXiv はできて 30 年たっていますが、図 3 の統計は 1991 年から始まっています。下から、青色が高エネルギー物理学、緑色が固体物理学、赤色が天体物理学です。これに水色を加えた下から四つが物理系です。さらに、紫が数学、オレンジがコンピュータサイエンスです。小さく見えるのが最近入ってきた統計など、広い意味での経済系です。

これを見て分かるように、元々の始まりである高エネルギー物理学はほとんど変わらない、減りもしなければ大して増えもしない状況です。それ以外の物理はやや増えています。数学も比較的増えています。近年増加率が大きいのは、何といってもコンピュータサイエンスです。2017 年は投稿の 22%がコンピュータサイエンスとなり、コンピュータサイエンスが顕著に増加しています。高エネルギー物理学は、相対的には arXiv の中のウエートとしては顕著に下がってきているということが分かります。とはいえ、物理全体で、5 割弱ぐらいは占めているという状況にあります。

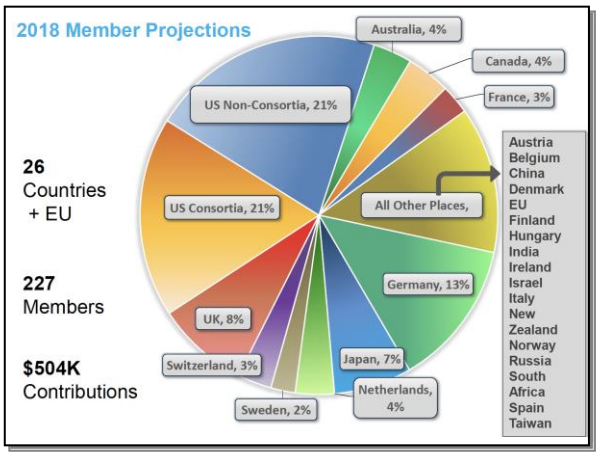
これが分野としての分析です。これは図の下に記載されたサイトで手に入ります。

図 4 は、arXiv のメンバーはどこの国から来ているかという分布です。US Consortia が、アメリカの大学等がコンソーシアムを組んで加盟しているメンバーです。US Non-Consortia がそれ以外の独立で入っているメンバーです。それらを合わせてアメリカで約 40%です。そしてイギリスなど主要なヨーロッパの国があ



（図 3）

（図 2）



（図 4）

り、アメリカとヨーロッパ以外には、中国やオーストラリアがあります。日本も 7%入っていて、US、ヨーロッパ以外では最大となっています。

現在、200 強のメンバーが加盟しています。このメンバーはお金を払います。arXiv は多く投稿されている機関にメンバーになることを依頼しています。東京大学に所属する人が arXiv にたくさん投稿していれば、東京大学にメンバーになってほしいと依頼します。ただし、arXiv のメンバーになるのは強制ではありません。arXiv はご存じのように誰でも、所属があろうとなかろうと投稿できます。ただ、財政のためにこのような仕組みを取っています。

どこから投稿されているかというと、メンバーの国と比例するように、アメリカが約 40%でヨーロッパは約 30%です（図 5)。中国はメンバー数に比べて投稿数が多く、3 番目となっています。これは課題です。

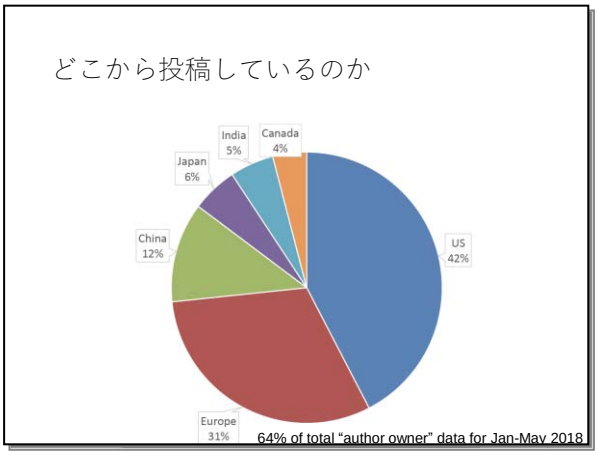
運営体制は、大きく、執行チームと運営チームと技術アドバイザーとなっています（図 6)。その他にアドバイザーのような形で、Paul Ginsparg がいます。彼は arXiv をつくった人で、今、コーネル大学のコンピュータサイエンス学科にいます。

執行チームの Oya Rieger さんはライブラリアンです。今年までは arXiv はコーネル大学図書館が運営するという立場を取っています。だから、執行チームのトップの Oya さんはライブラリアンなのです。

既にカレントアウェアネスでも報告が出たと思いますが、来年からこの体制が変わって、コンピュータサイエンス学科に移ります。われわれも心配したのですが、この運営体制そのものはほとんど変わらなくて、ただ、コーネル大学での責任を持つ部局が変わるということです。チームそのものが今、図書館の中にあるのですが、当面は物理的にも図書館の中にいるそうです。そういう意味では、ユーザの立場からですと、すぐ来年 1 月から何か変わるということはありません。

今まで、開発のことに関しては図書館が責任を持つという体制で、執行チームと図書館が並行にあって、複雑だったのですが、コンピュータサイエンスに変わったら、今度は執行チームが開発チームを管理するという、むしろシンプルになると説明されました。Paul Ginsparg のアドバイスを受けて、より IT システムとしての改善を図りたいということが一つのモチベーションとなったそうです。

ビジネスモデルは重要なところで、これはサービスである以上、お金がないと回りません（図 7)。一体どうやって回しているか、大きく言うと三つのリソースからできていて、まずはコーネル大学自体資金を出しています。現金も出しているし、Oya さんのようなコーネル大学所属の人に、人件費という形で出しています。それから、いわゆる財団系のもの、Simons Foundation 等が資金を日本円にすると 4,000 万円ぐらい出しています。あとは先ほど言ったメンバーから集めたお金が 5,000 万円ぐらいです。3 分の 1 ずつを負担しているようなイメージで動いているのが現在の仕組みです。総予算が約 1 億 5,000 万円です。



（図 5）

（図 6）

この金額が妥当かどうかというところは結構な問題なのです。arXiv を使ったことがある方は分かると思いますが、arXiv は実は結構、古くさいシステムです。それを今、ニュージェネレーションシステムに非常に時間をかけて移行しようとしています。歴史があるので、とても複雑なシステムになっています。LaTeX から入って、そのまま PDF が生成されるので便利なのですが、その分複雑なシステムになっていて、それがシステムをリニューアルするのにネックになっています。

今の話をまとめます。arXiv の分野としてコンピュータサイエンスが正式に入ったのが 2012 年ぐらいで、まだ 5 年ほどしかたっていません。それからさらに最近、統計や一部の経済系も入ってきています。先ほど表で示したように、arXiv は投稿数も増えているので、プレプリントサーバーを維持して、分野を順次増やしているという面においては arXiv は順調に成長しています。

また、arXiv がつくったプレプリントの文化およびビジネスモデルが、OA 時代を迎えて、分野を超えて高く評価されているということは大きいです。

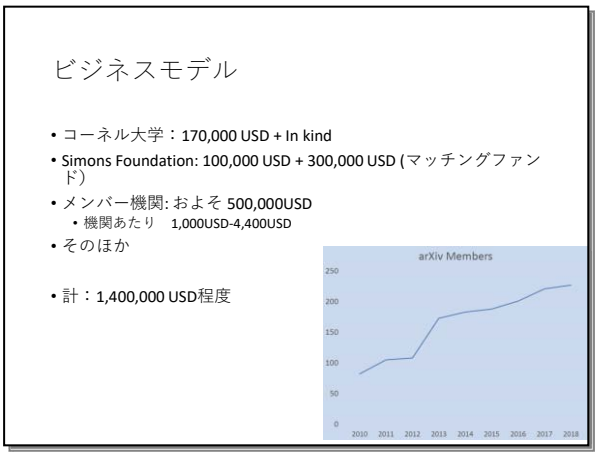
図 8 が最近の動きです。arXiv は 1991 年に誕生しています。そこから 20 年以上、大した動きはなく、2013 年ごろから急激にプレプリントサーバーが他の分野で増えてきました。PeerJ のプレプリントが出てきたのが 2013 年で、2016 年以降、立て続けに出てきて、2016 年に bioRxiv、engrXiv、SocArXiv、PsyArXiv が出てきています。2017 年には LawArXiv、AgriXiv、ChemRxiv なども出てきています。

このうち、後ろに xiv と付いているものは、基本的には arXiv をリスペクトした、あるいはインスパイアされたとそれぞれのサイトのページに書いてあります。arXiv のモデルをわれわれの分野でも使いたいと考えて立ち上げたと書いてあります。特に、赤いものが arXiv にインスパイアされたものです。

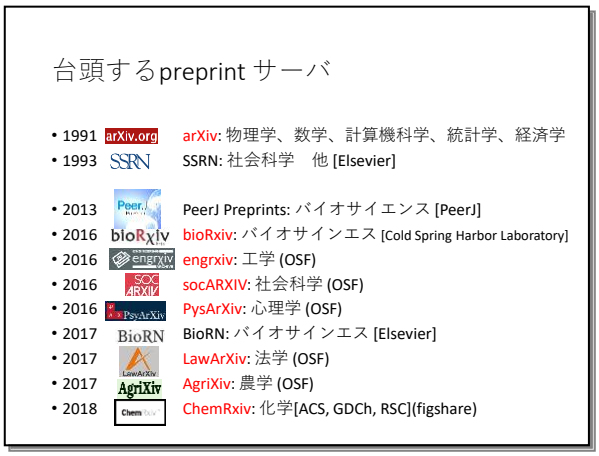
ここに違うもの、SSRN があります。これは約 2 年前にエルゼビアに買収されています。SSRN は元々、ソーシャルサイエンスのプレプリントサーバーでしたが、エルゼビアはこれをいろいろな意味で活用しようとしています。

その系列に当たるものが BioRN です。RN 系列のバイオ版をエルゼビアが立ち上げたということのようです。

また、同じ arXiv 系でも後ろに OSF と書いてあるのが、これは Open Science Framework で、Center for Open Science がつくっているプラットフォームに乗って、それを利用してサービスを提供しているプレプリントです。この五つは、マネジメントが違うだけでプラットフォームは同じものを使っています。それとは全く違う、インスパイアされたけれど、独自運営されているのは、bioRxiv、ChemRxiv です。これはアメリカとドイツとイギリスの化学学会が共同運営をしています。面白いのは、プラットフォームは figshare が提供するものを使っていると言っていることです。



（図 7）



（図 8）

2013 年、特に 2016 年以降にたくさんのプレプリントが生まれていて、ただ、多くは arXiv のつくった文化をむしろ自分の分野に取り入れたいという形で使われています。そういう意味で、arXiv が 30 年間行ってきた活動は大変高く評価されています。

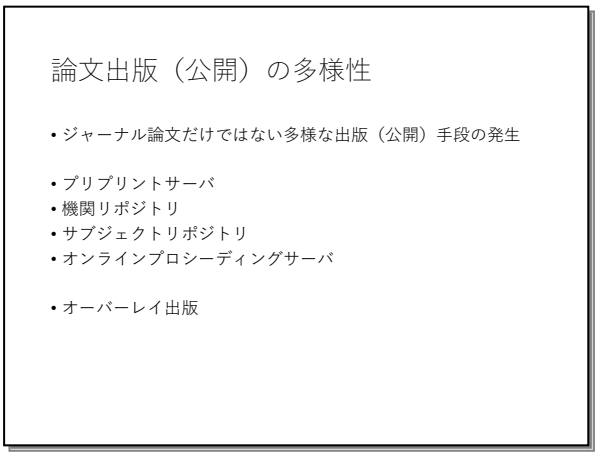
ところが、この 30 年で学術コミュニケーションが変わってきました。オープンアクセスの動きが強くなってきましたが、30 年前はそのような動きはなかったのです。プレプリントサーバーの役目は変わってきています。arXiv が変わった環境の中できちんと対応できるかということが、arXiv という組織の課題となっています。

プレプリントサーバーの新しい役割

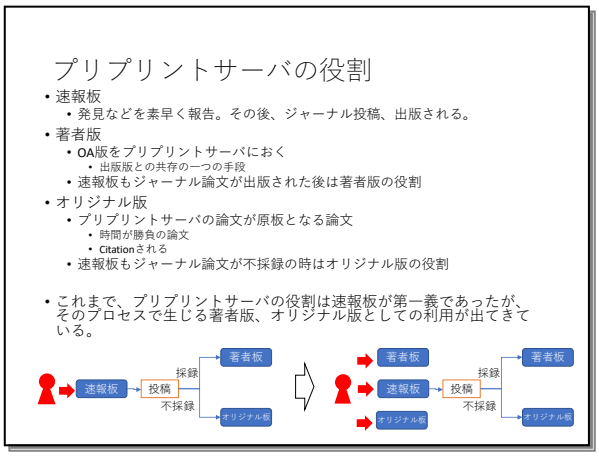
arXiv ができた当初、高エネルギー物理学では、新しい知見が発見されると、まずは関係する研究機関で情報を共有するという文化がありました。当初はファクスで送ったりするということがあったそうです。高エネルギー物理学の場合は、やっている研究所と大学が限られていて、世界中に拠点は数えられるほどなので、そういうことをやっていました。しかし、Paul Ginsparg が、それでは効率が上がらないから、一つサーバーを置いて、そのレポートを全部蓄積すれば楽だろうと考えて始まったのが arXiv です。大体の場合、そのレポートはいずれジャーナルにパブリッシュされるというものです。始まったころのプレプリントサーバーの役割は、比較的閉じたもので、あくまでも同業者が知らせ合うというサービスでした。

ところが、2017 年、2018 年においては、いろいろな論文の発表の仕方があって、その一部としてプレプリントサーバーが位置付けられるようになりました。これは、当初のプレプリントサーバーの役割と違う役割を持たせるようになったということです。だから、プレプリントサーバー自体の機能もそのようにならなければいけないというのが、今の課題です。具体的には、他の publishing（出版・公開）の方法との役割分担、あるいは他の publishing との連携、さらには他のオープンアクセスへの積極的関与ということが求められています。

多様な publishing のやり方が生まれてきました（図 9)。では、具体的に何が違ってきたかということですが、元々のプレプリントサーバーの役割というのは、まず速報版を出して、それがジャーナルに投稿されるというものです。採録されても消されることはないので、それは著者版として残ります（図 10)。採録されないと、結果的にオリジナル版になります。つまり、ジャーナルには公開されないけれど、arXiv には残るのです。これが元々の arXiv の運用方法でした。それが逆に、初めからもう投稿しないつもりで出す人が出てきたのです。このプロセスに関係なく著者版を置きたいというだけのモチベーションで arXiv を使う人も出てきました。つまり、速報版、著者版、オリジナル版が同じ arXiv の中に共存しているということになります。



（図 9）



（図 10）

特に arXiv は速報性が重視される分野で使われるようになってきました。図 11 は昔、北本朝展さんが紹介したものが面白かったのでもう一回引っ張ってきました。この論文が 2016 年 10 月 10 日に arXiv に載せられると、別の arXiv の論文が引用するのです。その論文の出版日が 2016 年 10 月 11 日ということで、前日の投稿を引用しているのです。

調べたところ、この二つの論文ともジャーナルには採録されていないようなのです。でも、そんなことは関係ないのです。Google Scholar でどれだけ使われているかを見てみると、例えば右の論文は 86 件引用されています。左の論文も 23 件引用されています。ですから、別にジャーナルに採録されるかどうかは関係なく、arXiv にあるだけである種の目的は達成しているのです。

（図 11）

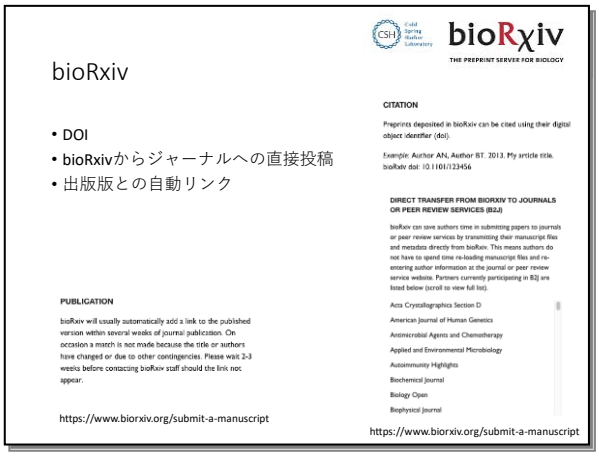
arXiv は OA に貢献しない？

もう一つは、ジャーナル、カンファレンス連携ということがあります。これは最終版とのリンクがあること、投稿時に連携することです。

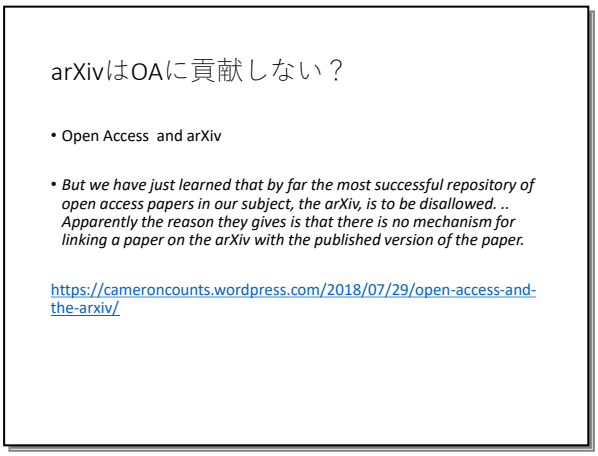
ある人が、arXiv は出版版とのリンクがないので、オープンアクセスの手段としてはカウントされないということを言っています（図 12）。実は、arXiv には明示的に出版版とリンクする仕組みがないのです。また、DOI もないのです。ここが問題ではないかということです。

後で坊農さんが説明する bioRxiv はきちんと DOI が付きます（図 13）。また、bioRxiv においてジャーナルに投稿するという仕組みが確立されています。従って、当然、投稿されると自動的に出版版とリンクが付くという仕組みを用意しています。

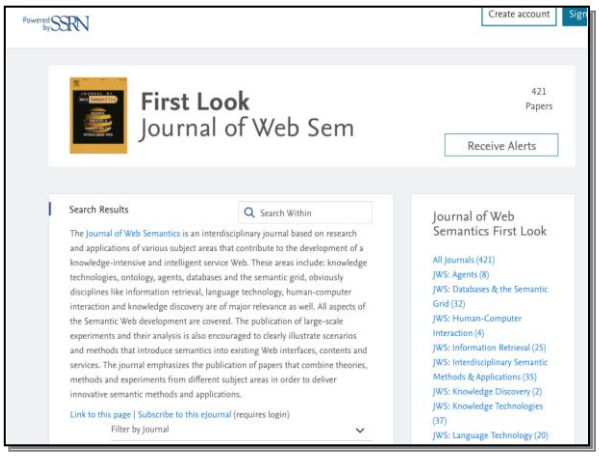
図 14 は私がエディターをしている『Journal of Web



（図 13）



（図 12）



（図 14）

Semantics』というエルゼビアが出しているジャーナルです。これは最近、投稿時は公開しないのだけれども、アクセプト時には著者版が SSRN に自動的に置かれるという機構を今度から導入すると言っています。今までのジャーナルはプライベートなサーバーに著者版を置いていたのですが、そうするとアクセスしづらいので、今までの 400 件の論文を全部含めて SSRN に置き換えたのです。

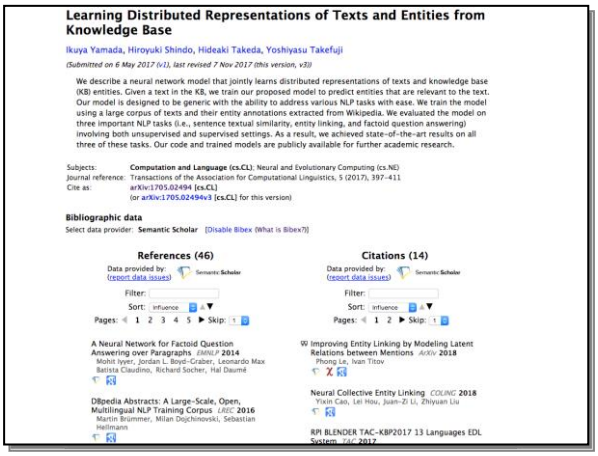
『Journal of Web Semantics』は、全くソーシャルサイエンスではなくてコンピュータサイエンスなのですが、なぜか SSRN に置くということです。これは恐らくエルゼビアがプレプリントサーバーを戦略的に使うための一つのトライアルです。こういうことが起きていて、これも著者版とのリンクです。

では、arXiv は今どうしているかというと、DOI 付与、メタデータの再設計に関してもまだめどが立っていません。それはテクニカルな理由です。システムが直せないというのです。メタデータがきちんとできないと言っていて、まだ見込みが立たないのです。

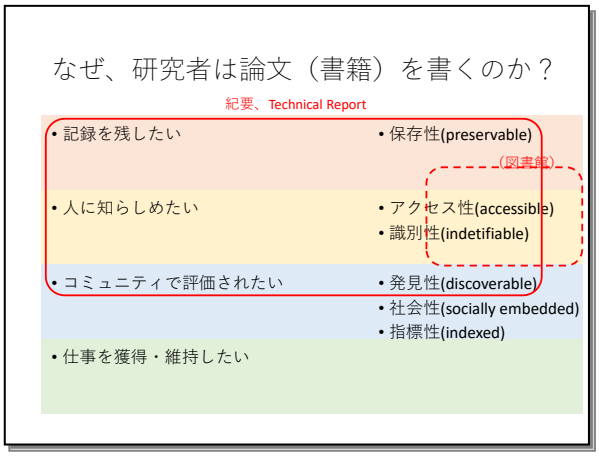
ジャーナル、カンファレンス連携に関しては、実験をしています。ACM と連携実験をしていて、まさに ACM のカンファレンスに投稿するときに、同時に arXiv に載せるという仕組みを今テスト中です。

今回の MAB にも ACM の人が来ていて、ACM および学会が arXiv とぜひ協力したいと。つまり、われわれはむしろ一緒に共存することが学会の使命だと思っているとはっきり言いました。ただ、arXiv 側がなかなかシステムの運用がままならないのが現状ということです。

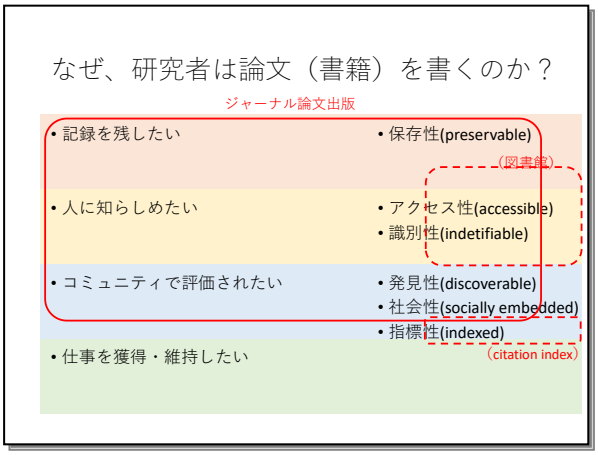
また、arXiv Labs ということで、少しだけシステム改善で、サイテーションが下に出るような仕組みを実験的に導入しています。皆さんもぜひ試してください。



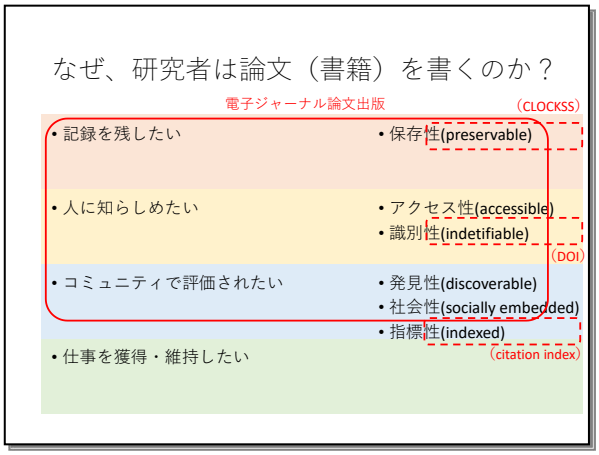
（図 15）



（図 17）



（図 16）



（図 18）

リファレンスとサイテーションが出るような仕組みができています（図 15）。

結論

なぜ研究者は論文を書くのか。それは、記録を残したい、人に知らしめたい、コミュニティで評価されたい、仕事を獲得したい・維持したい、これに尽きるのです。

17 世紀は、それを往復書簡で行っていました。それを書籍に書くと、人に知らせることができます。でも、評価されるのはなかなか難しいので、書評などがあります。

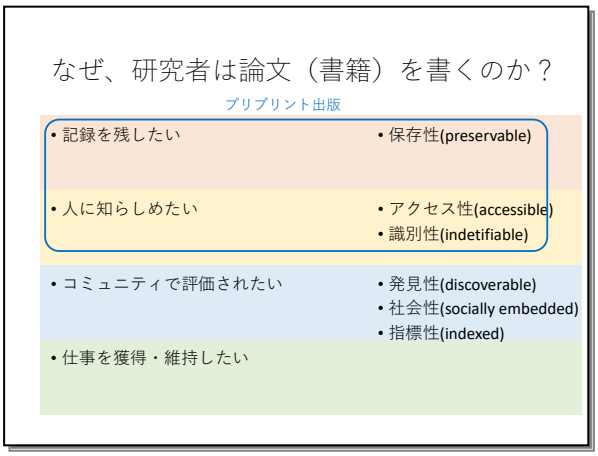
ジャーナルというのはある種の素晴らしいシステムで、記録、人に知らしめる、コミュニティに評価されたいというのをワンセットにして、非常によく回りました（図 16）。紀要は図 17 のようになります。電子ジャーナルは役割は紙版と同じですが、DOI などで、より識別性が良くなりました（図 18）。

では、プレプリントではどうかというと、コミュニティで評価されたいというところまで入ってきません（図 19）。

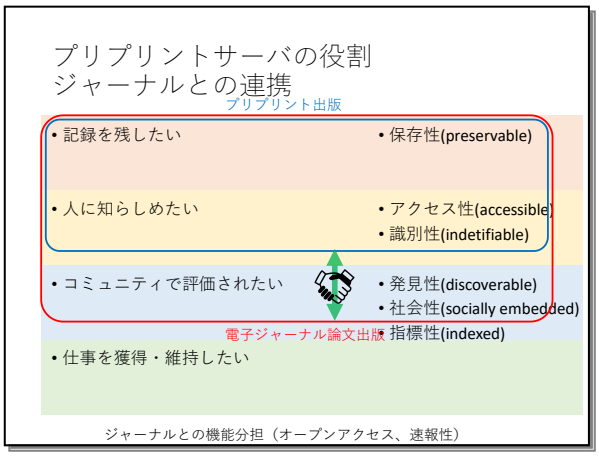
そのときに、青い部分と赤い部分で役割分担をしようというのが今日の午前中の話でもあり、プレプリントの新しい役割です（図 20）。評価されるといった部分は別の機能に持たせればいいのではないか。そういうことが今起きているのがジャーナル連携です。

一方、直接オリジナル版でいいではないかという人にとっては、むしろプレプリントサーバーの機能を強化してほしいというところがあります。機能を強化するということは、いろいろな指標性、アクセス性、識別性をプレプリントサーバー自体に持ってほしいという要求ということになります。

まとめると、プレプリントサーバーは求められる機能や役割が増えてきています。今回の MAB に出席して受けた印象は、arXiv はシステムはあまりモダンではなく、またマネジメントも比較的ゆっくりしているということです。ですので、現在の変化のスピードについていけるかという点を危惧しました。arXiv はサービス自身もサービスモデルも非常に期待されている一方、なかなか困難な段階になっていると感じました。



（図 19）



（図 20）