New Initiatives in Open Research

Clifford Lynch
Coalition for Networked Information, Clifford@cni.org

Lee Dirks
Microsoft Research, ldirks@microsoft.com

Follow this and additional works at: <https://docs.lib.purdue.edu/charleston>

An indexed, print copy of the Proceedings is also available for purchase at:

<http://www.thepress.purdue.edu/series/charleston>.

You may also be interested in the new series, Charleston Insights in Library, Archival, and Information Sciences. Find out more at: <http://www.thepress.purdue.edu/series/charleston-insights-library-archival-and-information-sciences>.

Clifford Lynch and Lee Dirks, "New Initiatives in Open Research" (2011). *Proceedings of the Charleston Library Conference*.
<http://dx.doi.org/10.5703/1288284314874>

New Initiatives in Open Research

Clifford Lynch, Executive Director, Coalition for Networked Information
Lee Dirks, Director, Portfolio Strategy, Microsoft Research

Clifford: Lee and I have both been involved in a number of things over the past two or three years that seemed to me to be gaining a considerable amount of momentum and I think are going to really change the landscape in significant ways. What I want to do quickly is to put a frame, a context, around these developments and then Lee is going to run through a series of example systems at blinding speed. We are going to try very hard to get through this quickly enough that we've got 10 minutes for questions and comments at the end. I'm at least going to focus my remarks primarily around scientific work. A great deal of what we're discussing applies to other kinds of academic scholarly communication, humanities in particular, but there are enough differences and nuances that I just want stay primarily focused on science.

Let me suggest that we have got some really serious problems right now with the system of scholarly communication and science. They are not the problems that have been so exhaustively documented and analyzed over the past decade or have more to do with economics. There are certainly economic problems in the system, but the stress points I want to talk about here are separate from economics, and they occur along several different dimensions.

One dimension is scale. It is easy to forget how the system of scientific publishing is getting bigger and bigger and bigger, and as the scientific enterprise grows globally at an alarming rate, it gets steadily worse. Demands on scientists for ever greater volumes of publishing to gain academic rewards of course don't help either. If you do the math you will find horrifying numbers, something like a scientific paper is published every minute or two. It means you're buried. It means everybody's buried.

There's a problem with speed. We are under constant pressure to make scientific research and discovery more effective, to make it move faster, to get more science per dollar, to accelerate medical breakthroughs. And yet in some areas the rate of progress, and the rate of transmission of

knowledge, seems to be at best holding constant, or even slowing. Most of the formal journal system is really slow. In some communication channels where we are achieving considerable speed, it is often at high cost. When you just look at the scale and speed challenges together, you realize that we have got huge problems with filtering, with reviewing work, validating work, finding colleagues, finding important papers. We also face a very weird and frustrating situation where at least one of the hopeful prospects for getting some handle on this enormous and ever-expanding scientific literature is through computation upon it. We see the very beginnings of that with things like Google Scholar, Microsoft Academic Search, some of the text mining experiments that have been done, but from a practical basis the obstacles involved in getting access to really substantive segments of the scientific literature are intractable. These are mostly trapped in silos right now, so this is another part of the system that we need to overcome and I think is particularly challenging one for everybody involved.

The last stress point, and this is the one that I think that we're going to spend the most time on here, is that there is a growing disconnect with practices and norms in scholarly work and the way the system of scholarly communication is operated. There has been a lot written about this. There is a wonderful book that actually Microsoft Research and Tony Hey over there put together, in memory of Jim Gray called *The Fourth Paradigm* which is available for downloading for free over the net. There is a very nice new book by Michael Nielsen, *Reinventing Discovery*. There are a mass of reports from the National Science Foundation's Office of Cyberinfrastructure and similar organizations in other nations. The bottom line here is that more and more science is data intensive, is computationally intensive, it relies on big data, it relies on complex software, it relies on large-scale collaborations among dispersed people. And there's a great focus on how data is managed, preserved, discovered, shared, reused and repurposed that's emerging, in part at the urging of the funding agencies.

Now, this is pretty mainline. I want to note that there are a cadre of people, many of them graduate students and young scholars, who really want to take this farther who are making an argument for a more fundamental change where data and experimental work is exposed very, very early in its life cycle. They speak of “open science” or “open research”. You can see systems like myExperiment, which is a platform that we don't have time to talk about here in any detail, which is a way of sort of putting up your lab work for inspection. When taken to the limit, these are ideas that I would say are still relatively out on the fringes of scientific practice, but the kinds of things I'm talking about here are much closer to consensus, the center of scientific practice.

Somehow we've got to get past the point of producing articles that are designed for print and just storing and delivering them electronically. A scholar from a hundred years ago would have no trouble with the genre, the format, the presentation of today's scientific articles. We can do better, just like we can do better than chalk and blackboards. And it's the same old presentation of science when the science has gotten much more complicated, where there are major issues about reproducibility, particularly in the context of very complex data and software platforms, where there are issues around building on other people's work—both facilitating this, and giving and receiving credit where it's due. We have certainly seen both scholars and funders recognize data as a primary input and primary output of scientific inquiry and one that needs to be reused and managed systematically. There is a very significant debate now about how to connect data with articles.

Those of you who saw MacKenzie Smith's talk this morning got some taste of that. There are models that go all the way from citation to data in ways that are not that dissimilar from citation to other articles, all the way through people who have visions of data literally interpenetrated with the article as a computational whole and the article being accompanied by tools that allow the visualization or analysis of that data. There is very interesting work going on in a number of quarters to realize that, including some quite unexpected ones. For example, you look at some of the things Wolfram is doing

these days with Mathematica, Wolfram Alpha, and large datasets you may be surprised.

There has been a whole series of workshops with an informal group of people that includes scientists, some librarians, and some information technologists: Beyond the PDF was one such workshop. There is a group called Force 11 in Europe that is in the process of drafting a very nice manifesto piece in this area, and just a couple of weeks ago, Harvard and Microsoft Research did a two-day workshop in Cambridge to look at current developments. Basically, the questions are about how we can get the tools to manage data and scientific workflow, to present these kinds of materials, and to interconnect them with the traditional scholarly literature.

This is not going to be an instantaneous change. As we all know, the Academy is very conservative and indeed some very interesting issues show up here about identifying important work and what role tools and mechanisms and measure developed to surface important work should play in the evaluation of scholars themselves. There are numbers of proposed measures now for trying to gain insight to scholarly impact. Personally I'm still a bit skeptical about these, but it surely is infinitely better than the sort of voodoo misuse of the bibliometrics that were designed for absolutely different purposes. At least it is asking the right questions in my view. We have questions, as I said, about data citation, about bringing together the work of authors as they author in very different media, and about simultaneously disambiguating the work of authors with similar names, so that we can have good underlying datasets to do computations that attempt to identify and bring together significant work.

Those are some of the threads that you will see over the next 10 to 15 minutes. And with that I'm going to pass it over to Lee with the view that in this area, a demonstration as worth a lot of words and he will at least flash up demonstrations of a number of systems. I want to stress that for every system he will show you there are three or four more we could've picked that are similar in objectives but that differ in interesting ways. There are also a number of significant ones that address other parts of the scholarly communication cycle that we don't have time for. But with that, over to you.

Lee: Thank you very much, Cliff. I appreciate that. So I'll echo the last point. By way of a very brief introduction I wanted say little bit about Microsoft Research. The team that I'm reporting to is part of the pure R&D division at Microsoft. It is about 1,000 people distributed in some labs around the world. My team is about 50 that report into Tony Hey, the aforementioned Tony Hey. Our team is responsible for collaborative research projects with academia. My specific area of focus is working in scholarly communication which is kind of part and parcel my interest in this specific area. I just wanted to provide a little bit of context there.

There are seven tools that I want to call out today, again, to make reference to Cliff's points. These are simply highlights. There are many tools that are out there, and these are several that I just wanted to point out and give a little specific detail about. Over the course of the last 18 to 24 months, these are tools that have really started to emerge, and the reason I'm highlighting this is because they're starting to get some momentum in this area. I am pointing these specific tools out because they interact in many cases, and are starting to reveal the tools that will build out the scholarly communication lifecycle. There are several common themes related to these: Many of these are open source, many of these are available for free, and almost all of them have some sort EPI that allows them to communicate and share information back and forth, which is critical. So we are starting to see these tools which are interoperable and can be integrated into one another. So this is a nice development to see. This is going to be very quick whistle stop tour though.

The first tool that I want to mention is VIVO (<http://vivoweb.org>). VIVO is actually built and launched at Cornell, I think back in 2003. I've heard the mantle of "a Facebook for scientists". That mantle has been given many times. I think that VIVO is probably the closest to actually legitimately holding that title. This was actually recently expanded based on some NIH funding that was given about two years ago. About seven institutions have started building this out and they're slowly building a broad network. But it is a way for scientists and researchers to log in, build a profile and be able to find others that are semantically connected via this open-source network to connect with one another, find

each other and facilitate research in that way. This has been great to see over the course of the last two or three years to see this community really come together. They are having annual meetings nearly every August that have been very successful and well attended. So, it is great to see this sort of tool that's being developed. I'm going to quickly get through these so hopefully you'll get an idea in the sense with the URL below to visit the tool yourself.

ORCID (www.orcid.org), which I believe MacKenzie mentioned a little bit earlier today, is an incredibly important development. This is an initiative that originally was started jointly, I think, brainstormed by Thomson Reuters, Nature Publishing, and several others. The two conveners of the original meaning were Thomson Reuters and Nature. It has rapidly grown to be, I would say, not just an industry-wide, but a very community-driven effort. It stands for Open Researcher and Contributor ID, and the effort is to build a system that all publishers at all libraries and all academics will be able to use in terms of assigning unique IDs to authors and contributors so that we can find one person. We can disambiguate people from one another and be able to assign citations and other attributions as necessary.

This has grown from an initial meeting where a handful of people got together, to 44 founding sponsors—those who are actually paying people that are contributing money to see that this development happens—as well as over 250 participating organizations having several meetings a year to brief the interested parties on this. I was able to go to the most recent meeting that was in Geneva where they were able to actually announce some availability, planned availability, of the APIs. In a kind of "precompete" way, all of the industry players have gotten together and are going to share information and work cooperatively, which is a good thing to see. This is not unlike the ways in which the industry came together to develop CrossRef, but, as you can see, they're going to have the plans for their query, and deposit APIs will be tested in the Fall. I think they're planning for a launch—I'm not sure if it's going to be called an alpha or a beta—anticipated in May of this year. I just went up to their website and grabbed a screenshot (Figure 1). This is a draft, or a prototype, but this is in essence the way researchers would be able to go

on and request an ID, and that is the ID that they would use for submissions with multiple publishers,

which would be a quick, easy way to find material.

Figure 1

Another project that I think has been growing and getting some momentum recently is Harvard's Dataverse Network (<http://thedata.org>). Simply put, it is a repository for data and is a way for scientists and researchers who need a place to store their data, version their data, and point to their data. It is giving them a method and a protocol for doing that, so it is actually coming out of the social sciences at Harvard, but this has been a way of propagating the system. As a researcher you're able to get an account, deposit large and small data sets, pull excerpts from them, share that with colleagues, store it there for preservation purposes, or share it, indeed, with a database or a publisher. The database or publisher actually have the ability to say that if you have a printed citation you can give a citation and it takes you back to a unique identifier and the URL for finding that data later. I think it's a fantastic tool, it's been very widely used across departments at Harvard, and we are starting to see it again as an open source project in that they are very willing to share the code. As a result, it is starting to develop and move to other institutions as well. I think it is critical that even though this came out of the social sciences, it is being used across domains. I think it's fantastic to see this made available for domains that may not have previously had a methodology for sharing or storing data.

Related to that—and I think, again, that Mackenzie might have mentioned this this morning—is DataCite (<http://datacite.org>), which is an effort that was born a little bit over two years ago. The TIB, the National German Science Library, as well as the British Library, came together to say that we really need to establish a protocol for sharing data and citing data, so they decided to leverage the DOI—the digital object identifier approach. This has been very, very well taken up; we're starting to see it grow. It started with originally just two members and I think as of the last two years they're up to about 15 members. Over a million DOI's have been assigned to data sets around the world. They're starting to build more metadata specifications, and they're starting to share technical infrastructure and cloud storage of these. So this is an important development, and it is significant that these tools are being made available to scientists so that they'll be able to move data forward.

A particular—and this may not be one that you heard of—fantastic new tool that was actually developed at a hack-fest about six months ago in the UK is something called Total Impact (<http://total-impact.org>). The two or three original developers of this pulled it together within about 24 hours, and they have continued to evolve it over the last four to six months. What this is doing is basically saying

we need to have a better view into the actual footprint that a researcher can develop. As Cliff was saying, it's not just a paper; it's not even just a data set. We're starting to talk about blogs; we're starting to talk about code, the images, and all the supporting research objects that might go with that object, as well as even tweets, Facebook pages, etc. that are related, to show the traffic and the generation of communication around that research object. There are also a few other things that are dabbling in this kind of alt metric space. Total Impact is one that is very cool. I have a screen shot here, and, as you see, there are different artifacts that can be tracked: published papers via various places, from Pub Med, from Mendelay, from anything that has a DOI, data sets from various databases, software, slides, etc. You will be able to login and actually enter your unique ID's—obviously when ORCID is availa-

ble you would put your ORCID ID in—but anything where you have a profile somewhere on the web, you are able to embed it in here and then over time link to articles or link to yourself overall as a profile.

This happens to be from Heather Piwowar (Figure 2), and these are just articles on the left and then some references so you start to see comments associated with that specific article from maybe Facebook, maybe Twitter, or maybe Mendelay. So you see the impact beyond citations and in a much faster time frame than the traditional two or three year lag. You start to see the immediate and total impact of that article, data set, or source code. So, it is a tremendous idea that is up and live, and I encourage you play around with it and also to consume information via an API.

Figure 2

Just this week DuraCloud (<http://duracloud.org>), an offering from DuraSpace, was officially launched. It has been in alpha and beta for the better part of the year, but they officially launched on Tuesday. It has 11 pilot partners that have been working and testing it. I think is a tremendous idea. My background is actually as a preservation librarian, and I was very pleased to see this initiative started. I've got a schematic of what DuraCloud does (Figure 3), and the idea is that at the very top of the screen would be your repository—your database—and then in the middle you see what your DuraCloud is doing.

Basically, it is saying that if you sign a service level agreement with us, we're going to back up your data in the cloud and provide services around that, and behind that DuraCloud then goes to any number of cloud providers behind the scenes. I think at present they are working with Amazon and Rack Space, and they're in alpha right now with Microsoft Azure. But behind the scenes you, as the repository owner, it doesn't matter necessarily, you sign an SLA with them that says you're going to keep my data, and you're going to give it back to me whatever I need. You can even have the ability to

say what continent you might want to sort on depending on what the service provider allows. It just takes that up contractually and gives a preservation. You can have multiple copies in multiple places. This is a fantastic model; I'm very encouraged to see that

DuraCloud is moving forward with it, so I encourage you look into this, and I think their plan is, in addition to just plain storage and backing up, they will offer additional value added services related to videos and text mining and things in the near future.

Figure 3

The one tool that I am a little bit more expert in is Microsoft's Academic Search (<http://academic.research.microsoft.com>). This is a search tool that has actually been around Microsoft for several years, but it was very focused on computer science papers, but in the last 18 to 24 months, we have started to expand it to the entire Academy, to all domains of research. We've grown from about 8 million articles, starting last December, to over 37 million publications now. We have actually 70,000,000 to 75,000,000 in the queue for processing, and so it's very data intensive processing to get them in, so we're getting about 5 or 10 million a month added to that with the goal of maybe somewhere around 100 to 150 million within 9 to 12 months.

This is a free service; I want to stress that this is free and open. We are working actively with publishers, with aggregators, with scholarly societies. We've signed content arrangements with more than 50 partners. No money is changing hands. This is all about getting access to the content and also providing access to you. Microsoft is actually getting the full text, doing entity extraction and natural lan-

guage processing on the text, and putting that out in a way that you can use it via our interface or actually consume it via a free API, so there is a way for anyone who wants to take this data and use it for noncommercial purposes to integrate it with their own data and with systems.

I'd like to give you a few screen shots of what academic search looks like. If you would run a search on Michael Nielsen (Figure 4), the author that Cliff was mentioning a moment ago, you would automatically see that this is a dynamic page. We've done all this extraction from these papers and pulled together, in his case, over 106 publications that were crawled and indexed that related to him. We've pulled that together and built this profile automatically so you can see his number of publications and citations, and we're calculating the G index and the H index for him. We're actually surfacing on his photos and building a profile around him that shows you his number of citations over time, and then you have the ability to click and go straight to a visualization. This is a way for you to see who Michael has co-authored with, or, on certain cases, who has cited him, so there is an easy digitalization.

You can actually walk the graph and follow links and follow other authors who are related to him. You have the ability to manage that profile and so you can actually go in and say, "Wait a minute, you don't have two of my papers, I would like to add that URL in", or, "Actually, the paper that you have assigned to me is not mine". So again, it's done automatically so you can have us pull that out or

merge authors or add things. You have the ability to edit and maintain a profile. You also have the ability as an author to embed this, so we give you little bit of script so that you can actually put this on your page of your library. When maintaining profiles for faculty at your college you can actually embed this in your local webpage.

Figure 4

Figure 5 shows what a publication would look like in Academic Search. So you can see there's a lot of information that we're surfacing: the abstract, the number of citations over time, and the various locations where we've sourced it. If it's from a publisher or if it's from an open access, we'll link to that. Down at the bottom, you'll see we're actually putting the citation context from those papers; we're actually showing you so you don't have to go and track down those 5 or 15 citations. We're actually surfacing the citation context right there on the page. On the left nav you'll see a keyword. You can

actually click on a keyword, and that keyword will take you to an abstract that will tell you about the definition of that keyword. The interesting thing on the graph here is we're actually able to show, based on our analysis for our corpus, when that term has been used over time. So you can actually see new terms, and when they came into usage, and when they've ramped up or ramped down. And then we provide a little bit of context again from these papers by starting to give you some various definitions from across the literature about what that term might mean.

Figure 5

Figure 6

Figure 6 shows what a journal would look like, so you can actually look at an all-up version of that Journal. This is actually listing all the publications, literally all of the articles that have come from maternal overtime, the citation count against those articles. We give you a year range for what we're able to have in the site and then you can actually look at the most cited articles throughout the history of that particular journal title.

You can even build profiles around not just authors, but around organizations, so this allows you look at the top publications, and the top citations, and the top authors from a particular institution so your institution would be represented in here. You can run the search and pull that back and find the top cited authors at your institution. You can even compare institutions. I want to stress, again, all of this information is available via a free API so you're able to consume this data and do your own rankings, for instance, if you're interested in doing so.

We have some interesting visualizations. This is one way of visualizing computer science (Figure 7). This is literature dating back to 1960, and you're able to see the subdomains of computer science and actually drill down and what you are seeing here is actually looking at hardware and architecture publications from 1992 to 2009. There were 95,000 publications during that period of time, and on the bottom you'll see those are the top cited authors in that timeframe in that domain. So it is a very quick and easy way to perhaps find co-authors, to find the top people in your field to maybe come give a talk

at a conference, to maybe work on a project with, give an award to, etc. It's a very interesting way to quickly delve in and find people. This is available publicly for noncommercial purposes, so there's more information about it on our website. One example is the Eigenfactor Team at the University of Washington, which is currently using the API and some of their services. They built a recommender service, a visualization mapping tool based on content that they are getting from our API, and several other services that I encourage you to investigate.

Figure 6

So, in closing, I want to point out that when you think about scholarly communication, you can think about it as working in a lifecycle: originally collecting data, doing your research, doing your analysis, and moving to an authoring phase—the publication and dissemination phase where you're sharing the data and maybe even writing an article, writing a book, giving a talk at a conference, writing your blog, and ideally storing, archiving, and preserving that information so the cycle continues.

I typically augment those four kinds of core steps with two other ideas: the need to collaborate with others in and across these four core steps, as well as the need for discoverability. In all the services

and initiatives that I've given you during this quick tour of, they all map exactly to that. So it is refreshing to see work going on in all of these areas, and, as you can see in several cases, the different tools can map in different sections in different areas. And there I mentioned the VIVO team is working with the ORCID folks, and the Dataverse folks are working with the DataCite Folks. It's tremendous to see the integration and interoperability that naturally happening. But, I thank you very much for listening to us and hope that you find some value in exploring these tools further. The landscape is definitely changing, but I am encouraged that we have some fantastic tools for us and for the researchers to work with. Thank you very much.