Preprints, Institutional Repositories, and the Version of Record

Judy Luther
Informed Strategies, jluther@informedstrategies.com

Ivy Anderson
California Digital Library

Monica Bradford
Science

John Inglis
bioRxiv

Judy Luther, Ivy Anderson, Monica Bradford, and John Inglis, "Preprints, Institutional Repositories, and the Version of Record" (2017). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284316717

Preprints, Institutional Repositories, and the Version of Record

Presented by Judy Luther, Informed Strategies; Ivy Anderson, California Digital Library; Monica Bradford, Science; and John Inglis, bioRxiv

The following is a transcription of a live presentation at the 2017 Charleston Conference.

Judy Luther: I'm Judy Luther. I have a background that pretty much covers all different sectors of the market. I started as an academic librarian. I was a library director for a period of time. I worked in sales to libraries. I worked for what was Thomson Reuters and is now Clarivate as head of sales, so I had a chance to talk to a lot of libraries at that point, and then for the last 20-plus years I've been consulting mostly with publishers, societies, helping them sort out their market-facing issues related to journals and books and anything else they've published, and that conversation is now turning toward content. I'm on editorial boards of journals. I have an MLS and MBA, so I bring kind of both perspectives to the table. I can hear both sides of a question or an argument.

What actually prompted this session this morning were questions that began to bubble up for me several months ago when I looked at the news that CrossRef had created a schema for DOIs for preprints. Some people had been registering them, but now there is an official schema for them and there is a growing number of preprint servers. I also know that librarians were sometimes assigning DOIs to content that they put in institutional repositories, and in some cases I wondered if it was indeed the author's submitted manuscript or the published version of the manuscript and to what extent that varied? And then I thought about the fact that for me the DOI, without my thinking further about it, meant that that was the version of record. Now that languaging seems to have arisen about the same time as we began to trust digital enough that we consider that the authoritative version and, in some cases, it held more content than the print version. It might've had colored content. The print might not have. It might've had additional files. It certainly could link to a lot of content. Today the digital pretty much is, if we have a version of record, it is the version of record. But to me that was the version that was distributed, it was archived, it was secure, it was what people paid for. And what did it mean to suddenly have DOIs on all these different types of content? My universe kind of went into tilt mode and I thought, "I'm not even sure I have the questions to ask at this point."

I have a very helpful panel who has come up with some very good questions. The one percolating for me is what do we do when we have more articles with DOIs? I think of it as the version of record but Herbert Van de Sompel has referred to it as a "record of versions," and I've been trying to get my head around that as well.

I've worked with some societies who have had to go through a very painful process of retractions and the reason they did that, and that meant disowning one of their own members because they fabricated data. It was very painful for them but they worked through it and it took several years, attorneys, a lot of unhappiness, but they did it because they thought they were keeping clear the authoritative version of the research in their discipline. Going forward, is that something that is important? There's a whole website called Retraction Watch. Are we going to continue to care about that? And if we have a continuous progression of all these different forms, do we know that scholarly publishing is on a continuum? But if it's all digital and we start tagging and identifying all of it, what do we keep? What do we cite? It makes my head hurt. So, at this moment I'm going to turn it over to John Inglis, who will introduce himself. Actually, I just want to say how happy I am to have the whole panel here. It's John Inglis from BioRxiv, Monica Bradford from *Science,* and Ivy Anderson from CDL.

John Inglis: Well, good morning, everyone. My thanks to Judy and the organizers for the opportunity to come to Charleston for the first time to this legendary event. So, I trust you'll be kind, which, given what my countrymen did to this town in 1780, is probably asking a lot. But, trust me, I grew up in Scotland so I wasn't responsible for that. Judy asked us to give sort of a potted bio to give you a sense of where each of us on the panel were coming from, so my background is in science. I spent some years in immunology research as a research assistant and a PhD student. I learned enough to know that I was not a natural investigator. I loved doing the experiments but I wasn't so confident about asking

https://doi.org/10.5703/1288284316717

the questions, and I got the opportunity to join *The Lancet,* where I learned really the fantastic skill set that professional editors can bring to their jobs and that group of people remain people that I'm hugely admiring of. I then, thanks to Elsevier, was given a wonderful opportunity to found a journal and I did that and some others there, and then Jim Watson of Watson and Crick invited me to come to Cold Spring Harbor for two or three years to build up the publishing activities that had been embedded in that research institution for, at that point, quite a long time. So, that was in 1987. So, clearly I have failed because I'm still there and still doing it and things have expanded enormously. We're in journals and we're in electronic books. Five years ago next week my colleague Richard Sever and I founded the preprint service for biology called bioRxiv and we are en route to setting up a complementary project called medRxiv, which I can talk a little bit about if you'd like to.

So, five years of running a preprint service has told us a good deal, and I thought I would share some of that with you. It's fair to say that preprints in biology were a cause of anxiety five years ago. Physicists and mathematicians had been working with preprints for 25 years, and 1 million articles on arXiv shows how effective a means of communication that was. Efforts had been made to start preprints in biology, which had not taken root, and we were aware that the anxieties revolved around first of all the possibility of being scooped as a scholar and secondly the fact that you might not get the paper that you've written published in the place that you want to publish it if you had put it on a preprint server. And I think five years on both of those anxieties were still present but are much less than they were. And what we have found is that the joy of sharing, this is reflecting what authors tell us, is the joy of sharing exceeds that of publication because the process of publishing in a journal is often long, torturous, and drawn out, and so being able to share your work instantly with a worldwide community has an enormous amount of appeal to the kind of scientists that we are working with, and yet a paper is still a vitally important part of that scientist's career progression. Preprints are no longer confined to biology and to physics. There are now over 30 services in different specialties, different disciplines, and I can't keep track of them all at this point. But they are all growing in various ways. Speaking exclusively about bioRxiv, we're currently running at over 1,200 manuscript submissions a month and that rate is growing. We have currently 17,000 manuscripts on the servers from a very large number of authors in many, many different countries. Revision is a feature of bioRxiv that I think we did not expect to see happen quite so frequently. About a third of the manuscripts are revised at least once, sometimes many more times, and some of that revision is in response to community feedback that comes via social media through blogs and Twitter and so on, but most particularly it comes, authors tell us, by personal contact with people whom they meet at conferences or who contact them by e-mail or by phone to discuss the work that they've posted. So there is a lot of momentum behind this preprint movement, if you want to call it that. However, there is no doubt that journals remain the preeminently important thing for scientists, at least biomedical scientists. Sixty percent of the papers that are posted on bioRxiv appear in some form in journals, and that may well be an underestimate, and the journals that publish these manuscripts are hugely varied from the most prominent to the less prominent, from the broad spectrum to the highly specialized. And of course, the question that is so interesting is what happens to the other 40%? And we might talk about that.

We were asked to discuss this concept of the "version of record" and I started from the premise that no scientist that I've ever talked to in 30 years of being at Cold Spring Harbor, and I'm embedded in a research institution, has ever used the phrase "version of record" in my hearing. So my guess is that they neither know and possibly don't even care what a version of record is. So, the question in my mind is why does this term persist? And I've asked a number of people. The more sort of radical elements in the scholarly communication ecosystem have told me that it is flat out a mechanism for subscription-based publishers to retain a stranglehold on scholarly communication. Well, okay, that's one particular perspective. Is it valuable to librarians? I hope that will come out in the discussion. I also asked my friend Louise Page, the publisher of *PLOS,* if a publisher that specializes in open access and CC -BY communication uses the term "version of record," and she said absolutely not. So, why don't we use some other term, like a "published Journal article" or a "publisher's version"? Because as Judy has said, we're now in an era where, thanks to digital technology, we can trace the evolution of scholarly output over many different stages, and what point therefore is that output intended to enter the scholarly record, and that's a question.

Preprints are taking on greater significance. They can be cited. They can be used to support your grant

application or your tenure committee evaluation, and then there's the vexed question of what happens with claiming priority. How is that done and on what basis does a preprint qualify? Then there is the other question that I'm very interested in is the fact that as the scholarly output progresses through its journey, it acquires additional information. It acquires commentary. It acquires possibly posting peer-review. There is a kind of conversation that takes place around that work. Do we capture that? If so, how? So, there are a whole host of challenges, which is what I think makes this session interesting and prompted me to say "yes" to Judy's request to participate in it, and I'm hoping to learn from you folks in the audience. I stole this phrase, this last phrase from a paper that was actually preprinted on BioRxiv earlier this year and published in one of the *PLOS* journals, written by Cameron Neylon and colleagues, and basically they pose the question should the version of record instead become the version *with* the record? In other words, trailing all these conversations that have taken place around that work as it progresses on this journey? Thank you.

Monica Bradford: Good morning, everybody. I'm Monica Bradford and I'm the executive editor of *Science* and the *Science* family of journals. And I'm coming here and sharing with you kind of more of the point of view of the editor, the person who is working in the peer review process, who is working with the authors to try to figure out the best way to vet their information and to present it. And my background has been in scholarly publishing at nonprofits for more years than I would like to admit. I began with the American Chemical Society and my background is in chemistry, and I've been at AAAS working on *Science* for the last 28 years. I had the pleasure of working on that technology that Georgios mentioned that's 20 years old. We were one of the first journals to go on with Highwire and that was exciting times when the business models were really not discussed. It was just can we do this? Can we make it happen? And will people use it? So that was like the fun days and our focus was really on what can the technology do to make the research more accessible and actually match what was becoming very complex networks of information? Since that time, as you all know, lots has changed and *Science* has a very great institutional sales team and many of you are wonderful institutions that are using our content and that's what it's all about. It's probably more people are seeing and using our content than ever before. We've also developed a whole lot of online things that surround our content, including daily news. We've always had news in our journals. In our journal, we've always had commentary in our journals and now we even just have a much quicker, faster news component. We also have taken advantage of video, podcasts, you know, we realize that a lot of the research data now is captured and it makes more sense to have a video of how a cell is separated or doing things and we've tried to incorporate that kind of content into our digital version. As many of you know, *Science* launched an open access journal to be highly—our idea is that it would be highly selective to see if we could make that work, to maintain our quality and all of our different requirements related to reproducibility, authorship, conflict of interest, all those things, be selective and still do it within an APC that was reasonable, and that's an ongoing experiment. In addition, we have four research journals that are slightly more targeted. So, that's my background. I'm really coming less from the publishing business side and more from the author peer-review side, and how does technology help us and how do all the recent changes affect what we do and how we maintain quality in this environment?

Just to give you an idea of how our approach has been over time, the topic is institutional repositories, archives, and preprint servers and how all these things fit together with the version of record. Our policy at *Science* has been, you may not believe this, but is to really try to follow the community. We were an early adopter of arXiv and we thought that was fine. This was clearly something the physics community felt worked within their workflow. We do not want to do things that impede the workflow of the research. We want to support it. We want to be there and to play our role, which is the peer review and certification and verification, and also improving the accessibility and the understanding that someone may be able to pull from the research that we publish. So, that was really for a long time the only, arXiv the only preprint server that was really in our concerns, and it wasn't a concern, it worked fine. We had a citation style. Institutional repositories we slowly accepted. We've been green for many, many years and we allow posting of the accepted version in institutional repositories and we allow it at six months. All our research is free on our site at 12 months, so we support it because, again, we felt that this is where the research community was going. This is what their funders are mandating and this is what their institutions are requiring to capture the institution's output and record of scholarship. And we feel as ourselves as part of that community that we want to support these things.

BioRxiv was a little bit of a—we did a little bit more thinking about. I don't know how many of you would agree with this, but our experience has been that the biomedical research community is way more competitive in their behaviors than some of the other scientific scholarly fields, and we weren't sure that they were going to actually embrace this concept, but we attended the ASAP Bio meetings and were convinced that there was a need within the community for biomedical scientists to be able to show progress on their research as it related to their career development, if they were up for a job, if they were up for a grant, and their work, their one paper or whatever was stuck in a pipeline somewhere. We actually were convinced that this was a good thing. But here we are now and we just did a quick experiment, we did using Google we searched, looked at the first two pages of the results and we found 60% of our papers in multiple different versions whether it was the published version, the preprint, a postpublication version, and so there are a lot of different ways that people are now finding the content in all of its life, at different points in its lifetime, and we are wondering how do users decide which version to use? What do they value? That's kind of where we are now and how we're trying to think about how these things come together.

So, what's keeping me up at night? I think we're trying to understand what the purpose is for the different versions and how do they serve academic institutions? How do they serve the public? Are the needs for instruction and research different? Are the needs for a journalist different? How important is peer review? We put a lot of resources into peer review. Does the version of record imply certification through a formal peer review process? Does informal pre-or postpublication commentary carry the same weight? This is a very research intensive, not only for the journals that are running the peer review process, but for the scientists. The peer review process takes time and effort. Do we value it? I think the other things we're looking at is how as librarians, how does an author, how does the publisher signal to the user what version they are looking at and how does it fit into the life of the scholarly development here and the scholarship that is involved in this research? And at what point should the user feel confident about the validity or the usability of the content? *Science* is published by AAAS. We're supposed to be helping the communication of science. In what way can we make sure that all these various versions actually help us communicate science better and not just muddle or confuse people? And again back to what do people value? Is editing a presentation important? Maybe not if everybody's willing to just look at the preprint and move from there or is that just researchers? I mean, there are a lot of different people that use content at different stages. What do they value? And then just talking about the technology, I mean things are changing rapidly. I think eventually we're going to be able to very easily support a quote-unquote "living document," assuming that scientists are actually willing to put the time in to continue the revisiting of something that they think is done and finished. But, if we do have living documents, how does credit get assigned across the life of the document? Particularly when for such a long time journal publication has played an important role in evaluating scholarship of a researcher, and I think that is going to be changing. We need to talk about that. These are some of the things that I would love to talk to you about and hear what you are thinking and we can go from there.

Ivy Anderson: Hi everyone. I'm Ivy Anderson from the California Digital Library. I'm pleased to be here to talk with you about this topic. Many of you who know me from CDL know that I've been involved in licensing. My role at CDL is associate executive director and more well-known, I think, director of shared collections activities for the CDL and the University of California system, and much of my work is involved in licensing electronic content and electronic journals. We spent a significant amount of money on behalf of the University of California system with my colleagues at UC to help foster and preserve the scholarly record.

We also have a very strong focus on scholarly communication and scholarly communication transformation. That's been an important focus of my work, trying to transition our expenditures from licensing toward open access and other forms of scholarly communication. It's also a very strong value at CDL and at the University of California as a whole. Some of the other developments that we're engaged in around that, and I should say I'm here a bit under false pretenses because I don't actually oversee our repository services. Many of you may know my colleague, Catherine Mitchell, who probably should be on this podium talking about our institutional repository e-scholarship, which is our publishing and repository platform. It's a very important part of CDL's infrastructure and services. The University of California has an open access policy across the university, and CDL is the designated repository for the articles that are deposited as a result of our open

access policies. Personally, I've been involved in a number of transformative efforts in the scholarly communications and open access realm. I'm very involved in the SCOAP3 initiative where I chaired the governing council. I'm on advisory boards of a number of other open access initiatives and prior to coming to CDL, I worked at Harvard overseeing Harvard's licensing program, so I have a long history in licensing, and these are the transformations into what we hope will be an open access future.

I wanted to talk a little bit about e-scholarship and provide some context here. So, our e-scholarship repository is one of the largest ones in the country probably, and we support both Gold Open Access publishing, what we might call gray dissemination, so working papers, electronic theses and dissertations, as well as postprints that are deposited as a result of the UC open access policy. So, I'm going to be talking mostly in the context of our Green Open Access, our postprint deposit and how that relates to the version of record, but I also want to recognize that we actually support a continuum and a range of outputs of varying statuses of officialdom, if you will. We assign persistent identifiers to the postprints in our repository. We don't assign DOIs to them. We do assign DOIs to journals if our journals request them, but we otherwise assign other forms of persistent identifiers.

I did want to highlight a little bit at the bottom of the screen the kind of usage that our repository materials get. So, we've seen since our open access policies were enacted in 2012, nearly 1,000,000 downloads of the 45,000 articles that have been deposited under the open access policy from literally every corner of the globe, and in addition to those statistics, we have many anecdotes, many stories of people writing to us, graduate students, citizen scientists, thanking us for making this content available because they did not have access to it elsewhere. So, we have some real user stories about the value and the impact of providing this kind of access outside of the formal publishing system.

However, I think it is also important to acknowledge that institutional repositories and green deposit in particular face a certain number of challenges, so author uptake is not as high as one might like to see. Many institutional repositories aspire to be a comprehensive record of institutional output, but that depends on being able to actually capture all of that institutional output, and our track record is not terribly good in that regard. CDL uses the Symplectic Element System to harvest metadata and push that out to our faculty authors to help them with the deposit process and that has increased our uptake significantly, so we're seeing much more uptake since we've created some automated tools that ease the process for authors but absent those tools, self-driven author deposit does not have the kind of uptake that one might like, so whether we can really realize that aspiration of being a comprehensive repository of institutional research is really dependent on being successful in that.

Other issues, the versions that are deposited often don't link to the published version whether one calls it a version of record or not. So, again, because we harvest metadata via Symplectic from a number of sources such as Web of Science and so forth, we do in that case capture DOIs from that metadata and we are able to create that linkage, but material that is deposited just independently by our authors may not have those DOIs assigned, so we may not be able to link to those versions of record. And of course another issue with the postprint world, if one aim of this system is to facilitate transformation in scholarly communication, it's unclear how we're doing that if we're maintaining a parallel system to the existing publishing regime. So, there are some questions there.

But, at the same time, we should acknowledge the aspirations of institutional repositories. There is an aspiration, I think, to evolve from a secondary dissemination system to becoming a primary dissemination system very similar to preprint servers and the official publication stream. So I mentioned trying to comprehensively capture institutional output but also this notion of evolving to a primary dissemination system is one that I think animates certainly many of my colleagues in the library community. Can these repositories serve as a foundation for institutional assessment and faculty assessment? Will that help with deposit rates? Can it more transformationally serve as a foundation for formal peer review and publication overlay services? And there is a very active vision in our community, as many of you all know, about trying to develop overlay of peer review systems on top of repository services. When we think about what our faculty want, we certainly have many faculty who do aspire to a vision that I think is beginning to be realized more in the preprint server community of immediate publication that can happen without the kind of pipeline delays that the current publishing system involves. So how can we support that through the work that we do in our libraries?

So the questions that I bring to this; we do have multiple versions today. We have the postprint versions in our repositories and of course versions such as arXiv and now the many other preprint services that are developing. What is the impact of those multiple versions on the metrics that we use for evaluation both in libraries and in institutional evaluations, faculty evaluations, and so forth? I'll just take an example of usage. If we think about arXiv, for example, we know from some of the research that arXiv has recently done and also some of the work that we at SCOAP3 have done with arXiv that a great deal of usage remains on the arXiv platform even postpublication, and there is a perceptible drop in access on arXiv once something is published, so you can see what happens at the point of publication and the fact that the formal publication begins to capture a great deal of that traffic, but it's not all of the traffic. So there is a significant amount of usage that's happening on these other platforms similarly in our institutional repositories. There are 1 million downloads of those articles. When we in libraries evaluate journals for retention decisions, we're not capturing and we're not seeing that usage, even though those articles are in fact part of the publication stream. So how do we factor in the fact that we are not capturing all of the usage that in fact we use to make decisions in the COUNTER statistics that we get from publishers? I think we don't think about that a whole lot.

On the other side of the scale, some libraries are now beginning to look at the availability of Green Open Access whether it is in preprint versions or other forms of postprint dissemination on decisions about cancellation. Can we cancel a journal if a significant or a sufficient percentage of the content is available through other mechanisms? And that is beginning to be a very active area of study for a number of institutions, particularly as organizations like ImpactStory and OA-DOI make it possible that the work that one scientist is doing, for example, make it possible to actually capture a lot more data about the existence of all of these different versions.

Another question: how can we better link green deposits, our postprint deposits to published versions and other related outputs? So, how important is this to authors, to readers, to libraries, the issue of retractions, errata and corrigenda? If we don't have links to the official versions, are we in fact in danger of not capturing all of the changes that are happening to that scholarly record? Is there a way to aggregate citations' usage altmetrics in a way that will make that information more useful to authors and to others who rely on that information?

And then another question. As preprint servers take hold and publication becomes more continuous, I think both John and Monica talked about this, which versions do we in libraries need to preserve as part of the scholarly record and how do we do that? What's important to capture for preservation purposes?

And then finally some of the larger strategic questions about institutional repositories as a whole. Is this kind of postprint deposit, is this really a transitional mechanism at a waystation toward direct open access, and institutional repositories will at some point no longer really be needed for the purpose of postprint deposit because open access will solve the dissemination problem? Preprint servers will solve the open dissemination problem, or will that mechanism persist in some disciplines that will find it very difficult to transition to open access? Again, we are finding that the dissemination that we offer through repositories is providing some real value to a significant global community, so we don't want to withdraw that until it no longer becomes necessary. If it is transitional, however, should we maybe not worry so much about the versioning issues because they will eventually solve themselves as the world resolves to a more open access publishing stream?

And then that other question. Can institutional repositories and/or preprint servers in fact fulfill that promise of serving as a primary dissemination mechanism upon which formal peer review and publication services are layered? Can that bring down cost? Can the speed of dissemination of research bring down cost at the same time and help the academy retain and regain control of the scholarly communication stream? What would be gained if that development took hold? What would be lost if that development took hold, or is there a function that formal publication is serving now that is still needed and that would really be lost if preprint servers and institutional repositories became the only mechanism for dissemination? So, again, I look forward to discussions and comments.