A Simpler Path to Public Access Compliance

Howard Ratner
CHORUS

David Crotty
Oxford University Press

Jack Maness
University of Denver

Judith Russell
University of Florida

Howard Ratner, David Crotty, Jack Maness, and Judith Russell, "A Simpler Path to Public Access Compliance" (2017). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284316714

A Simpler Path to Public Access Compliance

Presented by Howard Ratner, CHORUS; David Crotty, Oxford University Press; Jack Maness, University of Denver; Judith C. Russell, University of Florida

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

Howard Ratner: What we're going to talk about this morning is all about how a pilot ran last year with CHORUS, which I am the executive director of CHORUS, and CHORUS is all about helping researchers comply with funder mandates. Last year we actually started a pilot between some publishers, and we'll talk to some publishers here, and universities with CHORUS from, basically, we started in the summer of last year and ended in the spring of this year, and arguably we've had a very successful pilot but I'm going to let them tell you all about it.

One of the things that I found particularly interesting about working on this pilot is CHORUS is all about working in a community and a community effort. We are a not-for-profit organization and we want to solve these things, so the best way to do these things is actually getting together, asking questions, bringing in some technology, which is where CHORUS comes in, and try to solve the problem. This morning we're going to hear from David Crotty, the editorial director in charge of journals policy for OUP. He will be speaking first. He'll be followed by Jack Maness, the associate dean at the University of Denver Library, and then Judy Russell, the dean of University Libraries for University of Florida will round it out for us. Each one of them will do approximately 10 or 12 minutes of talk and then we'll have an open mike at the very end for questions, and hopefully we'll get that all done within about 40 minutes. So, I'm going to kick it off and not waste any time. David.

David Crotty: Thanks, Howard. I am from OUP but I'm sort of speaking today with my CHORUS Board of Directors hat on. So, as more and more regulations on research outputs are imposed, we're seeing a continuing increase in the burden that is placed on the academic researcher. Given the high number of degrees that are awarded and the very low number of tenure-track faculty positions that are made available, research careers have basically become something of a buyer's market. We see universities sort of continuously increasing the demands that they make of their research employees. Researchers are required to do more and more beyond their actual research as sort of administrative tasks and teaching requirements are piled on. In the sciences, even tenured faculty function something like freelancers basically required to pay your own way through securing grants. If you want a salary or employees you're going to have to pay for these things yourself while the university rents you some lab space. And if you don't like it there are hundreds of people just as smart as you in line behind you for your job who would be happy to do all that extra stuff.

Now on top of that we're piling on a lot of new requirements. If you are doing a research project you need to take the time to preregister your experiments and go through some level of peer review before you have even done anything. So, clinical trials, for example, have to be publicly registered. As you start to do your research, some feel that you should be continuously making each result public, again taking time to write up each incremental piece, post it online, have it reviewed, and be part of a discussion around it. When you have completed the project, you then need to make early drafts of your write-up data public, and monitor and respond to any comments, and you need to publish the actual paper with all the hoops that one has to jump through to do that. Then you have to make the data behind the paper publicly available. You have to help others use it. If you really want to drive reproducibility, you have to write up and release your methodologies. Now, societal impact is increasingly seen as important, so now you have to become your own publicist. You have to promote yourself and the work via social media, and then at the same time people may be talking about you and your paper via postpublication peer review systems, so you have to monitor those, respond to any questions, any criticisms, and then of course on top of all that you have to comply with your institutional, national, and funding agency's policies around public access. So, you have to figure out what those are, and in a recent study more than half of researchers did not know their funder's access policy. You have to figure out what the right version of the paper is to post, you have to figure out where it goes, you have to figure out under what conditions and at what time,

https://doi.org/10.5703/1288284316714

and then likely every paper has multiple authors, for multiple institutions, with multiple funding sources, so you have to do this for multiple policies. So, that's a huge amount of time we are asking researchers to devote to things that aren't really what they want to do, which is research. Nobody goes into science or history because they really love bureaucracy and they really like filling out forms. So, further, if we see the purpose of all this research as benefiting society, then every second that we take a researcher away from the bench means slower progress. So, a lot of institutions recognize this and so they try and shift that burden off the researcher, which sounds like a great idea unless you are the research administrator or the librarian who is tasked with this work, which turns out to be both complicated and expensive.

ROARMAP now lists more than 880 national funding agency organizational and institutional policies toward providing public access to research papers. Each policy is a little different. Each has variable requirements and then again research is increasingly collaborative, so most papers have multiple authors with multiple sources of funding and multiple institutions and multiple countries of origin, so the number of possible permutations of 880 variables is close enough to infinity that most calculators can't even figure it out for you.

In recent years, surveys of academic libraries found that an average of slightly over four library employees devoted at least 10% of their time on open access initiatives. In its first year the RCUK open access policy saw a staggering amount of administrative costs for a fairly low level of compliance, but it's important to monitor these things because without careful monitoring and enforcement, even a mandate becomes an empty promise. We know that researchers are overburdened. We know they're short on time and anything they don't have to do they won't do. So when people come up with goals and they create these new rules to achieve these goals, very often little thought is given to compliance, so you end up with a policy that is either toothless and everybody just ignores it or a policy that is strictly followed but at great expense and great effort. The question then is, how do we ease that burden and make it easier for every stakeholder in the chain to deal with these complex requirements? There are different ways to approach this, but one advantage that publishers have is that we're sort of starting at the source, at least for the research papers. We know what is being published as it happens, and we can shape those publications to better meet compliance needs, but we're dealing with scale. Again, way too many requirements, too many institutions, too many funders to sort out by hand. So, just as we have done with pretty much every other aspect of our lives, as complexity increases we turn to automation and we take advantage of the sorts of approaches used to handle big data sets. We can no longer handle compliance paper by paper, so we need alternative ways to process it. And how we do that is through persistent identifiers or PIDS. I assume you are all familiar with the DOI, the digital object identifier. It's been around since 2000. It's a way of tagging an object and in this case an electronic document or journal article. The DOI for an article remains fixed over its lifetime, whereas things like the location of the article, the URL where you can find it may change. So, the DOI gives us a permanent identifier for the article that we can plug into our system for automating compliance, so, what else do we need to know? We need to know who wrote the article, to identify the authors. So, for that we increasingly have ORCID, the open researcher and contributor ID. ORCID creates a unique permanent identifier for each individual researcher, so now we can identify the individuals behind the paper and we can try to meet their compliance needs. How do we know what those needs are? Through what used to be called FundRef, now the CrossRef Open Funder Registry, which lists around 15,000 unique funding agencies so we can associate the paper with its funders, and then that lets us know what requirements have been placed upon it. So, the combination of those three things gives us what we need for a basic system. But to make a really effective system we could use some more information, and work is in progress on things like institutional identifiers. Where was this work done? Is it from the University of York in England, York University in Toronto, or York College in Pennsylvania? There is also a tremendous amount of development going on for licensing identifiers. Is this work under copyright? Is it an open access paper under a Creative Commons License? Is this the published version of record? Is this a preprint version? Is it the author's accepted manuscript version? So each identifier that we provide gives us more data and we can make more effective systems for automation.

We have used this idea of automation via persistent identifiers to power CHORUS. U.S. federal funding agencies require funded authors to make a version of their articles publicly accessible within 12 months of publication. Eight of those agencies and now the Japanese Science and Technology Agency and the

Australian Research Council is now in a pilot, but the others have officially signed on as partners to use the CHORUS system to drive compliance. Essentially the idea was let's take all of this infrastructure that we have already built and put it to use for public access rather than making the funding agencies build their own expensive systems, potentially taking money away from funding research.

So, basically we build this CrossRef Open Funder Registry into our article submission systems and the author identifies their funding sources as they submit the article to the journal, and that adds a funder tag to the article's metadata. Now some journals, by the shovel there, still mined their information from the text in the acknowledgments or the funding section of the paper, but the CrossRef-approved vocabulary is a better system because it removes any ambiguity in agency name whether people throw a comma in there, or an "of," or things like that. It can vary quite a bit. But once we have that tag, the article then is automatically made freely available in the journal at the appropriate time based on that particular funder's requirements. So, no manual intervention is needed by the author or the publisher. This information is also used to drive discovery and to monitor compliance. So, we've built our own search tools as examples, but more importantly we have an open API where anyone can tap into the data and enhance their search tools. And then a really important point about CHORUS that I don't think enough people are aware of is that every paper that goes into CHORUS is permanently archived so we take a copy, we put it into one of these permanent dark archives, CLOCKKS or PORTICO, so that if for some reason in the future the free version that the journal was supposed to make available becomes unavailable, this archive version comes to light and this absolutely ensures that perpetual public access is guaranteed.

Using our identifiers, we build dashboard tools for member publishers and for funding agencies, so that gives a quick sort of "at a glance" way to track compliance. So a funder can drill down, down, down to the individual article level, see what articles have been published that list its funding and when are they supposed to become freely available, and then check to see well, have they actually really become freely available? With these dashboard tools, we're now working with a number of university libraries on building tracking tools for institutions, and that is what you are going to hear a little bit more about from our other speakers on the panel.

And to sum up, basically we know that time is a researcher's most precious commodity and that they likely won't consistently do anything that they don't absolutely have to do for funding or career advancement. Automation is key. There are too many researchers, too many papers, and too many policies to do this by hand. Persistent identifiers are important and much work remains to fully establish them as standards and to fully implement them in our systems. These open standards benefit us all, so if you're not already doing so, I strongly encourage you to familiarize yourself with them and to do what you can to help drive their uptake. All right. Thanks.

Jack Maness: Well, thank you David and Howard and thanks to all of you for coming here this morning. I know there's a lot of wonderful sessions here in Charleston and I'm glad you chose to come here. So, I arrived at the University of Denver after about 10 or 11 years at the University of Colorado, Boulder in February. I joined when the CHORUS project was maybe midstream, something like that. The week I was there, the head of our IT department left. These are not related occurrences, I assure you. I did not chase her out. She took an AUL position somewhere else. But this gave me kind of an opportunity to ask some ignorant questions like, "What are we doing with CHORUS? What are our goals with it and what's the goals of the larger projects?" So if there's anything that I think I would like you to take away, it would be that you're going to see some differences between what Denver has done and what Florida has done and some similarities. These are due to scale and scope between the two institutions but also kind of the institutional context in which we are operating. This is a project that has given us data and allowed us to leverage it as we see fit.

A little bit about that institutional context in Denver. We are the oldest private institution of higher education in the state of Colorado. I sometimes like to point that out to my Boulder colleagues. We have them beat by a dozen years. We're still 100 years behind the College of Charleston, but pretty old for something in the West. We are an R2 institution, so we have a high research activity, but we're going to see some scale differences between us and Florida. About 11,000 students split down the middle between undergraduate and graduate students. Graduate programs are pretty heavily weighted toward professional programs. We have a lot of lawyers and CEOs that come out of the institution. We like to say that we are a liberal arts environment at a research institution. I think that is pretty true.

We have a beautiful urban campus surrounded by tree-lined neighborhoods of South Denver that kind of grew up around us. We have 700 full-time faculty, 500 part-time, and they're pretty productive. So, we are a research institution with about 500 articles indexed in the Web of Science annually.

If you look at our historical expenditures for research, they've been rising over time. These are again going to be kind of a fraction of what you see at Florida, but we are on the rise in a lot of ways. The associate provost for research likes to point out that the steepest rise most recently is when she took her current job. The previous rise before that was when she joined the institution and we really began seeing research expenditures when she was born, and if you knew her, I think she deserves a little credit for that. So, we need to take these mandates seriously. Even though we've got a lot of professional programs, we are a growing research enterprise and we want to keep track of federal-funded research and where the papers go.

There are two big challenges that I think you'll all be really familiar with but I would say are maybe a little pronounced at the University of Denver, and David alluded to some of these. There's a lot of faculty confusion and these might be pronounced at Denver, given the scale and the type of research we conduct, but they get really confused over different mandates at different levels, when there's authors from different institutions on the same paper and then identifying those acceptable versions for deposit. They're sort of endlessly confused about that. And then we have a bit of a challenge in populating our institutional repository. I don't think that's unique to us. It's a high-touch process. We run Digital Commons and we have dedicated a staff member to managing that repository only in the last year or so, and she is already incredibly busy. It's not a big priority for a lot of faculty members, and again they get confused about the rights and so we are interested in an automated process and thus our involvement with CHORUS. We only have two and a half programmers and their time is dedicated to a lot of other things, so we didn't want to throw them into some big harvesting or compliance tool.

We have chosen to focus on the dashboard that CHORUS has provided us and the data that comes with it and analyze where our faculty are. This is what the dashboard looks like for us or it looked like in May as we wrapped up the first round of the project. We had about 72 publications that we identified as results of federally funded research; 11% of them we could verify were publicly accessible on publishers' websites. About a fifth of them had reuse terms available, and then 95% or so had archival access through PORTICO or CLOCKKS or the like. Couple of things jumped out at this. We thought there probably should be more than just those 72 and that only 11% of them were available through the publishers' sites. Those are the two data points that sort of jumped out immediately. That was rectified over the summer. Sometime in July CHORUS made some adjustments to author affiliation data coming out of SCOPUS and we jumped into the 200s, 250s articles that we are tracking. That's what we think is probably right and accurate and we feel happy moving forward with that. It jumped up in terms of reuse terms available. The archive access went down a little bit and then the publicly verified access on the publisher site was about 14.5%.

In May, though, we became really interested in this unknown aspect of what was accessible on the website. I think we now know what the metadata problems were, the CHORUS people do, why we didn't have some of the information that we needed there, but we became really interested in this. And so we kind of decided to exacerbate our own problems with high-touch process and we decided to go analyze some of this ourselves. So, here's what we found. Out of 71—and one article must've been missing, I don't know why it's not 72, but we looked at 71 articles—62% of them were openly available on one platform or another. Now, it's important to point out that that doesn't necessarily mean that's compliant with the mandates, right? A lot of them dictate where it needs to deposit and what version it needs to be, but they are open and they're out there somewhere. Seventeen of those 44 are available in multiple places and there are 66 different iterations of those 17 articles in these places. So, that creates versioning problems, we think, and probably other problems. And that leaves 29% of those articles are not open in any way. That doesn't necessarily mean that it's not compliant. They could be embargoed and should not be available.

One thing that jumped out at this, and this is may be self-apparent in some ways or something that you all may be familiar with, is that there is a lot of self-archiving happening at DU. We didn't necessarily think that would be the case, again, given the kind of scope and type of research we conduct. We didn't think a lot of our faculty were doing this, but they are. If you look at that pie chart, the upper right light blue, those are the 29% that are not available

openly. The 8% is through the publisher. Again, that is the May data. Down in the lower right, that is what is supposed to be happening, that's PubMed, that's NIH-funded research available through PubMed Central, and the rest of the graph, some faculty member or graduate student involved in the research has taken some agency to post these articles somewhere, in Archive and the like, and then ResearchGate of course is one of the bigger ones at 18%.

We were fascinated by this and we decided we wanted to know a bit more about it and look into who is doing what where. We looked at access by platform by academic department to see if we can find trends and basically we didn't. The only trend that we could see is that there's a lot of red in every row, which means there's a lot of not openly available across all of the departments, so even though there is a lot of self-archiving happening, it's not like happening in some particular discipline or group of disciplines. Mathematicians and the physicists of course are posting in ArXive but not really at the rate you would expect them to or that I would've expected them to. A couple of other interesting notes I guess is the biological sciences do have a lot of that PubMed access but not exclusively and, again, it's, well, the Institute for Healthy Aging really seems to like ResearchGate, and then the Graduate School of Social Work has a lot of not open access and they're actually responsible for 29% of our research expenditures. We don't know if those are federal funds, so that's what we've got to look into.

What does this tell us about DU scholars? We're not entirely sure yet. We're going to go talk to them. We're going to conduct semistructured interviews with, I think, about 42 faculty members. This is based on some work I did at University of Colorado, Boulder where we talked to scientists in various disciplines about how receptive they would be to library involvement in research data management, and we're going to try to create personas or themes around demographics or disciplines, and who's archiving where, and what are they trying to get out of it, and how can we help, and how can we move them toward a more compliant sort of self-archiving behavior?

Beyond that, we now know what articles should be open. We want to work with the Office of Research in integrating some of this data with our IR, with faculty reporting and research profiles. We didn't use these metadata in our institutional repository. We see it is something that is both a content source and a promotional tool. We are open to doing that and moving forward, and I guess the point is that it's been really useful even for an R2 school. It helps us understand our faculty environment, faculty behaviors, overcome some of these challenges and work toward a more automated process in the long run. Thank you.

Judith Russell: Okay. So, as Jack said, we had two very different types of institutions and that was, I think, very important for us and for CHORUS that they had an R2 private, a large research-intensive public. One of the important things was trying to be sure that this project and this service would work for institutions of different sizes and type. And, as you can see, we are a very research-intensive university and very proudly just got ranked number nine among the *U.S. News & World Report* public university rankings, so something we have been striving toward and we're very pleased to see.

This is a university that is so large that with 53,000 students and thousands and thousands of faculty and even more researchers that there was no way for the libraries to manage this except with an automated process. One of the questions I get asked most often is how does CHORUS help and why would we go to CHORUS? Howard had actually developed a version of this slide and I'd seen it at a presentation that he did about CHORUS a number of years ago. He was talking in the early stages about how they were working with the publishers and the funders to put this together, and I went up and introduced myself and said, "Look at the people on this slide. It's a three-legged stool. At some point you also need to be including us, and when you're ready, call me." And he did. So we got started on this project and we are very pleased to have the engagement with them.

So, for us, why CHORUS? This bubble chart was developed for us and it shows the, I think, the top 14 or 16 publishing families in which our researchers publish and you can see that most of them are members of CHORUS, and so the idea for us that we could work with CHORUS and get a huge percentage of the some 8,000 articles that our faculty publish every year covered by one system in terms of gathering of information was really important to us, and so that sort of made CHORUS extremely attractive to us as a participant. And why would they choose us? Well, we are a large and very diverse population, so we have a fair amount of research in a lot of different fields and we are publishing in a lot of different disciplines, so we are very good cross-sectional representation of content for them.

These were our objectives. As we started to identify articles, as Howard said, or as David said, that we wanted to be sure that we were isolating for this pilot the articles that were federally funded and then that they would work on a dashboard that would help us track what had happened with those articles. Were they yet compliant? And then we really wanted to be working collaboratively on our campus with our Office of Research and parts of our faculty. We also looked to the idea that we might get to some additional kind of discovery options. Jack mentioned the possibility of importing this data into our IRs. We did this as a very rapid project and so we did not pursue that particular option but really hope that will come out as we continue forward.

I'll show you some screens from our dashboard and you'll see the similarities. This is the way ours looks and you can see that we had the same kinds of issues in terms of the volumes. We did have a surprising amount of compliance, I thought, given past history. But, there were definitely metadata issues that we have been addressing through this process to try to better understand what do we have to do with the data that is being brought in, in order to make sure that we get good and reliable results? This gives a picture of how the publishing split among the agencies and, as Jack showed, also kind of what are the ones where there's issues for us, where there are still unknowns to be resolved? So, part of the pilot was to identify how good are our sources? What happens is we pull information and compare and contrast information from different sources so that we can make the process better as we go forward. This one, which is very small and probably hard to read, gives some indication of the kind of data that is available, and one of the things that is very important to us is all of this data can be viewed on the dashboard and sorted in different ways and presented in different ways, but it can also be exported. And so we're looking at this as something that we will integrate with the systems at our Office of Research to help them reduce the number of articles that are noncompliant and that they have to follow up on and they are very interested in this. We didn't test that yet in the pilot but we have been briefing them on it and sharing the results with them and they're very interested and enthusiastic, as are the individual colleges and departments, about our ability to give them relevant information because the Office of Research is making it very clear that it is the responsibility of the dean, the department chair, and the PI to be doing compliance, and while they're certainly going to monitor it and be very conscious of it, there is also an expectation that they will make these decisions locally. So, again, just some different views that come up. They've been very helpful in working on the dashboard with different ways that we wanted to parse the data.

So, at the end of the pilot, those of us who were participating, both the publishers and the libraries and the universities, were asked specific questions, which you see here on this slide, about what we felt about the pilot, had to rank the level of importance against those initial goals, which I shared with you. The pilot definitely helped to identify gaps in compliance with public access mandates, and it highlighted the need to reach out to authors and work directly with the Office of Research and Scholarship on our campuses to close gaps. We felt that the expectation that the pilot would help facilitate institutional compliance was very much fulfilled and that it will definitely reduce manual processes, but that more time is needed both to perfect the data and to implement integration with it into the compliance functions at our institution.

The universities as a whole, both of us, stated that the pilot provided very useful metadata but we did not choose to ingest it into the institutional repository at this stage, but we do expect to do it later and as we move into more of a production mode. Also, although ORCID IDs and grant IDs are provided in the dashboard and in the reporting, the data sets were not available, and we're very interested in linking to data sets as another element of this, and we were also interested in exploring whether the universities might be able to actually enhance the metadata by, once we have validated or if we have ORCID IDs that are missing from the metadata of the publisher, can we actually pass information back up and enhance the metadata that is then available to everyone? So, that's another thing that we hope to look at going forward.

There's certainly more to be done to explore how information complements internal reporting, but it has definitely helped with the discussions with our Office of Research and we've been very pleased with the progress. I've got for my final slide here just a quick summary of what we expect to do in the future, so we are very interested in making sure that we capture the DOIs from the agency repositories. You heard David talk about the importance of those in terms of a permanent tracking of the item. We want to also gather those DOIs where we can for related data sets. We do want to be able to select articles, not just based on the name of the

institution, there's a lot of institutions with the words "University" and "Florida" in their name, so just aggregating that data has been a little bit of a challenge. So, we really want broader and more accuracy in that selection. We obviously would love to have more publishers. You saw the range of publishers that our authors are using, and the more publishers that come in, the more valuable the data will be. And this was acknowledged from the beginning, that the short-term focus would be on federal funding sources, but eventually we really want to look at all of the articles by our authors, regardless of funding source, and to be able to track private foundation, corporations, other funding sources as well as the federal. And as I say we want to be able to pass information back up and enhance the metadata so that it remains available more broadly to the entire user community.

This was a statement that I made at the end when were asked to make comments and I thought it was worth sharing with you because I think it is a really good summary of what the process was. First of all, it was a very rapid process and everybody really took that to heart and we met very frequently. It was a very open and engaging process. There was a great deal of back-and-forth and exchange of information. There was a lot of sensitivity to what was the data we needed? What were the formats we needed it in? How did we label things so we had a common understanding of what we meant by the different elements in the data? I think that it is a very true statement that it will only become more valuable as additional publishers participate and as we can integrate it with our administrators and researchers. So, I've been very grateful to the CHORUS folks for working with us and for allowing us to participate in this, and I'm very hopeful that the work that we have done has set a stage now for many other institutions, hopefully some of you, to join in this and benefit from it as well.

Howard Ratner: So, before we take some questions, I just wanted to mention something. I mentioned that the pilot actually ended in the spring. We're actually doing a soft launch literally this week of our new institution dashboard service. I'm not here to show it to you or pretend that I have all of your information quite yet, but we are soft launching it now and if any of you are interested, come up to the podium here and we'll be happy to explain it to you and talk about the terms, which aren't incredibly expensive. Let's say it tops out at 5K, so is really very reasonable. But we've got 10 minutes of questions. My speakers did an amazing job of keeping time, so thank you very much, and if anyone has a question please do step up to the mic and we'll be happy to cover your questions first.