The Secret Life of Articles: From Download Metrics to Downstream Impact

Carol Tenopir, Professor, University of Tennessee

Lorraine Estelle, Director, Project COUNTER

Wouter Haak, Vice President Research Data Management Solutions, Elsevier

Meg White: Welcome. This is part of our Neapolitan sessions here at the Charleston Conference. You are in ballroom number two for The Secret Life of Articles, so if that is not where you’re meant to be, now is the time. My name is Meg White, I will be your moderator for the session, and I’m going to quickly turn it over to our wonderful panel of speakers. Just to remind you, if you have questions we are going to do our best to reserve some time at the end of the talks for those questions. Don’t be shy, we will invite you to approach the mic and let’s really get as much out of these three wonderful folks who are here to share with us today by giving them our questions at the end of the session. So, with that, we are going to get started. It’s exactly 10:30, and I will turn it over to Carol for The Secret Life of Articles.

Carol Tenopir: Okay. Thank you, and I know that in the earlier session, the acoustics were a little problematic. I think this room is a little better, but if you have trouble hearing me or if I am swallowing words or wander away from the microphone as I sometimes do, raise your hand at me or make a sign and I will remember to come back. For those of you who are regular Charleston attendees, and I know there are quite a few, some of you may remember that last year we introduced this “Beyond Downloads Project” at a session and we were in the middle of collecting data. The project is complete now, at least the research aspect of it is complete, so I’m going to present some of the research, at least some of the key takeaways from the project on article sharing. We are really looking at the idea of the impact, but more at the instances of sharing, so I am kind of digging deep into article sharing. I’ll give you the results of the research project, and then Lorraine and Wouter are going to tell you about some of the practical implications from each of their perspectives, and some of the next steps that both COUNTER and Elsevier are doing. So, I am doing the research part and they are doing the “what happens next” or “what are the implications.”

I do need to start by acknowledging that this was a wonderful team across kinds of organizations and an international team, and I want to thank all of the team members. There are some others that are here: Hazel Woodward is in the front, and I think Anthony Watkinson is walking over here, and those of you that I don’t see right now, apologies, but we had a really quite wonderful team working together. So, at the “Beyond Downloads Project,” our real motivation is what we wanted to know what are Project COUNTER reports missing? What are those downloads reports missing in terms of use? And we really wanted to look at this issue of how much do scholars share articles once they’ve downloaded something? How do they share? What do they share and how much does that measure—is there a way to measure total use of an article, not just downloads, which we’ve gotten down really well thanks to COUNTER and lots of the extensions and things that libraries and publishers have done. But the assumption is that only tells part of this story now that we know that there’s a lot of sharing going on. Are there ways to calculate and measure that sharing and what is—is it possible to look at a more complete measure of value because a lot of our applied measures of value of library collections, applied measures of journals or articles, are based on amount of use, and with sharing if we’re not measuring that complete use, how do we measure value in a more complete way? So those were kind of the motivations.

Another motivation was to initiate discussion of these issues, and I think that has certainly succeeded judging from the size of the crowd last year and this year as well, and the discussion will
continue, I’m sure. We did use several methods in the research portion of the project, and I’m going to talk about where we started, which was the focus groups and interviews. We did focus groups and interviews in both the US and the UK. We had a total of 29 participants, academics but ranging from senior academics to PhD students so we had a good mix of ages. We had less of a mix of subject disciplines. We are really science- and social science-heavy, so I’m not going to give you a lot of conclusions about humanities scholars. We are really focusing on sciences and social sciences among our populations. The focus groups helped inform an international survey, which we had responses from 69 different countries; about a quarter of them were from the US, but the other three quarters were spread out everywhere, and yes, we really did have 1,000 responses. So two days before it was supposed to close we had 996 or something, and I said “let’s keep it open until least we have 1,000.” As soon as they got to 1,000 we closed it because the math is really easy when it’s 1,000. So, it looks a little odd. We don’t have 1,000 to every question, but we do have 1,000 responses. Again, there is a good range, excluding humanities—we had just a few humanities, but it is a good range. About a third are from sciences and the rest are engineering, social sciences, so that is our group that we’re looking at.

Good age range; they range in age from 18 to 83. The average age was 48 and about a quarter are 39 and under, so they do skew a little on the older side in terms of 40 and over. As you can imagine, we recorded in the interviews and the focus groups, and the survey had an incredibly wide range of opinions and behaviors when it comes to article sharing. Different motivations, different behaviors, different opinions, etc. I’m going to give you a feel of some of that range but also a feel of the things that I think are the main messages when you look at—or the main takeaways when you look at all of these results and what do kind of the average or indicative sorts of behaviors and opinions tell us.

So, five key takeaways from my part. The first is that sharing of scholarship is a means to an end, and the end is not the sharing. The end is the scholarship or the science or getting the work done. A little bit about function or purpose or motivation of the sharing drives the form or format of what is shared. The third is that e-mail seems to still be the most popular or is the most popular number one choice, but version matters, and I will show you that in a second. The library is still key when it comes to initiating the sharing, but we’ll look and see how the library fits in there.

Okay. So first, means to an end. This is a word cloud from the question how do you feel about your work being shared with others? And you’ll notice there are some very interesting words in here like happy, good, free, fine. We had 690 opened comments in the survey, and I must share that researchers and teachers and scholars like to share. It makes them feel good. They like to have their material shared and they like to share the works of others. The vast majority of the comments are positive; only a few expressed reservations and most of the reservations came with, “It is my work being shared. I want to make sure that the proper citation is included so that I can credit for it.” So, it isn’t negatives about the act of sharing. Very few comments in terms of worrying about whether it is legal or not legal, or whether it is appropriate or not—“it makes me feel good to have my work shared and it makes sense.” Again, we have a lot from both the focus groups and the open-ended comments. We have a lot of really rich quotations and things that show what people are feeling, and, again, scholars on why they share: they find it useful to further their scientific and academic discovery, to facilitate collaboration when it’s their work, to further their own work. Sharing is often initiated by others who are interested in an author’s work and ask for something, and an author will share it or they ask for something of somebody else and they’ll share it and again; this message is that sharing is a natural part of scholarship. They share materials for good reason. It is part of what they have always done; it is just easier now and they look for the ways that make it easy.

The second most common reason is if somebody asks for a copy of an article, either of their own article or of somebody else’s article, the person who is asking has a reason to believe that the person they’re asking of may have a copy, and the
idea is they’re trying to get around a pay wall or if they don’t have access. So this idea of helping a colleague fulfill an information need was mentioned by many. I share because one of my colleagues doesn’t have a good collection or does not have access or the cost per article is too high, so I’ll share whether or not it is strictly legal. And it was interesting in the focus groups to have people say, “Oh, I just thought it may not be legal but, what the heck, I’m going to do it anyway.” There are a couple messages here: convenience, fitting in with their work, and helping others are all more important than the letter of the law. We also asked about embargoes, about what are reasonable embargoes, and the longer the embargo the less reasonable people thought, although there wasn’t much—there weren’t many fans of any length of embargo, if I put it bluntly; but with the longer embargoes we got a lot less people saying that they agree.

Scholars who work in a research group are more likely to share their articles or the works of others than those who work alone. So, many of the scholars say that their sharing will increase, or they expect it to increase, or it has increased over the last few years. If you put that together, they think it will increase, plus the tendency of work to be done more often in research groups, the tendency of more co-authors on papers—this idea that sharing is going to increase anyway but research groups are going to increase, collaborations increase; you put those two together and I think you can pretty safely say that the instances and the amount of sharing will increase into the future. And add that to sharing becomes easier; I think that is a pretty safe assessment that people are going to share and find the need to share and find ways to share.

That leads me to the next point that is function drives form. We assume that sharing meant sharing the full text, but it doesn’t always mean that. It became clear fairly early on that there are two main methods of sharing. One is to share just the citation and most people do that; they say when that researchers who work in their group or who they’re sharing with have easy access to the articles, so then they would likely share a link or a citation. Sharing the full document in PDF is still the favored form of sharing; sharing of full text is most common when access may be an issue or when it’s easy to do. And so we found both. What’s interesting about this when we are talking about calculating amount of sharing is when you share a citation, your downloads may go up. When you share the full text your downloads may go down, so there’s a little bit of balancing going on here. The instances of sharing full text as reported were still higher than the instances of sharing citations, so it doesn’t quite balance out but there is a bit of bouncing. One thing drives usage statistics up; one thing drives usage statistics down. When we ask more details about their most recent shared link or reference, we found that when they’re sharing by e-mail they said they shared an average of about 11 times, 10 to 11 times, per download, so for every download it’s about 10 more times the link or reference so that might make it go up. For teaching, learning management software is the preferred method of sharing and that’s slightly higher, about 14 times per download. We asked also about full text, what’s the average time of share, or how many times did you share the last full text you downloaded, and it’s about the same. They share via e-mail most commonly again about 11 times, and then full text is more likely to be shared on research social networks for example about 11 times.

So, not surprising now after what I just said, that e-mail is number one but what surprised us is how much it is number one. You know, academics are not always the first to change. They do change if things are easier, and right now for most of them e-mail is the fastest and easiest. Remember sharing is not the main purpose. Sharing is to get to the main purpose of learning or helping their research, so if it is easier to share for research, for e-mail, they’ll do it. About 74% said that e-mail is their preferred method of sharing, followed by cloud services like dropbox and then internal networks. Notice that things that you would expect, systems designed specifically for sharing, don’t make those top three; those come afterward. Both general social networks and research social networks—doesn’t mean that they’re going away but a lot of the comments expressed the idea that there are so many choices
right now I’m not sure it’s just easier to send somebody something when the choices settle down in the research and social networks, and that becomes easier then we expect these to get more popular. It doesn’t mean that they aren’t using them, because clearly they are—it’s just that we asked, “What’s your preferred method?”

For teaching, e-mail again by far number one but in here number two is learning management systems sharing with students and then internal networks number three. Again the research sharing networks, both general and specific, come after those. However, if you look at the number of times you share and your preferred method versus the number of people you think are reached, you can’t just do a calculator so the number of people who were reached becomes in the hundreds or thousands people estimate depending on the method, so if they want to reach hundreds or thousands, research social networks are going to reach a lot more people than e-mail will, for example, which is a more one-to-one sharing—general social networks in the hundreds as well as internal networks for more people. So, if you think of, instead, one download equals 10 shares, 10 times I share, then I might share 10 times if it reaches hundreds, and I have to say times hundreds if I’m going to look at reading, so 1 x 10 x 100 is potential readership or potential people that this particular article is reaching, so the numbers get bigger depending on the method quite quickly.

Version matters. There are some people in this audience that might not like to see this but when sharing articles, either your own work or the work of others, which version do you prepare to share? Guess what? It is the final published version. We had a lot of people who commented that it gets too confusing to have various kinds of preprints, accepted manuscripts, and all of these other versions. I want to share the one that will be cited and the one that we know people will find again, so the final published version is what people prefer and want to share overwhelmingly of both their own and others’ work. And again lots of comments about this—this first comment—a main point is that I want to share the version that has the publisher and journal stamp on it. I want to show that it is the final version, and I particularly like the second quote, “I’m a little fish let them come for me.” There was a fair amount of this kind of like, “I’m just doing my job. It’s not having that much of an impact and I’m going to do it, just come if you will.” There’s another quote I wanted to read that’s on here. But, content providers in general have to realize that the relationship between seller and buyer has fundamentally changed in the last 20 years. In order to get someone to pay you for something they can get without your permission, you need to (1) make the process of obtaining your product legally at least as effortless as the process of obtaining it illegally, and (2) make it so cheap that it’s not worth pirating. Some very opinionated but perhaps wise scholars here and perhaps since this project was done, some of you are aware of the hashtag “#icanhaspdf,” which is this kind of copyright freely, share copyright papers, and the article it describes. The first one that I saw, academics have found a way to access insanely expensive research papers for free, so there is this whole pirate bootleg kind of community.

So the last is the library—is key, sort of. We asked about the average number of downloads for your last research project and for your last teaching term. Put these together, assume two terms per year, and you put one project you have, maybe 115 or 120 downloads. Where those downloads come for research the library and publisher is still the number one, teaching is still number one but less so, so this is where the downloads are coming. All of these others are coming from other places and just because the library downloads is a source of download doesn’t mean that is the complete source of count of use, because we asked when you share it, when you return to that article, when you reread an article, most of them do not download again. Only 10 to 12% say they download again. What they do is they go to a saved copy. People are saving and so those download counts don’t get to rereading, and they don’t get to sharing, except for those few that links are shared.

Okay, so, let me wrap up my part and then turn it over to the others. Again, sharing is a natural kind of behavior of science—that download counts
underestimate the true amount of use of articles but sharing is done because it helps disseminate results, support research, and helps people share in ways that are convenient, and they share because it is convenient. They don’t want to stop and think, “Should I be doing this? How do I do this? I just want to do it.” And any policies or measures from publishers I think must fit those preferences and the likely behavior, because you don’t want to have punitive kinds of policies. You want to be part of that research conversation and not seen as an enemy but seen as a partner, and, again, I think downloads are really just the tip of the use iceberg. Let me turn it over now to Lorraine Estelle who is the new director of Project COUNTER.

Lorraine Estelle: Good morning, everybody. Before I start I would just say that I’ve been with COUNTER since June and I’m speaking to slides which are not my work; they are the work of my predecessor, Peter Shepherd, who was the last director of COUNTER, so I just want to make clear that I’m not plagiarizing his work. I’m giving him credit for it. So, in terms of COUNTER, what we have been looking at is calculating sharing metrics. Is it possible and what are the approaches that can be taken? And clearly we could see from the project that there was a very wide range of sharing mechanisms and all of them different, not comparable, and clearly we’re in an area where there are no standards yet. And we also found from the project that the data collected is likely to go out of date very quickly. Two approaches can be considered when we’re thinking about calculating sharing metrics. The database approach—employing usage data from publishers and citation data where such data is comprehensive and reliable. So there are limits there because that data has to exist. The other approach, survey-based approach, is using an online community as a basis for monitoring sharing behavior over time. Again, we have to think about how frequently that would have to happen and therefore how feasible that would be. So, really just looking at those problems in a bit more detail, data on article sharing by formal and informal methods will be difficult to collect reliably owing to the large number of channels used and as I say the lack of standards at the moment. The mix of sharing methods used varies from discipline to discipline and from year to year. And if we were to use a databased approach, really it would seem to make sense to supplement that with the surveys from time to time.

So we then really come to the question is a sharing calculator possible? I think we can say that it is difficult, it’s early days, and there are problems to address and in the project we can see that there are outliers that really skew the results, and demographics matter a lot and make a big difference. So, an approach that we have considered here is a range of sharing with a lower—using the lower range and upper range, and here in the table you can see this approach and I think it is very useful looking at the range across those different categories that Carol was telling us that she found that the researchers have been using and again. We can see there right at the top e-mail is the winner for sharing with the highest range. So that may be one approach that we could build on.

Carol, again, pointed out demographics do matter and it will be no surprise to hear that the younger researcher, the less formal their methods for sharing. And, again, I think what Carol said the researchers of the slightly more mature years—more formal sharing—they post class but when they do they have greater reach. And of course disciplinary differences, social scientists share more via social media so they have really embraced it, the social scientists. So when we think of the original question of thinking about a calculator, can it be done? Well any estimate on a number of times an article is shared per download depends on many things. There are lots of things going on there. An alternative to the exact calculator might be as we said that range of sharing for those different categories, and this range could in turn be combined with download numbers to estimate a level of factual posts downloaded usage.

I am, when we’re talking about sharing, also going to use this presentation as an opportunity to highlight two of the projects that are going on in this sort of area of COUNTER working with CrossRef. So, we have the DET and the DUL, and if you get them confused, well you’re joining me
there. I get them quite confused. So the DOI event tracking (DET) initiative really looks at creating an open common infrastructure to track activities around DOI’s, recording user content interactions, so articles but also other items at the DOI level. The distributed usage logging (DUL) initiative—again really very similar—similar infrastructure but a peer to peer system for the exchange of usage data on user content interactions. So, for example, you might have the exchange of data from a publisher’s platform at an institutional repository, and there you would see the use of an article in those two places. So, I think those initiatives are something we’re working on and which will also give us that wider picture. So, thank you very much, and now I’m going to hand it over to Wouter, who is going to tell you about his project.

**Wouter Haak:** Thank you. I guess hopefully everybody can hear me and you probably immediately note my accent. I’m not American. I’m Dutch, so I’m sorry for that already. I’m also a product person, so that means that my job is to actually translate insights into products that lead to results. That’s my life, that’s what I like doing, and I particularly like doing that online. So, I’m not going to talk about policy implications. There’s actually another discussion later today that will discuss how sharing affects changing publisher policies. I highly advise you to attend that session as well if you’re interested in that. I’m going to talk about products and researcher behavior and what this may mean for you as a librarian from my perspective being a publisher. So, let me start with Mendeley because as you may know Mendeley was acquired by Elsevier a couple years back, and Mendeley is a collaborative sharing platform. Let me just give you some metrics of what sharing on Mendeley is doing. So, sharing is growing. As you can see from the start of Mendeley sharing has grown tremendously, double-digits every year. So, obviously Mendeley is a popular tool, and it is growing in popularity, but aside from that it is really the group, and this is just one picture—this is groups. They’re growing tremendously, and a Mendeley group is a group where you invite others to privately share references and full-text documents with others. It is very popular as you can see. And then the articles they’re sharing—that is growing even faster. So as you can see the number of articles shared on Mendeley with other researchers is growing tremendously. Now, to Carol’s research, this actually is the tip of the iceberg because most researchers share by mail. Now, our challenge is we actually wanted to make it so easy to share with others on platforms like Mendeley, and actually I’m broadening the scope here because it could be Mendeley, it could be Papers, it could be ReadCube, but of course our focus is on Mendeley. We want to make it so easy that researchers will prefer to share on collaborative platforms that we can then use to improve their research and to help you as a librarian.

So, what I would like to talk about is three actions we’ve done. So, first of all, we’ve created a “My Research Dashboard” where we now feed back to the researcher what’s happening with their article. Secondly, we’re now testing in institutional dashboards, where we’re giving back to you as a librarian metrics about what’s happening with the articles including sharing and collaboration, and thirdly we’ve joined forces with other publishers to create a movement to kind of outline principles, what constitutes simple, seamless, and good sharing, and all three I want to highlight quickly.

So, first of all, let me just show you some screenshots of the “My Research Dashboard.” So, and I guess maybe in the back you can’t completely read this, resolution is always an issue, but basically this tells the researcher how many views, how many citations of the articles that he has published with Elsevier and non-Elsevier. This is the full world of your research, and why is this important? Particularly, early-career researchers have a hard time kind of establishing how they’re doing because they need to wait for the first citation or they need to judge by the impact factor of the journal how well they’re doing, as you know the first citation could take as long as two or three year, and the impact factor of the journal is a gross kind of collection of an aggregate result. It doesn’t say anything about your success as a researcher on that particular research. So, that is why we’re starting to feed back these signals immediately after you publish your article to get
earlier signals about how you’re doing, who’s reading you, and what they are doing with your article—and that’s also why sharing is so important.

So, this is feedback from Mendeley. For example, this article has not just been downloaded and viewed many times but it’s also been added to a Mendeley library 548 times, and by the way, out of those times that people added to their library, 16 people decided to share it in their group. Now, that is a very strong signal and very strong, let’s say, following of what happens with articles. First you glance at it, then you read it, then you deep read it, and then you decide to share with your fellow collaborators. This is what we’re doing with the sharing. We know that sharing is important, so therefore we thought we need to feed this back to the research community. When we presented this to a group of librarians last year, the first question we got is, “Hey, can we get this?” We were like, “Yeah, why not?” So, that leads me to the institutional dashboard. We believe that you as librarians should get better controls of what’s happening with your authors, your readership, and what’s happening around your institution, so we’re creating a—this is actually a markup of a prototype that is now live at a couple of universities we are working with. We are creating a prototype where we again give you feedback about usage and statistics and stories about what’s happening at your institution. But sharing and collaboration is important, so, for example, this a picture of the collaborations that happened at your institution internationally in the last three years, so we’re trying to zoom out from the article and researcher level to an institutional level and help you understand what’s happening with your authors.

Having said that, this is explaining why we believe it’s really important to feedback back sharing to researchers, to institutions, because it’s happening and it’s increasing as the research showed. Now it’s an incomplete picture, so, for example, we can’t track e-mail usage and we can’t, we as Elsevier, can’t for example track what’s happening with our documents on, for example, ResearchGate or Academia. But we’re talking to those platforms and we’re talking to other publishers, and what we said is actually, “Sharing isn’t bad. Sharing is actually a good thing,” usually, and that is actually also what your research showed. Usually, researchers share amongst fellow researchers that are actually—we like them to read those articles, so we don’t care that much about entitlements in research collaboration groups. We actually want to enable that as seamless as possible, so what we did is we got together with a group of publishers—and actually I happily see a couple of them sitting in this room—driven by the STM to create principles for sharing for social collaborative networks (SCNs), and this is difficult, right, because what is an SCN? We said well, probably, let’s use a broad definition. We want to apply principles for sharing of articles on social collaborative networks, and what we did is with that group and together with the STM we came up with a set of initial principles and we started a consultation, and that was back in March, and we asked the whole community including you for feedback, and actually a lot of feedback came so what happened is 50—actually I think it was more than that—52 people gave feedback from librarians, institutional repositories, SCN’s, other publishers, and actually also we got a lot of response privately. All of those responses were analyzed and we actually said, “Okay, that is actually really valid feedback.” So, we changed the principles based on that consultation. And based on that, we have now come up with a new set of what we call “voluntary principles” because anybody can sign up to it where we say first of all publishers actually facilitate the dissemination and discovery of their authors’ scholarly articles. That may sound very basic but is actually really important. It means that we actually want to help authors share their work and researchers share their insights with their collaborators.

Secondly, we say sharing should be allowed within a research collaboration group. It is really important. Sharing should be allowed. There should be no doubt about that. Any publisher that signs up for these principles allows sharing in private collaboration groups. Those groups can be of subscribers and nonsubscribers, so we’re actually okay with researchers that are entitled to share with other researchers that are not entitled.
It’s very important. That may sound very trivial; it is actually—a lot of the feedback that we got was, “That’s happening all the time already,” but I did talk to some librarians and they’re worried to endorse some sharing tools because they’re like, “Well, we know that sometimes illicit sharing is happening, and actually through these policy changes that we’re going through we are actually allowing this.” So, we’re allowing you to endorse sharing and collaborative tools and to help your researchers do this. We even say, look, it can be researchers sharing with other researchers and it can be researchers sharing with the wider public. That’s okay as well. What we do say is that commercial research is different. Commercial researchers from highly paid organizations—there’s different tools, different mechanisms to facilitate that. There are already several platforms facilitating that. We do want to kind of follow those mechanisms. The only thing we’re asking back in return is that those platforms that endorse this as well, we’d like to follow and track the usage. We would like to work with COUNTER and CrossRef to track the usage and sharing going on. Why? Because we want to feed that back in those dashboards that you’ve seen and in the reports to institutions. We want to use their signals to improve our recommendations engines.

I was really impressed with the speech this morning by James O’Donnell with the Jedi sword who said your task as a librarian should be to help researchers with discovery. Well the same holds for us. We’re actually in that boat together. Our engines, our discovery engines, are power-fed by the user signals. That is why this is so important that we track those user signals because our recommendations are worth nothing if we don’t understand what’s happening with the articles, and it’s our responsibility to give good recommendations to make researchers read less because they need to read the proper stuff.

So, we are actually working together on standards with all of the standards organizations to make this more of a standard mechanism, and we’re changing our policies to reflect these principles. So far only, I should say only, actually in publisher time frame it’s actually amazing—publishing is not that fast. Three publishers have changed their policies, so Nature, Springer, Elsevier, and Wiley, but that’s a large portion of articles out there. So, I talked to other publishers. I hear rumors that more publishers will follow so I think this is a good thing. So what I’m really asking you as a library community is do you feel comfortable with this? Do you think this is the right direction? Do you feel that you are supportive of this? If so, I would really encourage you to endorse this movement because we want to enable sharing. We want to help sharing. You can help me by endorsing this or leaving your feedback on the STM Association site. I’m sure this presentation will be shared at some point, but I think the simplest way to do it is if you Google for STM principles or STM consultation, you’ll also land on this. The importance of this I can’t underestimate because research sharing is increasing. COUNTER and CrossRef are actively making this happen. What I would like from you is the support so that we can clearly communicate to all of the platforms and publishers out there that this is important. So, please help me with that. Thank you.