The OA Interviews: Judy Ruttenberg, ARL Program Director for Strategic Initiatives/Co-Director of SHARE

RICHARD POYNDER

27th October 2017

To go directly to the Q&A please click here

When the open access movement began it was focused on solving two problems – the *affordability* problem (i.e. journal subscriptions are way too high, so research institutions cannot afford to buy access to all the research their faculty need), and the *accessibility* problem that this gives rise to.

Today, however, there is a growing sense that what really needs addressing is an *ownership* problem. Thus where in 2000 The Public Library of Science petition readily acknowledged publishers' "right to a fair financial return for their role in scientific communication" (but sought to "encourage" them to make the papers they published freely available "within 6 months of their initial publication date"), today we are seeing calls for research communication to become "a community supported and owned enterprise" outside the control of publishers (see also here).

The key issue today, therefore, concerns the question of who should "own" and control scholarly communication, and more and more OA advocates are concluding that it should no longer be traditional publishers.

This change of emphasis is not surprising: as legacy publishers have sought to co-opt open access and bend it to their own needs, it has become clear that, since it is leaving legacy publishers in control, OA is insufficient on its own – because for so long as publishers remain in control the *affordability* problem that drove the calls for open access will not be solved. (More on this theme here).

What gives this issue greater urgency is a new awareness that legacy publishers are looking to leverage the control they have acquired over scholarly content to dominate and control the data analytics and workflow processes/tools that are emerging in the digital space – a development that could usher in a new generation of paywalls, and lock the research community into expensive proprietary services.

Issues

This then is the *ownership* problem facing the research community. How is it playing out in practice? The interview below with Judy Ruttenberg, Co-Director of SHARE, surfaces the issues well I think.

SHARE (the SHared Access Research Ecosystem) was launched in response to a 2013 memorandum issued by the US Office of Science & Technology Policy (OSTP) directing Federal agencies with more than $100M in R&D expenditures to "develop plans to make the published results of federally funded research freely available to the public within one year of

publication and [require] researchers to better account for and manage the digital data resulting from federally funded scientific research."

SHARE was an expression not just of librarians' conviction that publicly-funded research should be freely available, but an assertion that it should be universities that provide access to it. As Ruttenberg puts it below, SHARE was founded in the belief that "university-administered digital repositories should be the mechanism by which federal agencies provide public access to funded research, most of which is conducted in universities."

In its 2013 concept document SHARE suggested that this could be achieved by depositing publicly-funded papers into institutional repositories (IRs) and making them freely available by "adopting a common, brief set of metadata requirements and exposing that metadata to search engines and other discovery tools."

This it added, would "federate existing university-based digital repositories, obviating the need for a central digital repository and leveraging the considerable investments already made by universities and their libraries over the last decade."

Interestingly, SHARE proposed that publishers help do this. As part of the process, the concept document suggested, journals would "submit an XML version of the final peer reviewed manuscript (including the abstract) to the PI's designated repository". Failing that, SHARE added, the author could submit "the final peer-reviewed and edited manuscript accepted for publication (including the abstract) to the PI's designated digital repository."

The document added, "Upon ingest of the article, designated SHARE repositories will make abstracts and metadata available to commercial search engines (e.g., Google, Google Scholar, Yahoo, Bing, etc.) and other discovery tools."

As a result, it said, global access would be provided to "the corpus of digital repository content, both full text articles as well as the associated data sets", and thus meet the OSTP requirements.

Publishers, however, had a different scenario in mind. As the incumbent distributors of published research, they take the view that they are the natural providers of access to scholarly research, and should continue playing that role even in an open access world. To that end, they responded to the OSTP memo by launching CHORUS (Clearing House for the Open Research of the United States). As Ruttenberg puts it below, "CHORUS was launched as (and in my understanding remains) a mechanism for publishers to maintain control and stewardship of publicly funded research articles by opening select articles to public readership in the context of the overall paywalled journal."

In other words, CHORUS and SHARE were competitive. True, both groups tried to downplay this, but that is how the two initiatives were rightly perceived.

Be that as it may, observers were quick to point out that publishers had a distinct advantage, not just because they already own and operate a scholarly publishing infrastructure, but because they are the originators and owners of the all-important Version of Record (VoR) of scholarly papers. These two things mean that (in theory) publishers are able to make any article they publish open access by simply flicking a switch on their platform. As such, they argued, they can provide public access without any duplication of effort being required. And

by collaborating with one other through CHORUS, they added, they could provide funding agencies with an "information bridge" to enable them to monitor compliance.

In retrospect, we can see that SHARE would always have struggled to compete with publishers in meeting the requirements of the OSTP memo, not least because although most universities have installed an institutional repository, they have struggled to fill them with the target content, and they have struggled to make them interoperable (which would be essential for SHARE's proposal to deliver on its promise). It has not helped that repositories have been seriously underfunded by their institutions.

As *Library Journal* pointed out in 2013 (quoting former repository manager Dorothea Salo), even making "'a relatively simple networking of the existing ragtag gaggle of institutional repositories,' let alone 'a highly complex re-visioning of how the entire research academy deals with digital materials,' on its proposed timeframe, given current IR software and staffing" would prove a nigh impossible task.

And so it proved. SHARE was not able to compete with publishers as it had envisaged. For this reason, it has had to regroup and reimagine itself.

Meanwhile, CHORUS is winning round funders. True, some agencies haven chosen to partner with PubMed Central for hosting purposes (i.e. NIST), rather than rely exclusively on publishers. Others have chosen to use their own platforms (e.g. DOE). Nevertheless, last year it was reported that six agencies had signed up to use CHORUS, and we can surely expect that all mandated funders will link to the VoR on the publisher's site even if they host the Authors Accepted Manuscript (AAM) themselves (as the Department of Energy decided to do). And they will all surely have to rely on the Elsevier-donated FundRef if they want to be alerted when a paper they have funded is published.

Challenges

To get a clearer understanding of why publishers won the battle let's review the three main challenges SHARE faced in more detail. First, despite the long-standing existence of the Dublin Core Metadata Schema, developed in 1995 to allow web resources to be described, and despite the 2002 OAI-PMH initiative, which developed a dedicated metadata standard for repositories to enable them to interoperate, effective networking of IRs has remained more promise than practice, not least because many repository managers implement their metadata in a haphazard, inconsistent and lacklustre manner.

Second, authors have proved highly resistant to depositing papers in their institutional repository, despite a plethora of OA policies requiring or demanding that they do so. For this reason, the task generally falls to librarians. Since librarians are not the authors, this can be a very difficult task. For instance, to deposit a paper on behalf of an author a librarian will need to know when and where a paper funded by an agency has been published and, if so, which agency funded it. It can be extremely difficult establish this information if you are an intermediary, especially if the author is unwilling to co-operate.

At one point, SHARE considered creating a central portal where papers could be deposited offsite, and then routed to the relevant repository. But this was eventually considered impractical, not least because it was felt that universities would be uncomfortable with such a

plan. In any case, those responsible for the deposits would have faced the same problems of mediated deposit outlined above and below.

Third, intermediaries need to have the necessary rights to deposit papers on behalf of an author. Since authors submitting to subscription journals are required to transfer ownership of their work to publishers as a condition of acceptance these rights invariably belong to publishers. And since publishers believe that it is they who should be providing public access most prohibit anyone else from providing access to the VoR. They also impose tight restrictions on if, when and how non-final copies can be deposited in a repository. Indeed, many publishers are reluctant to provide any help at all to libraries wanting to deposit a paper (whatever version) in the ragtag gaggle of institutional repositories.

It is no surprise, therefore, that SHARE's plan proved impractical, and it had to rethink its approach. As Ruttenberg puts it below, "SHARE determined that its initial goal was premature because of the weak position of IRs."

Notification system

By the time it formally launched, therefore, SHARE had had to rework it plans. In its new incarnation it was focused on creating a "free, open, data set about research and scholarly activities across their life cycle."

In other words, SHARE had had to conclude that it would after all be necessary to create a central digital repository – one, however, designed not to host research papers but metadata about them. The metadata records then link out to the source. Below Ruttenberg describes SHARE as "an open database of millions of records harvested from some of the largest repositories and registries of scholarship, including outside the US."

By building the database, adds Ruttenberg, SHARE is able to provide "a notification system [made] out of harvested, normalised metadata from open scholarly repositories."

We will come back to the word "notification", but why the reference to "normalised" metadata? This highlights the fact that the quality of repository metadata remains poor, and in order to make it useful SHARE has to put it through a process of "remediation" after harvesting.

Since this work cannot all be automated, the remediation process includes what Ruttenberg refers to as a "social" element (from which I infer that it requires human labour). And while SHARE has come up with an interesting crowdsourcing approach for doing some of the work, it must surely remain an expensive and time-consuming process.

So, if the original purpose of SHARE was to provide public access to research funded by the OSTP-mandated agencies what is its purpose today? Why go to the expense of aggregating and normalising all this metadata?

Clearly one role SHARE can play is as a discovery tool. It is possible, for instance, to conduct searches on the circa 40 million SHARE records directly from its search interface. As such, it is one of a burgeoning number of scholarly search engines.

At the same time, SHARE has also widened its brief and now aggregates other types of scholarly output as well as published papers – e.g. preprints, conference papers, research data, and what Ruttenberg refers to as "contextual materials" and "related assets". Importantly, points out Ruttenberg, SHARE's trawl encompasses scholarship that lies "outside of traditional discovery systems."

And with help from its partner the Center for Open Science (COS) SHARE has created a public API so that other applications can be built on top of the database.

Using the API, for instance, SHARE plans to offer universities branded dashboards so that they can ask what Ruttenberg calls the "very basic question" of "what research is produced on this campus?"

This is where the concept of "notification" comes in. As well as pulling information from university repositories, SHARE can push it to them. This is useful because in the digital world the many, varied research outputs and associated material produced during a research project will likely settle in different parts of the network. This could see multiple versions of papers (drafts, preprints, postprints, VoR etc.) distributed around the Web, along with research data and the growing range of other types of digital content created in the slipstream of the research process. The logic of the SHARE dashboard is that it can bring these all together in a single view.

This is not public access

We should not doubt that SHARE is creating valuable and useful tools. However, we need to remember that SHARE's *raison d'être* was to provide barrier-free access to publicly-funded research. As a result of it having refocussed, it now offers something rather different, something OA advocates might argue is of lesser value.

Remember, for instance, that SHARE's discovery service (and any dashboard created from the service) contains bibliographic records not source documents or files. And the latter may, for a number of reasons, be on closed access. They might, for instance, be behind a login wall, or sitting behind a publisher's paywall – see, for instance, the records here, here and here in the pilot dashboard SHARE has created for the University of California San Diego (TritonSHARE). If you don't have a subscription to view these documents, you will be required to pay a one-off access fee.

Moreover, when I tried to access some of the documents I received the message, "We're sorry, there's been an error resolving this DOI. Please try again later" (or sometimes just a timed-out blank screen). This reminds us that online aggregation services aren't always able to deliver on their promise.

Some of the records in SHARE are also surprisingly uninformative – lacking, for instance, a publication date, or any kind of abstract.

Similarly, when I searched on my own name directly from the SHARE search interface, many of these records also linked to documents sitting behind a publisher's paywall (see here, here, and here for instance), or they linked to print publications (e.g. here and here). The publisher-hosted online documents, by the way, are 24 years old and still not freely available

– reminding us that OA does not liberate historical data. Note also that access to these articles costs, variously, £20 or $36. This is not public access, and it is certainly not open access

Of course, SHARE does also link to full text documents, but I am not sure what percentage of the records in its database do so. Below Ruttenberg says that a little more than half link to research articles, 1.5 million to data sets, 2 million to preprints, and 3.4 million to conference papers. However, she does not specify how many of the circa 20 million papers that SHARE links to are full-text open access documents.

Whatever the numbers, it seems fair to say that SHARE does not link to enough freely available full-text content to have any significant impact on the *accessibility* problem, or to prevent publishers from continuing to charge extortionate prices for the services they provide (thus perpetuating the *affordability* problem). Moreover, given publishers' determination to ensure that their platforms remain the primary source of published research, and given that the vast bulk of the research corpus remains toll access (TA), it would surely be naïve to assume that things will improve any time soon.

But what is key here is that one of the main reasons SHARE had to clip its wings was due to an *ownership* problem. This *ownership* problem is not just a product of authors routinely transferring the copyright in their work to publishers, but because the research community has over time outsourced most of its publishing activities to external organisations, often for-profit organisations. Today, the Top Five scholarly publishers (all of whom are for-profit) not only own a disproportionate amount of the research corpus but the infrastructure of scholarly communication too.

This suggests that if the research community wants to solve the *affordability* and *accessibility* problems it is going to have to solve the *ownership* problem first. And that will surely require building (and maintaining control of) its own infrastructure (in addition to ceasing to give away its research to publishers). For this reason, some OA advocates are now arguing that publishers need to be by-passed all together. One way to do this is for researchers to create post-publication systems on top of the burgeoning number of preprint servers being set up on platforms like COS' Open Science Framework (OSF).

Meanwhile, the long-term goal of organisations like SHARE, OSI, and a number of similar community organisations, is to build an alternative infrastructure around IRs. It remains to be seen, however, whether such initiatives will prove achievable or affordable. (Although a few ideas have started to circulate – e.g. here).

Too little, too late?

In the meantime, it has become clear that publishers are desperately seeking new market opportunities, and one new market they have set their sights on is that of monetising data generated by the process of scholarly communication and information sharing. Amongst other things, this will see the introduction of new paywalls. This will not be welcome news for a research community that has spent the last 15 years trying to do away with them. However, preventing it presents a very real challenge, not least because the sheer quantity of research papers the big publishers have acquired (and continue to acquire) puts them in a very powerful position.

Why do I say this? Because by controlling the scholarly communication infrastructure, by hosting a large number of research papers, and by ensuring that access to those papers takes place primarily on their platforms, publishers can capture and leverage valuable user-generated information. With this information they can create a whole new generation of products to sell to the research community. And that, presumably, is why publishers are now going after for-profit competitors like ResearchGate, who have spent the last decade or so also amassing large numbers papers on their sites. It is important to note here that whether the hosted papers are OA or TA does not affect the value that can be extracted and monetised; it is enough that a platform hosts a large amount of content, has a lot of users, and has the necessary technology to capture usage data. We should also note that these will primarily consist of data generated by the research community itself – an exploitative business model reminiscent of the way in which publishers assembled large databases of journal content freely given to them by researchers in the first place, and then sold the information back to research institutions on a subscription basis.

In addition, publishers are seeking to build, acquire and control the digital research workflow tools that are emerging, not least by buying smaller companies like Hivebench. The potential that publishers see in these new markets is amply demonstrated by Elsevier's decision to rebrand itself as an "information analytics business". (See also).

As Alejandro Posada *et al* point out on the G.A.P. website, "the rebranding [of legacy publishers] into data analytics has entailed an active process of acquisition of the existing research infrastructure as well as the development of new platforms surrounding the knowledge production cycle. We argue that this is possible because of a leveraging on their already disproportionate ownership of content."

Indeed, some are now predicting that the new market for workflow tools and data analytics is set to become a duopoly controlled by Elsevier and Springer Nature.

As noted, these developments are not good news for the research community. However, they will be especially galling for librarians who have spent the last decade reimagining their role in the networked world. Their conclusion: they need to move beyond their traditional role focused on "holdings" to one that extends "from licensing published content to managing workflow and research outputs". This idea is neatly encapsulated in Lorcan Dempsey's mantra that "workflow is the new content".

One implication of this could be that where repositories were initially viewed as places to deposit artefacts, they will come to be seen more as an integral part of the research lifecycle infrastructure, one in which pointing and linking to scholarly assets takes precedence over hosting them. It is this view that would seem to have informed SHARE's change of direction, all be it out of necessity. Thus, instead of being the location for research papers, IRs will link to and aggregate scholarly outputs and the various processes that take place during a research project. Or as Ruttenberg and her ARL colleague Elliott Shore put it in a letter to the *Chronicle of Higher Education* last August, libraries future role should be to "support scholarly workflow at all stages of the research life cycle, including preservation and stewardship of research outputs".

The problem is that if publishers are intent on colonising the workflow and paywalling parts off librarians' envisaged new role would appear to be threatened. As Ruttenberg and Elliott put it in the above cited letter, "platforms and business arrangements that lock in scholarly

content and data about scholarly process make stewardship of that content – research libraries' core mission – impossible."

The fear must be that librarians (and the wider research community) have lost the content wars, and could be about to lose the platform and workflow wars too. This points to the challenge the research community faces, and again suggests that in order to solve the *affordability* and *accessibility* problems it will be necessary to first solve the *ownership* problem.

Given the size and nature of the challenge we must wonder whether SHARE will prove to be too little, too late. Time will tell, but we must hope not, since the success of initiatives like SHARE, COS, COAR, OpenAIRE, and LA Referencia looks to be essential if the *ownership*, *affordability* and *accessibility* problems are to be solved. The good news is that these organisations have begun to collaborate and co-operate together, and there is a growing sense that it is now essential to build a public infrastructure "open to researchers from everywhere."

For librarians the frustration must be – as it always has been in OA matters – that these are not wars they can win on their own. Yet the vast majority of researchers remain oblivious and/or unconcerned about them – not least because they are rarely the people who have to pay for the expensive content and/or services that publishers provide.

For their part, university bureaucrats appear to be too focused on buying publishers' products to help them "manage" their workforce and boost their international research rankings to consider the larger ecosystem in which they operate.

Unfortunately, the people who ultimately pay for publishers' expensive products and services are non-cognisant taxpayers and/or hapless students (who face ever rising tuition fees to help pay these bills). Neither of these groups would appear to have the necessary knowledge and/or power to intervene. Likewise, the competition authorities do not seem to understand the issues, or perhaps they simply don't care.

The interview begins …

P: As I understand it, SHARE was founded in 2013 in response to the OSTP memorandum. Is that correct?

JR: Yes. The Association of Research Libraries (ARL) welcomed the directive and was very proud of the advocacy work its members and staff did to support it. SPARC Executive Director Heather Joseph, who as you know was a leader in bringing this policy to fruition, announced the memorandum to the ARL Board of Directors, which was meeting that day, February 22, 2013, in DC. There was a champagne toast to this collective achievement.

In the couple of years leading up to the directive, ARL and the Association of American Universities (AAU) had a joint task force on scholarly communication that was already looking at the possibility of either a shared open access repository or a distributed network of open access repositories. There was general support among ARL, AAU, and the Association of Public and Land-grant Universities (APLU) that university-administered digital repositories should be the mechanism by which federal agencies provide public access to funded research, most of which is conducted in universities.

After the memorandum was issued, a small group of ARL deans and directors, along with several of us on the senior program staff, and David Shulenburger, then a senior fellow at APLU, drafted the first concept paper for SHARE, which stood for "SHared Access Research Ecosystem." That concept paper advanced the distributed repository perspective, along with a set of basic requirements that repositories would have to meet in order to serve as a network of public access sites.

This paper was drafted several months after the directive was published, and it posited a very aggressive development timeline (12-18 months to build out the network) in order to be a plausible mechanism for agencies to meet this new mandate as they considered their own implementation plans and policies. There was no funding for SHARE at this point.

After the initial draft of SHARE was circulated publicly, we gathered advisors from ARL, AAU, APLU, and the Coalition for Networked Information (CNI), and with the help of a

consultant, ARL got a $50,000 planning grant from the Alfred P. Sloan Foundation to scope this vision of a network of university repositories and how it could serve as a federal agency partner for public access deposit.

There were two conditions that the participating associations believed were essential to something like SHARE succeeding: 1) agencies would need to claim sufficient rights over funded research to enable libraries to collect manuscripts and data in an automated way, rather than by author deposit; and 2) every scholarly manuscript arising from a grant and submitted for publication to a scholarly journal would need to include the award identifier, PI number(s), and the digital institutional repository in which the article would reside post-publication.

Unfortunately, there was no uptake to make these requirements mandatory, either on the federal funding side or among universities as conditions of acceptance of that funding. That left our distributed, library-based repository network in the position of figuring out both how to network the repositories (a significant metadata mapping exercise in its own right) and how to insert itself into the funded research workflow and facilitate deposit in a mostly unfavourable IP and licensing environment.

With respect to the latter, this was a challenge repository managers and scholarly communications librarians were familiar with. The SHARE Steering Group, which included library deans, association leadership and staff, a CIO, a vice president for research, and a provost, considered a proposal to build a deposit portal that would route files and their associated metadata to intended repositories, but ultimately believed that universities would be reluctant to rely on a third-party website for something as high-stakes as compliance with a federal mandate.

Two active use cases

RP: So, what is SHARE today, and what is it hoping to achieve?

JR: The objective we launched with—by which I mean the objective that earned ARL its first round of project funding ($1 million) from IMLS and the Sloan Foundation in 2014—was to create a notification system out of harvested, normalized metadata from open scholarly repositories. The idea was that SHARE would tackle the first necessary step—metadata aggregation—and build a product that would provide immediate value as a feed that libraries and research universities could monitor in order to identify articles and data associated with their institutional researchers (and funded by federal grants), no matter where they were deposited.

We hoped that product and the SHARE initiative in general would galvanize the community and provide insight into gaps repositories would have to address with respect to making their metadata interoperable and actionable, and with respect to what they collected in the first place. SHARE began to talk about the lifecycle of research output, not just articles and datasets. We would subsequently address metadata remediation, and ultimately content aggregation—that was the vision and its initial and ongoing first steps.

The second round of grant funding, from Sloan and IMLS, was to grow the database of aggregated metadata by adding more repository providers, link related assets (or scholarly works), build out the open API, and investigate the funded research workflow in three prototypical universities to determine how data from SHARE could improve local knowledge of and stewardship of research output.

What is SHARE now? It's an open database of millions of records harvested from some of the largest repositories and registries of scholarship, including outside the US. SHARE has a public API that we (and others) use to build applications on top of the dataset.

The two active use cases for the SHARE dataset and API are 1) to power discovery of scholarship that resides outside of traditional discovery systems—such as preprints and research data; and 2) to power a local, institutional view of research. For use case #1, we recently got a grant from the National Endowment for the Humanities to work with digital humanities scholars to investigate how SHARE can be deployed as a discovery system for that kind of work in the humanities. Use case #2 is being prototyped by the Center for Open Science (COS) with the UC San Diego (UCSD) Library in TritonSHARE, a dashboard that allows users to explore research being conducted at UCSD.

RP: As you indicated, SHARE was an initiative of the Association of Research Libraries (ARL), the Association of American Universities (AAU) and the Association of Public and Land-grant Universities (APLU). I do not think that the Center for Open Science (COS) was involved in SHARE at that point, but is now. What did COS bring to the party? Are the AAU and APLU still involved?

JR: Yes, that's right too. When we received funding to build the notification system, we solicited proposals from individuals and groups to build the technology. COS's was the best such proposal we received (out of several excellent candidates), and the partnership they proposed—rather than a role as a contractor—was born. AAU and APLU, while founding partners in the initiative and original members of the Steering Group, are not actively involved in managing SHARE at this point.

COS brought the Open Science Framework, developers, interns, a commitment to open source software development, and a mission-aligned organization working on opening up scholarship. Jeff Spies, co-founder and Chief Technology Officer of COS, is now my co-director of SHARE.

Partnership

RP: How does the partnership work? What is SHARE's corporate structure? Does it have a unique non-profit status, or is it structurally part of ARL?

JR: SHARE operates as a partnership between ARL and COS. It is not a unique non-profit. ARL has been the grant administrator for funding to this point and has contributed program staff, visiting program officers, and administrative support to SHARE. COS maintains the

technology and, similarly, contributed administrative support. An operations team comprised of ARL and COS works both with the community (currently a Stakeholder Committee) and the development team to direct SHARE activity toward its active use cases.

RP: So how are decisions made? Are they voted on and decided by the Stakeholders Committee?

JR: Day-to-day decisions are made by the Operations Team. We consult on strategy with the Stakeholders Committee, as well as the leadership of ARL and COS as partner organizations with their own governance. We also consult with our funders, who have a good cross-ecosystem vantage.

RP: It strikes me that the Stakeholders include employees of Elsevier (since both Mendeley and bepress are on the Committee), and employees of Holtzbrinck Publishing Group-owned companies like Digital Science and Symplectic (Holtzbrinck also owns Springer Nature I think). Does this mixing of for-profit and non-profit stakeholders create any issues?

JR: Thank you for pointing out something confusing and somewhat outdated on our website, which we have since changed. We initially assembled a long list of "stakeholders" including anyone (individual or organization) who agreed at the outset that SHARE was a good idea. They were basically endorsements and included commercial service providers, especially in the OA implementation space such as Symplectic and in the repository community.

But our Stakeholder *Committee* (formed in May of this year) is listed here, and includes a small group of library deans and directors who have agreed to direct SHARE strategy and contribute to SHARE's development.

RP: You mentioned a number of start-up grants. How is SHARE funded today? And what are its current and anticipated revenue sources? Do you think it will need to engage in regular funding rounds in order to be sustainable?

JR: SHARE has received a little more than $2.2 million in grants as well as in-kind contributions from ARL and COS. For SHARE to be sustainable, it's going to have to expand to include contributions of money and/or developer time from other groups that want to use the infrastructure.

We just wrapped up a year-long cohort program of SHARE Curation Associates—library professionals that worked on their own repositories and on improving SHARE. The question of sustainable funding is also why we formed a Stakeholder Committee earlier this year. The committee is working with the SHARE Operations Team on developing a long-term plan for community ownership and sustainable contributions to SHARE.

In particular, we have been reviewing SHARE's documentation and will make a concerted effort for library developers to become code contributors. But that takes time, and we, as well

as our participating libraries, may need additional grant funding in the short and medium term.

RP: So you have received $2.2 million in total, which includes $50,000 from the Sloan Foundation in 2013, $1 million from IMLS and the Sloan Foundation in 2014 and $75,000 from the NEH in 2017. Who provided the rest?

JR: Sloan and IMLS.

RP: Can you say something about SHARE'S funding going forward? From what you say, I assume that you envisage a model similar to arXiv (in which libraries are invited to make voluntary institutional contributions), or does SHARE plan to sell products and services in order to generate revenue?

JR: This question is really a big part of the Stakeholder Committee's remit. But yes, since SHARE's aim has always been to be part of higher education scholarly infrastructure, an arXiv model (institutional contributions to its sustainability) is a good place to start.

Speaking now as an ARL Program Director, we need new, scalable, collective-funding models for public goods content and infrastructure in scholarly communication. With respect to infrastructure, it succeeds when people not only use it but come to depend on it existing and being in good shape—like moving goods to market on a public road. We want to see tools and services built on top of SHARE—to aggregate specialized content like data, for example, or analytical tools for doing meta-scholarship. In that scenario, we might (collectively) envision a contribution model for infrastructure.

Here's another example: I sit on the SocArXiv Steering Committee. SocArXiv will need sustainable funding for the development of its community of reviewers, editors, and authors. But since we're using free public goods technology to run the service (OSF Preprints), it makes sense to contribute some percentage of funds raised, or revenue, toward the maintenance of that technology. This is a much bigger question than how to fund SHARE specifically, in other words.

At this time, SHARE has no plans to sell products or services.

SHARE vs. CHORUS

RP: SHARE was announced shortly after news that publishers planned to create the publisher-based CHORUS service. SHARE and CHORUS were viewed by the research community as competitive, but I think SHARE has tried to present them as complementary services. What are the respective roles you expect SHARE and CHORUS to play going forward, and in what ways are they likely to prove competitive/complementary?

JR: SHARE is focused on strengthening the role of open repositories to preserve and steward modern scholarship in the most expansive interpretation of what that means—contextual materials as well as completed products. SHARE is supportive of, and provides infrastructure

for, new kinds of scholarly communication, beyond the article and certainly beyond the journal. I remain taken with and committed to Herbert Van de Sompel's concept of a "record of versions" in networked scholarship, rather than a version of record. That's how I see SHARE.

CHORUS was launched as (and in my understanding remains) a mechanism for publishers to maintain control and stewardship of publicly funded research articles by opening select articles to public readership in the context of the overall paywalled journal.

In order to do that, CHORUS has done great work in improving scholarly metadata—including the proliferation of ORCID IDs and funding information on published papers. SHARE, and everyone, benefits from more robust, open metadata on the network, which CHORUS is contributing.

RP: When SHARE was launched there was some scepticism as to whether it ambitions could be realised, certainly in the timescale envisaged. One of those sceptics was Dorothea Salo, who expressed her doubts to* Library Journal*. Do you think that SHARE has proved the sceptics wrong yet?

JR: I don't know. Dorothea Salo's comments were totally fair, and she knew the repository landscape very well. And not to split hairs, but that June 2013 paper was a concept statement, not a launch. SHARE had no funding at that point and no governance structure, not even a Steering Group.

To the extent that we're still talking about that concept statement and whether SHARE competes with CHORUS, and not what SHARE has accomplished since our actual product launch in spring 2015, then I'm afraid the answer to your question might be no.

RP: I am thinking that the real point here is that publishers have won the battle over content, in so far as public access to the papers subject to the OSTP Memo will in the main be provided via publishers' sites rather than IRs?

JR: I don't share the perspective that "content" = "papers subject to the OSTP Memo." That's rather the whole point, right? Scholarly communication is evolving (as is the scholarly record), research communities are recognizing the value of contextual materials, understanding that there are multiple contributions to a research project (see development of the CRediT Taxonomy, for example), and that the final paper is just that.

The fact that preprints have exploded, that concern over research reproducibility has captivated entire disciplines and is helping to fuel such critical efforts as software preservation in the past few years—all of those forms are content.

Lorcan Dempsey has been saying that "workflow is the new content" for several years, and I agree with that. We should absolutely be worried about enclosure and lock-in of that workflow and advocate for more open platforms and processes so that libraries can preserve the full scholarly record.

But even in the realm of publicly funded papers, there is more to this equation than commercial publishers and university IRs. There's PubMed Central, a solution for more than half a dozen agencies to provide public access per OSTP, and there are other agencies that have their own repositories.

You published an excellent interview with CNI Executive Director Clifford Lynch just last year about the state of IRs, and CNI subsequently held two Executive Roundtables about the future of IRs this spring—several years after the 2013 SHARE concept paper. Is investment in IRs sufficient? What is their core value proposition? Should they be local to institutions, or more centralized? None of these are settled questions now, and certainly weren't then.

But universities are still free to embrace policies that allow them non-exclusive rights to archive their research output, broadly defined. That would still enable some kind of automated solution to aggregate that output in either a distributed or centralized manner.

RP: So how would you say SHARE's objectives and ambitions have changed since 2013?

JR: I think I've covered a lot of this in the questions above. ARL remains committed to an open scholarly communication system, and SHARE is contributing to that by exposing and linking, in an aggregate dataset with an open API, highly distributed scholarship on the web. The Open Science Framework (OSF), with which SHARE is associated through COS, provides a platform to integrate the tools of open scholarship, along with an environment conducive to research stewardship—including metadata control, versioning, and provenance tracking.

The work SHARE/COS is doing with UC San Diego Library—an extensible prototype using the SHARE API to expose institutional research activity in customizable ways—is an exciting direction for SHARE. It recognizes the value of scholarly metadata and the potential for institutions to bring a data science approach to metadata to ask scholarly questions of it.

That can only be done in an open environment, or by paying enormous licensing fees. Declan Fleming and Jeff Spies presented this work at CNI this spring and there is great interest among other institutions in building on it.

Initial goal was premature

RP: Can we take a moment to allow me to check my understanding so far: When SHARE was launched it was envisaged that institutional repositories would be the mechanism for providing public access to research subject to the OSTP Memo. This would be achieved by universities posting faculty papers in their IRs and SHARE would then build the infrastructure needed to allow the content in those IRs to be aggregated on a distributed basis. The assumption was that funders and/or universities would retain the necessary rights to allow papers to be made freely available in IRs and the necessary funder and location data would be attached to them in the process.

This however did not prove possible. SHARE then envisaged creating a deposit portal to allow papers to be deposited centrally and then routed to the relevant IRs. As this too did not prove possible, SHARE decided to create a "notification system." By harvesting metadata from the many possible locations where papers might be located SHARE would be able to alert universities to any new research outputs or "events" relevant to them. It was also decided to focus not just on articles, but datasets and other outputs. And the current position is that SHARE has created a central metadata database and is piloting a scheme to allow the creation of a local institutional view ("branded dashboard") for universities that will allow them to access the slice of the database that is relevant to them. The first example of this is the one created for the University of California, San Diego here.

Have I understood correctly, and can you say something about the other research outputs SHARE is interested in?

JR: I think that's basically it. With respect to retention of rights, I'm not sure it was an assumption that either agencies or universities would make those claims or mandate particular metadata; more that the founding associations of SHARE understood both of those to be necessary conditions for success, particularly within the timeframe originally proposed.

With regard to the branded dashboards—UCSD's motivation was to be able to answer the very basic question, "what research is produced on this campus?" UCSD's concern was with the distributed nature of research output and how the university could have knowledge of UCSD-related content that wasn't deposited into its (relatively mature) data repository or into the UC system-wide article repository. They saw in the SHARE API an opportunity to concentrate on building a tool that would serve the institution, without having to collect the data (or metadata) at the outset.

RP: In summary, SHARE has, in effect, abandoned its initial goal of taking on the task of enabling IRs to provide access to the research papers covered by the OSTP Memo. Instead, it is creating a bibliographic database that points to papers, some of which may be in IRs, but many of which will be hosted on publisher web sites, and many of which will be behind a paywall? This is what you mean by "exposing and linking"?

JR: The breakdown in SHARE's aggregated assets is as follows, out of 40 million records: a little more than half are articles. You can also find 1.5 million data sets, 2 million preprints, 3.4 million conference papers, and more.

By exposing I mean bringing into an aggregate data set that which was only discoverable through its own system, and by linking I mean bringing together component parts of a whole work or related intellectual works.

SHARE determined that its initial goal was premature because of the weak position of IRs. So we endeavoured to do something that was both possible and that was intended to strengthen IRs, both with respect to greater exposure of IR assets through aggregation, and by promoting stronger open access policies at the university level. In 2015, AAU, APLU, and

ARL sent a memo to its member universities outlining steps they could take on the licensing front that would assist them with public access policy requirements.

And in the meantime? We've seen the growth of SciHub and ResearchGate on one hand, and we've seen scholarly communities moving forward with thoughtfully built open disciplinary repositories on the other. And we have many IRs trying to figure out their future. Green OA? ETDs? Multimodal digital scholarship? Are they for preservation, discovery/access, or both? In any of these scenarios, I would argue that a system that can connect services together will be necessary for a robust open-repository network.

RP: We have learned that one of the main problems of aggregating IR content is lack of standardised metadata. As I understand it, various standards have over time been developed – e.g. Dublin Core and OAI-PMH – but these have often not been implemented correctly, or in a consistent manner. You talked earlier about "normalized metadata" and "metadata remediation". Would I be right in thinking that once SHARE has ingested all the metadata exposed by IRs it is then having to rework them in order to create standardised data, which is presumably essential if effective discovery is to be provided? If I am right, what exactly does this involve, and how labour intensive is it?

RJ: You are correct. Repositories have done well to provide standard support for harvesting Dublin Core via OAI-PMH, and with other standards developed like ResourceSync, OAI-PMH remains the most widely supported mechanism for harvesting metadata. However, despite this standardization it is indeed not being used in a consistent manner. Usage of free-text fields such as 'dc:description' and 'dc:rights' diverge tremendously across sources.

However, most problematic is that OAI-PMH endpoints each often provide different sets of fields (versus inconsistently using the same fields). This can be due to configuration, but most often is due to the unavailability of metadata. The majority of endpoints also use Simple Dublin Core versus Qualified Dublin Core, thereby further limiting the richness of the metadata.

SHARE has employed several strategies to remediate the metadata: both automated and social. With widely variant uses of metadata, SHARE decided to create a custom harvester configuration for each data provider. This then allows us to map incoming fields to a common set of attributes in our schema. We also pull metadata from multiple sources and then cross-reference records via common attributes such as title, author, and DOI.

It is most helpful when data providers provide unique identifiers like DOIs, but automated mechanisms often cannot completely determine whether records are matched without some kind of review.

So, we also employed social mechanisms to improve metadata, most notably our pilot Curation Associates program. In 2016–2017, we worked closely with our Curation Associates to assess, reconfigure, and improve harvesting mechanisms with their respective repositories as well as improve metadata flowing downstream from sources like DataCite and Crossref.

Avoiding lock-in

RP: Can you say something about the API you have developed, how it works, and what benefits it provides?

RJ: The API(s) can be used in a few different ways, both for pushing and pulling metadata to and from SHARE (documentation here). As an alternative to harvesting mechanisms like OAI-PMH, SHARE provides a Push API that is used by a variety of data providers to push data to SHARE for ingest into its data set (i.e., index).

Different from OAI-PMH the SHARE Push API accepts data as JSON-LD already formatted and mapped to the native SHARE schema. Its REST-based operations include create, update, and delete operations on records, and it is most notably being used by many Samvera/Fedora based repository systems.

Secondly, SHARE provides another REST-based API for searching and accessing information within the SHARE data set. This can be used as a simple one-time query or request of records through the SHARE web portal or through a scripting language like Python (or any other language that supports REST http requests).

In addition, the API can be used to power fully functional applications independent of the core SHARE system. A prime example is TritonSHARE, the application I mentioned that has been developed jointly between the UC San Diego Library and the Center for Open Science. It is a research activity dashboard using SHARE data that only communicates to SHARE via its API.

RP: One form of research output that I assume you are taking a great deal of interest in now is the preprint, if only because your partner COS is getting increasingly involved in helping set up branded preprint servers. Preprints, of course, are open access, and so publishers cannot (I assume) claim any ownership or control over access. Since you are helping to build branded research dashboards for universities I would think a natural development would be to create branded overlay journals. Is that something you envisage doing?

JR: I think ARL and COS are interested in preprints because scholarly communities are embracing them as a step toward greater openness, including in fields that do not have a culture or tradition of sharing work before formal publication. SHARE was already indexing prominent preprint services, and then COS made technology available freely to interested communities—now up to 13. Building a preprints discovery platform was a natural fit.

With respect to overlay journals, I certainly hope to see these develop out of the preprint services themselves as they determine the rules of the road for moderation and peer review, and organize editors and reviewers within their disciplines.

Helping launch these is another potential partnership between ARL and COS, insofar as ARL can organize library funding and other resources, and COS is providing the technological infrastructure for so many services. But the overlay journals themselves will likely come from the research communities themselves.

RP: After looking at the Declan Fleming and Jeff Spies presentation you pointed me to another thought has occurred to me. In that presentation Jeff Spies says, "we need to collect and make use of campus analytics before big, for-profit publishers sell them to us and lock us in like we are with text". Am I right in thinking that while the research community has struggled to wrest back control of publicly-funded research from publishers, there is a hope that it may be able to win the next battle – that is, control of workflows and analytics? The challenge here has been well mapped out by Roger Schonfeld here, here and here, and the dangers if it happens expressed here.

As I understand it, the issue is that as publishers move into the workflow and analytics space they are creating expensive new tools that could lock the research community into a new form of paywall. What is needed is to create an alternative model of a public goods infrastructure. As you put it recently in a letter to* The Chronicle of Higher Education*: "Simply put, platforms and business arrangements that lock in scholarly content and data about scholarly process make stewardship of that content—research libraries' core mission—impossible." How big a threat is this, and how confident are you that the research community can avoid it?

JR: I absolutely agree that the research community (including libraries) needs to invest in viable public goods open infrastructure to conduct the core business of the university, including scholarship. At the same time, we need universities to sign strong licenses for any proprietary services they do use to develop, store, compute, or curate content.

Such licenses must ensure that content (including analytic content generated in the course of use of a product) is retrievable in a useable, non-proprietary format at the conclusion of the agreement. Otherwise, that content and the customer are locked-in.

RP: Thank you for your time, and good luck with SHARE.
