SHARE (SHared Access Research Ecosystem) is a partnership between the Association of Research Libraries (ARL) and the Center for Open Science (COS). It is building a free, open, data set about research and scholarly activities across their life cycle.

It is gathering, cleaning, linking, and enhancing metadata with the help from community members. This quote encapsulates part of SHARE’s mission.

While many people were already involved in the SHARE community, the community wanted to include more people with a variety of skills and experience who deal with institutional repositories. They developed the SHARE Curation Associates to involve more librarians in the project and to teach the skills necessary to work with SHARE data. I am one of the curation associates, as are my 3 collaborators on this project.

This recent article will give you even more information about SHARE.

Before I was a curation associate, we asked SHARE to harvest our repository’s metadata. This was a simple process, which largely involved providing information about our OAI feed for simple Dublin Core.

The process is slightly different now; the form is still simple but it seems to have a few more options, such as using something other than oai_dc or using an API instead of OAI. Simple Dublin Core is preferred because no repository specific customizations are needed.

The process to register is simple so I encourage you to do this.

SHARE is currently getting data from 153 sources, with the bulk coming from CrossRef. If you have unique content in your repository, especially content without DOIs, registering will ensure your publications are fully integrated with everything else. Even if your content has DOIs, your local metadata may be better and it can help you identify works by local authors. (I should also mention that a different project is working to identify your own institution's content in SHARE and then create a view so you can better analyze local publications or to find things for your repository. Being involved in SHARE will mean you will learn about these projects as well.)

All this sounds great—until you remember SHARE is harvesting simple Dublin Core, which lacks nuance, granularity, and data. At the Curation Associates meeting it was clear our data could be harvested so much better using qualified Dublin Core. This is an example of the same item in SHARE and in our repository. However, repository specific harvesting would only be a sustainable prospect for the OSF programmers if we had consistent metadata across at least a group of repositories. It was from these discussions that I started collaborating with 3 other curation associates who used the same software, Digital Commons. One example of the problems with our simple Dublin Core one is that in Digital Commons it lacks the DOI.
I should mention that Digital Commons is a hosted product so we don’t have complete control over it. However, Bepress allows us considerable flexibility across our separate series and between libraries. There are some fields that have specific uses in the system, so customers generally use them consistently, and there are of course defaults that we can use if we haven’t specified something else. This flexibility is good since there are series that have very specific needs. However, the lack of standardization of metadata across US institutional repositories in general makes any such aggregation using qualified Dublin Core challenging. Because there are a lot of customers using this platform, if we can collectively map data to specific fields, the programmers could make a Digital Commons harvester to take advantage of this additional metadata. I also want to be clear that none of this is meant to be bashing Digital Commons. We have good relationships with the company and believe that by working together, customers and vendor, we can have a better product. Because it is a hosted product we see a real opportunity in standardizing data needed by SHARE. I also want to stress that many of the issues apply to institutional repositories using different software.

Last fall we began to review standard fields and Dublin Core mapping provided by Bepress for their customers. We compared these to several other schemes. During the course of this work we learned from the OSF programmers that the closer we could get to DataCite’s scheme the better it would be for them; they have found this scheme to be the most complete.

In April, Lisa Palmer presented a poster about our project at ACRL. I am covering some of the same content. Since that time, we are clarifying our recommendations and have begun to share this with the other curation associates and OSF programmers. We will then share this more widely—to Bepress, other Digital Commons customers, and the IR community as a whole. We have discovered quite a few issues with the data and our understanding of the fields, as well as gaps in what we had been doing.

Authors are very difficult due to the flat nature of Dublin Core. Our OAI does not include affiliation, which means this is not getting into SHARE. Affiliations of co-authors or of people publishing in one of journals would help other institutions find their content. Digital Commons does not yet have ORCID IDs implemented, but will soon. When they are implemented, we will want not only affiliations but ORCID IDs to be shared. We may need a solution beyond qualified Dublin Core to contribute this information to SHARE. We are hopeful that Bepress will be able to implement a nested structure like in DataCite that accommodates affiliation and identifier. Contributors have always been a bit difficult in Dublin Core since they can perform any number of roles. We are hoping there could be a means to assign a specific role to a contributor or creator, much like a relator code in a MARC record.

Types are another challenging area due to the lack of consistent terminology being used in various places. There are several options, all of which lack options that reflect in our repositories, but two particularly good ones to look at are from COAR and JISC. Type is also used to help facet the results in SHARE which makes it even more important to use consistently across many platforms. It is a bit of a moving target as SHARE is considering if there are additional terms that should be included in its schema. One term SHARE uses that we haven’t
used is preprint — we have been using the NISO recommendations for Journal Article Versions (NISO-RP-8-2008), but should probably add a preprint document type. In addition to these problems, Digital Commons has an extra challenge because of a functional use of the document type field.

[slide 16] Some of our challenges are due to changes in the broader metadata landscape. For example, rights are not being displayed in SHARE so for this specific project, modifications to how we deal with them are not needed. On the other hand, if we are talking as a group about what our metadata practices ought to be, we should discuss Rights Statements for Cultural Heritage institutions, use of URIs, and if the rights should be split into a rights holder and date copyrighted or remain a single field. Restrictions on access are another important component of the rights field and this does not seem to be handled consistently by Digital Commons customers. The software deals with embargoes well, but could include that information more clearly in the metadata. In addition, many institutions have items or collections restricted to campus users, such as their theses and dissertations. We need this information to be easily included in our OAI metadata in a consistent manner.

[slide 17] In some cases, there is not a technical issue as much as a repository metadata issue. In Digital Commons, by default the name of the repository is used as the publisher, but this can be replaced with a standard value or a free text field if requested. Publisher is a required field in DataCite as it is an integral part of the citation. For the content we publish, I have tried to always include a publisher. However, in our preprint/postprint/version of record series, we have not included the publisher in a specific publisher field. I don't think we are alone in this. We do of course include a citation to the journal site, but this may not identify the publisher. As a former serials cataloger I am very aware of how variable publishers can be over time for a single journal which may be why I haven't included publisher. However, I know some repositories do include this information. And then the question is who is the publisher of the pre- or post print? Is it the institution or repository?

[slide 18] Another type of challenge are fields that we honestly don't understand how best to use, such as Source. We have seen a recommendation that the source or suggested citation of an item (e.g. journal's name, volume and issue of an journal article) should be in <dc:source>. This again is outside the specific SHARE project and instead falls into the category of determining general repository best practices that we should all be following.

[slide 19] This project has been enlightening regarding our own metadata. Collaborating with people from our institution has been incredibly helpful in considering issues broadly. I hope we will be able to share our draft recommendations by ALA. If you are interested in seeing them, please let me or Lisa, Jo or Emily know and we will be sure to send you a link.