Giving datasets context: a comparison study of institutional repositories that apply varying degrees of curation

At this year's 13th annual Digital Curation Conference, Amy Neeser, Program Manager for Research Data Management at UC Berkeley, co-presented on a research study comparing four academic libraries’ approaches to curating the metadata of dataset submissions in their institutional repositories: no curation, pre-ingest curation, selective curation, and post-ingest curation. The findings were 1) the metadata elements varied greatly between institutions, 2) repositories with more options for authors to contribute metadata did not result in more metadata contributed, 3) pre- or post-ingest curation did not significantly result in more metadata contributed, and 4) data sets submitted to a repository with pre- or post-ingest curation more often included documentation. The outcome of this study will help institutions determine if curation has the expected benefits of research data discoverability and reusability.

Amy will present the findings of the study she helped to conduct and document at this session of Research IT's Reading Group, and discuss its implications for how UC Berkeley supports its researchers' data management needs.

When: Thursday, 3 May 2018 from 12 - 1pm
Where: 200C Warren Hall, 2195 Hearst St (see building access instructions on parent page).
Presenting: Amy Neeser, Program Manager, Research Data Management (Library / Research IT)

Please review in advance of the meeting:

- Giving datasets context: a comparison study of institutional repositories that apply varying degrees of curation, Koshoffer, A., Neeser, A.E., Newman, L., Johnston, L.R. (2018)
- Slides presented at the Digital Curation Conference in February 2018

Presenting: Amy Neeser, Program Manager - Research Data Management

Attending:

Aaron Culich, Research IT
Andrew Wiedlea, LBNL
Anna Sackmann, Library
Chris Hoffman, Research IT
Cody Hennesey, Library
Meggan Leavitt, ETS
Nicholas Tripcevich, ARF
Patrick Schmitz, Research IT
Rick Jaffe, Research IT
Ron Sprouse, Linguistics
Scott Peterson, Library
Steve Masover, Research IT

Slides (PDF)

Amy used to work at U Michigan, which is why this talk will focus a lot on that institution's repository & use cases.

A lot of libraries are investing in Institutional Repositories (IRs), and in operations -- this study began as an effort to begin assessment of whether the investment is worthwhile.

FAIR: data curation qualities -- Findable, Accessible, Interoperable, and Reusable. This study focused on the first two, findable and accessible.

Institutions with less curation required fewer fields to be filled out on data deposit. At Michigan, found in UX investigations that if a researcher came upon a form with lots of fields to fill out they tended to throw up their hands and give up. Fewer barriers to submission was judged a better approach. At Minnesota, Cody suggested, no need for required fields when 6 curator/librarians were on-hand to fill in blanks.

Documentation of data not a requirement for any of the four institutions IRs.

Keywords are a contested topic in libraries -- question whether they are useful in light of full-content indexing.

Curation resulted in more documentation (esp. README files). Researchers not motivated to document unless it was a requirement.

Other institutions expressed interest in joining the study when this paper was presented: UCSF, U Tennessee, U Colorado Boulder.
Steve: RE: future studies, is there any ability to study the qualities of metadata and documentation, as well as relationship to curation, for data sets that “get reused”, whether occasionally or frequently.
Amy: Difficult to assess reuse. Downloads don’t really tell. Citations is about all you can do.
Cody: Downloads do say something about findability.

Chris: Quality of metadata studied?
Amy: Difficult to measure. Word count -- do more words mean “better”? Actually, found instances where that wasn’t the case.
Chris: Constant formats might be a way to approach that question.
Ron: Linguistics; OLAC; metrics evaluated in OLAC, could look at how that’s done. Questions like whether a controlled vocabulary is used.

Aaron: Identifiers -- focus on DOIs -- did you look at any other identifies, like ARCIDs.
Amy: None of the repositories used other than DOIs, so didn’t examine.

Aaron: In curation process, README/documentation existence was a measure?
Amy: Yes. Some data sets had multiple types of documentation.
Aaron: Interested in comparison to GitHub, in which the README.md is so core to description of what’s in the repository.
Cody: Wonder if some of the collections or repositories are light on metadata because the same information is included in the README.

Rick: Back to overall question -- value of IR investment. What’s the value to a university of a repository?
Amy: We were looking at value of investment in curation, not so much investment in the IR itself.
Rick: So to the matter of the five research questions in this study, I wonder which best correlate with value.

Steve: Is structured vs. unstructured data a better feed to search engines that people might be using to find data sets.
Anna: Interested in how people discover. Some communities: tweets with hashtags. Is a common dataset discovery through papers that cite a dataset?
Chris: Is there a difference between findability of literature and findability of data sets?