ISSN 2161-3974

JeSLIB 2021; 10(3): e1202
https://doi.org/10.7191/jeslib.2021.1202

Journal of eScience Librarianship
putting the pieces together: theory and practice

Full-Length Paper

Responding to Reality: Evolving Curation Practices and Infrastructure at the University of Illinois at Urbana-Champaign

Hoa Q. Luong, Colleen Fallaw, Genevieve Schmitt, Susan M. Braxton, and Heidi Imker

University of Illinois at Urbana-Champaign, Urbana, IL, USA

Abstract

Objective: The Illinois Data Bank provides Illinois researchers with the infrastructure to publish research data publicly. During a five-year review of the Research Data Service at the University of Illinois atUrbana-Champaign, it was recognized as the most useful service offering in the unit. Internal metrics are captured and used to monitor the growth, document curation workflows, and surface technical challenges faced as we assist our researchers. Here we present examples of these curation challenges and the solutions chosen to address them.

Correspondence: Hoa Luong: hluong2@illinois.edu

Received: April 1, 2021 **Accepted**: June 4, 2021 **Published**: August 11, 2021

Disclosures: The authors report no conflict of interest.

JeSLIB 2021; 10(3): e1202
https://doi.org/10.7191/jeslib.2021.1202

Abstract Continued

Methods: Some Illinois Data Bank metrics are collected internally by within the system, but most of the curation metrics reported here are tracked separately in a Google spreadsheet. The curator logs required information after curation is complete for each dataset. While the data is sometimes ambiguous (e.g., depending on researcher uptake of suggested actions), our curation data provide a general understanding about our data repository and have been useful in assessing our workflows and services. These metrics also help prioritize development needs for the Illinois Data Bank.

Results and Conclusions: : The curatorial services polish and improve the datasets, which contributes to the spirit of data reuse. Although we continue to see challenges in our processes, curation makes a positive impact on datasets. Continued development and adaptation of the technical infrastructure allows for an ever-better experience for the curators and users. These improvements have helped our repository more effectively support the data sharing process by successfully fostering depositor engagement with curators to improve datasets and facilitating easy transfer of very large files.

JeSLIB 2021; 10(3): e1202
https://doi.org/10.7191/jeslib.2021.1202

Introduction

The National Science Board's 2011 report stated, "A core expectation of the scientific method is the documentation and sharing of results, underlying data, and methodologies," and data sharing is considered as a "critical national need" in (National Science Board 2011). In this report, the National Science Foundation (NSF) also announced its requirement to have a data management plan included for each grant proposal submission. In 2013, a memorandum from the Office of science and Technology Policy (OSTP) indicated efforts to increase public access to research data generated from federally funded projects (OSTP 2013). In practice, data sharing has shifted from sharing among personal networks to making data publicly available. Posting data on personal or project websites is no longer adequate, but because not all disciplines have established disciplinary repositories, there is a demand for institutional repositories (Heidorn 2011).

Along with federal mandates, there are also incentives for data sharing/data publication. For example, there is the potential for increased citation through the creation of additional citable objects such as data and software (Gewin 2016). The availability of data repositories has been shown to help biological scientists develop community norms around data sharing (Kim and Burns 2016).

These developments led to changes at the University of Illinois at Urbana-Champaign. Established in 2014 under a call from the 2013 campus Strategic Plan, the Research Data Service (RDS) realized a goal of this plan by developing data publishing infrastructure, called the Illinois Data Bank. In the strategic plan for 2018-2023, the university continues to emphasize the importance of "responsible data sharing practices throughout the institutional and constituent lifecycles" (Strategic Plan 2018). The Illinois Data Bank maximizes the public access to unrestricted research data created by Illinois researchers by centralizing, preserving, and providing persistent and reliable access to the data. All datasets are provided with timely and professional curation to ensure each dataset's completeness, understandability, and accessibility in the future.

In 2016, the article "Overly Honest Data Repository Development" provided a holistic view of the development process for the Illinois Data Bank (Fallaw et al. 2016). The paper addressed why certain software, features, and elements were chosen. Going into its fifth year of operation and with increasing numbers of users, we see more acceptance of the Illinois Data Bank as a place to share research data. We use this paper as an opportunity to reflect on our data curation practices, describe the process of how addressed the frequency of versioning datasets through addition of a new feature, and discuss technical improvements and challenges we face as we continue to strive to meet the needs of our researchers.

Illinois Data Bank Background and Summary Statistics

The RDS was funded by the Office of the Vice Chancellor for Research & Innovation at the University of Illinois at Urbana-Champaign and has its home in

JeSLIB 2021; 10(3): e1202
https://doi.org/10.7191/jeslib.2021.1202

the University Library to leverage the established digital preservation and repository services expertise (Fallaw et al. 2016). While the unit offers different data management services, for the purpose of this paper we only focus on the data curation service provided to all datasets submitted to the Illinois Data Bank, which was launched in 2016 to support research with high degrees of transparency and professionalism.

The data and data creators must meet those requirements to deposit data:

- The data must be in the final stage and not expected to undergo revisions.
- Since the Illinois Data Bank is a public access repository, all sensitive, confidential, or other legally protected information must be removed from the data before its deposition.
- At least one data creator must be affiliated with the University of Illinois at Urbana-Champaign.
- Finally, the data creator must have permission from all creator(s) and/or copyright owner(s) to publicly distribute the data.

The goal of the Illinois Data Bank is to make published data available to anyone. The Illinois Data Bank is a front-end web application and developed on top of the Medusa digital preservation repository to leverage its preservation functions (Fallaw et al. 2016). This strategy allows smooth integration, supports robust management and preservation of data, and allows the flexibility to create new features to support the need of researchers.

Since launching, the number of published datasets and downloads has increased over time (see Figures 1 and 2). As of December 2020, the Illinois Data Bank held 300 published datasets from various disciplines on our campus, with over 193,700 downloads, and another 140 datasets in the draft stage. The datasets are from fifty-three (53) different units across campus and are categorized into five different subject areas (Figure 3).

At the end of 2019, the RDS completed a five-year review. The review consisted of a survey and in-depth interviews of researchers. In the survey, we asked if the participants have used the RDS and which service was found to be the most useful. The majority of respondents had used the RDS and identified the Illinois Data Bank as the most useful service. New depositors and returning depositors are split roughly evenly, with 58% (n=175) "new depositors", defined as those publishing their first dataset with us and 42% (n=125) "returning depositors", defined as those who have published at least one dataset before.

JeSLIB 2021; 10(3): e1202
https://doi.org/10.7191/jeslib.2021.1202

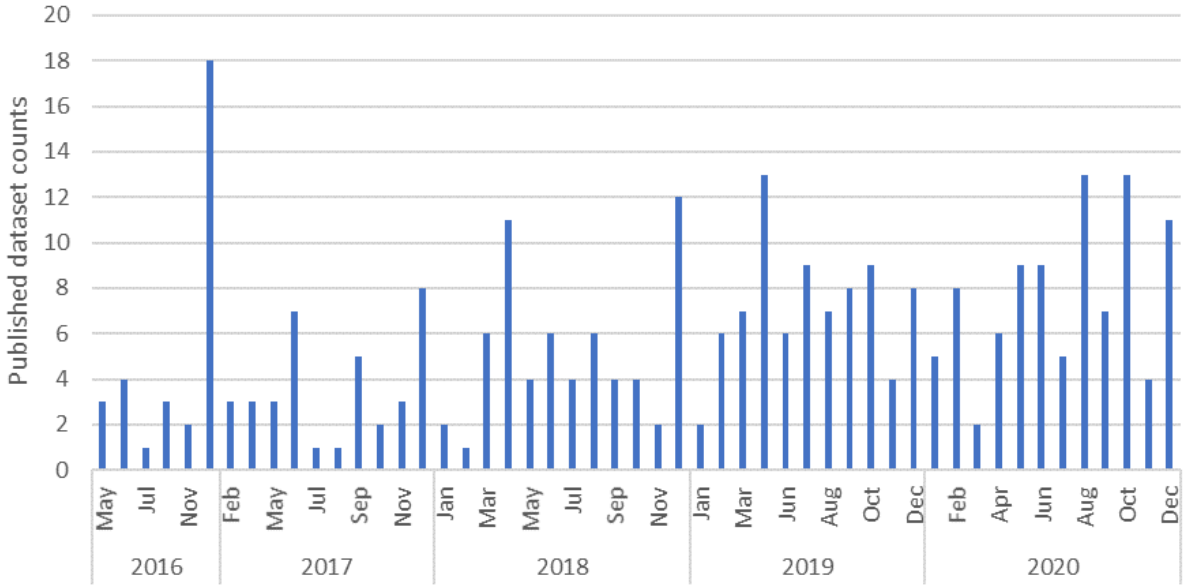

Figure 1: Number of datasets published per month in the Illinois Data Bank, from 2016 to Dec 2020, with an average of 6 datasets per month.

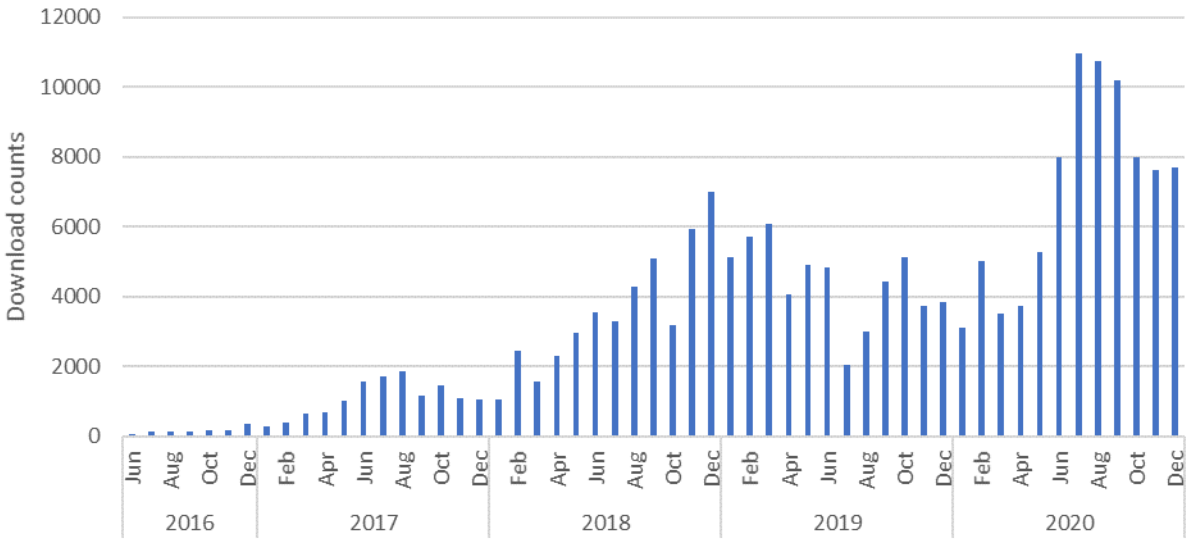

Figure 2: Total download counts (2016 to December 2020) is 193,788.

JeSLIB 2021; 10(3): e1202
https://doi.org/10.7191/jeslib.2021.1202

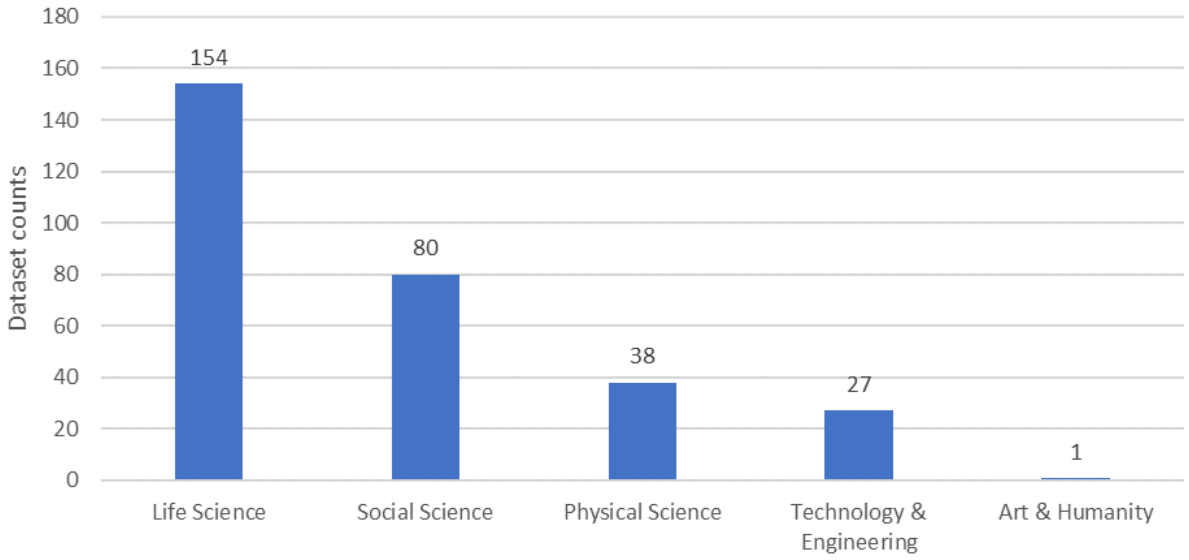

Figure 3: Categorization of the published datasets

Section 1: Our workflow for data curation

This section describes our curation workflow for the Illinois Data Bank and the challenges we face.

The Illinois Data Bank provides ongoing, professional curation services to all the deposited datasets. Curation largely follows the Data Curation Network CURATE steps[1]. Additionally, Illinois Data Bank curators ensure links to related materials are present and represented accurately in the DataCite metadata schema. The curation services happen at the initial deposit and throughout the duration of the dataset's life in the Illinois Data Bank. For example, additional related materials are added to datasets as they are reused and referenced by the depositors or others.

Currently, there are two pathways that researchers can choose when submitting their dataset: post-ingest curation (publish then review) or pre-ingest curation (review then publish); the latter is a new feature implemented in 2018. Despite which option is chosen, the curator receives an email notification from ticket system to start the curation process.

In general, we first check metadata to ensure the minimal requirements (mandatory elements to register with DataCite) are accurate, including: dataset title, license, and corresponding author. Next, we review documentation (e.g. README.txt, if available) or/and the dataset description to understand the dataset thoroughly. The depth and length of the curation process heavily depends on the size and/or complexity of the dataset deposited and how well it is documented. For

1 Data Curation Network CURATE step: https://datacurationnetwork.org/outputs/workflows

JeSLIB 2021; 10(3): e1202
https://doi.org/10.7191/jeslib.2021.1202

example, some datasets from our Crop Sciences department required 6-8 hours to curate due to the number of files in different data formats and high-level documentation describing individual files. On the other side of the spectrum, we also have datasets containing only one or two tabular files which require less than an hour to curate. On average, we spend almost 2 hours (1.89 hours) to review each dataset in the Illinois Data Bank.

Datasets that request pre-ingest curation remain in the draft stage, thus the curator must log in into the system to access the dataset. For published datasets, to avoid triggering the download count for the dataset during curation, we have implemented a file management mechanism that is available to curators only. For curation, we review as many of the files in the dataset as possible. Datasets that contain large numbers of similar files or multiple large files (e.g., >20GB), we check a subset of the files rather than every single file.

The next step is to take note of any missing information and gather suggestions. Collaboration is a key and Heidorn (2011) suggested that a data curator needs to collaborate closely with data creators to understand the data and be able to identify applicable standards and best practices for each dataset. Fortunately, as part of the University Library, RDS is able to leverage not only the library's technical infrastructure, but also the expertise of preservation staff and subject specialists, who have a strong understanding of the practices in their domain area. The expertise of our small RDS staff cannot possibly cover all of our campus research disciplines, thus relying on functional and subject specialists, such as the Library's GIS Specialist or the Librarian liaison to the depositor's department, helps to guide our curation process. After performing the general check on metadata and files, the curator decides if a particular dataset would benefit from having a review from the liaison for the depositor’s department and reach out to them. An example of this collaboration is when we received a dataset in zoology. After the review, the curator reached out to the subject liaison for this area with questions regarding the dataset. At times, our initial curation questions turn out to be common practice in this field, thus no changes needed. Without the help of the subject specialists, we may have requested unhelpful changes to the dataset, potentially frustrating the depositor and reflecting poorly on our curation practices. This working relationship provides a mutual benefit for both RDS and subject specialists; the dataset receives expert curation while subject specialists can strengthen their relationship with their department.

As Johnston shared in an OCLC blog in 2020, hiring curators skilled in all file formats with diverse backgrounds to properly curate the data for reuse is nearly impossible for institutions. Following Heidorn’s suggestion (2011) about cross-institutional collaboration, in 2016, Illinois became a member of the Data Curation Network, a cross-institutional network with a shared-staff model to support open data. When a dataset is deposited in a format that we lack the local expertise to review, we send this dataset to the Network. For example, we currently lack local expertise in MatLab and NetCDF or images in CZI formats, so datasets containing these file types are sent to the DCN for review/curation.

JeSLIB 2021; 10(3): e1202
https://doi.org/10.7191/jeslib.2021.1202

When providing feedback to researchers, we try to limit our recommendations to the three most important actions to improve the understandability and re-usability of the dataset. Our goal is to provide actionable suggestions without overwhelming busy researchers. All suggestions from curators are optional, and depositors are free to accept or reject those suggestions.

Our detailed curation workflow is captured in the comic figure below:

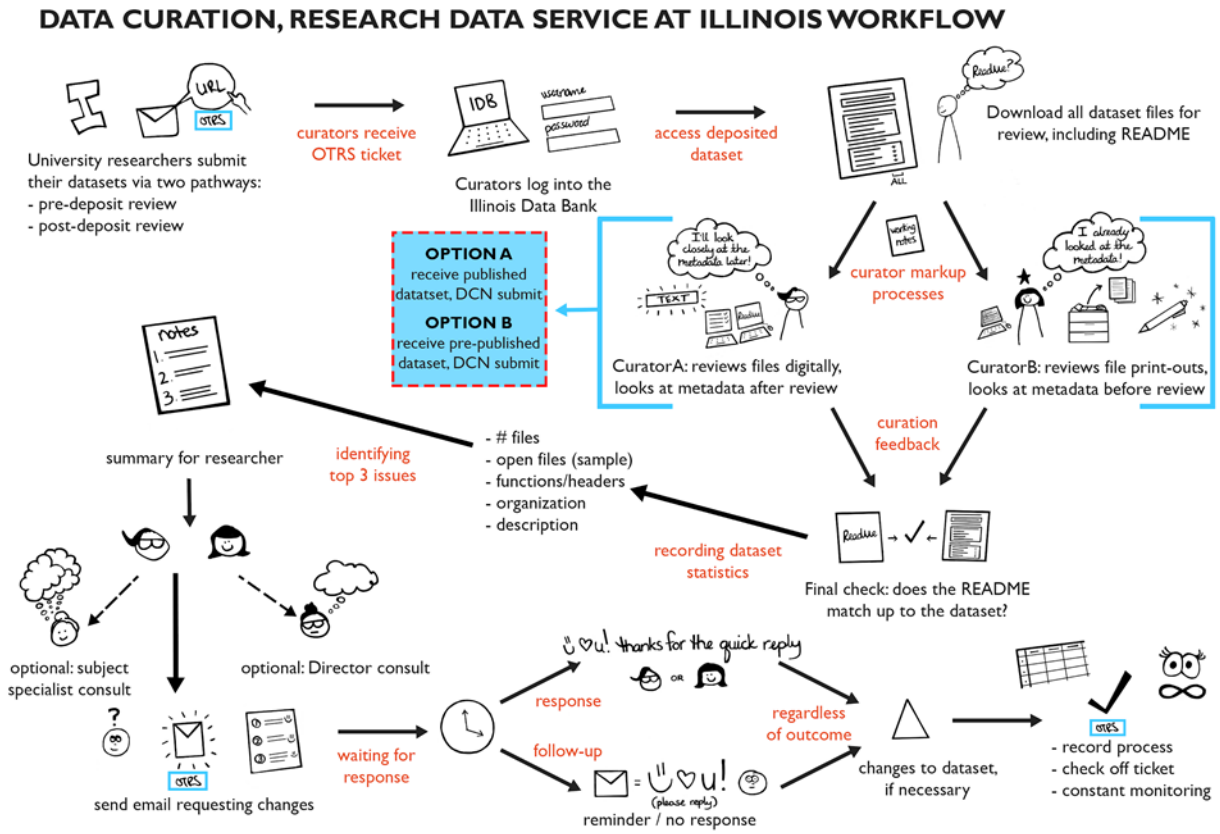

Figure 4: Our data curation workflow, as first presented at Data Curation Network All-Hands in meeting 2018.

For all 300 datasets, issues found during curatorial process are categorized into three different levels:

- Major curation actions required – this includes requesting that the depositor add documentation; missing files/attributes/values were found which requires correction(s) from the depositor; and/or suggestions that the depositor add data in a more preservable format along with any proprietary formats, if applicable. As shown in Figure 5, 46% (n=137) of datasets fall into this category.
- Minor curation actions required – this includes curators fixing typos in the record; addressing metadata by adding more clarification, funders, or keywords to enhance discoverability; and/or provide more descriptive title.

JeSLIB 2021; 10(3): e1202
https://doi.org/10.7191/jeslib.2021.1202

As shown in Figure 5, 24% (n=72) of our datasets are in this category.

- Basic curation – this means the dataset is in good shape and the curators have no suggestions to make. As shown in Figure 5, 30% (n=91) of the datasets are in this category.

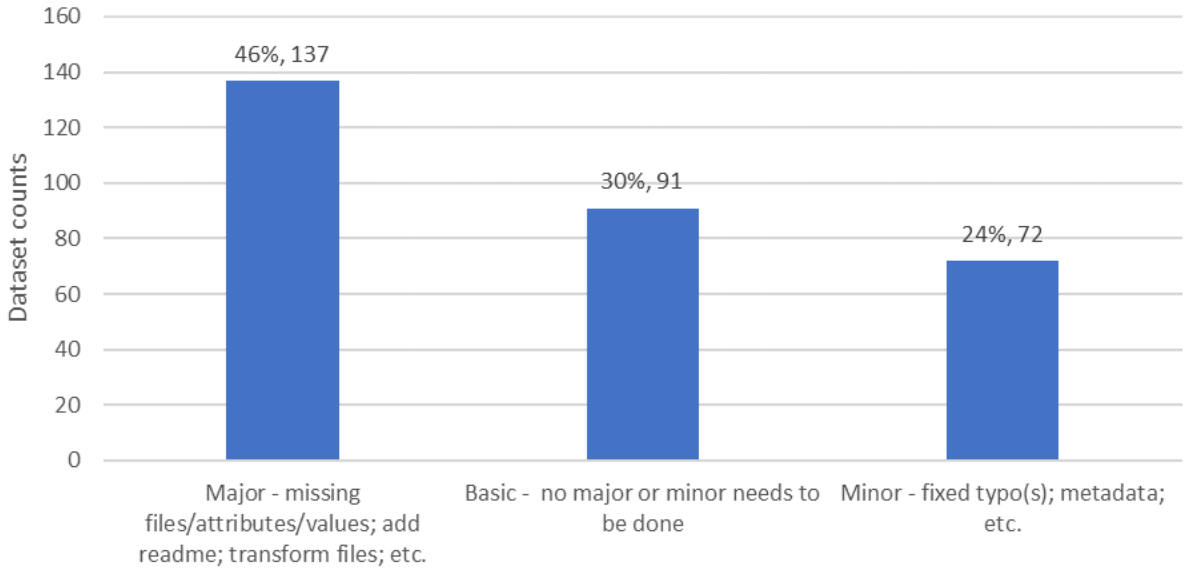

Figure 5: Distribution of datasets by level of curation action required.

We do encounter several challenges during curation. For example, since we began tracking responses rates in June 2018, approximately 22% of datasets had no response from the depositors. This poses a difficult obstacle because we do not apply any major changes to the datasets without their permission.

To assist our researchers who may lack the time to address curator recommendations, we've adopted the practice of our colleagues at other universities and attach partially completed documentation (as a Readme.txt, using Cornell University's template[2]). The information is manually filled out by the curator based on the metadata in the Illinois Data Bank and leaves blank sections that require input from data creators. Researchers are generally willing to complete the documentation when we do this.

In the next section, we describe the process of developing a pre-publication feature that provides an opportunity for the curator to provide feedback while the dataset is still in draft stage.

2 Cornell University's README.txt template: https://data.research.cornell.edu/content/readme

JeSLIB 2021; 10(3): e1202
https://doi.org/10.7191/jeslib.2021.1202

Section 2: Balancing the researcher needs and curation benefits.

The Illinois Data Bank is a self-deposit platform. This means the depositor is free to upload their files and publish the data at their convenience without waiting for approval. The datasets are curated after publication by our curators, which we refer to as post-ingest curation. This process allows depositors to obtain their DOI in a timely manner as we find that they work whenever the time permits and often under very tight deadlines. For example, we have had datasets published in the middle of the night, on weekends, over breaks, and even on major holidays. However, per our repository policy, any changes to files within a published dataset requires creation of a new version of the dataset, which is less than ideal (Fallaw et al. 2016). In two years of operation, we observed some confusion among depositors about versioning, and the process of versioning is time consuming for both depositor and curator. Although the unit offers consultations before the submission process, we rarely received requests for dataset review before its publication. Initially, 19% of datasets resulted in versioning with the majority of them requiring major curation action. To be more proactive in assisting depositors in preparing the dataset in a proper manner and to minimize the chance of needing a new version, in 2018, the unit decided to create a feature that allows and encourages depositors who are not in a time crunch to request a review prior to dataset publication. This feature is called "pre-publication review."

The process started with a team meeting where we presented the idea to our programmer, which required that we decide on the workflow and how the new feature would be connected to our current ticket system. Understanding how critical communication is to a pleasant user experience, we chose language carefully and tested iteratively to ensure that it was clear and concise enough that researchers would fully understand the options and ideally be persuaded to choose the recommended option. After a month of refining the language, we came up with the final result which is shown in Figure 6, and the feature was implemented in June 2018.

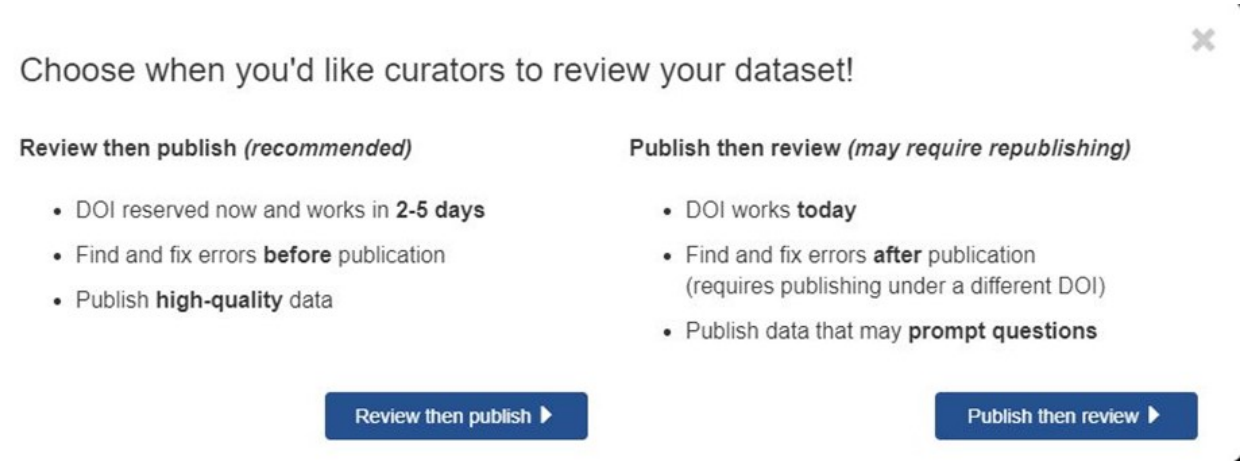

Figure 6: Screenshot of the prompt which shows options for the depositor to choose from before they move forward with their submission.

JeSLIB 2021; 10(3): e1202
https://doi.org/10.7191/jeslib.2021.1202

Since its implementation, 143 (70%, n=205) of the datasets have been reviewed (as of December 2020) using the "review then publish" feature. The adoption rate suggests depositors are open to working with curators to improve their datasets, which is a success in itself. We have also observed a drop in the need for versioning. Currently, 100% of datasets that opted for pre-publication reviews have not required versioning to resolve curatorial issues (although some versioning has been done to satisfy subsequent journal reviewer comments). During pre-publication review curators recommended major changes for 54% (n=77) of datasets. Recommendations were sent to depositors and addressed prior to publication; thus, versioning was avoided for these datasets.

We have observed that pre-publication review also has an impact on documentation. Documentation is known as the heart of data management. Missing documentation or insufficient documentation is one of the barriers in data reuse. In 2011, Tenopir et al. conducted an international survey to study data practices and perceptions of scientists. The survey of 1329 respondents revealed the lack of awareness about the importance of metadata or documentation among the scientific community. After that baseline survey, they did a follow-up survey and published it in 2015. The results indicated more interest and willingness in data sharing among scientists, along with increasing in concerns over data being misinterpreted. Kervin et al. (2013) also found that not providing enough metadata to enable others to interpret and potentially reuse the data was one of the common errors found among researchers when publishing their data for sharing and reuse (Kervin et al. 2013). This emphasizes the importance of documentation in data reuse. Tenopir concluded in their study that having a proper documentation would potentially ease those concerns (Tenopir et al. 2015). Creating the relevant metadata is as essential as presenting datasets themselves (Kowalczyk 2011).

We speculated that depositors would benefit from pre-publication review to help prepare their datasets before publication, specifically through improvements in documentation.

Since documentation is a good indicator, to check our hypothesis, we compared the number of documentation files provided before and after the time the "pre-publication review" feature was implemented. To do this, we began explicitly tracking to see if documentation was added after the curator's suggestion for all datasets deposited after November 2019. This new metric is not strictly accurate for past datasets since we could not always tell if documentation was added at the curator's suggestion. Therefore, this comparison serves as an estimate.

In 95 datasets deposited before the "pre-publication review" feature was available, 32% (n=30) were deposited without documentation and 19% (n=18) resulted in versioning with 33% (n=6) versioned specifically to add documentation. Of the 205 datasets published after implementation of the "review then publish" option, 70% (n=143) were reviewed prior to publication, and documentation was recommended/added for 54% (n=77) of the pre-publication review datasets. As

JeSLIB 2021; 10(3): e1202
https://doi.org/10.7191/jeslib.2021.1202

mentioned above, none of those datasets required versioning due to curatorial issues. In comparison, of the 62 datasets that did not undergo pre-publication review, 52% (n=32) were published with major curation issues, including inadequate documentation, and 12% (n=4) were later versioned in order to add/ correct documentation.

Because new pre-publication review lets us catch missing documentation beforehand, we not only have more datasets published with documentation, conveniently, we also have fewer datasets with documentation-related versioning. As mentioned above, lack of specific curator recommendation metrics prior to June 2018 prevents firmer conclusions, but our results suggest that our depositors benefit from the pre-publication review feature, and that the curation process improves datasets before they become publicly available.

Section 3: Technical Challenges

Our researchers increasingly need to share large, complex datasets. This calls for reliable, highly available, cost-effective, scalable storage accessible to computation resources. We have been drawn to the promise of these features in cloud services with some success, but also with painful bumps along the way. Whether on premises or in the cloud, we manage challenges related to large file transfer for ingest, curation, and access. Other challenges across platforms involve extracting technical metadata about the contents of archive file types such as zip and tar to support use and curation of complex dataset directory and file structures.

Cloud infrastructure

When Illinois Data Bank launched, all storage and web service resources were hosted and administered on-campus. This worked, but we saw room for improvement in reliability, streamlined system administration, scalability, and flexibility. Unexpected downtime in storage system availability required curators to spend time remediating incompletely processed datasets since processing steps depending on the storage system may have been disrupted.

For a myriad of reasons beyond the scope of this paper, the file storage utilized at the time of Illinois Data Bank's launch had been designed to be integrated with a compute cluster for active research projects, not to support highly-available web applications with uncertain and variable storage needs. We wanted storage designed to be more highly available, and to scale better with storage needs, which we expected to grow but it was not clear to us how much or when.

An additional element of the environment building momentum toward migration to cloud architecture was that our university's centralized technical services group was brokering a deal with Amazon Web Services (AWS) and encouraged us to take advantage of the potential benefits they were evaluating. This aligned with trends in the library community (Goldner 2010).

JeSLIB 2021; 10(3): e1202
https://doi.org/10.7191/jeslib.2021.1202

After a year of exploratory pilots and adaptation to cloud infrastructure, on February 18, 2019 we switched the production instances of our digital library storage and web applications from on-campus platforms to Amazon Web Services (AWS). The reliability improvements from the migration were as significant as we hoped, ending the storage-system related curation headaches of tracking down and remediating processes that may or may not have happened during unexpected disruptions.

Large File Transfer

While the migration to cloud infrastructure provided many benefits, this change also brought new challenges, particularly in the area of large file ingestion.

In our experience, researchers with data files greater than 50GB particularly value our free-to-them institutional support of sharing their data in our repository, but support at that scale calls for techniques beyond basic web forms. In offering to support sharing of up to 2TB per researcher per year, we understood that we would need to support alternative ingest workflows.

Soon after the Illinois Data Bank's launch, a data file ingest API and sample python client were developed, which worked by incrementally adding chunks to a file on a filesystem, which became the go-to solution for command-line transfer of files from research computing clusters. One of the adaptations from an all on-campus infrastructure vs. cloud object storage and web servers involved modifying the API for data transfer. Despite our efforts to optimize the API, in field conditions, our researchers started encountering frustrating reliability and transparency issues with files larger than ~100 GB using the updated API and sample client adapted to our cloud architecture. On top the aggravation of failed transfer attempts, this led to rushed, disruptive portable hard drive transfers.

Enter Globus, a transfer service, familiar to many researchers who routinely transfer large files. Globus users put the files in a bucket or a filesystem some other way, then expose them through a Globus endpoint. These users can then transfer files between endpoints using Globus, and then manipulate them using other tools in the same location. This was key to its utility for our purpose. Although setting up Globus for use by Illinois Data Bank required significant organizational and technical investment initially, the investment has produced very useful results.

While the pain that prompted us to integrate Globus into Illinois Data Bank related to transfer in, once it was set up, we also offered transfer out, which has become increasingly important as Illinois Data Bank hosts larger datasets. When a dataset landing page is served, Illinois Data Bank checks if the files are available in Globus. If available, the interface offers a link to the Globus File Manager page for that dataset. If problems arise from Illinois Data Bank users attempting to download very large datasets using other methods, non-Globus downloads may be disabled for those datasets, where "very large" means whatever size causes

JeSLIB 2021; 10(3): e1202
https://doi.org/10.7191/jeslib.2021.1202

problems. Through implementation of Globus, we've been able to improve the experience for our end users and also minimize some of the challenges curators (and programmers) have faced while trying to work with large datasets.

Facilitating the curation of archive file types

Many of the largest and most complex datasets include archive type files, such as .zip, .tar, or .7zip, which can make these datasets challenging to curate without some way to provide an overview of the contents. When our infrastructure was entirely on-campus, we pointed file analysis tools at the files on the filesystem that was used for preservation and access. However, with the cloud infrastructure set up now, in order to offer a listing of the contents of the archive files in the user interface and for curation analysis, we had to adjust to extract the files to filesystem storage and traverse the resulting tree.

As development has settled into the AWS ecosystem it became apparent the archive extraction process could be transitioned into a serverless solution. Research archive deposits into the Illinois Data Bank are large, infrequent, and require increased compute power for short bursts of time making the extraction process ideal to utilize a serverless architecture.

After a period of researching options, implementation, and testing we devised a solution that could utilize existing AWS products without the need to maintain any new technologies ourselves. In short, we now utilize AWS's Fargate Elastic Container, Elastic File System, Simple Storage Service, and Elastic Container Repository for file transfer, decompression, and traversal and AWS's Simple Queue Service to communicate with the Illinois Data Bank. Ultimately, this allows curators and users to readily view the trees of archived files within the Illinois Data Bank interface. This is similar to the functionality prior to our migration to the cloud, and we were able to realize our ultimate goal, which was to develop mechanisms that were intuitive for our curators and users.

Conclusion

In this paper, our goal was to reflect on our curation services. We provided examples of how we use our internal metrics to monitor the growth of the Illinois Data Bank, highlighted improvements to our curation workflow, and described the technical challenges and the solutions we've used to help us to offer high-quality curation services. As technology evolves, our services must also evolve. In order to do this, we implemented new features to reduce the curation workload, migrated to cloud infrastructure for cost efficiency and reliability, implemented Globus for uploading and downloading large files, and adjusted our system to accommodate easier curation of archive file types. As we reflect on these examples, we note that any changes to the curation process, whether it be workflow adjustments or migration to cloud infrastructure, cause a cascade of adjustments that have to be made in order for the system to function as intended and expected. Internal metrics have been crucial in helping us prioritize the

challenges where improvements will be of enough benefit to justify the effort invested. Given their value so far, we expect that we will continue to add to and standardize these metrics to help us monitor and assess our curation services.

Acknowledgements

The authors wish to thank Daria Orlowska for digitizing the data curation workflow comic (Figure 4).

References

Braxton, Susan, Colleen Fallaw, Hoa Luong, Daria Orlowska, Ashley Hetrick, Kyle Rimkus, Bethany Anderson, and Heidi Imker. 2018. "Should We Keep Everything Forever? Determining Long-Term Value of Research Data." Poster presented at: iPres2018, Boston, USA. http://hdl.handle.net/2142/91659

Fallaw, Colleen, Elise Dunham, Elizabeth Wickes, Dena Strong, Ayla Stein, Qian Zhang, Kyle Rimkus, Bill Ingram, and Heidi J. Imker. 2016. "Overly Honest Data Repository Development." *Code4Lib Journal* 34(2016-10-25). https://journal.code4lib.org/articles/11980

Gewin, Virginia. 2016. "Data sharing: An open mind on open data." *Nature* 529: 117–119. https://doi.org/10.1038/nj7584-117a

Goldner, Matthew R. 2010. "Winds of Change: Libraries and Cloud Computing." *BIBLIOTHEK Forschung Und Praxis* 34(3): 270–275. https://doi.org/10.1515/bfup.2010.042

Heidorn, P. Bryan. 2011. "The Emerging Role of Libraries in Data Curation and E-science." *Journal of Library Administration* 51(7-8): 662–672. https://doi.org/10.1080/01930826.2011.601269

Johnston, Lisa. 2020. "How a network of data curators can unlock the tremendous reuse value of research data." *OCLC (blog)*. https://blog.oclc.org/next/data-curators-network

Kervin, Karina E., William K. Michener, and Robert B. Cook. 2013. "Common Errors in Ecological Data Sharing." *Journal of eScience Librarianship* 2(2): e1024. https://doi.org/10.7191/jeslib.2013.1024

Kim, Youngseek, and Sean C. Burns. 2016. "Norms of Data Sharing in Biological Sciences: The Roles of Metadata, Data Repository, and Journal and Funding Requirements." *Journal of Information Science* 42(2): 230–245. https://doi.org/10.1177/0165551515592098

Kleidl, Marius. March 7, 2016. *Tus.io (blog)*. https://tus.io/blog/2016/03/07/tus-s3-backend.html

Kowalczyk, Stacy, and Kalpana Shankar. 2013. "Data Sharing in the Sciences." *Annual Review of Information Science and Technology* 45(1): 247–294 https://doi.org/10.1002/aris.2011.1440450113

National Science Board. 2011. "Digital Research Data Sharing and Management." Accessed February 17, 2021. https://www.nsf.gov/nsb/publications/2011/nsb1124.pdf

Office of Science and Technology Policy. 2013. https://web.archive.org/web/20160304043850/https:/www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Office of the Provost. n.d. "The Next 150: Strategic Plan for 2018-2023." The University of Illinois at Urbana-Champaign [online]. Accessed 16 February 2021. https://strategicplan.illinois.edu

Office of the Provost. n.d. "The Illinois Strategic Plan." The University of Illinois at Urbana-Champaign [online]. Accessed February 16, 2021. https://strategicplan.illinois.edu/2013-2016/goals.html

Stuart, David, Grace Baynes, Iain Hrynaszkiewicz, Katie Allin, Dan Penny, Mithu Lucraft, and Mathias Astell. 2018. "Whitepaper: Practical Challenges for Researchers in Data Sharing." figshare. https://doi.org/10.6084/m9.figshare.5975011.v1

Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff and Mike Frame. 2011. "Data Sharing by Scientists: Practices and Perceptions." *PLOS ONE* 6(6): e21101. https://doi.org/10.1371/journal.pone.0021101

Tenopir, Carol, Elizabeth D. Dalton, Suzie Allard, Mike Frame, Ivanka Pjesivac, Ben Birch, Danielle Pollock, and Kristina Dorsett. 2015. "Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide." *PLOS ONE* 10(8): e0134826. https://doi.org/10.1371/journal.pone.0134826