code{4}lib JOURNAL

Issue 34, 2016-10-25

ISSN 1940-5758

Overly Honest Data Repository Development

After a year of development, the library at the University of Illinois at Urbana-Champaign has launched a repository, called the Illinois Data Bank (https://databank.illinois.edu/), to provide Illinois researchers with a free, self-serve publishing platform that centralizes, preserves, and provides persistent and reliable access to Illinois research data. This article presents a holistic view of development by discussing our overarching technical, policy, and interface strategies. By openly presenting our design decisions, the rationales behind those decisions, and associated challenges this paper aims to contribute to the library community's work to develop repository services that meet growing data preservation and sharing needs.

By Colleen Fallaw, Elise Dunham, Elizabeth Wickes, Dena Strong, Ayla Stein, Qian Zhang, Kyle Rimkus, Bill Ingram, and Heidi J. Imker

Background

While data have long underpinned research findings, routine and widespread access to and preservation of research data has not been emphasized until recently. The current interest in research data stems from the desire to enable greater research transparency, increase the return on investment of government-funded research through data reuse, and demonstrate by a number of long-standing, discipline-specific data archives where data can have significant long-term value within and beyond their domain interests.
Although the concept of research data as an important research output has gained traction in many disciplines, the state of data availability varies dramatically. Today data may be made available as separate files that are supplemental to a journal article, which is a highly flexible, albeit wildly variable, sharing mechanism. However, supplements are not always allowed (Maunsell 2010; Borowski 2011). Even when data are shared in supplemental files, the data are frequently entombed in PDFs, which are not an ideal format for reuse and do not lend themselves to easy extraction of the data. Furthermore, supplemental files are not always persistently available (Williams 2016), are subject to size limitations (Kenyon and Sprague 2014), and are rarely independently discoverable from the article itself (Carpenter 2009, Imker 2016). Some communities have established discipline-specific repositories, such as the Inter-university Consortium for Political and Social Research (ICPSR; https://www.icpsr.umich.edu/icpsrweb/), GenBank (https://www.ncbi.nlm.nih.gov/genbank/), and the Protein Data Bank (PDB; http://www.rcsb.org/pdb/home/home.do). These resources are tailored to specific communities and data types and serve both extremely well through establishing and/or enforcing standards, providing quality control, and aggregating like-data. While discipline-specific repositories are a critical component of the data ecosystem, their specificity means they are unable to accommodate the extraordinary diversity of research data that exists. Finally, in recent years general-purpose repositories such as figshare (https://figshare.com/) and Zenodo (http://zenodo.org/) have emerged and represent a midpoint on the flexibility continuum. Such repositories do require minimal, but not discipline-specific, metadata. On one hand, the required minimal metadata is an improvement over the unstructured nature of supplemental files, but advanced curation is not supported within these general-purpose repositories due to their scale and scope. Furthermore, their long-term funding models are uncertain, suggesting that their missions will be more susceptible to change in the face of sustainability. Therefore, although such resources are a welcome addition to the landscape, we can be less certain about their ability to retain and preserve deposited data.

It's clear that there are gaps in the current solutions for making data available, but it's also clear that resources are emerging and evolving. The Office of Science and Technology Policy (OSTP) memo of 2013 (Holdren 2013) and subsequent implementation plans from federal agencies indicate that we can expect some development of data sharing mechanisms from federal agencies in the future (Scholarly Publishing and Academic Resources Coalition 2016). However, murkiness abounds as to what those developments will look like; data inventories, a data "commons," and/or partnerships with third party services were all mentioned in agency implementation plans. While the vagueness of the agency plans make it difficult to envision what the agencies may develop in the future, we note that agency-developed repositories were universally proposed for public access to research articles. Despite years of library-led institutional repository (IR) development, at the time of writing, of the 16 federal agency plans, only the US Department of Energy mentioned IRs as a solution for public access to research articles (US Department of Energy 2014). It is not clear if the agencies' lack of acknowledgement toward IRs signifies a barrier with regard to roles and perspectives about what modern libraries do or if agencies feel the need to show an independent, and possibly individualistic, commitment to the public access mandate. Regardless, if history were to repeat itself, in five or ten years we can anticipate that agencies will develop their own data repositories with associated requirements that researchers deposit their federally-funded data within those specific agency-lead resources; thus, we have prepared ourselves for the chance that our efforts to build a data repository may be a short-term, stopgap solution.

Overall Strategy

While the national "big picture" remains nebulous, University of Illinois at Urbana-Champaign researchers are coming forward in search of a data publishing platform that fills the gaps left by current options. This leaves us in a predicament. How do we support the needs that researchers have today, yet plan for an unpredictable future? At some level, the research prowess of an institution is proportional to its research output, and institutional support of such output is increasingly expected. As more research receives funding, more research publications and related data will be produced. Therefore, an institutional data repository serves as a parallel to our other research collections and also provides a "home" for research data with an otherwise ambiguous fate.

The Univeristy Library at University of Illinois at Urbana-Champaign (referred to throughout as "Illinois Library") already maintains several digital repositories, including the Illinois Digital Environment for Access to Learning and Scholarship (IDEALS; https://www.ideals.illinois.edu/). IDEALS primarily serves as a document repository for articles, reports, theses, dissertations, presentations, and in some cases, small datasets. IDEALS launched 10 years ago and has since amassed >85,000 items with >22,000,000 item downloads. Accessions and downloads are steadily increasing over previous years, and we are comfortable saying that IDEALS is successful in the goals of being persistent, collecting a variety of scholarly materials, and being adopted across campus. Since 2012, the Library has worked to develop Medusa, a preservation repository that centralizes preservation for all of the Library's digital collections (Rimkus 2015). The Medusa digital preservation repository provides an enduring storage and management environment for the digital content collected or created by the Library. It also serves as a backbone infrastructure platform for building public-facing access to the Library's digital materials via various web applications.

We considered many options during initial planning for data preservation and access and attempted to draw on our current repository development and operations experience to assess those options:

1. Customize our current DSpace IR (IDEALS) to better support datasets
2. Adapt our preservation system (Medusa) and build an integrated web application for dataset ingest and access
3. Adapt our preservation system (Medusa) and integrate with a separate existing platform, e.g. DSpace, Fedora/Hydra/Sufia, CKAN, Dataverse, figshare for Institutions, etc.

Each of these options has pros and cons. Ultimately, our goal was to centralize as much of the repository functions as possible, as we have found that maintaining different repository systems is taxing on resources; thus we rejected option 3. Additionally, because we are uncertain about data in terms of long-term expectations, value, preservability, and scalability, we also wanted to make sure to be pragmatic in our commitments and plan for change. While we could have tried to work within the DSpace platform to create additional customization in IDEALS, it would have required a change in policies and functionality that would likely be orthogonal to IDEALS' current success as a document repository originally intended for sustained, long-term preservation. Building on DSpace also would have meant customizing the underlying software, which is outside of our control and likely to change. Over-customization is already a problem with our IDEALS instance of DSpace; further customization to accommodate research data would exacerbate the difficulty. In light of this, we rejected option 1. Therefore, we elected to go with option 2 and develop a web application that interacts directly with Medusa in order to leverage Medusa's preservation functions. This approach would allow us to benefit from the long-term preservation functions of Medusa while managing data-specific policies and commitments in a separate web application. By building the "data repository" as a front-end web application on top of the Medusa preservation platform, smooth integration would enable robust management and preservation of data deposits and at the same time support the Library's ongoing approach to digital stewardship. We elected to name the application the "Illinois Data Bank" because it is not an acronym [1], relays clarity of purpose, is easy to remember, and no one objected.

Funding Model

The Illinois Data Bank (http://databank.illinois.edu) is developed and operated primarily by the Research Data Service (RDS; http://researchdataservice.illinios.edu), a campus-wide program that provides the Illinois research community with the expertise, tools, and infrastructure necessary to manage and steward research data. In an ongoing commitment, the RDS was funded in 2013 by the Office of the Vice Chancellor for Research at the University of Illinois at Urbana-Champaign, but the RDS has its home in the Illinois Library to leverage the established professional LIS, digital preservation, and repository services expertise within the Library. Along with developing and operating a repository for public access to Illinois research data, primary services offered directly through the RDS include knowledge of policies and best practices for the preservation and sharing of research data, advising on data management planning and implementation, and providing data management training. RDS funding supports four full-time staff: a director, two data curators, and a repository developer. While the developer is devoted entirely to the Illinois Data Bank, the former three have responsibilities that cover the entirety of the RDS's services.

Team Construction

Although the technology stack sometimes takes center stage, a repository encompasses much more than lines of code, a server, and some storage. Indeed, policies, workflows, adoption and adaptation of a metadata schema, and user interface design are all critical components of repository development work. Development of the Illinois Data Bank therefore required the input from multiple people with expertise as summarized in Table 1. All parties participated in regular brainstorming and discussion of requirements, features, challenges, and potential solutions, although participation ebbed and flowed depending on the stage of development. RDS staff and their developer adopted an Agile development process with weekly sprints documented in a JIRA ticketing queue. The developer, the RDS director, data curators, and CLIR postdoc met weekly to discuss features and progress regarding the interface and curation functions, with others participating on an as-needed basis. For particularly challenging features, such as determining the complex workflows involved in supporting embargoes, the data curators worked to suss out the functional requirements ahead of time in an attempt to present the developer with a fairly mature vision of the feature. The team iterated upon most features throughout the process of articulation, implementation, and testing. This resulted in a tight coupling of the repository developer, who otherwise works directly within the Library's repository development team, with the RDS team that would be primarily responsible for promoting and operating the Illinois Data Bank as a service on campus. We found this overall structure to be effective because it did two things: 1) created an environment where everyone represented in Table 1 was mutually invested in the success of the project but did not feel overly taxed with unnecessary meetings and 2) ensured that no one role was isolated from others. While occasional (but relatively few) issues surfaced with this team structure, they arose fairly quickly because of the interconnectedness, and were likewise resolved quickly by bringing needed parties together.

Table 1. Staff breakdown during initial Illinois Data Bank development (May 2015 – May 2016).

Staff	% of Time	Primary Duties During Illinois Data Bank Development
Repository Developer	100%	programming, technical implementation, testing
RDS Director	30%	direction, develop cohesion, gather researcher input and feedback, testing, arbitration
RDS Data Curation Specialist	30%	metadata schema, policies, copyright, curation workflow
RDS Data Curation Specialist	30%	environmental scans, embargoing, ticketing, curation workflow
Repository Services Manager	10%	direction, technical oversight
Metadata Librarian	10%	cross-walking of metadata schema, controlled vocabularies, metadata content format, implementing DataCite metadata schema 3.1
CLIR Postdoc	10%	testing, documentation
Information Design Specialist	10%	usability design, usability testing
Preservation Librarian	5%	direction, technical oversight

Technical Infrastructure

Illinois Repository Suite Overview

In developing the suite of repository services within the Library, the repository development team adopted the microservices architectural style of designing and building software applications to support nimble and sustainable infrastructure. A suite of services emerged over short, iterative development cycles, based entirely on the business needs of the librarians. This established an ongoing relationship between those at the library, including the RDS team and repository development team, and the software in production continues to evolve through continuous refactoring as new user needs and requirements, known as "user stories" in Agile parlance, arise.

One user story in particular is most responsible for the overall design of repository services at the Library. Above all, the preservation librarians wanted file-level access to all content inside the repository. They rejected the common practice employed by most popular repository solutions of restricting access to content to a set of (often Web-based) APIs. Likewise, they required that files ingested into the repository would not need to be renamed or obfuscated and that folder names and directory hierarchy would be persisted within the repository, unmolested. This requirement would allow preservation staff to use their existing suite of digital preservation tools directly on the files in the repository via standard Portable Operating System Interface for Unix (POSIX) APIs; they would not be too beholden to the development team to build new, or adjust existing, custom APIs in order to accommodate inevitable changes in preservation tooling. To accomplish this, the base of the Medusa repository infrastructure is a General Parallel File System (GPFS), which provides robust, high performance, petabyte-scale, clustered file storage that can be mounted by any ordinary computer and operated on by preservation staff without the need for additional tools or training.

The decision to base our repository architecture on a regular POSIX file system led to an important overarching strategy of "meeting users where they are," rather than requiring retraining or onerous adjustments to established workflows as a bar to adopting our software. This strategy has influenced several other design decisions, often characterized by tradeoffs between features and ease of use, but in most cases both goals were achieved. For us, the focus on meeting users where they are has resulted in a smoother rollout and wider adoption of Medusa compared to other comparable library software products developed in recent years.

Digital Preservation

The heart of Illinois' repository infrastructure is the Medusa digital preservation repository (https://medusa.library.illinois.edu/; https://github.com/medusa-project) which provides long-term retention and accessibility of its digital collection (Rimkus and Witmer 2016). In development of Medusa, the Library's goal was to closely integrate the Library's digital production, preservation, and access services in order to establish a more efficient, sensible, and sustainable digital library program.

Medusa is developed and managed locally by the Illinois library's repository development team, and features a web-accessible interface for collection management and initiation and tracking of preservation actions, such as format tracking, on-demand integrity checking, and storage of preservation metadata along with depositor and Illinois Data Bank-supplied content. Medusa's storage infrastructure, established through a partnership with Illinois' National Center for Supercomputing Applications, consists of a copy of every file replicated daily across two distinct campus nodes, both on spinning disk, and a third copy of every file backed up and stored outside of Illinois.

While Medusa content consists primarily of digitized and "born digital" materials such as books, manuscripts, photographs, audiovisual materials, scholarly publications from the Library's special collections, general collections, and repositories, it was determined research data entrusted into Library care should also be ingested into the Medusa digital preservation repository. As of 2016, Illinois Data Bank constitutes a sixth source of content in Medusa.

Illinois Data Bank Construction

The web interface component of Illinois Data Bank is a Ruby on Rails application (Fallaw 2016). The repository development team uses Ruby On Rails to support rapid development. Functional prototypes provoked and inspired meaningful feedback within weeks of the developer's start date.The interface continues to evolve responsively to feedback. Externally, Illinois Data Bank integrates with the DataCite Metadata Store (DataCite 2016) through Purdue's EZID service (Purdue University 2016).

When a depositor confirms an intention to publish, the web application requests a DOI (digital object identifier). Within a pre-registered prefix, the EZID API supports generation of a random DOI or a specified DOI. Illinois Data Bank specifies a DOI that is largely opaque, but encodes version information. After EZID returns a DataCite DOI, Illinois Data Bank sends a message to Medusa to initiate ingestion into the digital preservation system. Building on an approach Medusa uses for other functionality, the messages are sent using Advanced Message Queuing Protocol, specifically using a RabbitMQ server. Messaging supports effective integration, while allowing independent development. Custom selection of files for zipping and downloading files in published datasets, which are stored in Medusa, is supported with a distinct web service called Medusa Downloader.

Policy Framework

In order to effectively manage the research data stewarded by the Illinois Data Bank, we developed a robust set of policies that articulate the obligations of the Illinois Data Bank service and its variety of users. After nine months of drafting, revising, consulting, and fine-tuning, we debuted the set of policies, procedure, and guidelines when the Illinois Data Bank soft-launched in May of 2016 (Table 2).

Table 2. Illinois Data Bank Policies, Procedures, and Guidelines.

Policy Title	URL
Access and Use Policy	http://hdl.handle.net/2142/91041
Accession Policy	http://hdl.handle.net/2142/91042
Deposit Agreement	http://hdl.handle.net/2142/91043
Preservation Policy	http://hdl.handle.net/2142/91044
Preservation Review, Revision, Retention, Deaccession, and Withdrawal Procedure	http://hdl.handle.net/2142/91045
Withdrawal Guidelines	http://hdl.handle.net/2142/91616
Preservation Review Guidelines	http://hdl.handle.net/2142/91617

Preliminary Policy Efforts

To begin policy development work, we looked to the policies that govern IDEALS (Shreeves 2014) as a solid starting point because they had been vetted by University stakeholders and cover the policy areas previously identified as being important when establishing a repository for University scholarly outputs. In addition, we looked to peer repositories' policies (as acknowledged in each of the policies listed in Table 2) and reviewed policy requirements for achieving Data Seal of Approval certification (Data Seal of Approval 2016). The Data Seal of Approval (http://www.datasealofapproval.org/en/) policy requirements were useful to us even though we do not have an immediate goal of obtaining certification because they helped us guide our policy framework toward one that reflects that of a trusted repository. Review of established policies and Data Seal of Approval documentation led to the understanding that the Illinois Data Bank needed policies that governed the deposit, preservation, access, preservation review, and withdrawal of datasets. The policies needed to dictate what is expected of three major players in the Illinois Data Bank workflow: the depositor, the repository, and the accessor.

Ad Hoc Review Group

After developing initial drafts, we engaged with a wide variety of campus stakeholders to form an Ad Hoc Review Group for the policies (Table 3). Since the Illinois Data Bank serves all of the Urbana-Champaign campus, researchers' decisions to share data in the Illinois Data Bank could potentially impact a number of offices and roles on campus, so it was important to see that many voices were represented in the development of the policy framework.

The Ad Hoc Review Group undertook three rounds of revisions and met three times. The revision process involved soliciting in-line revisions and comments from each member. The feedback gathered ranged from suggested grammatical revisions to questions that challenged us to make operational and policy changes. We compiled feedback between each round of review and highlighted major questions and concerns as discussion points at subsequent Ad Hoc Review Group meetings.

The perspective brought by the variety of representatives from all over campus was invaluable to seeing that the Illinois Data Bank policies are adequately balanced in terms of the roles and responsibilities expected of depositors, the repository (and thus the University), and accessors. An additional benefit of working with the members of the Ad Hoc Review Group was that it provided the opportunity for in-depth discussions about the intentions and functionality of the Illinois Data Bank with professionals on campus who are connected with University researchers and may be advising researchers regarding data sharing needs in the future.

Table 3. Participating university staff and units in Illinois Data Bank policy review.

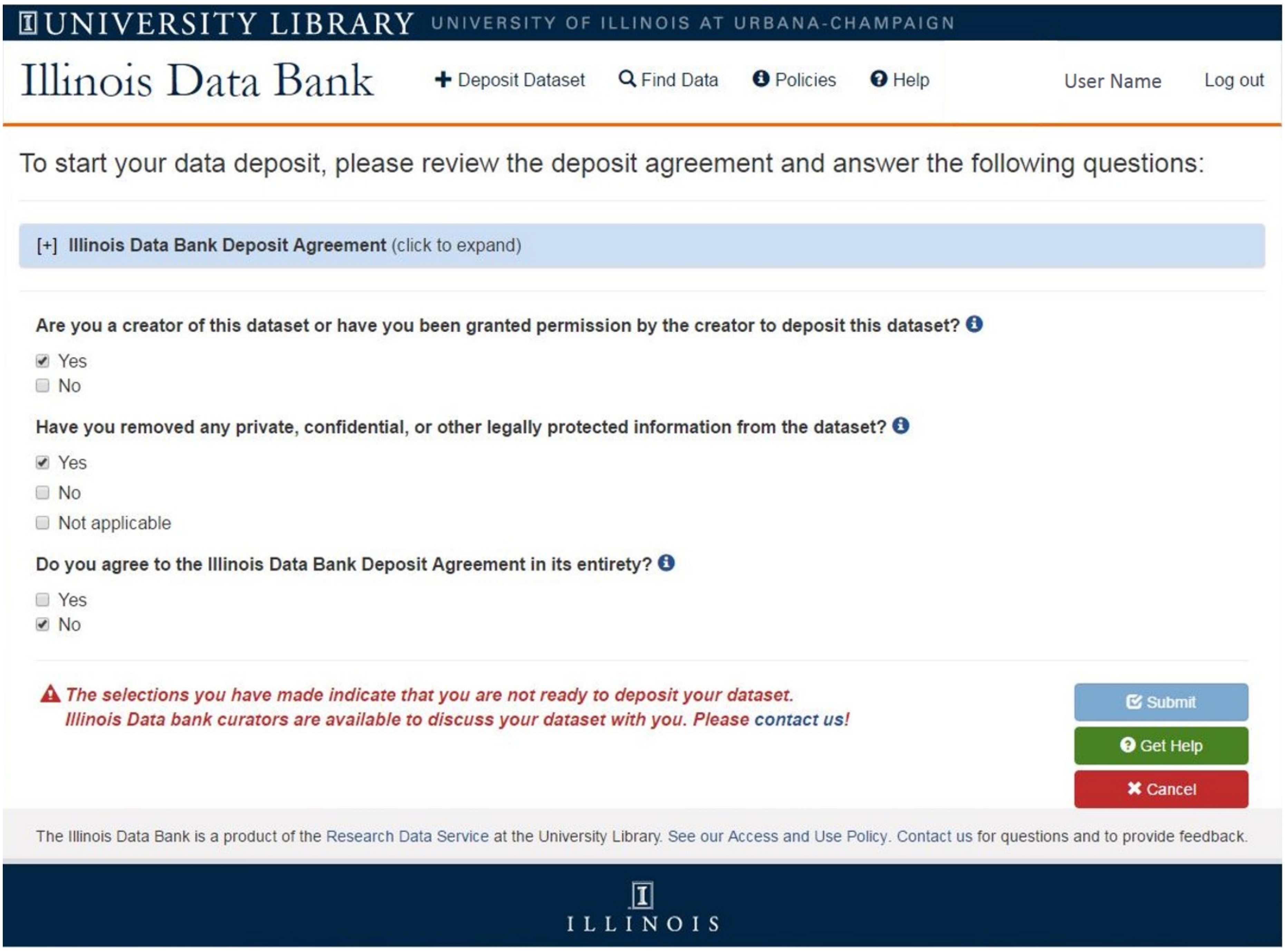
Staff Role	Campus Unit
Director, Research Data Service	University Library
Data Curation Specialist	University Library
Senior IT Security Engineer	Technology Services
Export Compliance Officer	Office of the Vice Chancellor for Research
Research Ethics Officer	Office of the Vice Chancellor for Research
Deputy CIO of Research IT and research leader at National Center for Supercomputing Applications	Technology Services; National Center for Supercomputing Applications
Director, Institutional Review Board	Office for the Protection of Research Subjects
Associate Vice Chancellor for Research and Director of Sponsored Programs	Office of the Vice Chancellor for Research; Office of Sponsored Programs
Director, Records and Information Management Services	Records and Information Management Services
Senior Associate University Counsel	University Counsel
Senior Technology Manager (1)	Office of Technology Management
Senior Technology Manager (2)	Office of Technology Management
Assistant Dean for the Graduate College	Graduate College
Associate University Librarian for Research	University Library
Manager, Scholarly Communication and Repository Services	University Library
Engineering, Physics and Astronomy Librarian	University Library

Interplay between Policy and Repository Development

The creation and refinement of the Illinois Data Bank policies occurred in tandem with the development, testing, and refinement of the Illinois Data Bank features and interface. The team was flexible and adaptable to the adjustments that ended up being required as the policy and functional elements of the Illinois Data Bank changed together.

For example, user testing of an early version of the Illinois Data Bank, which loaded the Deposit Agreement as a big block of text, indicated that many depositors were not likely to read the Deposit Agreement. From a policy perspective, there were particular elements of the Deposit Agreement that we felt were important to highlight for the depositor: the fact that depositors must either be a creator of the dataset or acting on behalf of a creator of the dataset and that all sensitive information, if any, must be removed from the dataset. The team worked together to develop an interface solution that would streamline the Deposit Agreement and require depositors to address these specific elements.

Thanks to our ongoing communication with the Ad Hoc Review Group, the Deposit Agreement workflow was refined over time to incorporate a workflow block (i.e., a depositor cannot move forward with their deposit when "No" is selected for any question) and text that recommends depositors seek help from the RDS when they have selected "No" for any question (Figure 1). Furthermore, these conversations inspired more sophisticated use of the deposit agreement by prompting us to design the system to create a plain a text version of the completed deposit agreement with the timestamp of submission and depositor information, which is both saved within the dataset metadata package and also linked in the deposit confirmation email sent to all authors.



To start your data deposit, please review the deposit agreement and answer the following questions:

[+] **Illinois Data Bank Deposit Agreement** (click to expand)

Are you a creator of this dataset or have you been granted permission by the creator to deposit this dataset?

☑ Yes
□ No

Have you removed any private, confidential, or other legally protected information from the dataset?

☑ Yes
□ No
□ Not applicable

Do you agree to the Illinois Data Bank Deposit Agreement in its entirety?

□ Yes
☑ No

⚠ *The selections you have made indicate that you are not ready to deposit your dataset.*
Illinois Data bank curators are available to discuss your dataset with you. Please contact us!

The Illinois Data Bank is a product of the Research Data Service at the University Library. See our Access and Use Policy. Contact us for questions and to provide feedback.

Figure 1. Deposit agreement showing page structure and an alert when a depositor has not indicated appropriate responses.

Another example of the Ad Hoc Review Group inspiring development work is that it became clear that it would be necessary to incorporate a mechanism for temporarily and permanently withdrawing access to files and/or dataset metadata. The conversations leading up to our decision to implement this capability centered on a number of potential problems that could emerge as the Illinois Data Bank began to accept submissions. Worst-case scenarios were articulated and listed as examples of compelling reasons for withdrawal in the Illinois Data Bank Preservation Review, Revision, Retention, Deaccession, and Withdrawal Procedure (see link in Table 2 for details). While one of the goals of the RDS is to circumvent many of these potential issues through educational efforts, it is likely that not every depositor will have had direct interaction with us before depositing. Thus, having the means to withdraw a dataset is important in case any of the withdrawal scenarios do occur.

With a new policy decision—the Illinois Data Bank has the right to withdraw datasets for a compelling reason—new system capabilities had to be developed in response. The RDS developer worked with others on the repository development team to implement support for the withdrawal of files and metadata in the Illinois Data Bank. We developed a tiered approach to escalating withdrawal actions to include temporary suppression, permanent suppression, and deletion of files and metadata. To accomplish this, new dataset and file properties were added to control how elements of display presented to system users in various roles, such as curator, depositor, and guest. For the extreme cases identified as having a legal or ethical need to delete even archived (non-public) versions of files from the preservation system, the repository development team architected potential system changes to support automated withdrawal of files from the preservation system, but did not implement prior to launch because the expectation is that the need for complete deletion would be a rare-if-ever occurrence. Instead, a manual process was identified, and we shelved system implementation of an automated process until more need arises. We then worked on developing the curator-only interface for managing file and metadata temporary and permanent suppression and internal procedures for determining when to implement such measures (see Curator Interface vide infra). The hope is that the need for this feature is likewise scarce but having it functional and available at launch was important to ensure that if a questionable deposit was identified, it could be immediately and effectively managed.

Policy Maintenance

Recognizing that the data publishing and scholarly communication landscape are constantly in active evolution, we will review all Illinois Data Bank policies on a quarterly basis during the Illinois Data Bank's first year (May 2016-May 2017). In subsequent years, the policies will be reviewed as needed but at least annually. Staying attentive to the work of, and working in coordination with, other Library teams, such as the units who manage IDEALS, the University Archives, and Medusa, and campus groups such as the Records & Information Management Services and Technology Services, the Illinois Data Bank intends to maintain the currency and robustness of its policy framework well into the future.

User Interface Design

One of the driving goals of the interface design was simplicity. We wanted to make sure that people depositing their data clearly understood what was happening with their deposit at every stage, and that the depositors wouldn't need to struggle through a difficult process regardless of how complex the behind-the-scenes decisions had been. The process of continually refining the interface while also adding new functionality meant that traditional usability assessment techniques needed some acceleration to fit smoothly into the rapid development cycle.

Jakob Nielsen's usability recommendations suggest batches of 5 users at a time, all experiencing the same interface, followed by modifications, and then another batch of 5 users (Nielsen 2000). However, that process would have put an unacceptable delay in the development cycle. Instead, the team began with heuristic analysis, generally performed one to two Morae-based usability assessments (https://www.techsmith.com/morae.html) per week, and analyzed the feedback in a way that integrated how recently the issue was encountered along with how many people had encountered it and how critical we assessed it to be. (More recently encountered issues meant that either the problem had not been resolved or the problem had been introduced.) Some other concerns were assessed with A/B click testing and term recognition assessments with the aid of the Optimal Workshop software suite (https://www.optimalworkshop.com/).

There were three major usability assessment cycles between October 2015 and April 2016, with ongoing and significant interface changes (often implemented the same day or week as the test) in response to user feedback. Sometimes the initial insights revealed by heuristic analysis and user testing were right on the mark. For example, many web-based software users have been trained to expect access to personalized views and controls by clicking on their name in the upper right quadrant of a screen. So many users displayed this behavior that we added a personalized view of their own datasets in that area, even though it hadn't been in the original development plan.

Other times we came up with seemingly-great ideas that didn't pan out in actual testing. The idea of the home page containing a visual display of the stages of deposit had appealed to us even in early stages of development; however, usability testing of a graphic designer's rendering of that idea took an unexpected turn. All of the usability testers who saw that graphic on the homepage tried to click on the graphics as though they were interactive buttons despite conscious efforts to avoid a graphic that invited interaction. Because the graphic design arrived very close to launch, the team decided it was better to fall back to a less-informative general data visualization rather than add new site functionality and interactions to accommodate the unintended trigger of expectations.

In some cases, user behavior patterns required an interface change that wasn't technically necessary but was behaviorally necessary to prevent confusion. The Illinois Data Bank is configured to save a user's progress at every step, so that closing a window would store an in-progress draft without content loss. However, despite the presence of a Google Drive-like message that the draft was saved, users never felt comfortable simply closing a window during testing; they needed the explicit reassurance of a clearly labeled "save" feature. While several other labels were tested over the course of development, "save and continue" and "save and exit" were the phrases that reassured data depositors that they knew what would be happening to their data at the next step.

In other cases, fairly small interface adjustments had very large results. The words "continue" and "confirm" by themselves didn't look as "clickable" to our users as "continue >>" and "confirm >>" did. Similarly, items that users perceived as headlines were suddenly made visually interactive by adding "[+]" to the start of the phrase. We made several adjustments to the tooltip "(?)" icon in terms of size, color, and placement in order to make it noticeable but not too distracting.

The interface transformation between October and April was dramatic. The policy agreement was streamlined from dozens of paragraphs to key checkpoints; the deposit process shrank from several pages to one; the upload system progressed from one file at a time to several methods of uploading, including from Box, a University-supported file storage service that provided larger file upload methods than the web browser alone would reliably support. New and potentially complex features such as data embargo and funder identification were also added at the same time the interface was being simplified. The usability testers at the end of April had several more tasks to accomplish, but their tests were progressing more quickly because they experienced less confusion thanks to the usability improvements introduced throughout the development cycle.

By adapting the usability testing size and frequency to the Agile development cycle and scheduling small but frequent rounds of feedback, we were able to test ideas quickly and adjust the interface accordingly. The experience was almost like digital sculpture, carving away complexities and refining the visual interface, testing new ideas, and responding in near-real-time to the users' experiences.

Metadata Schema

Metadata Schema Development

To establish the Illinois Data Bank descriptive metadata, we assembled a "metadata sub-team" comprised of an RDS data curator and a University Library metadata librarian, with additional feedback from the rest of the Library Metadata Services and the Illinois Data Bank development teams. The metadata team used the DataCite Metadata Schema 3.1 as the starting point for developing the Illinois Data Bank metadata (Ammann et al. 2011) because every dataset deposited into the Illinois Data Bank is assigned a DataCite DOI.

While researchers and journals may desire DOIs for the sake of persistency and/or branding, we also saw a major discoverability advantage in Illinois Data Bank metadata being transmitted back to the DataCite Metadata Store as a potential indexing source for data harvesters. A major shortcoming of an institution system can be isolation, and we deemed it imperative that data deposited into the Illinois Data Bank be discoverable in the larger data ecosystem.

In the process of determining what information would be captured when a researcher deposited a dataset, the metadata team sought to harmonize the mandatory DataCite elements and local institutional priorities and also find a delicate balance between "pain-free" and "enough" metadata for depositors. The DataCite Metadata Schema 3.1 defines eighteen properties, five of which are mandatory: Identifier, Creator, Title, Publisher, and PublicationYear. The Illinois Data Bank automatically generates several pieces of metadata: the DOI, version number, and publisher, the latter always being "University of Illinois at Urbana-Champaign." In addition to the requisite DataCite information, the Illinois Data Bank also requires:

- License/rights information
- Email address for each author
- Contact Person
- Date Available (i.e., when the dataset is available to the public)
- Version

All of the information collected by the Illinois Data Bank is not necessarily sent to DataCite. For example, the email addresses for each author are collected for institutional purposes only and are not sent to DataCite. The Illinois Data Bank sends the following information back to DataCite:[2]

- Author information, including the ORCiD, if applicable
  Corresponding Author
- Title
- Item identifier
- Funding agency name, including the Open Funder Registry identifier, if applicable
- Publisher
- Publication year
- Keywords
- Date available to the public in the Illinois Data Bank
- Version number
- Rights information, including Creative Commons identifier, if applicable
- Dataset description
- Resource Type (always Dataset)
- Related item information, including identifiers, relationship type, and the type of related resource, if applicable

Techniques such as drop down menus, auto-generated metadata, and ORCiD integration enable the collection of consistent, reliable, and robust metadata that adheres to best practices as well as laying the foundation for future work. The metadata team strove to keep the amount of free-text metadata entry to a minimum in an effort to simplify the process for depositors and mitigate errors. For example, the author list features an optional ORCiD ID lookup search for each author name. If an author has an ORCiD, the lookup will auto-populate their name and ORCiD number once selected in the tool's search results, thus limiting the possibility of typos.

The Illinois Data Bank requires a corresponding author to be designated for each dataset, which is indicated by selecting a radio button next to the author's name in the deposit interface. Other fields use drop down menus to collect information, including: license, publication delay, funding body name, and related resource type. The menu options for both the License and Funder fields include terms from existing controlled vocabularies. A depositor can opt to publish their dataset under CC0 or CC-BY licenses; selecting either of these options will populate the metadata using the appropriate Creative Commons identifier. A depositor may also choose the 'Other license' option. However, the depositor must upload the license language as a .txt file, which is enforced by the system in order to complete a deposit. The 'Funder' field is similar in that the menu includes a pre-populated list of options and an 'Other' free-text option. The pre-populated list consists of the top ten most common grant funding agencies at the Urbana-Champaign campus. The terms from this list are drawn from the Open Funder Registry taxonomy hosted by CrossRef (Crossref 2016), and the metadata includes respective DOIs from the same registry.

The implementation of related identifiers and our discontent at implementation has been described in a forthcoming publication (Dunham et al. 2016) so the discussion here will be brief. Of the twenty-five possible relationType terms in DataCite 3.1, we chose to use only six. This decision was based on analysis of current usage by other DataCite data centers, close reading of the documentation and other supporting materials, and personal communication with members of the DataCite community. Due to the complexity of defining relationships between items, depositors indicate that a related item exists by selecting the type of related resource from a short list, the values of which are: Article, Code, Dataset, Presentation, Thesis, or Other. The choice of 'Other' generates a free-text box for users to input their own resource type. When a depositor populates the related materials section, the data curators enhance the dataset record by selecting the appropriate relationType and checking that the related identifier is entered properly. We decided that being able to link related items reliably is a true value-add capability of the Illinois Data Bank, and therefore we are willing to commit to the curatorial time required.

Metadata Views

Currently, there are three user-facing ways to view item level metadata in the Illinois Data Bank. The first is the default HTML dataset landing page. It is also possible to view the dataset metadata as XML, by appending '.xml' to the browser URL, which dynamically generates the XML record. This valid DataCite XML record contains only the information that is sent to DataCite for inclusion in the DataCite Metadata Store. Finally, adding '.json' to the browser URL produces a JSON object. In addition to the metadata transmitted to DataCite, the JSON view includes timestamps for any updates made to the descriptive metadata, but does not include the complete changelog information at this time.

Future Metadata Work

Many of the decisions made for the initial iteration of the Illinois Data Bank lay the foundation for future work, including migration to DataCite Metadata Schema 4.0. For example, we had already made the decision to record all author names in 'name parts', i.e., two separate boxes for the given and family names a researcher chooses to publish under and transmit a concatenated version to DataCite[3]. Name parts are not currently supported by DataCite 3.1 but will be supported in 4.0 (DataCite Metadata Working Group 2016).

The metadata team has finalized the current metadata application profile for the Illinois Data Bank (Dunham and Stein 2016) and is monitoring the DataCite DCAP, mapping documents, and DataCite SPAR Ontology work. Mid-term plans include making the Illinois Data Bank metadata available as linked data. This may take shape by adding schema.org semantics to the HTML dataset landing pages as microdata, as well as by serializing the metadata into triples and making it available via a triple store. This linked data application profile is being designed to align with the University of Illinois Digital Library metadata application profile and will integrate into Library-wide RDF efforts.

Challenging Features

Licensing

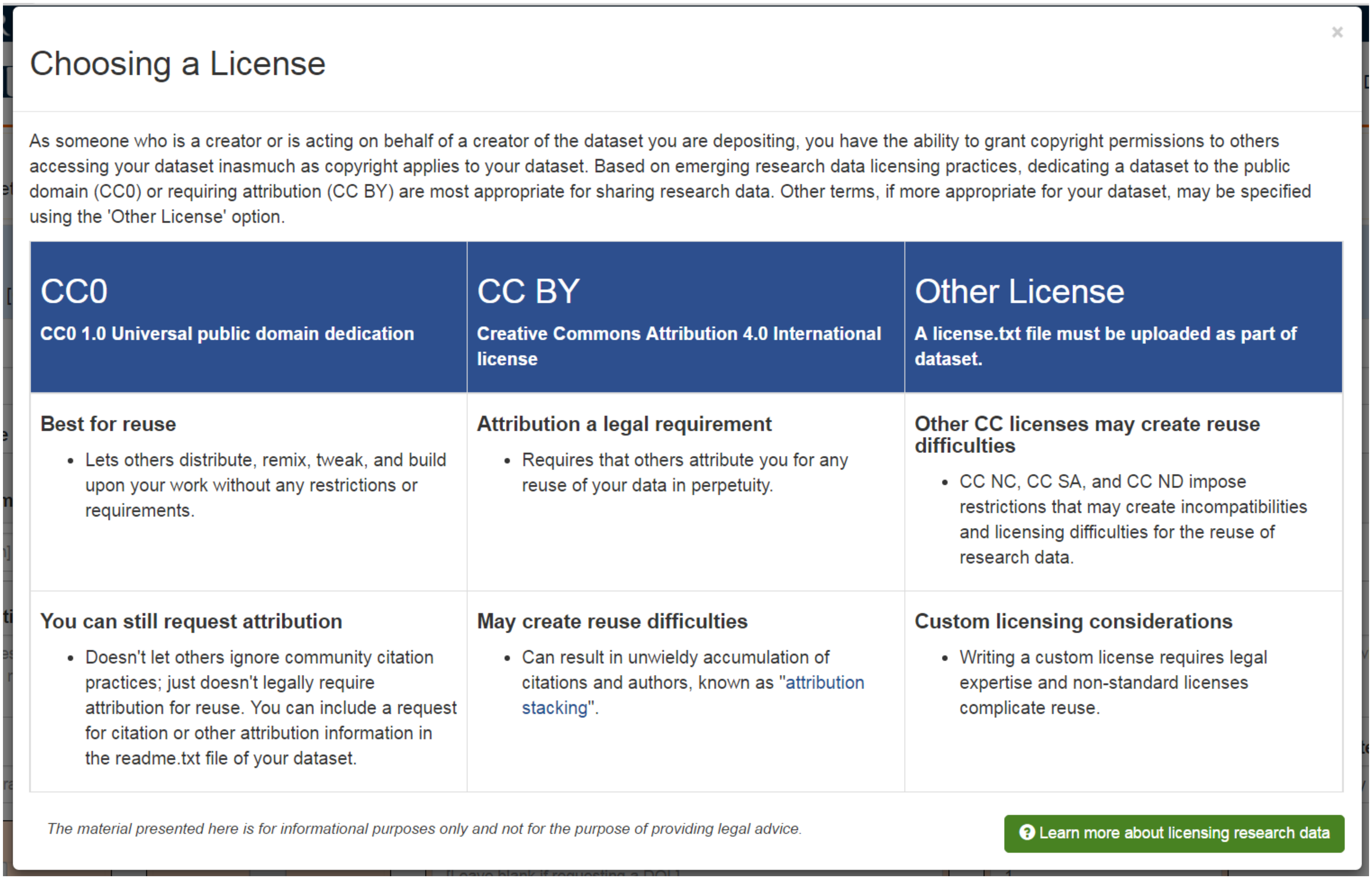
We researched the current state of data licensing (Ball 2014; Carroll 2015), and ultimately put a great deal of thought and discussion time into developing the license section of the Illinois Data Bank interface. Complex questions threaded throughout our considerations of licensing datasets, such as whether data can be copyrighted, who owns data created by researchers at the University, and what role an institutional data repository should play in promoting open licensing and legal interoperability of research data. Our research and discussion resulted in determining the following guiding principles for thinking about licensing in the Illinois Data Bank:

- Public datasets are more efficiently shared and reused when the rights of the accessor are clearly communicated.
- Public datasets are more efficiently shared and reused when any rights associated with them are waived, by using, for example, a CC0 public domain dedication.
- Attribution (i.e., dataset citation) would ideally be compulsory in scholarly communications and should therefore not need to be a legal requirement associated with a dataset.
- Attribution is really, really important to scholars (Kratz and Strasser 2015)
- Dataset creators, as rights-holders for any rights associated with their datasets, should be empowered to manage the rights associated with their datasets in whatever way they desire. The role of the Illinois Data Bank is to enable data sharing, not dictate the ways in which datasets must be shared.
- Dataset depositors, as potential data accessors and reusers themselves, must comply with licensing already placed on source data.
- Many would-be depositors do not gain familiarity with issues related to copyright and licensing throughout their academic training.

With a foundation in these principles, we came to the following decisions about the Illinois Data Bank's handling of licensing datasets:

- To support the clear communication of accessor rights, license field is required for datasets deposited in the Illinois Data Bank.
- To balance our desire to promote barrier-free openness of public data and our commitment to respecting the intellectual property rights of dataset creators, the dropdown selections represent our preferred licenses and support the right of the depositor to assign his or her own terms on behalf of all dataset creators. The depositor is presented with a dropdown menu with three options: "CC0 1.0 Universal Public Domain Dedication (CC0 1.0)," "Creative Commons Attribution 4.0 International (CC BY 4.0)," and "Other License (license.txt must be uploaded as part of dataset)."

  To support depositors throughout the self-deposit process, help content about licensing
- research data is provided and seeks to strike a balance between being informative and digestible for those new to issues associated with copyright and licensing (see Figure 2 and https://databank.illinois.edu/help#license).



Choosing a License

As someone who is a creator or is acting on behalf of a creator of the dataset you are depositing, you have the ability to grant copyright permissions to others accessing your dataset inasmuch as copyright applies to your dataset. Based on emerging research data licensing practices, dedicating a dataset to the public domain (CC0) or requiring attribution (CC BY) are most appropriate for sharing research data. Other terms, if more appropriate for your dataset, may be specified using the 'Other License' option.

CC0 **CC0 1.0 Universal public domain dedication**	CC BY **Creative Commons Attribution 4.0 International license**	Other License **A license.txt file must be uploaded as part of dataset.**
Best for reuse • Lets others distribute, remix, tweak, and build upon your work without any restrictions or requirements.	**Attribution a legal requirement** • Requires that others attribute you for any reuse of your data in perpetuity.	**Other CC licenses may create reuse difficulties** • CC NC, CC SA, and CC ND impose restrictions that may create incompatibilities and licensing difficulties for the reuse of research data.
You can still request attribution • Doesn't let others ignore community citation practices; just doesn't legally require attribution for reuse. You can include a request for citation or other attribution information in the readme.txt file of your dataset.	**May create reuse difficulties** • Can result in unwieldy accumulation of citations and authors, known as "attribution stacking".	**Custom licensing considerations** • Writing a custom license requires legal expertise and non-standard licenses complicate reuse.

The material presented here is for informational purposes only and not for the purpose of providing legal advice.

Figure 2. The licensing tooltip dialog box provides explanations of licensing options in the Illinois Data Bank. Click image for larger view.

Research data licensing is one of many areas that we will be consistently monitoring as norms and guidelines emerge across disciplines. Additionally, we expect to adjust help content and data management training efforts in response to questions and feedback we receive from depositors.

Delayed Release, a.k.a. Embargoes

The impetus of the Illinois Data Bank, in response to changing expectations from funders and publishers, was to develop an open and fully public repository, and we had no interest in supporting data on an inaccessible dark server. Not only did the idea of "dark data" seem misaligned with an increasingly open world, adding the ability to deposit a dataset not intended for immediate release would complicate workflows both for the depositor and the service. Considering our very intentional effort to keep the system light and intuitive, adding this complexity ran counter to our objective to keep the Illinois Data Bank simple to use and operate. As we grudgingly considered the embargo feature, a stream of immediate decisions to make quickly came to mind. How long could an embargo last? Will depositors need to embargo the whole dataset, including the metadata, or just the data files? Could datasets accidentally published prematurely be backed out into an embargoed state? Are we going to police whether an embargo is warranted? Is "embargo" even the right word? Despite our desire to avoid the issue entirely, in the end adding embargo functionality within the Illinois Data Bank was needed to support the following scenarios:

- Timing of article publication vs. data publication: we wanted to support the ability to include a dataset DOI in an eminent article publication, but understood that either the researcher or the publisher may not want the data to be publicly available prior to the article. [4]

- Other externally determined timing issues: e.g., as an implication of a thesis embargo, external data provider contractual agreements, recalcitrant collaborator, etc. Or, the worst scenario of all: a situation we didn't anticipate and therefore couldn't support.

Initially, we hoped we would be able to offer selection of a single date that would represent the point when the files would become public. This date would be permanent at the time of deposit, and an option to embargo the metadata would not be offered. From an interface standpoint, this was a simple and clean method of offering support. However, we knew that IDEALS offered both file-only and metadata and file embargos, and that both options were used. [5] Furthermore, as we continued to attempt to define the specifications for the embargo function, it became obvious that we didn't understand what our depositors needed, expected, or would even recognize, nor were there any external authorities or governing bodies that we could turn to for best practices for data specifically (in contrast to data licensing, which has been the focus of several high-level initiatives and working groups). Therefore, we elected to carry out a (perhaps overly) short Optimal Workshop survey of the embargo interface design and associated wording.

The response rate was small, with 7 respondents from an extremely convenient convenience sample. [6] All responded either positively (Very likely – 3; Somewhat likely – 2) or ambivalently (Unsure – 2) that they would need to embargo deposited data, which verified the need for the functionality to begin with. Respondents additionally reported that they would not be certain of the final release date for the embargo at the time of deposit. Both file-only embargoes and metadata embargos (along with the files) were desired. Respondents were also almost evenly split between use of the word 'embargo' (3 respondents) versus 'release date' (3 respondents) where we defined the word/phrase as indicating 'a situation in which data is not publicly available until a certain time.' These survey results, even though small and highly informal, revealed a fundamental issue with data publication at this time: researchers are unsure what to expect. Indeed, we realized the following complications were at play:

- Depositors are equally likely to relate to "embargo" and "delayed release"

- Depositors may legitimately not know the exact data release date they need

- Depositors may need to delay the release of just the files or both the files and the metadata

With this feedback in mind, we endeavored to create a system that would account for those complications. Because respondents were split between 'embargo' and 'release date' as the clearest wording, we chose a somewhat inelegant but straightforward approach: use both labels interchangeably (see Figure 3).

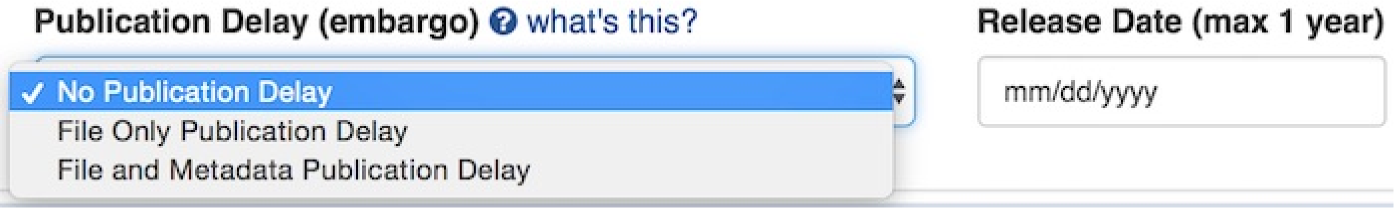
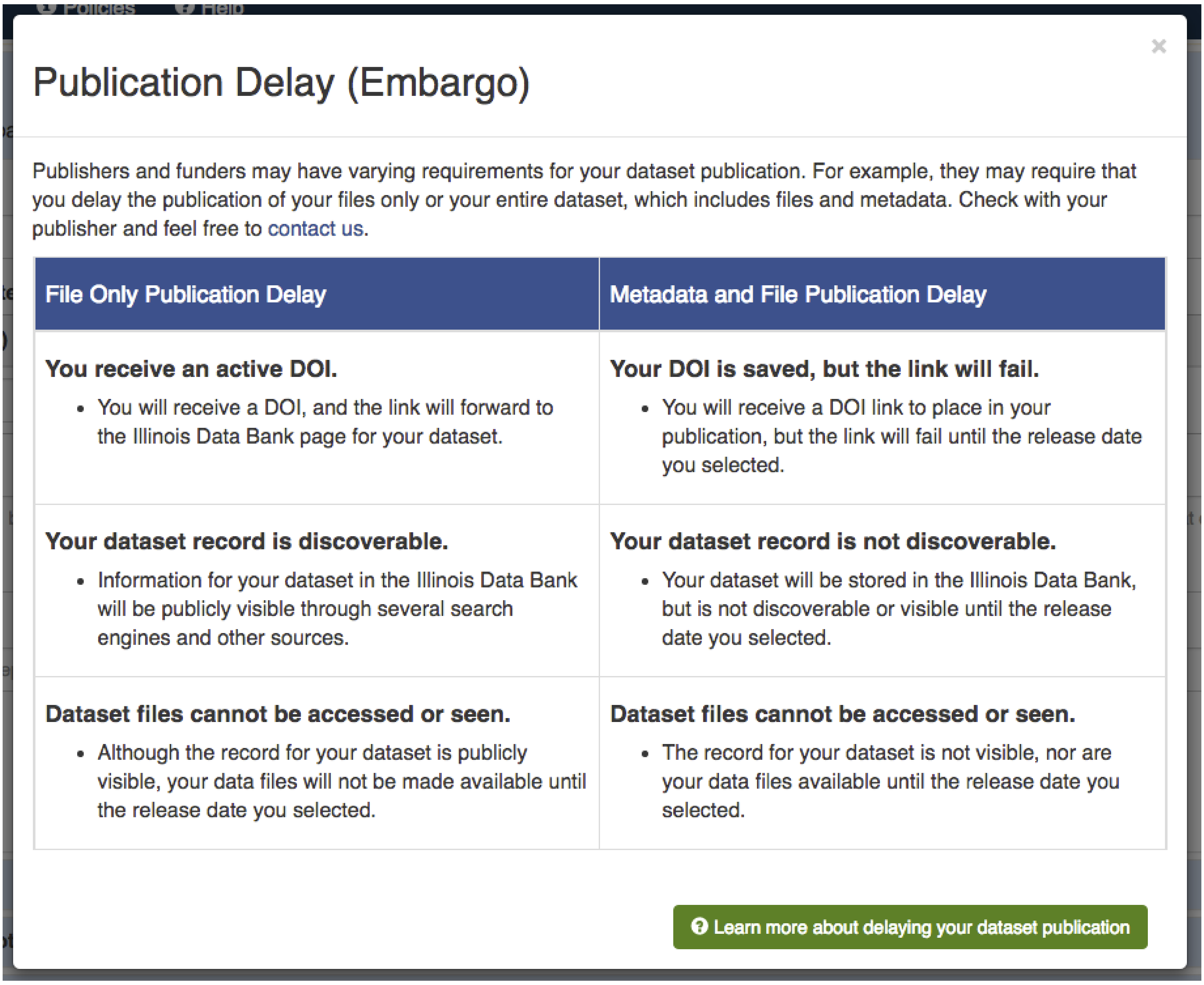

Publication Delay (embargo) ? what's this?
✓ No Publication Delay
File Only Publication Delay
File and Metadata Publication Delay
Release Date (max 1 year)
mm/dd/yyyy

Figure 3. Embargo options available to data depositors in the Illinois Data Bank.

To combat the issue of an indeterminate embargo date, we allow depositors to adjust the release date of embargoed datasets, and they are allowed to change the embargo state to a more public state at any time (vide infra). We established a maximum embargo period of one year, as this was the most common time frame indicated in the agency public access plans. However, curators have the authority to select any date, so that internally we have the ability to accommodate edge cases should they arise.

Once we decided to include embargo functionality, and with an understanding of the dangers associated with inadvertent publication, keeping depositors well informed of what was happening with the release of their dataset was vital. Similar to the licensing tooltip dialog box above, a table with core information was placed in a "what's this?" help link. This allowed us to include more detailed help information while keeping the deposit interface clean (Figure 4). The information provided in this area focuses strictly on the implications of embargo states from the perspective of the depositor and what may or may not be visible. As another way to mitigate any potential misunderstanding of embargo states, we also made sure that messaging within the deposit workflow was tailored to the specific embargo state the depositor selected (Figure 5).

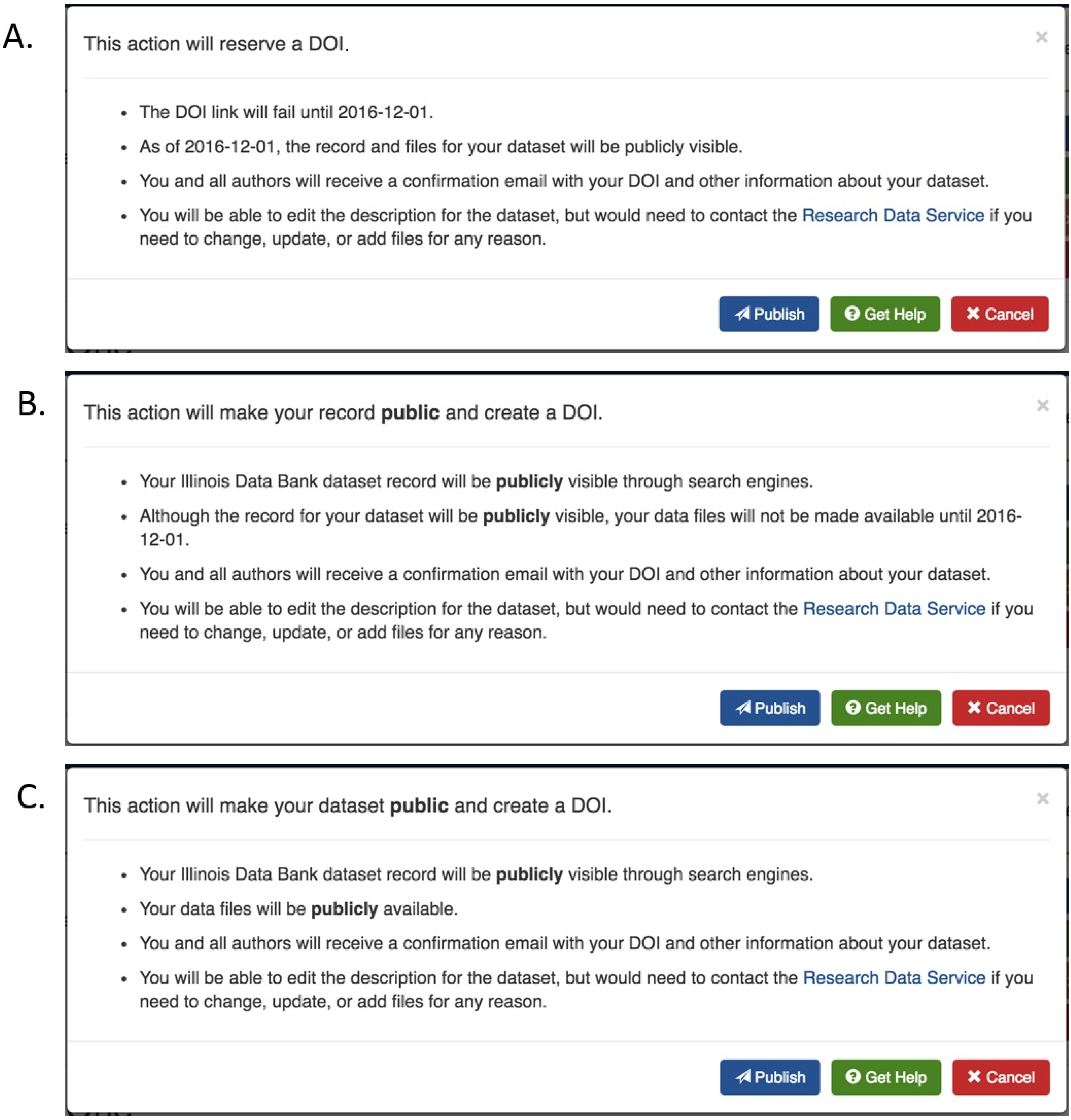


Publication Delay (Embargo)

Publishers and funders may have varying requirements for your dataset publication. For example, they may require that you delay the publication of your files only or your entire dataset, which includes files and metadata. Check with your publisher and feel free to contact us.

File Only Publication Delay	Metadata and File Publication Delay
You receive an active DOI. • You will receive a DOI, and the link will forward to the Illinois Data Bank page for your dataset.	**Your DOI is saved, but the link will fail.** • You will receive a DOI link to place in your publication, but the link will fail until the release date you selected.
Your dataset record is discoverable. • Information for your dataset in the Illinois Data Bank will be publicly visible through several search engines and other sources.	**Your dataset record is not discoverable.** • Your dataset will be stored in the Illinois Data Bank, but is not discoverable or visible until the release date you selected.
Dataset files cannot be accessed or seen. • Although the record for your dataset is publicly visible, your data files will not be made available until the release date you selected.	**Dataset files cannot be accessed or seen.** • The record for your dataset is not visible, nor are your data files available until the release date you selected.

Figure 4. The embargoing tooltip dialog box provides explanations of embargo options and ramifications.



This action will reserve a DOI.

- The DOI link will fail until 2016-12-01.
- As of 2016-12-01, the record and files for your dataset will be publicly visible.
- You and all authors will receive a confirmation email with your DOI and other information about your dataset.
- You will be able to edit the description for the dataset, but would need to contact the Research Data Service if you need to change, update, or add files for any reason.

This action will make your record **public** and create a DOI.

- Your Illinois Data Bank dataset record will be **publicly** visible through search engines.
- Although the record for your dataset will be **publicly** visible, your data files will not be made available until 2016-12-01.
- You and all authors will receive a confirmation email with your DOI and other information about your dataset.
- You will be able to edit the description for the dataset, but would need to contact the Research Data Service if you need to change, update, or add files for any reason.

This action will make your dataset **public** and create a DOI.

- Your Illinois Data Bank dataset record will be **publicly** visible through search engines.
- Your data files will be **publicly** available.
- You and all authors will receive a confirmation email with your DOI and other information about your dataset.
- You will be able to edit the description for the dataset, but would need to contact the Research Data Service if you need to change, update, or add files for any reason.

Figure 5. Confirmation prompts for the final step of publication are tailored to specific embargo scenarios. Panel A shows messaging if both the data files and the metadata record are embargoed. Panel B shows messaging if only the data files are embargoed. Panel C shows messaging if no embargos are selected.

While we were able to sort out the user interface such that our usability testers were able to navigate selecting an embargo with relative ease, the complications did not end there. Supporting detailed embargo options hit a classic stumbling point in design where the desired behavior seemed easy to explain while the technical implications were nearly the opposite. This was due to the nature of DataCite DOI states and the mixed nature of enabling embargo of just the files or of the files and the metadata. While the former renders a “published” DOI, the latter results in a “reserved” DOI. This is where a DOI is issued to the depositor for use within a manuscript but without the metadata being made public. The internal names for each of these states were transient in our conversations and changed right up until soft-launch of the Illinois Data Bank. We each differed in our personal preferences for these names and used things like invisible, partially visible, fully invisible, standard embargo, invisible embargo, public, etc. Linguistic drift was evident among all team members at different stages until the meanings became so muddled that development conversations were internally incomprehensible. This was the point when a specific meeting was called to come up with the final wording. Rather than attempting to assign code names or other single terms, team members opted to start with the most verbose options and trimmed them until the final labels remained. The final states are below:

Draft: when a record is saved within the Illinois Data Bank, but nothing has been communicated to DataCite. In this state:

- depositor may change the release date within the one year maximum and may also change the embargo state
- depositor may delete the dataset altogether at will

File and metadata embargo: when the dataset has been formally deposited into the Illinois Data Bank, a DOI is reserved, but the Illinois Data Bank dataset landing page and the DOI redirect fail. The depositor is provided a DOI suitable for inclusion in a publication, but they are told that link should be expected to fail. In this state:

- depositor may change the release date within the one year maximum or change the embargo state to either “No Publication Delay” or “File Only Publication Delay”
- depositor may not delete or “unreserve” the dataset without contacting the RDS

File embargo: when the dataset has been formally deposited into the Illinois Data Bank, and a DataCite DOI has been minted. While the landing page and the DOI work, the list of files is replaced with a message stating that the files are under embargo until the specified release date. In this state:

- depositor may change the release date within the one year maximum or change the embargo state to “No Publication Delay”
- depositor may not delete dataset nor can they “unpublish” the metadata without contacting the RDS

Metadata and files published: when the dataset is deposited and both the files and metadata are publicly available. This is the state where no embargos are in place. In this state:

- depositor cannot retroactively impose an embargo without contacting the RDS
- depositor may not delete dataset nor can they “un-publish” either the metadata or the files

Corrections and Versioning

One rule that appears to remain universal regardless of research domain or perspective is that the word “final” almost never actually means that something will never change again. Datasets are not immune to the need to make changes or clarifications. Instead of viewing file changes to published deposits as anathema and attempting to avoid the conversation entirely, we took the stance that changes, updates, and additions need to be transparent and sustainable to support. Indeed, as a stopgap method of hosting data until the Illinois Data Bank was live, the Library ingested small datasets into IDEALS. One of these datasets underwent two changes during the Illinois Data Bank development, giving the team experience with the very real consequences of not building in support for these tasks. Therefore, we understood that we would need to balance the need to enable corrections and the need to keep the integrity of the scholarly record. While support for versioning was marked as essential, a minimal curator-mediated workflow was adopted for the near term. An elegant, depositor-driven approach was something we decided could be tabled until we had better knowledge of the range of correction and versioning needs that would arise after the initial release.

As we considered how to implement versioning, we attempted to align our curator-mediated practices with that of the rest of the repository community. A cursory review of existing repositories was conducted during December 2015 to determine archetypical methods of supporting dataset versioning. [7] The 70 top DOI-issuing DataCite data centers (by volume) were reviewed and informally coded on how they handled versions. Only a few of these repositories seemed to represent dataset versioning on DOI landing pages. We summarize a subset of the repositories in Table 4 below.

In short, we found that most repositories appeared to handle versioning slightly differently, but we noticed two overarching approaches emerged: one which focused on the representation of the version on the dataset landing page and one which focused representation of the version within the dataset’s metadata. Both web and metadata aspects were important for the Illinois Data Bank, but we believe if we focused on metadata first, we would be more readily able to improve the web interface in the longer term.

Table 4. A subset of DataCite DOI-minting repositories reviewed for versioning practices.

Repository	Version type?	New DOI for major?	Old versions available?	Version to DataCite?	Uses version relation type?	Change log?
CDL.DIGSCI (figshare)	Major	No	Yes	No	No	No
CDL.LTERNET	Major	Yes	Yes	No	No	No
CERN.ZENODO	None; manual new entry	Yes; new entry	Yes	No	Supports	Manual
GESIS.ICPSR	Major	Yes (visible in DOI)	Yes	Mostly	No	Yes
GESIS.UBHD	Major on landing, major.minor to DataCite	No	Yes	Initial, but no updates	No	Yes
BL.OXDB	Major	No	Dataset landing page only	No	No	Limited
CDL.PEERJ (Document repository)	Major	Yes, visible in DOI	Yes	Yes	No	No
BL.YORK – Prospero (Document repository)	By date on dataset landing page, major in DataCite	No	Yes	Initial, but no updates	No	Yes
CDL.DATAVERS	Major.minor	No	Yes	No	No	Yes
TIB.TUB	Major	Yes	Yes	No	No	Yes
GESIS.ARCHIV	Major	Yes	Yes	Yes	Some	Yes

Ultimately, we decided that there were two distinct types of changes that we wanted the Illinois Data Bank to handle: changes to files associated with a deposit and changes to metadata values once deposited. We were only concerned with tracking changes to the published versions of datasets and files. No changes to drafts are stored, and depositors are free to change everything up until the point of publication. In this context, publishing is considered to be the act of receiving a reserved or published DataCite DOI. Changes to the files, including file level deletions and additions, are not allowed as soon as DataCite receives information about the dataset. After attempting to combine what we knew of best and actual practices from the repository community, the team determined several operating principles for our initial launch:

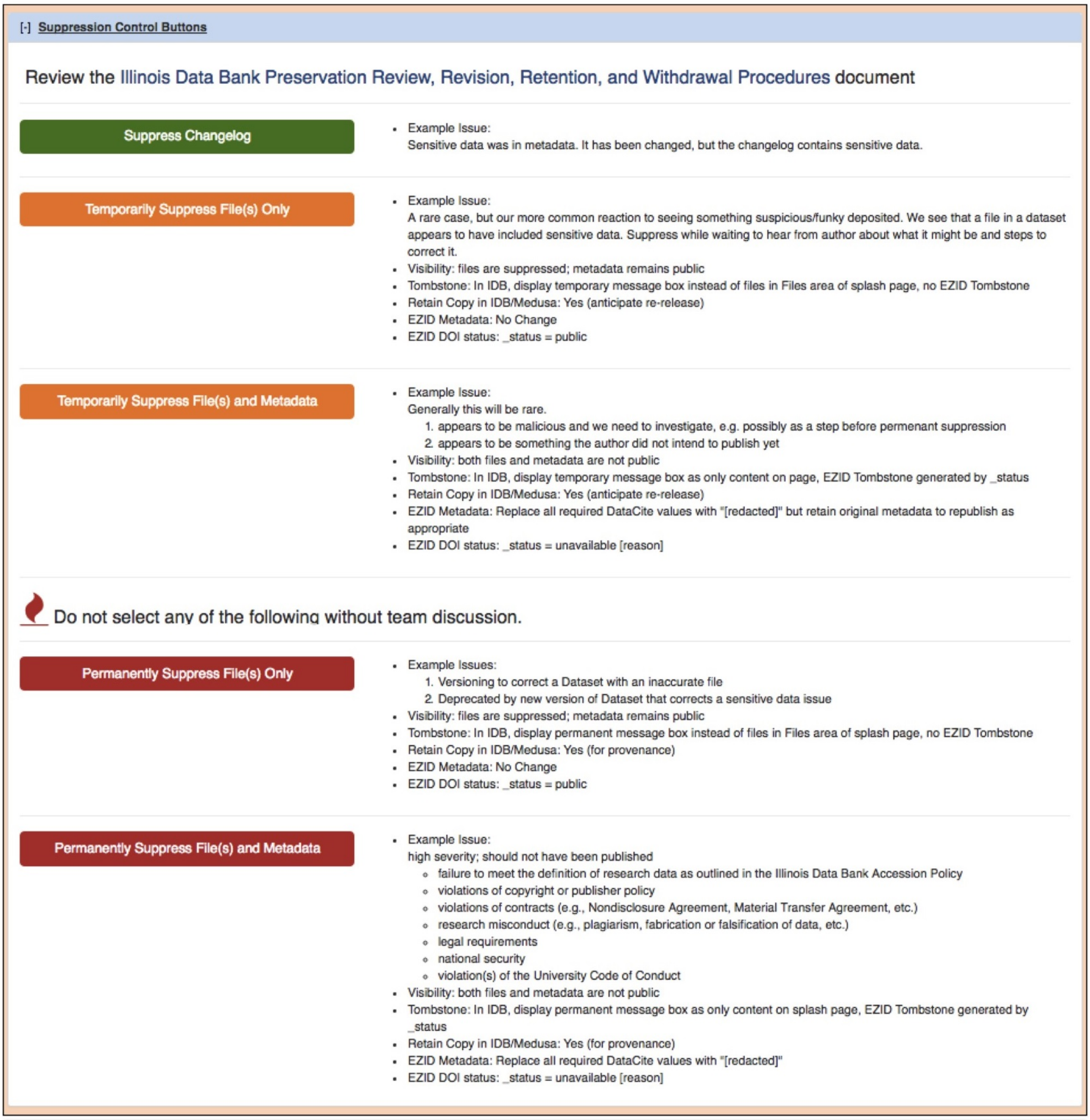
- Any changes to files within a published dataset require creation of a new version.
- New versions will receive their own dataset landing page.
- New versions will receive a new DOI, and the version number will be present in the DOI suffix.
- Version numbers will start at 1 by default and be represented as whole numbers only with no additional granularity. [8]
- Versions will be tracked in the metadata using the relatedIdentifer element and transmitted to DataCite.
- Version numbers will be reported and easily discoverable on dataset landing pages.
- Changes to the descriptive metadata within the Illinois Data Bank do not result in creation of a new version.
- Changes to the descriptive metadata will be tracked and reported on the dataset’s landing page within a publicly available changelog.

The descriptive metadata record, while we consider it to be a part of a dataset, has the potential for many and ongoing changes, e.g., addition of keywords or related material as the dataset is discovered and reused. Tracking changes to metadata values is straightforward for the system, so we determined that simply making the changes human readable and transparent was sufficient. Operationalized, this meant adding a changelog table to the bottom of the dataset page to report any post-publication metadata changes. Each copy of the updated metadata is also saved inside the system folder for each dataset, which can be made available upon request. Because we store snapshots of the metadata record, a disaster recovery procedure can recreate the entire dataset package (metadata records and files) from information contained within the associated Medusa folders.

Versioning of datasets for any needed file change has created more debate within our team. As a hard and fast rule, it is simple to implement, ensures absolute consistency in how file changes are handled, and provides a high-level of transparency to accessors. However, it does require that relatively minor changes, such as updating a ReadMe file with the dataset’s official citation with the DOI or correcting a URL, escalates to creating an entirely new version of the dataset, which may not be ideal.

Curator Functions

Within the Illinois Data Bank interface, a curator-only view was developed prior to launch. Currently, curator access is determined by a hard-coded list of identifiers in the application configuration files. Within the curator-only view, curators are allowed to edit any metadata fields for a given record and are also allowed the ability to override the one year embargo limit. Special curator-only fields are also displayed, which allow the ability to add a version number other than “1”, import a DOI if the record was created elsewhere, as well as add and assign related identifiers. Curator access also allows for the ability to toggle between a “depositor view” and a “curator view” to assist in demos and bug testing. Perhaps most importantly, curators are able to control the visibility of dataset changelogs, metadata, and files such that access can be temporarily or permanently suppressed. Based on policy discussions described above, we considered this feature to be important to have in place prior to launch. The interface for the suppression area provides verbose and explicit helper text to assist in (hopefully extremely) rare, “high-pressure” scenarios, when curators may need to make decisions quickly and confidently (Figure 6).


[-] Suppression Control Buttons

Review the Illinois Data Bank Preservation Review, Revision, Retention, and Withdrawal Procedures document

- Example Issue:
  Sensitive data was in metadata. It has been changed, but the changelog contains sensitive data.

Temporarily Suppress File(s) Only

- Example Issue:
  A rare case, but our more common reaction to seeing something suspicious/funky deposited. We see that a file in a dataset appears to have included sensitive data. Suppress while waiting to hear from author about what it might be and steps to correct it.
- Visibility: files are suppressed; metadata remains public
- Tombstone: In IDB, display temporary message box instead of files in Files area of splash page, no EZID Tombstone
- Retain Copy in IDB/Medusa: Yes (anticipate re-release)
- EZID Metadata: No Change
- EZID DOI status: _status = public

- Example Issue:
  Generally this will be rare.
  1. appears to be malicious and we need to investigate, e.g. possibly as a step before permenant suppression
  2. appears to be something the author did not intend to publish yet
- Visibility: both files and metadata are not public
- Tombstone: In IDB, display temporary message box as only content on page, EZID Tombstone generated by _status
- Retain Copy in IDB/Medusa: Yes (anticipate re-release)
- EZID Metadata: Replace all required DataCite values with "[redacted]" but retain original metadata to republish as appropriate
- EZID DOI status: _status = unavailable [reason]

Do not select any of the following without team discussion.

- Example Issues:
  1. Versioning to correct a Dataset with an inaccurate file
  2. Deprecated by new version of Dataset that corrects a sensitive data issue
- Visibility: files are suppressed; metadata remains public
- Tombstone: In IDB, display permanent message box instead of files in Files area of splash page, no EZID Tombstone
- Retain Copy in IDB/Medusa: Yes (for provenance)
- EZID Metadata: No Change
- EZID DOI status: _status = public

- Example Issue:
  high severity; should not have been published
  - failure to meet the definition of research data as outlined in the Illinois Data Bank Accession Policy
  - violations of copyright or publisher policy
  - violations of contracts (e.g., Nondisclosure Agreement, Material Transfer Agreement, etc.)
  - research misconduct (e.g., plagiarism, fabrication or falsification of data, etc.)
  - legal requirements
  - national security
  - violation(s) of the University Code of Conduct
- Visibility: both files and metadata are not public
- Tombstone: In IDB, display permanent message box as only content on splash page, EZID Tombstone generated by _status
- Retain Copy in IDB/Medusa: Yes (for provenance)
- EZID Metadata: Replace all required DataCite values with "[redacted]"
- EZID DOI status: _status = unavailable [reason]

Figure 6. Suppression options are available only to Illinois Data Bank curators. Anticipating rare use, descriptive help text is provided to assist curators. Click image for larger view.

External to the Illinois Data Bank, we have developed a series of guidelines and processes to curate deposited datasets, but these processes are not automated. We expect manual workflows will work sufficiently well enough during the initial rollout of the Illinois Data Bank when a smaller number of datasets are submitted and curator workloads are still manageable. However, if deposit into the Illinois Data Bank becomes routine for Illinois researchers, we would be quickly overwhelmed considering our campus' publication output exceeds 6,000 works a year (personal communication, W. Mischo 2015). In that case, additional development will be required to improve the curation workflow.

Reception, Known Limitations, and Future Directions

We launched the Illinois Data Bank in May of 2016 and spent the following summer engaging with researchers to build awareness and initiate deposits. We also fixed inevitable bugs and attended to more urgent value-adds that emerged, such as the need for elegant reversion to read-only secondary copies when our primary storage goes down for maintenance. Overall, reception to the platform has been positive. An interesting secondary use of the Illinois Data Bank that we did not anticipate is that it serves as an educational tool. A faculty member and a technology transfer manager independently shared that they thought the Illinois Data Bank has great utility to educate researchers about policy issues such as data ownership and data re-use.

Despite good feedback and having what we consider to be a relatively smoothly running and robust product, we know that there are several areas that could use additional refinement or implementation. We maintain a "parking lot" of features or improvements that have either occurred to us or we have already heard from depositors. Some of these are alluded to above (e.g., metadata and versioning improvements), but a few others include better support for large data files [9] and folders; developing tools to query, subset, or explore data in place; and more flexible access control mechanisms to enable transfer of ownership or multiple dataset editors.

Our development process reinforced the value of evidence-informed decision-making and considering the ripples our work may have into the external systems we rely on. We would have been able to develop the Illinois Data Bank more swiftly if we had not considered and (attempted to) address the potential consequences of how Illinois Data Bank metadata would be represented in DataCite—and to subsequent data indexers. Given the grander mission to make the research data products discoverable, understanding the value of aligning our design choices to maximize discoverability has benefitted our end product and provided us with important examples of why the Illinois Data Bank is valuable to our campus community. We understand our privilege in having the equivalent of several full time staff members dedicated to this development process, so we hope that by openly presenting the struggles and rationales behind some of our decisions, those facing data repository development projects will be able to see in our choices strategies to adopt, modify, or reject.

We decided to share our experience in developing the Illinois Data Bank in the spirit of the Twitter hashtag #overlyhonestmethods, which documented researchers' frank (and frequently funny and/or sad) admissions on the surprising hows and whys behind research. Although we did this somewhat tongue-in-cheek, the truth is never far away, and transparency is an ideal we all should aspire to. If we were to be truly honest, we would need substantially more text, but we are already teetering on 10,000 words and can hardly expect readers to indulge us farther. [10] In closing, we'll simply ask the question: "Have we developed a resource that will serve our campus, researchers, and research communities well?" Perhaps Twitter user @luispedrocoelho answers best with "We have no idea, but this seems reasonable #overlyhonestmethods". [11]

Acknowledgements

We thank Medusa developer Howard Ding for many helpful discussions and development of Medusa components, and Helenmary Sheridan for discussion and sharing of information on IDEALS.

Notes

[1] Although internally we immediately lapsed into calling it the "IDB."

[2] For example, see http://data.datacite.org/10.13012/B2IDB-4900670_V1

[3] Sometimes this changes, which is why we encourage the use of personal identifier schemes like ORCiD.

[4] And data release upon publication is commonly mentioned in funding agency public access plans, including NIH, DOT, NOAA, and NASA, etc.

[5] Although exactly to what extent was not something we knew at the time.

[6] Meaning we emailed a few of the RDS-friendly researchers we knew and asked for their input and willingness to send along to others. Therefore, we have no idea what our sample size actually was, but we did have representation of people with backgrounds in informatics, LIS, GIS, engineering, life sciences, and physical sciences. Nearly all (6 of 7) had published a journal article before.

[7] We note that in the time since figshare has changed its versioning practices, but we report here what we observed at the time of comparison with other repositories.

[8] Notably missing from this discussion is handling the version numbering system. An analysis of version numbers for DataCite datasets indicated a lack of community consensus of the more advanced and detailed numbering systems (Wickes 2016).

[9] Especially considering we've committed to 2TB per year per Principal Investigator (PI).

[10] So feel free to contact us.

[11] https://twitter.com/luispedrocoelho/status/288677624279093249

References

Ammann, Noémie, Lars Holm Nielsen, Sebastian Peters, and Madeleine de Smaele. 2011. "DataCite Metadata Schema for the Publication and Citation of Research Data." DataCite. http://web.archive.org/web/20160825024134/http://schema.datacite.org/meta/kernel-3/doc/DataCite-MetadataKernel_v3.1.pdf.

Ball, Alex. 2014. "How to License Research Data." Edinburgh: Digital Curation Centre. http://web.archive.org/web/20160825025111/http://www.dcc.ac.uk/resources/how-guides/license-research-data.

Borowski, Christine. 2011. "Enough Is Enough." The Journal of Experimental Medicine 208: 1337–1337. doi:10.1084/jem.20111061.

Carpenter, Todd. 2009. "Supplementary Materials: A Pandora's Box of Issues Needing Best Practices." Against the Grain 21: 84–85.

Carroll, Michael W. 2015. "Sharing Research Data and Intellectual Property Law: A Primer." PLOS Biol 13 (8): e1002235. doi:10.1371/journal.pbio.1002235.

Crossref. 2016. "Open Funder Registry." http://web.archive.org/web/20160825033231/http://www.crossref.org/fundingdata/registry.html.

DataCite. 2016. "Metadata Store." http://web.archive.org/web/20160825020124/https://mds.datacite.org/.

DataCite Metadata Working Group. 2016. "DataCite Metadata Schema for the Publication and Citation of Research Data v4.0." https://schema.labs.datacite.org/meta/kernel-4.0/doc/DataCite-MetadataKernel_v4.0.pdf.

Data Seal of Approval. 2016. "The Guidelines 2014-2015." Data Seal of Approval. Accessed August 25. https://web.archive.org/web/20160825023007/http://www.datasealofapproval.org/en/information/guidelines/.

Dunham, Elise, and Ayla Stein. 2016. "Illinois Data Bank Metadata Documentation v1.0." University of Illinois at Urbana-Champaign. https://www.ideals.illinois.edu/handle/2142/91020.

Dunham, Elise, Elizabeth Wickes, Ayla Stein, Colleen Fallaw, and Heidi Imker. 2016. "Pre-Metadata Counseling: Putting the DataCite relationType Attribute into Action." In Curating Research Data, edited by Lisa R. Johnston. ACRL.

Fallaw, Colleen. 2016. Databank: Launch. Zenodo. 10.5281/zenodo.51311.
Holdren, John P. 2013. "Increasing Access to the Results of Federally Funded Scientific Research." Office of Science and Technology Policy. http://web.archive.org/web/20160115125401/https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

Imker, Heidi J. 2016. "Overlooked and Overrated Data Sharing: Why so many scientists are confused and/or dismissive" In Curating Research Data, edited by Lisa R. Johnston. ACRL.

Kenyon, Jeremy, and Nancy R. Sprague. 2014. "Trends in the Use of Supplementary Materials in Environmental Science Journals." Issues in Science and Technology Librarianship. doi:DOI:10.5062/F40Z717Z.

Kratz, John Ernest, and Carly Strasser. 2015. "Researcher Perspectives on Publication and Peer Review of Data." PLOS ONE 10 (2): e0117619. doi:10.1371/journal.pone.0117619.

Maunsell, John. 2010. "Announcement Regarding Supplemental Material." The Journal of Neuroscience 30: 10599–600.

Nielsen, Jakob. 2000. "Why You Only Need to Test with 5 Users." Nielsen Norman Group. March 19. https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/.

Purdue University. 2016. "Learn About EZID." http://web.archive.org/web/20160825020403/http://ezid.lib.purdue.edu/learn/.

*Rimkus, Kyle. 2015. "Medusa FAQ." Wiki. Medusa Digital Preservation Service. http://web.archive.org/save/https://wiki.illinois.edu//wiki/display/LibraryDigitalPreservation/Medusa+FAQ.

Rimkus, Kyle, and Scott Witmer. 2016. "Managing File Format Endangerment in Digital Preservation Repositories: An Evidence-Based Approach." In Proceedings of the 13th International Conference on Digital Preservation. Bern, Switzerland.

Scholarly Publishing and Academic Resources Coalition (SPARC). 2016. "Tracking and Understanding Federal Open Data Plans." Sparcopen.org. http://sparcopen.org/our-work/research-data-sharing-policy-initiative/.

Shreeves, Sarah. 2014. "Policy Documents – IDEALS." University of Illinois at Urbana-Champaign. https://www.ideals.illinois.edu/handle/2142/5.

US Department of Energy. 2014. "Statement on Digital Data Management." Department of Energy. http://science.energy.gov/funding-opportunities/digital-data-management/#Requirements.

Wickes, Elizabeth. 2016. "Version Values for DataCite Dataset Records." University of Illinois at Urbana-Champaign. https://doi.org/10.13012/B2IDB-4803136_V1.

Williams, Sarah C. 2016. "Practices, Policies, and Persistence: A Study of Supplementary Materials in Crop Science Journals." Journal of Agricultural & Food Information 17 (1): 11–22. doi:10.1080/10496505.2015.1120213.

About the Authors

Colleen Fallaw
Research Programmer at the University of Illinois at Urbana-Champaign Library. http://orcid.org/0000-0002-0339-9809
mfall3@illinois.edu

Elise Dunham
Data Curation Specialist with the Research Data Service at the University of Illinois at Urbana-Champaign Library. http://orcid.org/0000-0002-7562-1923
emdunham@illinois.edu

Elizabeth Wickes
Data Curation Specialist with the Research Data Service at the University of Illinois at Urbana-Champaign Library. http://orcid.org/0000-0003-0487-4437
wickes1@illinois.edu

Dena Strong
Senior Information Design Specialist with Technology Services University of Illinois at Urbana-Champaign. http://orcid.org/0000-0003-1834-9940
dlstrong@illinois.edu

Ayla Stein
Metadata Librarian at the University of Illinois at Urbana-Champaign Library. http://orcid.org/0000-0002-6829-221X
astein@illinois.edu

Qian Zhang
CLIR postdoctoral researcher with the iSchool and Research Data Service at the University of Illinois at Urbana-Champaign Library. http://orcid.org/0000-0003-1549-7358
zqian1@illinois.edu

Kyle Rimkus
Preservation Librarian and Assistant Professor at the University of Illinois at Urbana-Champaign Library. http://orcid.org/0000-0002-9142-6677
rimkus@illinois.edu

Bill Ingram
Manager of Scholarly Communication and Repository Services at the University of Illinois at Urbana-Champaign Library. http://orcid.org/0000-0002-8307-8844
wingram2@illinois.edu

Heidi Imker
Director of the Research Data Service (RDS) and Associate Professor at the University of Illinois at Urbana-Champaign. http://orcid.org/0000-0003-4748-7453
imker@illinois.edu

Subscribe to comments: For this article | For all articles