2018

Illinois Data Bank

FIRST YEAR REVIEW

COMPILED BY RESEARCH DATA SERVICE STAFF

Contents

Mission

The mission of the Illinois Data Bank is to centralize, preserve, and provide persistent and reliable access to the research data created by affiliates of the University of Illinois at Urbana-Champaign, such as its faculty, academic staff, and graduate students. The Illinois Data Bank is supported by the Research Data Service (RDS), which was formally adopted as a campus strategic priority through inclusion in the 2013-2016 Campus Strategic Plan. The RDS launched the Illinois Data Bank in 2016, as part of its mission to provide Illinois researchers with the expertise, tools, and infrastructure necessary to manage and steward research data.

Data sharing has long been integral to science and discovery. In 2013, the federal government formally recognized the importance of data sharing in the White House Office of Science and Technology Policy (OSTP) memorandum. The memo required all federal agencies that financially support research to also determine ways to make research data accessible to the public. In grant applications to the National Science Foundation (NSF), for instance, researchers needed to specify where their research data would live after their projects were complete, what restrictions on access—if any—would be placed on the data, and how access would be ensured for a period of time.

The OSTP memo marked a turning point in how research data was conceptualized. With the force of federal policy, research data moved from a personal or institutional resource, exchanged among colleagues or universities, to a national resource to be shared with peer and public alike. Data sharing became both ethical commitment to, and national investment in, the future of science.

The Illinois Data Bank provides the technical and human infrastructure necessary to manage and share research data produced by researchers at the University of Illinois at Urbana-Champaign. The Illinois Data Bank is optimized specifically for data, which is material that supports or contributes to research findings that someone else might want to use in order to re-analyze or extend those findings (e.g. text or csv files, chemical or biological spectra, data from text mining or topic modeling, etc.). As crucial complements to the technical requirements for data sharing, policies and workflows, adoption and adaption of metadata schema, and iterative user testing continue to be employed and developed on and around the Illinois Data Bank. By bringing together technology with policies that are both continuously refined by user need, the RDS is able to support research with high degrees of transparency and professionalism.

Policies

The Illinois Data Bank is a file-based institutional repository that is intended to provide maximum public access to unrestricted research data for the advancement of scholarship and the public good in ways that are consistent with the U.S. President's OSTP memo.

To this end, datasets deposited into the Illinois Data Bank are governed by a series of policies related to access and use, accession, deposit, preservation, and review and retention. These policies are available at https://databank.illinois.edu/policies.

Datasets accepted for ingestion into the Illinois Data Bank can thus be ensured to share a number of major characteristics. All datasets in the Illinois Data Bank are:

- Unrestricted

- Professionally curated and linked to associated works, such as journal articles, source code, or data deposited elsewhere
- Preserved and made available to the public for a minimum of five years

The strength of Illinois Data Bank's policies has already been the subject of a recent international study. The OCLC Online Computer Library Center (OCLC) created a four-part research report on how institutions across the globe are supporting researchers' growing needs for research data management. Dr. Rebecca Bryant, Senior Program Officer at OCLC Research Library Partnership, noted that the University of Illinois at Urbana-Champaign was chosen as the U.S. case study because,

> "transparency of [the Illinois Data Bank's] communications and policy documents made it ideal for study, and also because those publicly available documents revealed to us that Illinois was largely alone in considering and developing policies related to usage as well as preservation and deaccessioning." (R. Bryant, personal communication, March 5, 2018)

Both the Illinois Data Bank and the policies that govern its use are the responsibility of the RDS at the University Library.

Summary Statistics

The Illinois Data Bank provides Illinois researchers with a free, centralized location from which they can publish and preserve their research data. As an institutional repository, the Illinois Data Bank supports research now and in the future by making it easy for Illinois researchers to fulfill federal requirements to share data that has received federal support, and some journal publishers' requirements to share datasets as part of article publication. The Illinois Data Bank went live on August 29, 2016.

Before going live, the Illinois Data Bank experienced a "soft launch". During the summer of 2016, RDS staff identified researchers on campus who needed to share their data. These researchers provided a total of seven datasets to the repository, all of which have been available to the public since June 30, 2016. Between this time and August 29, RDS staff gathered and applied feedback from the soft-launched datasets, tested the system, and refined processes.

Since launching, the Illinois Data Bank has successfully supported research on two fronts: **use by researchers at Illinois** and **datasets as part of the scholarly communications ecosystem**.

Use by researchers at Illinois

In the first year the Illinois Data Bank was available, over fifty researchers relied upon the Illinois Data Bank and accompanying services to curate and publish their datasets.

Publications

Between August 29, 2016 and August 28, 2017, Illinois researchers published a total of **40** datasets in the Illinois Data Bank.

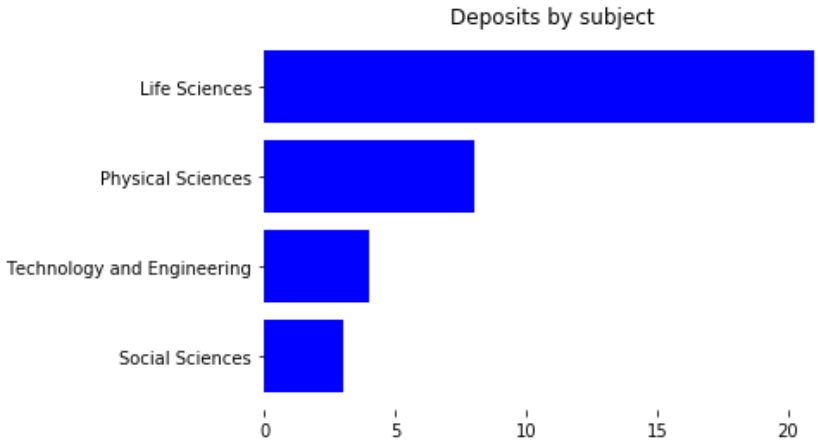
All datasets were provided with professional curation services, which include:

- Review of files for accuracy
- Review of accompanying metadata
- Links to related papers
- Creation of new versions

Curation services are provided at initial deposit as well as throughout the duration of the dataset's availability in the Illinois Data Bank. Three datasets (7.5%) particularly benefitted from initial review and were published in the Illinois Data Bank as second versions.

Subjects

Datasets published in the Illinois Data Bank in its first year came from a variety of subjects. The majority of published deposits (**58%**) were from the Life Sciences, while the fewest deposits (8%) were from the Social Sciences. As of yet, no deposits have been received for the Arts and Humanities.



Downloads

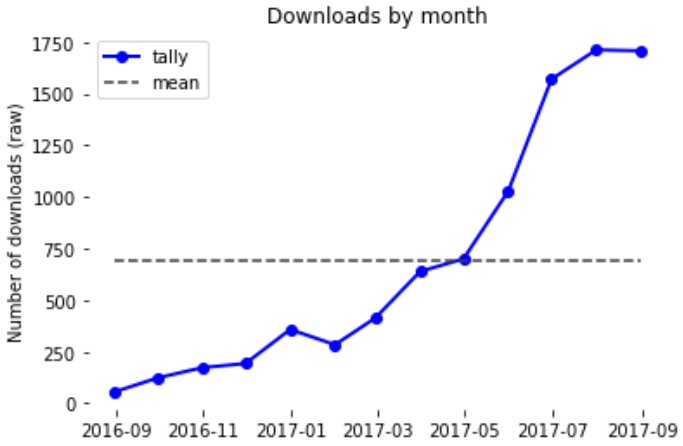
The Illinois Data Bank tracks download counts for datasets. As noted in the Illinois Data Bank's help guide,

> "[t]o mitigate possible over- or under-estimation of download counts, a dataset's download counter will increment up by one when one or more of any associated files are downloaded or viewed. However, only one download instance will be counted per IP address per calendar day. This means that a single computer downloading a dataset's files multiple times in the same day will only be counted once." ("How can we help you?", 2016)

Given this method for amassing downloads, in the first year of the Illinois Data Bank alone, datasets deposited into the Illinois Data Bank were downloaded **8,948** times.

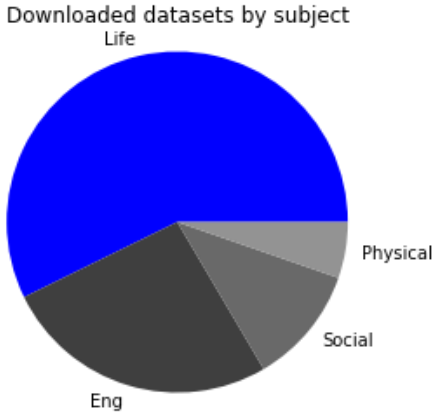
To determine downloads/dataset over the course of the year, download data was first filtered to fit within the specified date range of August 29, 2016 through August 28, 2017. The resulting set was aggregated by digital object identifier (DOI), and the accompanying download tallies were summed by month and then year.

In its first year, the Illinois Data Bank experienced steady increase, with downloads/views of datasets more than doubling between April 2017 and July 2017. Downloads peaked for the year in late summer 2017, with **55%** of downloads for the year occurring between the months of June and August. This steady increase may be attributed to increased awareness of the existence and usefulness of the Illinois Data Bank among researchers.

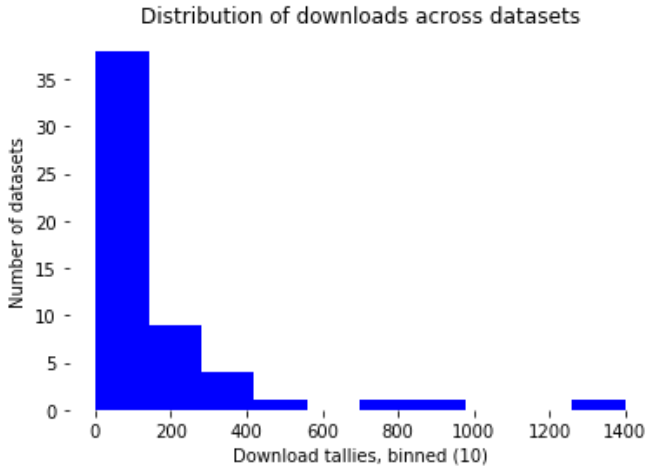


On average, each dataset published in the Illinois Data Bank during its first year was downloaded **162** times.

Datasets in the Life Sciences ("Life") category were most heavily downloaded during the first year of the Illinois Data Bank; downloads from this category accounted for **57%** of all downloads. Life Sciences was followed by Technology and Engineering ("Eng"), Social Sciences ("Social"), and Physical Sciences ("Physical") in terms of download distribution by subject.

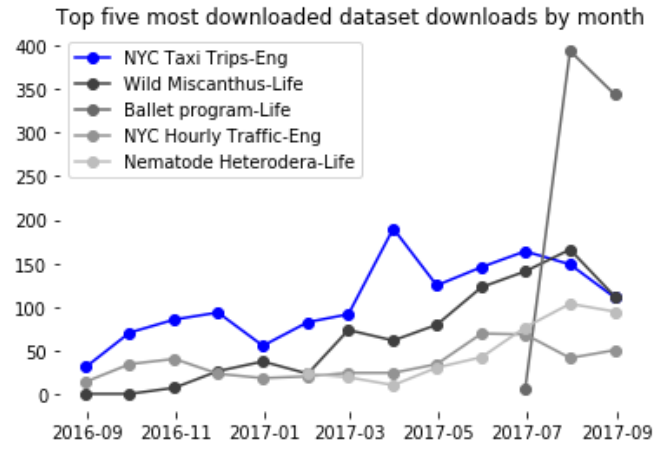


Between them, five datasets accounted for almost half of all dataset downloads (**43%**). These five datasets shared a combined download total of **3,875** times.



The five most downloaded datasets were almost evenly divided between Life Sciences (51.7%) and Technology and Engineering (48.2%).

All five of the most downloaded datasets experienced bumps in download rates at some point between the months of May and August, suggesting patterns in research that made the Illinois Data Bank a particularly valuable resource during this time. Future marketing efforts for the Illinois Data Bank may do well to increase outreach to researchers during the summer months.



The most downloaded dataset was from Technology and Engineering; it was downloaded 1,399 times over the course of the Illinois Data Bank's first year. The citation for this dataset is:

> Donovan, Brian; Work, Dan (2016): New York City Taxi Trip Data (2010-2013). University of Illinois at Urbana-Champaign. https://doi.org/10.13012/J8PN93H8.

The dataset was consistently downloaded throughout the first year of its availability in the Illinois Data Bank, though it experienced peak downloads in late March.

The dataset with the most downloads during any one month was from Life Sciences. It was downloaded 742 times over the course of the Illinois Data Bank's first year, with 393 (52.9%) of those downloads occurring just in July, one month after its publication in the Illinois Data Bank.

The citation for this dataset is:

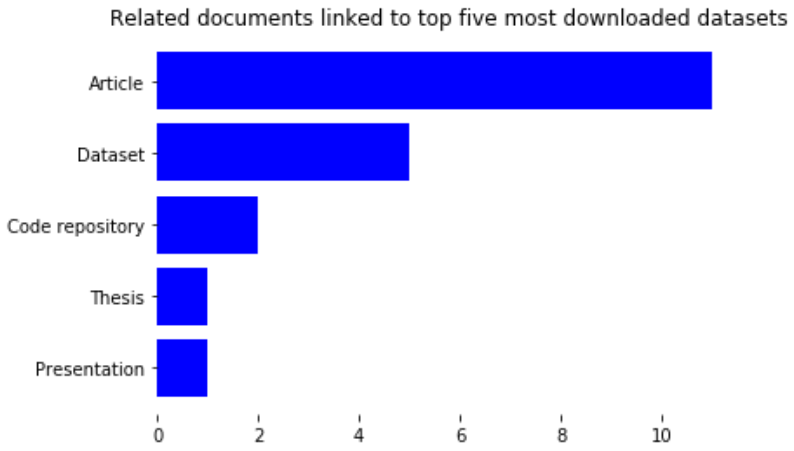
> Scheidler, Andrew; Kinnett-Hopkins, Dominique; Learmonth, Yvonne; Motl, Robert; Lopez-Ortiz, Citlali (2017): Targeted ballet program mitigates ataxia and improves agility in moderate-to-advanced multiple sclerosis. University of Illinois at Urbana-Champaign. https://doi.org/10.13012/B2Illinois Data Bank-6858418_V1

Datasets as part of the Scholarly Communications Ecosystem

Data live in an ecosystem where they are often related to content spread across the world; for example, journal articles that reside in publisher systems, data that reside in domain repositories, or scripts that reside in code repositories. RDS staff are committed to ensuring that reuse of published data is noted and displayed in the Illinois Data Bank as part of the dataset's metadata.

Illinois Data Bank metadata records provide information about datasets and their relationships to other research. RDS staff provide these curation services for the duration of a dataset's availability in the Illinois Data Bank. As such, when new articles or other scholarly work cite an Illinois Data Bank dataset using the DOI, curators add that relationship to the dataset's metadata, and those relationships (with associated links) are available to all visitors to explore.

The top five downloaded datasets in the Illinois Data Bank's first year have so far been linked to a total of **20** related documents. Of these documents, articles are the most prominent kind of document to be linked to a dataset downloaded often from the Illinois Data Bank (**55%**).



Four out of the five (**80%**) top downloaded datasets were linked to at least one related document.

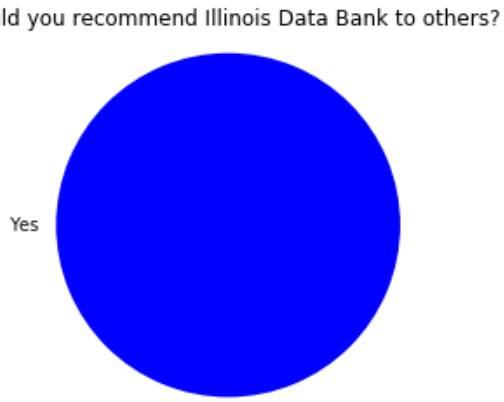
The New York City Taxi Drip Data dataset, for instance, was linked to 13 related documents (nine articles, one presentation, one dataset, one code repository, and one doctoral thesis). The second most downloaded dataset, "Ecological characteristics and in situ genetic associations for yield-component traits of wild Miscanthus from eastern Russia", was linked to five related documents (three datasets, one article, and one code repository).

Experience

Researchers have had positive experiences with the Illinois Data Bank and support staff.

Survey

According to survey results, **100%** of Illinois researchers would recommend the Illinois Data Bank to other people.



Other findings:

- The steps in the deposit process are between very easy and easy to follow (**100%** agree)
- The interface is between very easy and easy to use (**100%** agree)
- Contacting RDS staff is helpful (**87.5%** agree; 12.5% N/A)

Interviews

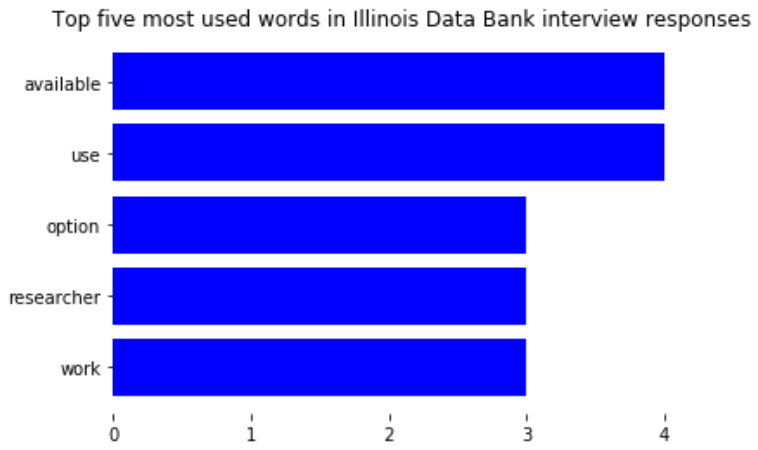
Researchers used the Illinois Data Bank for a variety of reasons, from the ease of sharing their work to ensuring that their data are housed in a safe place.

The most common reasons why researchers used the Illinois Data Bank are as follows:

- It was free
- It was easy to use
- It allowed them to share their data with other researchers

Four researchers, including the researcher who deposited the second most downloaded dataset, were asked why they decided to deposit their datasets in the Illinois Data Bank. By compiling their responses, removing common English words like “and” and “the” (also known as “stop words”) that do not significantly contribute to meaning, and summing the number of different remaining words, we were able to determine the top five most used words across responses. The top five most used words were as follows:

1. Available (23.5%)
2. Use (23.5%)
3. Option (17.6%)
4. Researcher (17.6%)
5. Work (17.6%)



Looking at each of these words in the context of their original sentences provides more insight.

In the case of the word “available”, researchers said that they used the Illinois Data Bank because it was both available to them at minimal cost and it allowed them to make their data available to others. As one researcher stated,

> “While I found many options **available** to store my data online, the cost was prohibitive for my tight graduate student budget and my project did not have funds to cover the cost of data storage. As I looked for more economical options I was surprised and happy to find that the U of I had an option that was free to students and faculty.” (emphasis added)

Another researcher was pleased that the Illinois Data Bank was easy to use because he needed to make his data publicly available in order to have his work published. As this researcher stated,

> “Many Scientific journals now require that your data be archived and available upon request if you are to publish your work with them. Because the university system is so easy to **use**, it was an easy choice to **use** Illinois Data Bank. The data is now safely stored and readily available for researchers to **use** at any time.” (emphasis added)

The Illinois Data Bank helped these researchers solve very present concerns like cost and changes in publication requirements. But it also served to help researchers plan for the future of their own work, as well as that of their fields. Perhaps most poignantly, one researcher used the Illinois Data Bank to bolster his work’s impact over time by sharing his data with other researchers. As this researcher stated,

> “By making data available to other **researchers** through the Illinois Data Bank we increase the potential impact of our work. Other researchers can use these data as part of reviews or syntheses, which allows the work we did to become part of a larger study, and may garner additional interest in our project. Moreover, depositing the data provides a safeguard against loss of the original data, and preserves it for future uses.” (emphasis added)

The full value of depositors’ data to the future of science and discovery is currently unknown. But because the Illinois Data Bank accepts multiple data types and preserves them for at least five years, researchers encountered fewer obstacles to depositing their data. Once they deposited their myriad data, researchers trusted that their data would be protected over time. By making it possible to deposit

many kinds of data and preserving those deposits, then, the Illinois Data Bank and RDS staff helped researchers become better stewards of their data in the present and for the future.

Indeed, the depositor of the second most downloaded dataset declared that the Illinois Data Bank perfectly suited her needs for a repository because it did not limit the kinds of data that could be deposited,

> "We liked that we could put essentially any information in there that we wanted, including the original handwritten field notes. Having that in the same repository as the genetic data, the phenotypes, the R scripts for data analysis - the paper explains everything from field observations to how to sequence the DNA. We have a broad range of data types all in one repository, which is great."

To access and use her research, other researchers do not need to search in multiple repositories, locations, or media. Instead, the Illinois Data Bank allowed this researcher to preserve the complexity of her project in one deposit, a service from which other researchers and, ultimately, scientific discovery will benefit.

As such, she will continue to rely on the Illinois Data Bank to store and make her data available to the world. "[W]e are going to be using [the Illinois Data Bank] in the future," she stated.

Preservation

Datasets deposited in the Illinois Data Bank are preserved using the Medusa digital preservation repository, a system developed whole-cloth at the University of Illinois Library. The RDS is committed to preserving and providing access to all research data published in the Illinois Data Bank for five years or longer.

The University Library manages Medusa storage in partnership with the National Center for Supercomputing Applications (NCSA). Medusa's storage infrastructure consists of:

- two copies of every file, replicated across two distinct campus nodes
- a third copy of every file backed up off-campus (currently in Amazon Glacier)

In the first year, Medusa registered the Illinois Data Bank's datasets which includes management of the 734 individual data files deposited by the researcher, plus handling of an additional 1,323 system files that are essential for provenance and preservation. All told, 638.56 GBs of single-copy storage and **1,915.5 GBs** total storage was required. The University Library is currently working on ensuring the future robustness of the preservation infrastructure; for example, by exploring the utility of Amazon Web Services (AWS).

Looking Forward

The RDS looks forward to continuing to support research at Illinois.

The Illinois Data Bank is expected to grow as more researchers become aware of the service through increased marketing and outreach, scholarly publication, and word-of-mouth.

Through the Illinois Data Bank, datasets produced by Illinois researchers are already enjoying more exposure. Preliminary analysis of download data from August 29, 2017 to February 15, 2018 shows that

download tallies of datasets published in the Illinois Data Bank are already at 6,646. In just the first half of its second year, then, the Illinois Data Bank is already reporting almost as many downloads as in its entire first year (**74%**).

We anticipate that this trend will continue.

References

Bryant, Rebecca, Brian Lavoie and Constance Malpas. (2018). Incentives for Building University RDM Services. The Realities of Research Data Management, Part 3. Dublin, OH: OCLC Research. doi:10.25333/C3S62F.

How can we help you? (2016). Retrieved from https://databank.illinois.edu/help#metrics.

Illinois Data Bank. (2016). Retrieved from https://databank.illinois.edu/.

Illinois Data Bank Policy Framework and Definitions. (2016). Retrieved from https://databank.illinois.edu/policies.