Should We Keep Everything Forever? Determining Long-Term Value of Research Data

Bethany Anderson[1], Susan Braxton[2], Elise Dunham[3], Heidi Imker[3], Kyle Rimkus[4]

1 University Archives, University Library, University of Illinois at Urbana-Champaign
2 Funk ACES Library, University Library, University of Illinois at Urbana-Champaign
3 Research Data Service, University Library, University of Illinois at Urbana-Champaign
4 Preservation Unit, University Library, University of Illinois at Urbana-Champaign

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

INTRODUCTION

The Illinois Data Bank's purpose is to provide University of Illinois at Urbana-Champaign researchers with a library-based repository for research data that will facilitate data sharing and ensure reliable stewardship of published data for a minimum of five years [1]. The Illinois Data Bank is intended to:

- *provide a mechanism for researchers to be compliant with funder and/or journal requirements to make results of research publicly available.*
- *promote the discoverability and use of open research data through a preservation and access solution that is trusted by researchers at the University of Illinois at Urbana-Champaign.*

Wisdom from the Archives Community

The archives profession has a long tradition of appraising and assessing the enduring value of records, and the University of Illinois Archives has a long history of appraising scientific and technological records. The idea that not every record can/should be preserved underlies archival appraisal [2,3] and digital preservation [4,5] processes. Archivists recognize the different contexts and factors that affect notions of value, retention, and long-term preservation of records. For data, concepts such as research impact, uniqueness, future utility, documentation, and use/citation, are important appraisal considerations [2,5,6,7], as are the technical characteristics of digital objects [4,6,7]. We have drawn from archival theory and practice and emerging data curation practice to create preservation review guidelines for the Illinois Data Bank.

Past Repository Experience

Our institutional repository, IDEALS, was established in 2005 to preserve scholarly output of faculty, staff, and students. Although IDEALS is file agnostic, and although data is within-scope, the promise of preservation within IDEALS is strongest for the simplest file types (e.g., PDF, CSV), and there are file size limits. Content in IDEALS is primarily textual, overwhelmingly PDF, with deposits usually comprising a single—and relatively small—file per record. In contrast, deposits into the Illinois Data Bank are expected to be non-textual, complex arrays of files in various formats, and potentially very large.

The Illinois Data Bank Commitment

The Illinois Data Bank commits to preservation and curation of deposited data (up to 2 terabytes per year with 3 replications) for a minimum of five years. Self deposit is allowed and encouraged, with intentionally low barriers to deposit. To ensure that we are able to fulfill our commitment to stewarding the deposited research data of the University of Illinois at Urbana-Champaign in an effective and scalable manner, the RDS has established a policy framework that includes assessment of the enduring value and viability of datasets deposited into the Illinois Data Bank. We have developed guidelines and processes for reviewing published datasets after their five-year commitment ends to determine whether to retain, deaccession, re-home, or dedicate more stewardship resources to datasets.

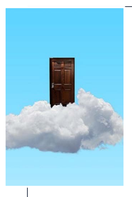
PRESERVATION REVIEW GUIDELINES

Reviewers	Criterion	Consideration
Curators / Librarians / Archivists	Cost to Store	What is the estimated cost of continuing to store the dataset?
	Cost to Preserve	What is the estimated cost of continuing or escalating preservation for the dataset (e.g., file format migration, software emulation, and/or enhancement of preservation metadata)?
	Access	What do download and page view metrics indicate about interest in the dataset over time?
	Citations	Has the dataset been cited in any publications?
	Associated Publication Citations	If the dataset supports the conclusions of a publication, has that publication been cited in any other publications?
	Restrictions	Does the dataset have any access or reuse restrictions associated with it?
Domain Experts	Possibility of Re-creation	Is it possible to create the dataset again?
	Cost of Re-creation	If it is possible to create the dataset again, what would be the cost of doing so?
	Impact of Study	Did the study that generated the dataset significantly impact one or more research disciplines?
	Uniqueness of Study	Was the study that generated this dataset novel?
	Quality of Study	Is the study that generated this dataset regarded as being of quality by domain experts?
	Quality of Data	Is the dataset of quality according to domain experts?
	Current Relevance	Is the dataset useful for addressing contemporary research questions according to domain experts?
Curators / Librarians / Archivists & Domain Experts	Availability of Other Copies	Is the copy in the Illinois Data Bank the only one?
	Understandability	Has the creator supplied sufficient metadata and documentation related to the dataset's creation, interpretation, and use in order to facilitate future discovery, access, and reuse?
	Dependencies	Are the software, computing environment, or other technical requirements for using the dataset known? If so, are they available?
	Appropriateness of Repository	Is there another trusted repository that, based on their collecting scope and user community, would be a better home for the dataset?

POSSIBLE REVIEW OUTCOMES

Escalate Curation / Preservation
Improve metadata and/or documentation, reformat/migrate files, emulate native software.
Data creators/primary contacts will be notified, consulted on escalation plans, and given reports on downloads and DOI activity.

Do Nothing
Files and metadata remain in the Illinois Data Bank as deposited.
Data creators/primary contacts will be notified of the decision to retain, and given reports on downloads and DOI activity.



Re-Home
If a more appropriate repository exists (e.g., disciplinary repository optimized for the specific type of data), RDS will work with data creator/primary contact to transfer data to that repository.
Description, with a link to new location, will be retained.

Deaccession
If full review indicates deaccession is appropriate, data creators/contacts will be notified and offered information about alternative storage options.
Description (with deaccession info) will be retained.

REVIEW INDICATOR

Because we anticipate it will be unsustainable (and possibly unnecessary) to perform an intensive, human-mediated, qualitative review of every dataset deposited into the Illinois Data Bank, we propose an automated "Review Indicator" to identify datasets most in need of review. There is a precedent for machine monitoring (e.g., checksums) in digital preservation in archival settings. Statistical sampling and risk analysis are known selection methods [7]. Automated triggers for re-appraisal based on technical characteristics have been explored [2], as has automation of the entire re-appraisal process [8].

We plan to use machine quantifiable measures of some of our established preservation review criteria that will be captured for Illinois Data Bank datasets. The proposed Review Indicator (***RI***) estimates a dataset's demonstrated value and preservation cost and alerts curators that deeper, human-mediated review is justified. The exact calculation of ***RI*** and the variables it will incorporate is currently under discussion, but we propose some function of downloads, relationships, file size, dominant file format, and (possibly) checksum or other file errors.

We hypothesize that datasets requiring above average effort for preservation offer the highest potential return on the investment of an in-depth, human-mediated review. This "cost" of retention triggers the review, while number of downloads and relationships of these "high-maintenance" datasets inform the focus of the review. We expect (*hope?*) to see a standard distribution for ***RI*** with the bulk of the datasets in the "no action needed" range and a fraction falling on either end of the spectrum requiring review.

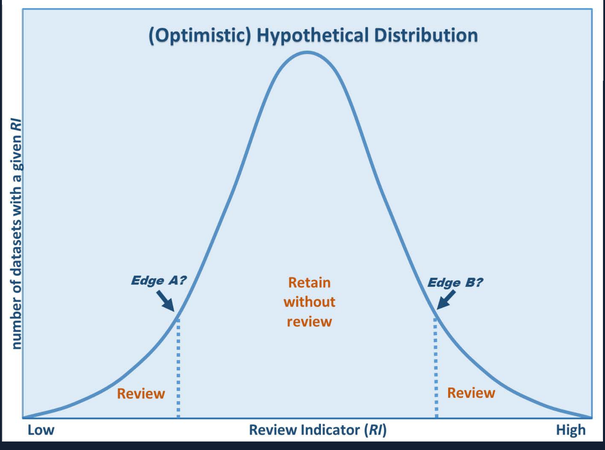
- *At the low end:* review is triggered by high preservation cost and lower evidence of use and relationship to other resources. The review will focus on whether deaccession is justified.
- *At the high end:* review is triggered by high demonstrated value coupled with preservation challenges. The review will focus on whether escalation of preservation is justified and possible.

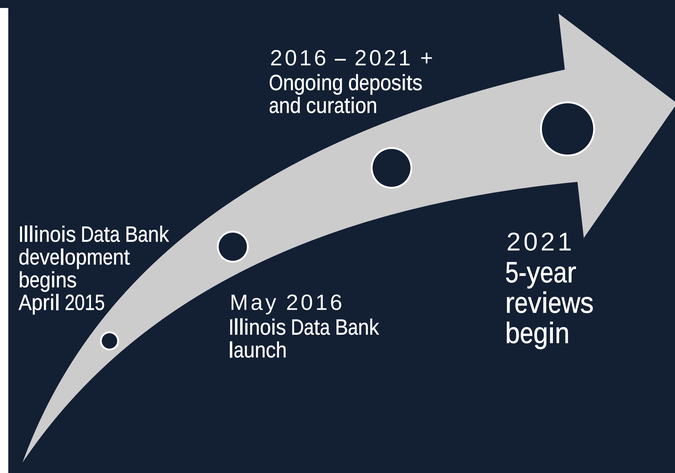
We propose this as a useful experiment for our service (and possibly for the the broader data preservation community) early in our data archiving efforts. The current number of datasets is manageable, so we have the opportunity to compare our proposed ***RI*** with human review to assess whether it reliably identifies datasets most in need of in-depth preservation review.

A formula under consideration is

$$RI = \frac{Downloads * Relationships * Format}{Bytes}$$

where format is represented with a scale that increases with a format's preservation "difficulty."





FUTURE CONSIDERATIONS

There are unknowns in the future of any service. The long-term viability of the Illinois Data Bank as a repository and a service for University of Illinois researchers will depend on its ability to assess, predict, and adapt. The Research Data Service will be closely monitoring its own processes and workflows as well as the complex landscape of digital repositories. Some important questions:

- Will the proposed "Review Indicator" be effective as a means of identifying resources needing review by teams of librarians, archivists, and domain specialists?
- Are there instances where the indicator fails to adequately indicate the need for more in-depth review?
- Are there machine quantifiable characteristics of datasets in the Illinois Data Bank that we have not already considered and that might be important?
- How much actual effort is involved in the in-depth review of a dataset in the Illinois Data Bank, and how does that effort compare to the cost of storage, file reformatting, or other preservation actions (or the cost of benign neglect of a valuable dataset)?
- What is the appropriate frequency for applying the automated indicator (every 5 years, every year after the first 5, or …)?

CONTACTS

Illinois Data Bank: https://databank.illinois.edu/
Research Data Service: http://researchdataservice.illinois.edu/
Email us: researchdata@library.illinois.edu
Find us on twitter: @ILresearchdata

Variable	vary size			vary format			vary relationships			vary downloads			Extreme low	Edge A?	Edge B?	Extreme high
Size, say GB	1,000	1	0.01	1	1	1	1	1	1	1	1	1	1,000	1,000	1	0.01
Format (1 "good", 10 "bad")	5	5	5	10	5	1	5	5	5	5	5	5	1	10	10	10
Relationships (must be n+1)	5	5	5	5	5	5	10	5	1	5	5	5	1	10	10	10
Downloads (must be n+1)	100	100	100	100	100	100	100	100	100	1,000	100	1	1	1,000	1,000	1,000
RI	2.5	2,500	250,000	5,000	2,500	500	5,000	2,500	500	25,000	2,500	25	0.001	100	100,000	10,000,000

REFERENCES

[1] Fallaw, C., Dunham, E., Wickes, E., Strong, D., Stein, A., Zhang, Q., Ingram, W., and Imker, H. Overly honest repository development. Code4Lib [under review].

[2] Haas, J.K., Samuels, H.W. and Simmons, B.T. 1985. Appraising the records of modern science and technology: a guide. Massachusetts Institute of Technology.

[3] Society of American Archivists, Technical Subcommittee on Guidelines for Reappraisal and Deaccessioning (TS-GRD). 2012. Guidelines for Reappraisal and Deaccessioning http://www2.archivists.org/sites/all/files/GuidelinesForReappraisalAndDeaccessioning-May2012.pdf.

[4] Harvey, R. and Thompson, D. 2010. Automating the appraisal of digital materials. Library Hi Tech, 28(2):313-322.

[5] UK Data Service. 2014. Collections Development Selection and Appraisal Criteria version 01.00 https://www.ukdataservice.ac.uk/media/455175/cd234-collections-appraisal.pdf

[6] Whyte, A. and Wilson, A. 2010. How to Appraise and Select Research Data for Curation. DCC How-to Guides. Edinburgh: Digital Curation Centre http://www.dcc.ac.uk/resources/how-guides

[7] Niu, J. 2014. Appraisal and Selection for Digital Curation. International Journal of Digital Curation 9(2): 65–82. doi:10.2218/ijdc.v9i2.272

[8] Oliver, G., S. Ross, M. Guercio, and C. Pala. *Report on Automated Re-Appraisal: Managing Archives in Digital Libraries.*" Glasgow: DELOS NoE, January 2008.

Image credits
Highsmith, Carol M. "Escalator at the Wilbur J. Cohen Federal Building, Washington, D.C." https://www.loc.gov/item/2013634384/
Allen, Shawn. "Couch potato" https://www.flickr.com/photos/shazbot/616633484/in/photolist-WupDN-j5MMLL-i9RgRb
Shepard, Nate. Door. https://pixabay.com/en/door-clouds-grass-perfect-beauty-1256751
Dietrich, Claire. "Return to sender."