Should We Keep Everything Forever (Reprised)?
Preservation review of research data in a repository as an art and a science

Susan Braxton, Colleen Fallaw, Hoa Luong, Daria Orlowska, Ashley Hetrick, Kyle Rimkus, Bethany Anderson, Heidi Imker

ILLINOIS

INTRODUCTION

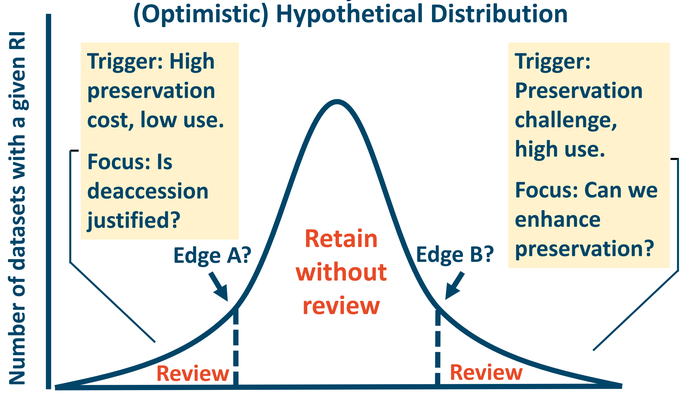
The **Illinois Data Bank** launched in 2016 as an institutional data repository offering a generous space allocation per researcher per year, no format prohibitions, and dataset preservation for at least 5 years followed by a robust review process.

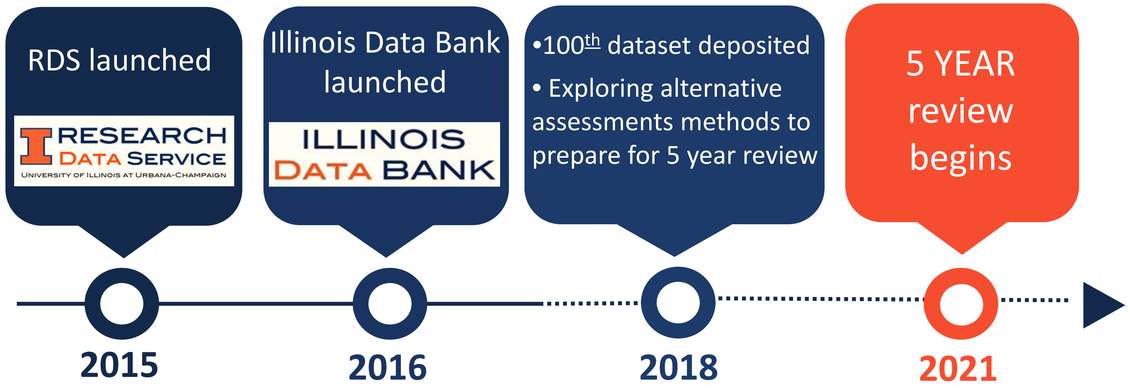
We proposed a **Review Indicator (RI)** to help focus human effort toward **retention**, **further curation**, or **deaccession**.

Hypothesis: A formula based on **downloads**, **relationships**, **size**, and **format** can reliably identify datasets that should be reviewed for possible deaccession or for further curation.

$$RI = \frac{Download * Relationship * Format}{Bytes}$$

(Optimistic) Hypothetical Distribution





METHODS

Sample - 113 datasets published as of 2018-07-02, including superseded versions. Datasets younger than 32 days were excluded from final analyses, leaving N=106.

RI Variable Details - File formats were scored, time normalization was applied for downloads and relationships, size classes were defined and used for some calculations.

Several versions of "RI" formula were tried, and datasets were also plotted according to risk and value.

RESULTS

RI Formulas - Formulas gave variable results, none reliably flagging for review. Size in bytes overwhelmed other variables, and were therefore replaced with size classes.

Risk vs Value Scatterplots - Impact of individual variables was better represented. Datasets falling in the high value/risk group are likely candidates for review, while low value/risk categorization seemed appropriate (included superseded versions).

CONCLUSIONS

None of the formulas appear to be better than human judgment or single variable evaluation.

Grouping datasets by relative value and risk was more effective than using a single formula, but selection based on single variables may be equally effective.

In all calculations, time normalization introduces error and raw bytes tend to swamp other variables.

DISCUSSION

Indicators can be simple:

- Downloads → deaccession review
- Format → preservation review

To inform 5 year review, assign format scores and account for actual content of container files during routine ingest curation process.

Concerns:

- New versioning post-preservation action
- Downloads as primary measure of value; relationships potentially more meaningful, but are not automatically recorded

RI Variables Analysis

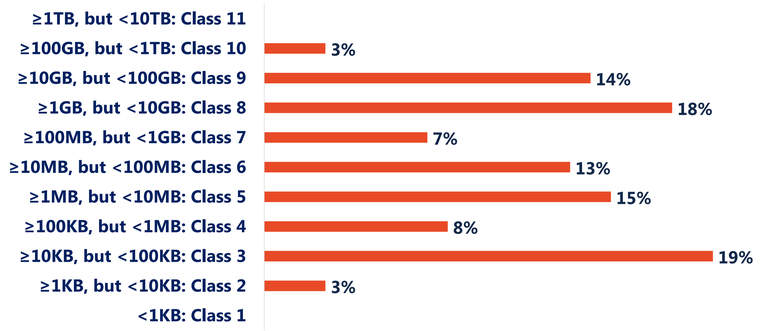
DOWNLOADS

- Dataset's download is counted when 1 or more files are downloaded/viewed
- Only 1 download instance is counted per IP address per day
- Range: 1 – 288 downloads/month (excludes superseded versions with suppressed files)
- Average: 26/ downloads/month, or <1 download/day
- Lowest downloads/month are superseded versions
- Highest downloads/month also a superseded version

SIZE

- Range: 1.318KB – 136.27GB
- Self-upload up to 15GB/file
- 83% of datasets < 15GB; 24 files in 12 datasets >15GB
- We explored both total bytes as well as size classes in Review Indicator formulas

Dataset Size Distribution



FORMATS

- 20 MIME types
- 41 MIME type/file extension combinations
- MIME type alone not indicator of format concerns ("text/plain" MIME type included many types of files)
- Dataset format score is the highest score of any file within the dataset
- More than half of the datasets analyzed have formats of possible preservation concern

Score	Files	Datasets*
1	1,054	35
2	126	16
3	412	62

*Of 113 published datasets as of 2018-07-02

MIME Type Assigned by Medusa	Extension(s)	Score*	Files	Datasets
text/plain	README, TSV, TXT, V2	1	872	59
text/csv	CSV	1	140	39
application/pdf	PDF	1	15	10
image/tiff	TIF, TIFF	1	2	2
text/xml	XML	1	2	2
application/vnd.oasis.opendocument.spreadsheet	ODS	1	3	1
image/bmp	BMP	1	2	1
image/jpeg	JPG	1	1	1
image/png	PNG	1	3	1
text/x-c	TSV	1	1	1
text/x-c++	TSV	1	4	1
application/vnd.ms-excel	XLS, XLSX	2	18	16
text/plain	FASTQ, PY, R	2	71	14
text/x-java	PY, R	2	21	6
application/msword	DOC, DOCX	2	5	3
application/zip	ZIP	3	86	29
application/x-gzip	BAM, GZ, RDATA	3	126	22
application/octet-stream	CZI, IMG, MAT, SAV	3	173	6
text/plain	IPYNB, M, MD5, NEX, TRE	3	10	4
application/x-rar-compressed	RAR	3	6	2
video/x-msvideo	AVI	3	2	2
application/x-7z-compressed	7z	3	1	1

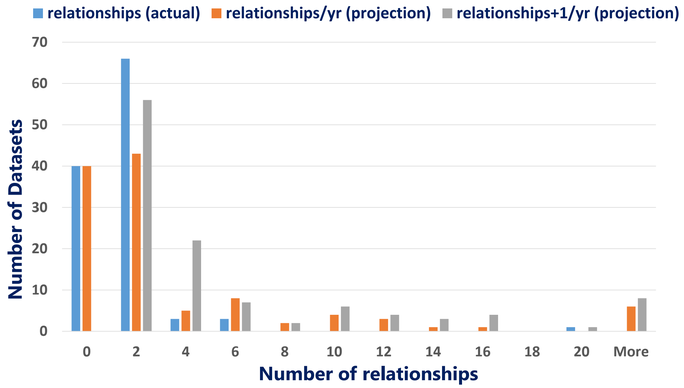
*Per Library of Congress(2017): 1- preferred format, 2- acceptable format, 3- format not recognized as preferred/acceptable

RELATIONSHIPS

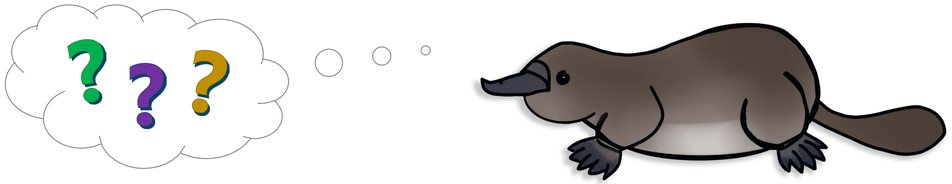
- 140 recorded relationships for 113 datasets published
- Most common relationship: IsSupplementTo
- Most common related material: Article
- 7 datasets have ≥2 relationships
- 1/3 of datasets have no relationships
- To avoid uninformative zero values, "relationships+1" were used in multiplication
- Normalized for time as relationships/year
- Graph shows frequencies of actual relationships and per year projections

Relationship	Count		Related material	Count
IsSupplementTo	86		Article	94
IsSupplementedBy	32		Dataset	18
IsCitedBy	33		Code	15
			Thesis	8
			Presentation	1
			Other	4
Total	**140**			**140**

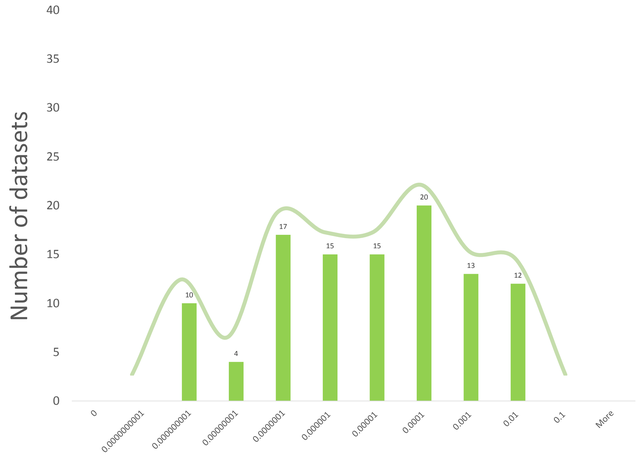
Relationship Frequencies as of 2018-07-02



Sample RI Formulas Comparisons



$RI\_2 = \frac{\frac{Downloads}{month} * \frac{relationship+1}{year} * formatscore}{bytes}$



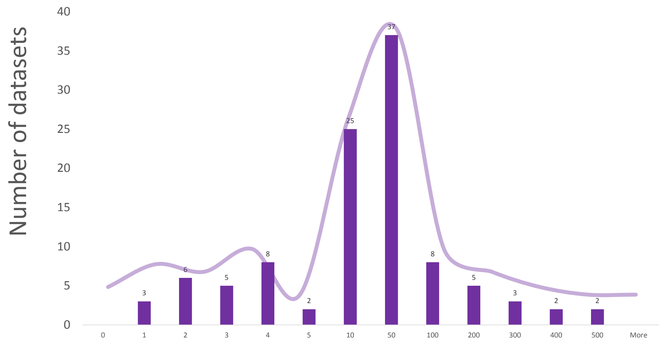
Low: Datasets with zero relationships; problematic formats; large size

High: Datasets with mostly preferred formats (opposite of predicted outcome--datasets on the high end do not require review); small datasets

Middle: Datasets with high downloads; includes some large datasets with problematic formats

Conclusion: RI_2 does not seem to highlight the datasets most needing attention. Size alone seems to overwhelm other factors.

$RI\_3 = \frac{\frac{Downloads}{month} * \frac{relationship+1}{year} * formatscore}{sizeclass}$



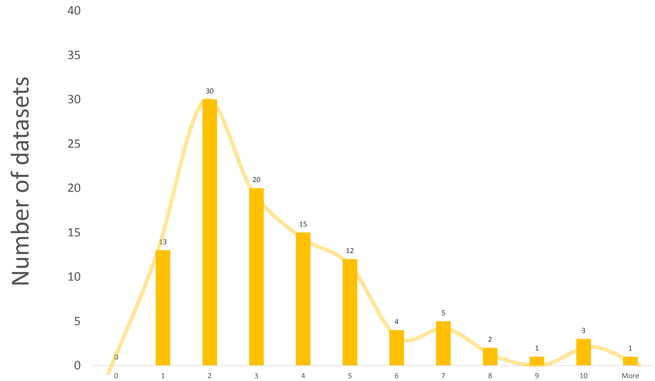
Low: Datasets with low downloads; zero relationships; size varies

High: Datasets with high downloads; many relationships; problematic formats; size varies

Middle: Mixed; includes large datasets with potentially problematic formats and high downloads

Conclusion: RI_3 does not reliably predict where review and curation efforts should be directed. Many datasets in the middle of the range are large and contained files of suspect preservability.

$RI\_6 = \frac{\frac{Downloads}{month} + \frac{relationships}{year}}{formatscore + sizeclass}$



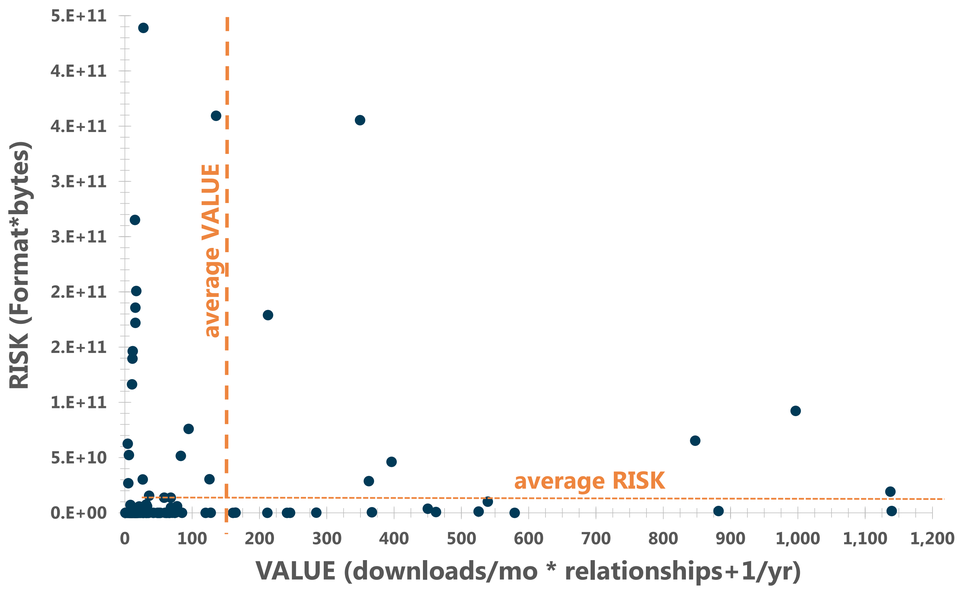
Low: Datasets with low downloads, zero relationships; size varies

High: Datasets with high downloads; many relationships; problematic formats; size varies

Middle: Mixed; includes large datasets with potentially problematic formats and high downloads

Conclusion: RI_6 does not separate datasets into 3 distinct categories. Format concerns and large sizes appear throughout the range, as do above average downloads and relationships.

Risk vs Value Scatter
(using raw bytes in Risk Calculation)

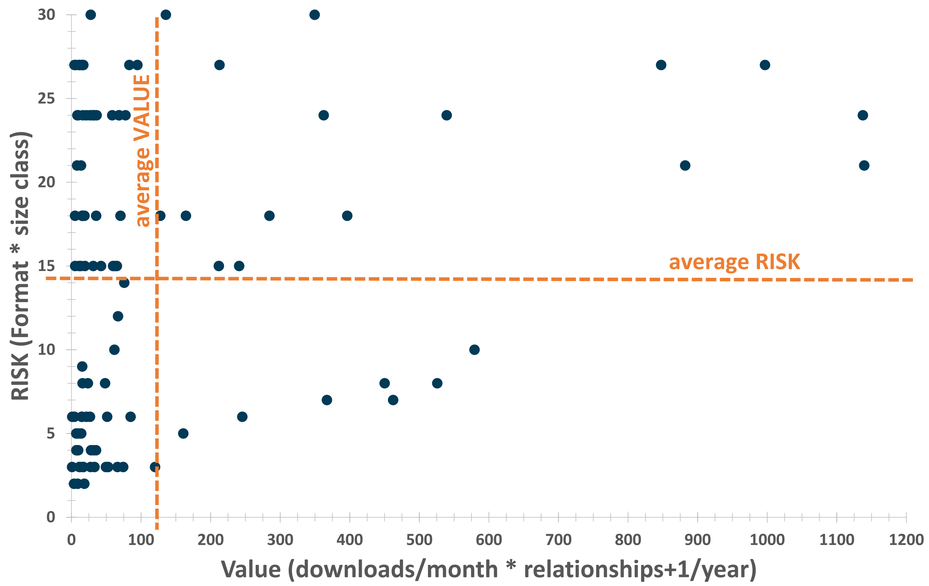


High Value, High Risk Datasets (N=6)
large, problematic formats, high downloads

Low Value, High Risk Datasets (N=12)
large, problematic formats, low downloads, mostly zero relationships

Low Risk, High and Low Value Datasets (N=88)
41 datasets containing problematic formats fell in “low risk” area

Risk vs Value Scatter
(using size class in Risk Calculation)



High Value, High Risk Datasets (N=16)
problematic formats, moderate to large size, most with very high downloads and multiple relationships

Low Value, High Risk Datasets (N=43)
problematic formats, moderate to large size, most with low downloads, and ≤1 relationships

Low Risk, High and Low Value Datasets (N=47)
all but one containing problematic formats fell in “high risk” area

GLOSSARY

Download (as defined by Illinois Data Bank): A dataset's download counter will increment up by one when one or more of its 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.

Medusa: University of Illinois Library's digital preservation management service for repositories including the Illinois Data Bank, our document repository (IDEALS), and the University Library's online and nearline digital collections. PREMIS records in Medusa represent stored digital objects and track their histories.

MIME type: Multipurpose Internet Mail Extensions type, a two-part identifier for file formats and format contents transmitted on the Internet (currently known as "Media type"). MIME types in Medusa are assigned automatically as files are ingested.

Relationship: A dataset's connection to another object, based on the DataCite properties relatedIdentifier, relationType. The relationType is limited to IsSupplementTo, IsSupplementedBy, IsCitedBy, and includes information about the object type (e.g., article). In the Illinois Data Bank, relationships to other objects are curated manually, usually by RDS staff using information supplied by the depositor or as a result of Google Scholar alerts on the Illinois Data Bank DOI prefix.

REFERENCES

1. Lambert, Dennis K., Winston Atkins, Douglas A. Litts, and Lorraine H. Olley. Guide to Review of Library Collections: Preservation, Storage, and Withdrawal, 2d ed. Lanham, MD: Scarecrow Press, 2002.
2. Texas State Library and Archives Commission. 2012. "CREW: A Weeding Manual for Modern Libraries. Austin, TX: Texas State Library." https://www.tsl.texas.gov/ld/pubs/crew/index.html
3. Reappraisal and Deaccessioning Development and Review Team. 2017. "Guidelines for Reappraisal and Deaccessioning." Society for American Archivists https://www2.archivists.org/sites/all/files/GuidelinesForReappraisalDeaccessioning_2017.pdf
4. Council, National Research. Environmental Data Management at NOAA: Archiving, Stewardship, and Access. Washington, DC: National Academies Press, 2007. https://doi.org/10.17226/12017
5. Anderson, Bethany, Susan Braxton, Elise Dunham, Heidi Imker, and Kyle Rimkus. "Should We Keep Everything Forever? Determining Long-Term Value of Research Data." In IPRES 2016, 13th International Conference on Digital Preservation, 3-6 October 2016, Bern, Switzerland, 2016. http://hdl.handle.net/2142/91659
6. Library of Congress. 2017. "Recommended Formats Statement 2017-2018." http://www.loc.gov/preservation/resources/rfs/RFS 2017-2018.pdf
7. Briney, Kristin. Data Management for Researchers: Organize, Maintain and Share Your Data for Research Success. Research Skills Series; Research Skills Series (Exeter, England). Exeter, UK: Pelagic Publishing, 2015.