DATA CURATION NETWORK

Data Sharing Readiness in Academic Institutions

by Lisa Johnston and Liza Coburn

Cite as: Johnston, Lisa R; Coburn, Liza. (2020-01-15). Data Sharing Readiness in Academic Institutions. Data Curation Network. Retrieved from the University of Minnesota Digital Conservancy, http://hdl.handle.net/11299/211358.

Back in 2017, several members of the Data Curation Network participated in the 2017 Academic Research Libraries (ARL) Data Curation Spec Kit, a survey asking 124 academic research institutions in the United States and Canada to self-assess their data repository and curation services (Hudson-Vitale et al., 2017a).

Three years later, institutional support for data sharing is as relevant as ever. Next month, the Association of Public Land Grant Universities (APLU) and the Association of American Universities (AAU) will convene a national summit in Washington, DC to address how public universities may increase public access to their research, particularly in light of funder and journal requirements supporting data reuse and research transparency.

So, we wondered, how has the academic landscape for data repository and curation services changed? To answer this question, we used website content analysis – a method that has been used successfully in recent years for examining the broader category of academic library research data services offerings (Yoon and Schultz, 2015; Kouper, Fear, Ishida, Kollen, & WIlliams, 2017) – to better understand data repository services in academic research libraries, building on the 2017 Spec Kit results.

Our approach

Of the 124 ARL institutions we chose to focus on academic institutions, and therefore excluded 10 civic libraries. For each of the remaining 114 ARL institutions we asked four research questions (see further details of our method and an analysis of the limitations of this approach at the end of this post):

1. Do they support data sharing via data repository services?
2. How many datasets did they hold as of January 2020?
3. What digital repository software platform was in use?
4. Finally, we compared our results with the 2017 SPEC Kit data (Hudson-Vitale et al., 2017b) to see how things might have changed in the last few years.

View the data

You can view our resulting dataset available as a google doc and as a downloadable version 1.0 (Johnston & Coburn, 2020). We would love to hear from repository owners about changes or suggestions for edits for an updated version (email us as: dcn-team@googlegroups.com)!

Our findings

The majority of ARL institutions support research data sharing with digital repository services. Collectively they have published at least 24,178 datasets. The number of dedicated data repositories offered across these academic institutions has grown over the last three years.

Fig 1. Academic Institutions with a research data repository (n=114). All the 114 academic institutions (100%) had an institutional repository (IR) of some kind that could potentially support data sharing (go team!). Additionally we observed that at least 50 institutions (44%) also had a dedicated data repository (either a standalone platform or a dedicated collection for data in the general IR) for showcasing the data shared by that institutions. We also observed that at least two more institutions were in the process of implementing a standalone data repository, in addition to their general IR.

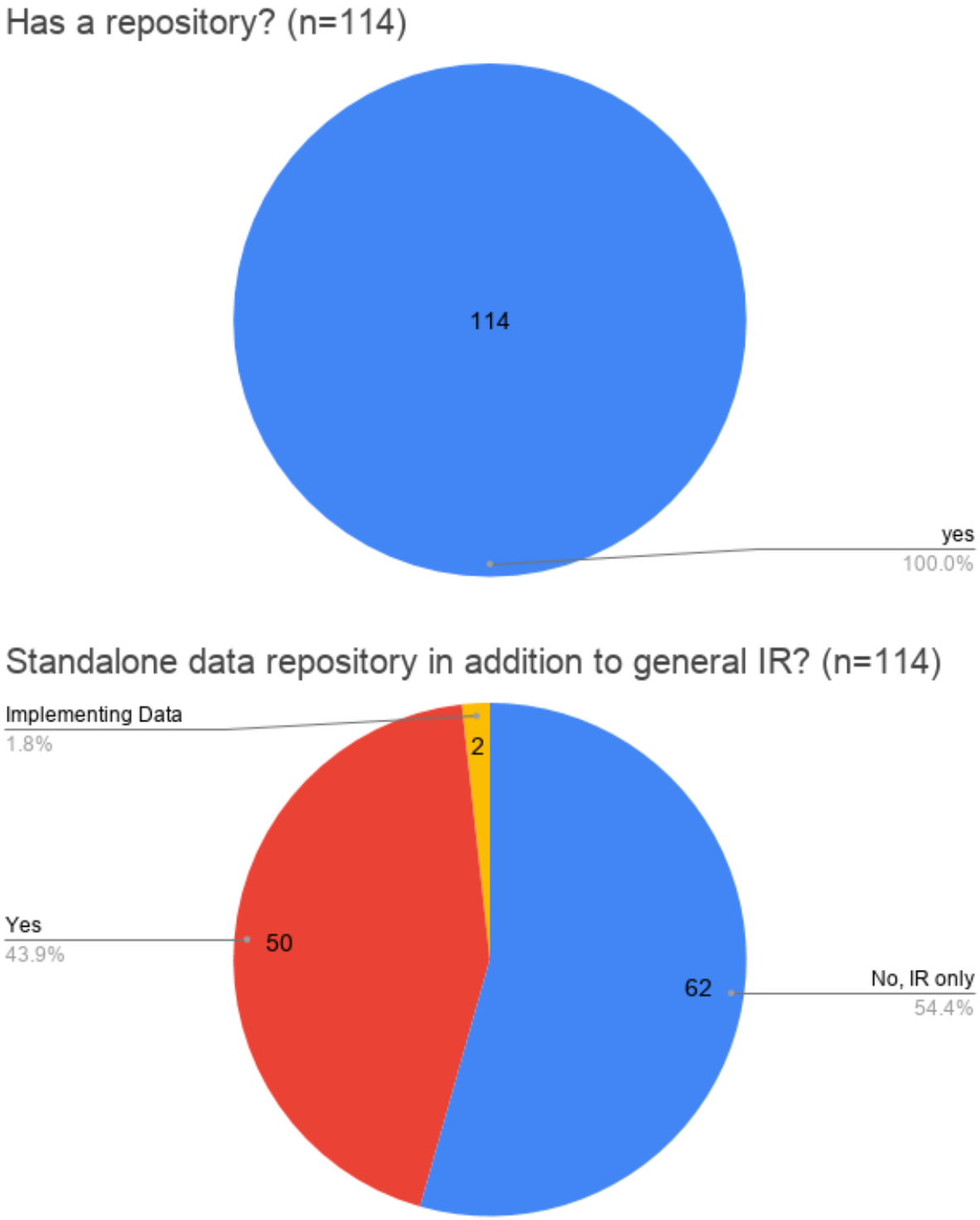
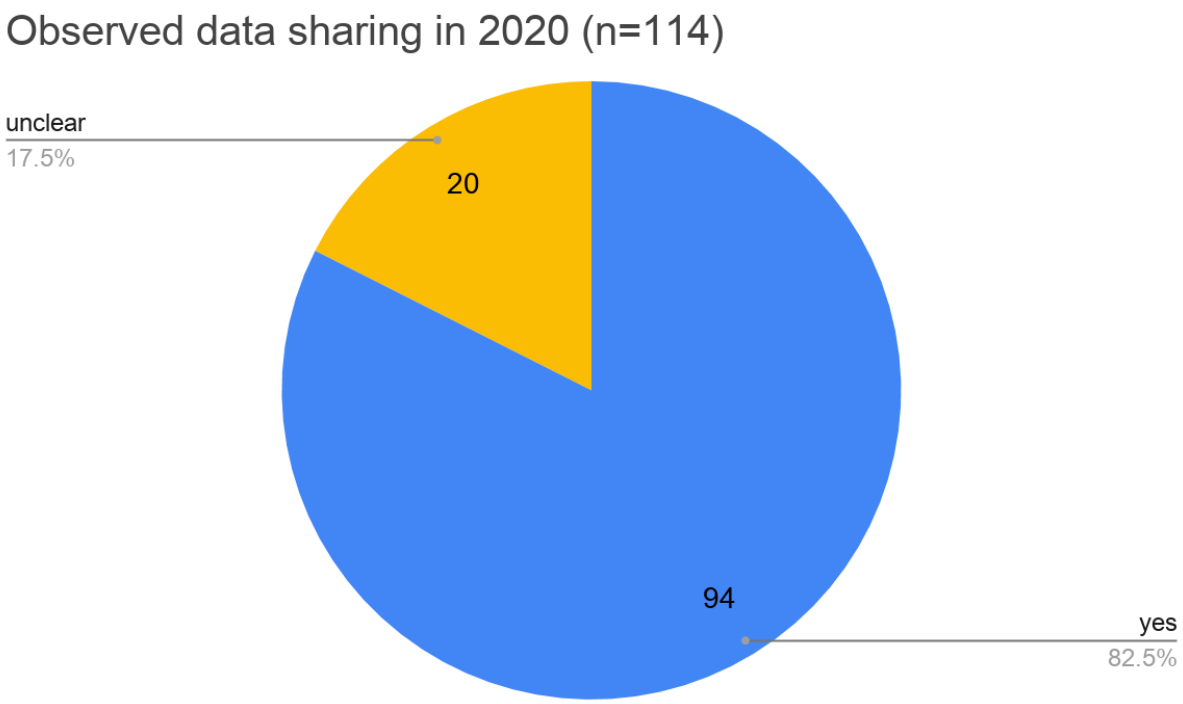


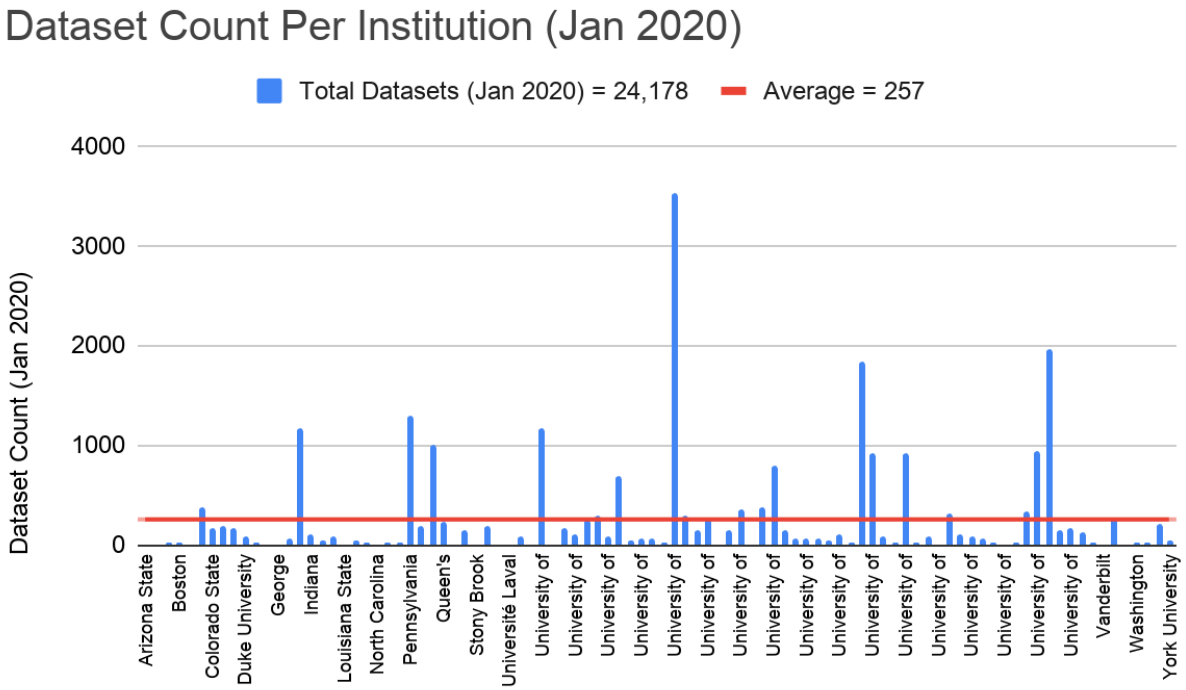
Fig 2. Academic institutions sharing research data (n=114). By looking at collection holdings and using record metadata fields (e.g, dc.type=dataset), we found datasets at 82% of the academic institutions (n=94). For 20 institutions, it was unclear if they contained data due to the lack of metadata differentiating a record as a dataset.



How we counted datasets. Any set of facts could be considered a "dataset" regardless of file format or subject area. In this study we counted a dataset if it was found in a standalone data repository or, in the case of a general IR collection, if it was labeled as "dataset" in the dc.type field.

Looking at the number of dataset holdings we had a few outliers and on further investigation it became apparent that this is not a foolproof approach as institutions do not count datasets in the same way. For example, a collection of 200 files relating to one another might be considered 1 dataset at Institution A, while Institution B might catalog each file as its own record and therefore hold 200 datasets. For example, 3,489 of the 3,532 records labeled as type = "dataset" at the University of Cincinnati were scanned index cards from a herbarium collection. While at the University of Minnesota, a camera trap dataset with over 6 million image files was cataloged as 1 dataset. Therefore it is hard to compare one institution to another. Another example, however, shows the reverse outcome of this approach. The University of Oklahoma had a "Research data" collection in their repository with 21,592 items. Only four of these were labeled as type = "dataset," while the rest were type = "stillimage" (and on spot check they appeared to us as scanned images). Here, we only counted 4 datasets for this repository (see more detail on our study limitations at the end of this post).

Fig 3: Number of datasets observed in each academic repository in January 2020 (n=114). We counted the overall datasets housed in each academic institution by totalling the number of datasets observed in the general IR and the dedicated data repository if available. We observed 24,178 datasets across 94 academic institutions, with an average of 257 datasets per repository (median = 80.5).

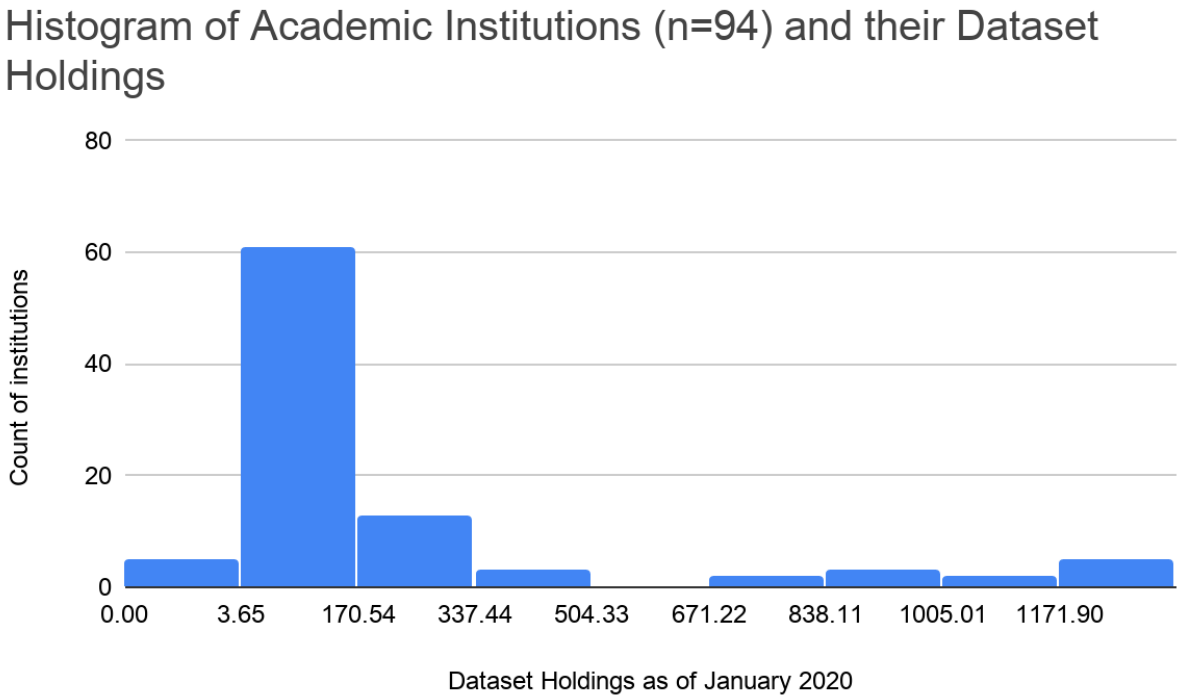


Click on image to view a larger version of this figure.

Fig 4: Data repository type and holdings observed (n=114). Looking at the different types of repositories observed, we see a large number of datasets housed in general purpose institutional repositories where departmental or individual collections might include a mix of articles, presentations, and theses in addition to data. Stand-alone data repositories on the other hand were branded specifically for data (such as a Dataverse instance or a unique collection within the larger IR).

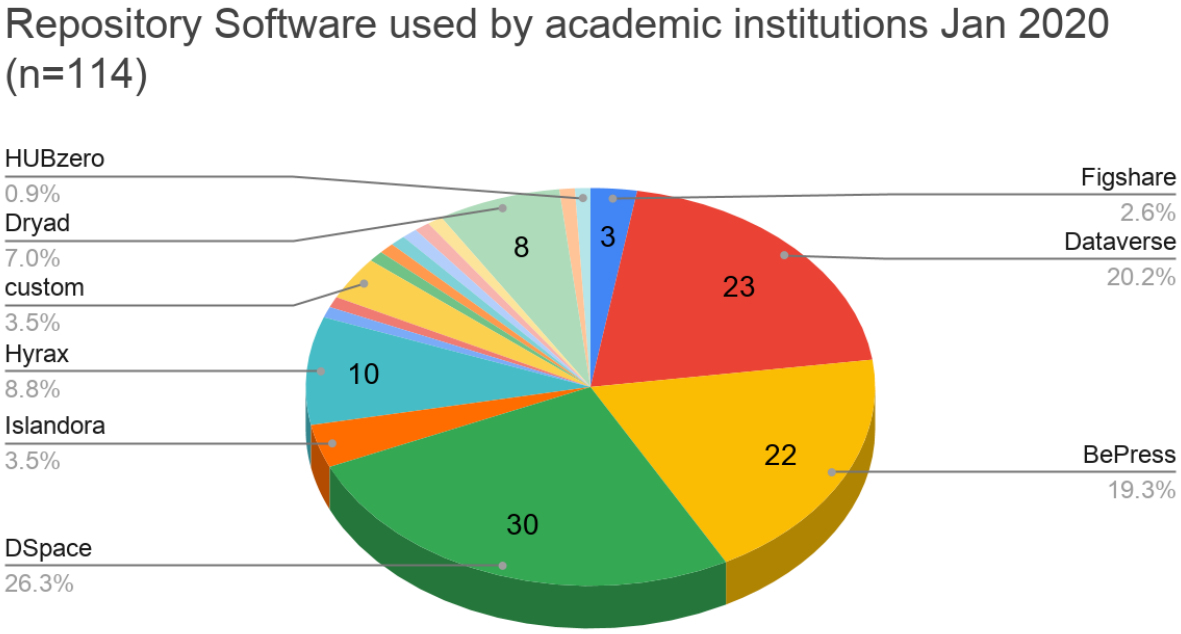
	Count of Repositories	Count of Datasets
General IRs	114	15,885
Standalone data repositories	50	8,293
Totals	**164**	**24,178**

Fig 5: Histogram of the data holding across academic repositories (n=114). The number of dataset holdings in each academic repository ranged from 0 to 3,532 datasets with the majority of institutional repositories holding 3-170 datasets.



Click on image to view a larger version of this figure.

Fig 6. Platforms used across n=114 academic research data repositories. All of the repositories reviewed use software platforms built for open access and discovery. Popular software used included DSpace, Dataverse, bePress, and Hyrax (Fedora/Samvera).



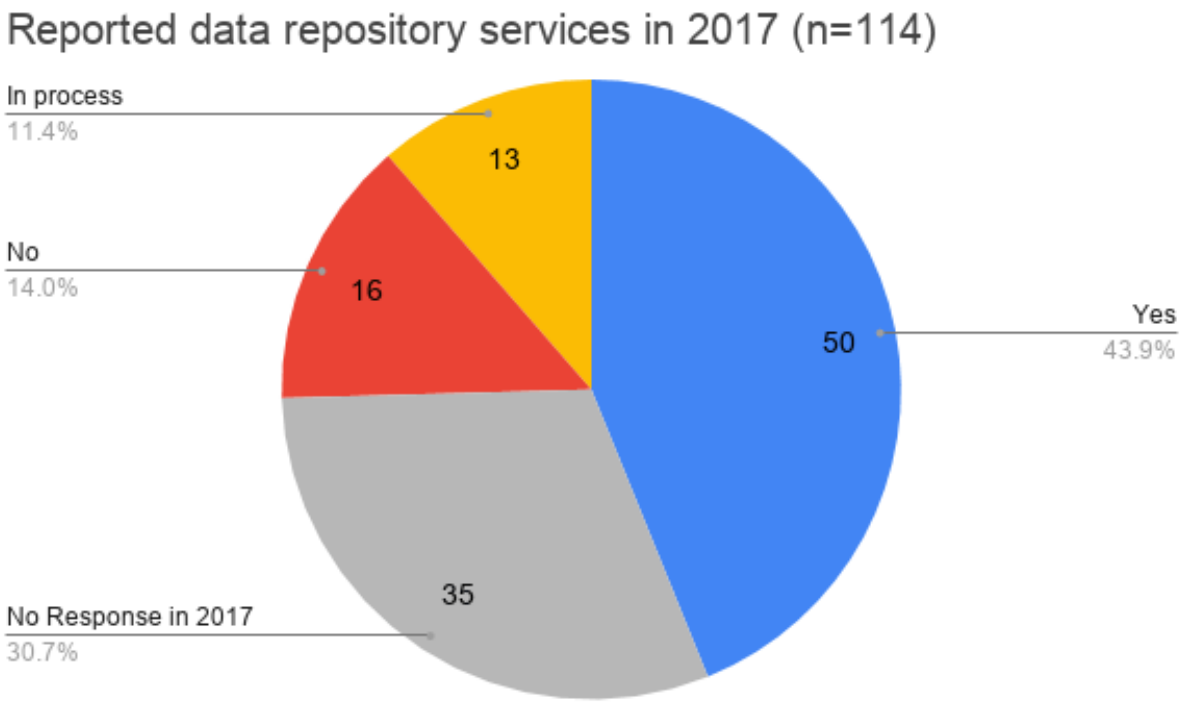
Platform	Count
DSpace	30
Dataverse	23
BePress	22
Hyrax (fedora/hydra)	10
Dryad	8
Islandora	4
Custom Solution	4
Figshare	3
contentDM	1

Platform	Count
Digitool	1
ePrints	1
Fedora	1
HUBzero	1
Rosetta	1
SobekCM	1
Sufia	1
TIND	1
unclear	1

Comparison of Today's Observations to those Self-reported in 2017

Our ARL Spec Kit survey captured data in January 2017, therefore our observations in this study (captured the first few weeks in January 2020) provide a three-year window for comparison. The results show progress in the number of datasets housed in ARL institutional repositories as well as the additional number of stand-alone data repositories launched.

Fig 7. Data Sharing Services 2017 vs 2020. We observed an increase in the number of institutions helping researchers openly share their data. In the 2017 ARL survey only 64% of responding academic ARL institutions reported offering data repository and curation services (50 out of 80 total respondents), yet we observed data shared in at least 82% of these same institutions in 2020. This uptick in data repository services among ARL academic institutions may reflect impressive growth in institutional support for data sharing, or it could be that academic institutions may not view their own digital repositories as suitable for research data, even when some researchers use them for this purpose.



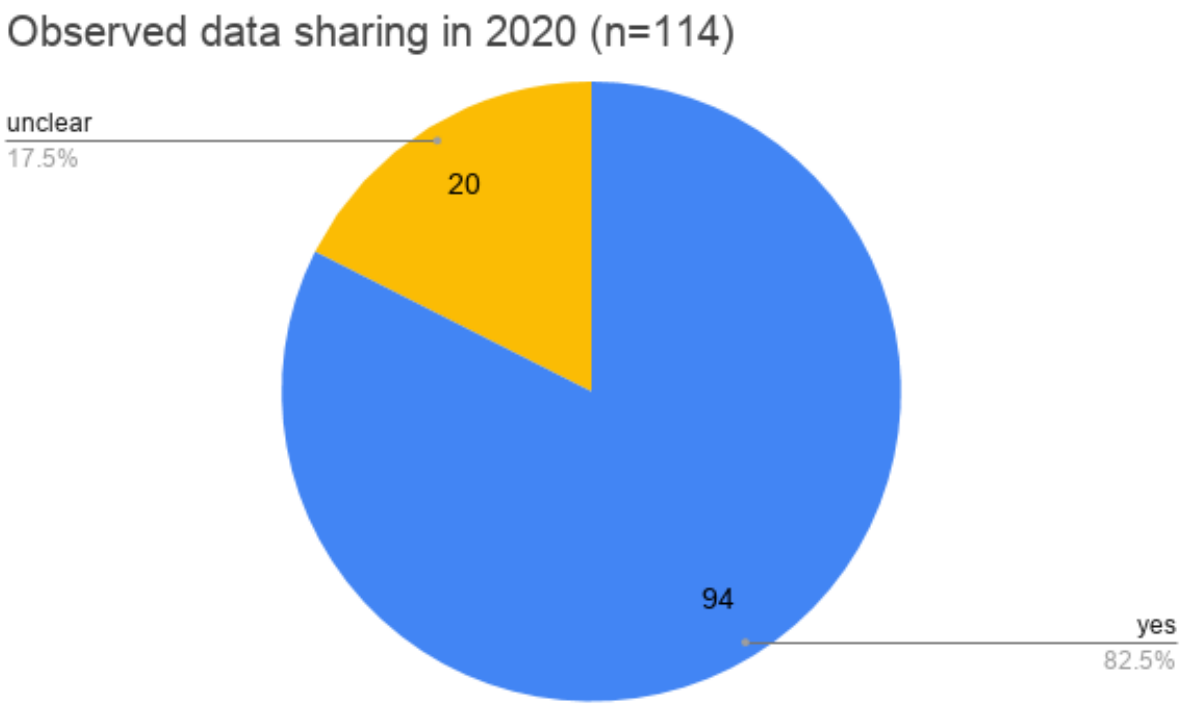
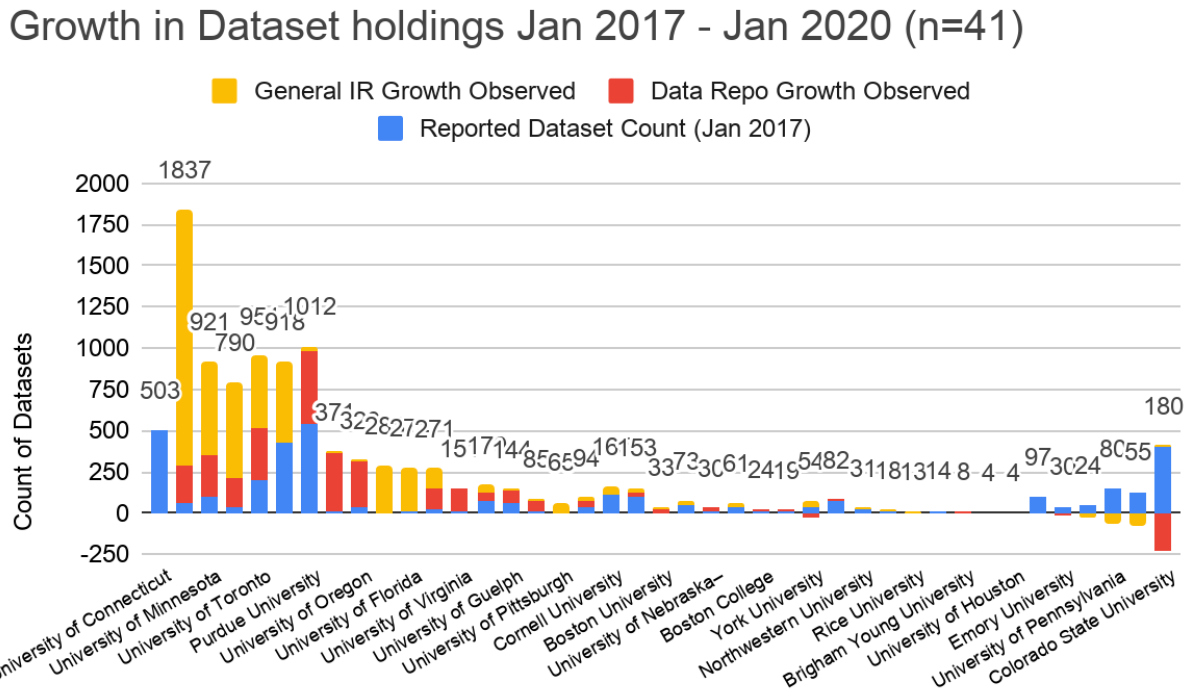


Fig 8. Datasets Holdings: 2017 vs. 2020. In 2017 43 ARL institutions reported the number of datasets in their repository. Three years later we observed an anticipated growth in dataset holdings in 34 institutions by an average of 77 datasets per year. However 8 institutional dataset holdings were found to *decrease* over time and 1 was unclear. A decrease in datasets is unlikely, rather this is likely due to our different interpretations for what counts as data. Two institutions are not shown on this graph due to the large observed decrease in numbers since 2017: Washington State University (-5106 datasets) and Vanderbilt University (-7437 datasets).



Click on image to view a larger version of this figure.

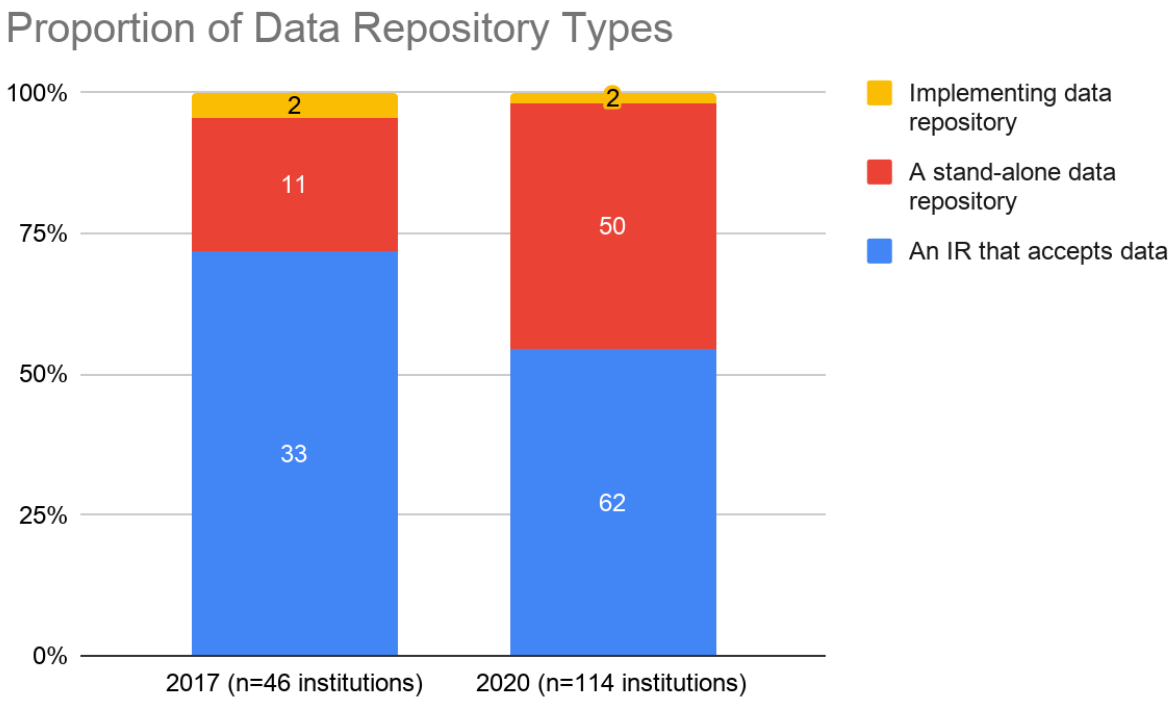
Fig 9: Data repository platforms: 2017 vs 2020. In the 2017 survey, 34 institutions reported their platform. Of those, 4 were observed to change platforms for data from what they reported using three years ago. These were

Institution	2017 (reported)	2020 (observed)
New York University	Custom	DSpace
University of California, Irvine	DSpace	Dryad
University of Illinois at Chicago	DSpace	Figshare
University of Tennessee	Custom solution	BePress

Fig 10: New Platforms. Across the 114 institutions observed in 2020, we saw 7 new platforms not previously reported in 2017. It is unclear if this demonstrates a growth in platforms used since the sample sizes were very different.

Year	Institutions Count	Total Platforms Reported (2017) and Observed (2020)
2017	34	10 = BePress, custom solution, Dataverse, DSpace, ePrints, Fedora/Hydra, HUBzero, Islandora, Rosetta, SobekCM
2020	114	17 = BePress, contentDM, custom solution, Dataverse, Digitool, Dryad, DSpace, ePrints, Fedora, Figshare, HUBzero, Hyrax (Fedora/Hydra), Islandora, Rosetta, SobekCM, Sufia, TIND

Fig 11. Standalone Data Repositories: 2017 vs 2020. In the 2017 survey, 46 institutions indicated the type of repository they used for data. At that time, 24% self-reported their repository as "A stand-alone data repository" while the majority (71%) classified it as "An institutional repository that accepts data." In 2020 we expanded our definition of a stand-alone data repository to also include a well-branded data collection housed within an IR. Therefore it is difficult to draw any conclusions from this comparison as the proportion of stand-alone data repositories rose along with our expanded definition.



Summary Conclusions

The results of this web site analysis of 114 academic institutions provides a relevant snapshot for how research libraries support researcher data sharing needs via digital repository services.

In this study we found:

- 100% of academic ARL institutions in our sample offered a digital repository. Additionally we observed that at least 50 institutions (44%) had a dedicated data repository that supports data sharing (either a data collection in a general IR or standalone data repository).
- Collectively, academic ARL institutions published at least 24,178 datasets as of Jan 2020. Datasets were found in 82% of the academic institutions (n=94). For the remaining institutions, it was unclear if they host data.
- There were 17 different software platforms observed in use by academic institutions for sharing data. This demonstrates a diverse variety.
- In 2020, we observed more academic ARL institutions supporting data sharing via their repository service than reported doing so in 2017 (83% up from 64%). This uptick in data repository services across ARL academic institutions may reflect positive growth in institutional support for data sharing. It could also be that academic institutions did not previously view their own digital repositories as suitable for research data, even when some researchers use them for this purpose.

Some recommendations

1. Academic digital repositories would benefit from a standard approach for how we label digital holdings as datasets. If useful guidelines already exist, we would love to hear about them!
2. There is a surprising amount of data found in general IRs, more so than we observed in the albeit newer approach of a stand-alone data repository. The pros and cons of these different approaches are worth exploring further!
3. We actually had a fifth question that we were unable to answer with this method of website analysis that asked "what data curation support do academic ARL institutions offer?" We rarely saw this type of information when reviewing the repository site. One example was Boston University that stated: "All data in the collection are curated to increase potential for access and are assigned a permanent Handle URL." We recommend that academic repositories consider including curation information to demonstrate to researchers that they can trust your digital repository as an appropriate place to share research data. A follow up survey may be a better approach to understanding how curation services have evolved.

Fine Print: More about our approach and the limitations of this study

Data collection was done by two people and verified independently. Institutional data repositories (IRs) were identified by searching two sources: 1) via Google.com with an institution's name and few different terms "data repository", "digital repository" or "institutional repository" and 2) via base-search.net which provided an excellent OAI-PMH metadata parser interface (BASE, 2020; Pieper & Summann, 2006).

This study aimed to count the number of original research datasets published in an academic ARL institution. Identifying dataset holdings was tricky. In the case of a dedicated data repository (e.g., a Dataverse instance), the total number of datasets is clearly displayed. In the more common example of an institutional repository that accepts data, we dug deeper looking for records labeled as a "dataset" vs. photos, articles, theses, etc. First, we browsed collection names and searched using a variation of search strings and keywords. Most repositories using the Dublin Core metadata schema could be successfully refined using the "dc.type" facet set to "datasets," "dataset," "data sets", or "data". Therefore, rather than relying on our own biases as to what "counts" as data, we counted the number of records labeled as dataset in the metadata. Next, results were spot-checked in an attempt to further verify the objects as research data, as opposed to an article describing a dataset. If the results could not be satisfactorily limited to records we objectively categorized as data, we labeled the number of datasets in the institutional collection as "unclear." Then we validated our findings with the base-search.net tool by directly parsing the OAI-PMH feed to type = dataset. This tool also provided us with the platform type.

Using this approach, our numbers are likely undercounting the total dataset holdings for this sample. Furthermore, geospatial data and other specialized data types may be stored in a dedicated data repository outside of our review (e.g., a GeoBlackLight instance), and as a result this study may significantly underrepresent the GIS datasets hosted by academic institutions. Datasets licensed by the library were not the focus of this study, though we may have unavoidably included some. Finally, it is difficult to compare the self-reported survey responses from 2017 with the observations made in 2020 due to the inconsistencies of interpreting what is data and how to count the number of datasets in a collection. In one case, an institution reported 7442 datasets in their repository in 2017 but on further inspection in 2020 it became clear that this number most likely represented all repository holdings including articles, reports, theses, etc. Our observation of the number of datasets in 2020 was much lower (7 datasets).

References

BASE. (2020). BASE (Bielefeld Academic Search Engine). https://www.base-search.net/.

Kouper, I., Fear, K., Ishida, M., Kollen, C., and Williams, S. C.. (2017). Chapter 6. Research Data Services Maturity in Academic Libraries. In *Curating Research Data, Volume One: Practical Strategies for Your Digital Repository*, (L. R. Johnston, ed.). Chicago: Association of College and Research Libraries, 153-170. http://hdl.handle.net/2429/61121.

Hudson-Vitale, C., Imker, I., Johnston, L. R., Carlson, J., Kozlowski, W., Olendorf, R., & Stewart, C.. (2017a). Data Curation. SPEC Kit 354. Washington, DC: Association of Research Libraries. https://doi.org/10.29242/spec.354

Hudson-Vitale, C., Imker, I., Johnston, L. R., Carlson, J., Kozlowski, W., Olendorf, R., and Stewart, C.. (2017b). Survey Data for SPEC Kit 354: Data Curation. Github. https://github.com/1heidi/dcn_spec_kit_data.

Johnston, L. R., & Coburn, E. Data supporting "Data Sharing Readiness in Academic Institutions" Version 1. Data Repository for the University of Minnesota (DRUM). https://doi.org/10.13020/2EVX-7A87.

Pieper, D., & Summann, F. (2006). Bielefeld Academic Search Engine (BASE) An end-user oriented institutional repository search service. Library Hi Tech, 24(4), 614-619. https://doi.org/10.1108/07378830610715473.

Yoon, A. & Schultz, T. (2017). Research data management services in academic libraries in the US: A content analysis of libraries' websites. *College & Research Libraries, 78*(7), 920-933. https://doi.org/10.5860/crl.78.7.920