# Factors contributing to repository success in recruiting data deposits

Michele Hayslett & Matthew Jansen[1]

## Abstract

What factors make data repositories successful in recruiting research data deposits from scholars? While quite a few studies outline researchers' data management needs and how repositories can meet those needs, few have assessed the success of various approaches. This study examines infrastructure for accepting data into repositories and identifies factors influential in recruiting data deposits.

## Keywords

data repositories, data deposits, deposit recruiting, best practices, marketing

## Introduction

Throughout the early 2000s, librarians have worked to build repository infrastructure, deposit workflows, access features, and support services to meet the needs of their audiences. Needs assessments and evaluation of research practices often informed the creation of repository features and services, and descriptions of these design efforts abound in the literature. However, few post-deposit assessments have been published to identify the most successful recruitment practices for datasets. Moreover, the content of many institutional repositories (IRs) is focused on textual deposits such as pre-prints and publications. Fewer repositories enable the preservation of datasets, and the literature yields very few assessments of data repositories.

Many questions can be asked about the work to recruit dataset deposits in repositories. What features and services have proven to be the most useful and attractive to researchers? Staffing, both overall and specifically related to data deposits, may logically have significant positive correlations with larger numbers of depositors, but does choice of marketing media likewise have a high association? This project addresses the over-arching research question: what factors are most associated with larger numbers of data depositors?

## Literature Review

As the field of data management has grown, a great deal of the literature has detailed the overall benefits of open data and more specifically the advantages an individual scholar would accrue from sharing their research. Many articles also focus on ascertaining investigators' needs and how to design repositories to meet those needs. These two pools of research overlap and offer repositories formative information with which to plan recruiting strategies for research materials generally, including data. Little is available in the literature, though, about the next step in the process: evaluation of how successful those early needs assessments and system designs have been, and what other factors are positively correlated with larger numbers of depositors. A variety of searches of the *International Journal of Digital Curation* between October and November 2021 ('evaluation success'; 'marketing'; 'effects deposit rate'; 'factors affecting deposit rates') yielded only one result related to the evaluation of repositories broadly. McHugh et al. (2008), describes the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA), a flexible framework for self-audit to assess 'demonstrable, and not just inferred, success (p. 135)' but does not actually discuss results of any specific assessment. While self-audit can provide valuable information for repositories and its results can be kept private, wider sharing of assessment metrics may yield valuable information for the broader community. This study seeks to address this gap.

### Data Sharing and Incentives

Many articles have reported on the benefits of sharing data, both as a public good that improves reproducibility and the overall quality of research, and for individual researchers, raising the profile of their work and increasing their impact (Gardner et al., 2003; National Research Council, 1985; Pienta, Alter & Lyle, 2010; Piwowar, Day & Fridsma, 2007). However, numerous surveys have found that while researchers often declared willingness to share their data, many barriers obstruct them from actually sharing (Fecher, Friesike & Hebing, 2015; Gardner et al., 2003; Lowenberg, 2017; National Research Council, 1985; Tenopir et al., 2011; Wallis, Rolando & Borgman, 2013).

Where compliance serves as an incentive for data deposit, researchers may sometimes only encounter a funder's compliance requirement when they apply for grant funding. Faniel and Connaway (2018) note that several librarians in their survey found their data management services being contacted just prior to grant proposal deadlines. This matches the authors' own experience with many researchers seeking help writing data management plans only days before a proposal deadline. IRs may be better positioned than some other repositories to assist researchers in this last-minute way, with low barriers to deposit. It might then follow that researchers find IRs more important for preserving data than articles. Bryant, Lavoie, and Malpas (2017) note, 'Given the extensive network of discipline-, consortial- and national-scale [research data management (RDM)] services, many institutions have scoped their local RDM service bundles to be complementary to, rather than parallel with, these external options' (p. 30).

Hudson-Vitale et al. (2017) found research libraries that provide data curation services viewed providing a persistent identifier as the most important data curation activity overall, but it remains to be seen if researchers agree with this. The authors have encountered quite a few depositors who did want to deposit their data specifically to obtain the persistent identifier assigned by the IR to insert in an upcoming publication based on those data. But what are repositories broadly (not just IRs) experiencing? What other services are repositories finding to be in demand? Does offering more advanced services like file or code review, encryption, and direct deposit via Electronic Lab Notebooks (ELNs) correspond with having more depositors?

### Needs Assessments and Repository Design

Quite a few needs assessments and pilot projects for repositories have been reported, covering everything from the benefits for depositors to repository technical infrastructure to metadata creation services (Abrams et al., 2014; Burton & Treloar, 2009; Hudson-Vitale et al., 2017; Mattern, Jeng, He, Lyon, & Brenner, 2015). Many note special challenges associated with archiving data. As early as 2008, Salo observed in her 'repository as a roach motel' article that the 'build it and they will come' approach to repository collection development with which many institutions started had not been successful and argued that different incentives than commonly cited ones like preservation and heightened impact were needed.

Plale et al. (2013) noted difficulties publications-based repositories can encounter in archiving data and explored how those challenges can be addressed through requirements, policy, and architecture. Minor et al. (2014) describe how pilot ingests of datasets directed the design and implementation of the UC San Diego Library's Research Data Curation Program. Borgman et al. (2016) explored cloud-based services as a data management solution for 'long-tail' research projects, that is, smaller projects with few resources for documenting and preserving their data. While these services were found to be useful for some basic tasks, they were insufficient for more complex needs such as development of specialized data tools and long-term preservation, needs that can be addressed in repositories. Peer and Green (2012) describe how the push to make more research open access is reflected in Yale's efforts to host an open access repository on open-source software with the goal of supporting research replication as well as re-use and instruction. Few of these case studies have published follow-

up reports on their particular results (although Peer and Green do note positive early feedback and outcomes from users), but there are some studies of outcomes generally.

Tillman (2017) surveyed self-deposit rates at 55 U.S. IRs and concluded, '…the short answer to the question 'is our faculty depositing?' is 'not really,' or the even more straightforward 'no.'…Everyone is trying. Few are succeeding' (p.13). But Tillman also evaluated other variables' correlation with high deposit rates: the age of the IR; the software on which the repository is built; and the outreach methods of the IR. (She cites several reasons why the software would be important, both from the standpoint of a faculty member's willingness to use [and re-use] the IR, and as an indicator of the administration's investment in the IR.) Her results are particularly revealing of the difficulties repositories face in achieving success. With regard to the age of the IR, she notes:

> It appears to take a minimum of two years on average for repositories to have even a 50% likelihood of getting at least a single self-deposit per month, with a much greater likelihood of success after five years. This time period allows for the IR to become an established entity on campus and for responsible parties to do a variety of outreach and build new strategies after failures in the > 2-year and 2–5 year ranges. However, as 66.67% of repositories that had existed for at least five years *still* had self-deposit rates of 20 items or fewer and a full 25% had 0 average monthly self-deposits, the comparatively positive correlation of success and age should not be considered a guarantee. Age correlates even more strongly with failure. (p. 14)

The results about software seemed to indicate it was important to high self-deposit rates but not conclusively why that was so:

> Institutions reporting higher rates of self-deposit are more likely to report the use of either a homegrown system or one that involves high levels of developer time and engagement. This factor may indicate that the institution is investing heavily in the repository, implying that it allots more staff time and effort for other activities that promote deposit. It may also indicate that the user interface for deposit in turnkey models does not promote self-deposit. (p. 14)

Finally, she notes that reports on IRs' outreach are also correlated 'strongly with both a strong deposit profile and the lack thereof. No conclusions, therefore, can be drawn about either in general (p. 14).' Tillman did not differentiate between publications and data deposits, though, and focused on self-deposit rates. This study focuses on data, in whatever way they are deposited.

### Marketing

Over time, as research norms change, more researchers will begin to deposit their data if such practices become the accepted cultural norm. Outreach is one way for libraries to reinforce such a cultural shift. Bryant, Lavoie, and Malpas (2018) note:

> There is an evangelistic aspect to…[educational data management] outreach—although researchers may not be ready to deposit data at the time the outreach occurs, they are at least made aware that data management services are in place to support them when needed…Successful outreach program, can over time, cultivate the demand that will help establish [RDM] as a critical piece of scholarly infrastructure (p. 16).

Indeed, many universities, especially in Europe, are targeting graduate students, seeding cultural change in the next generation of researchers. The University of Groningen (2013) requires its doctoral students to make their data 'available for further research,' although exemptions are possible for 'compelling reasons' (p. 13). To propose a similar directive be implemented at the Delft University of

Technology (TU Delft), Dunning (2017) compiled policies from Groningen and seven other universities in the U.K. and the Netherlands: Utrecht, Leiden, Twente, Bristol, Southampton, Bath, and Manchester. Most of these stated their policies as an expectation, without making deposit mandatory. TU Delft decided their doctoral students '…starting from 1 Jan 2019 will have to share data unless they have a compelling reason not to' (A. Dunning, personal communication, May 9, 2019). Their policy states the expectation of this covering 'all data and code underlying completed PhD theses,' and that they be 'appropriately documented and accessible for at least 10 years from the end of the research project …' (p. 7, TU Delft, 2018).

Many academic repositories' outreach efforts also aim to change faculty attitudes, of course. Otto (2016) describes the outreach message Rutgers used with the '…primary objective…to fully inform faculty so that they were motivated to make the open access choice …' (p. 11). She goes so far as to quote Thomas Jefferson in his description of the Declaration of Independence, that their objective was 'to present [to the 'tribunal of the world'] 'the common sense of the subject, in terms so plain and firm as to command their assent.' (Jefferson, 1825)' (p. 11). She concedes, though, '…evidence of its own efficacy remains, for the most part, unavoidably anecdotal' (p. 4). Several projects are assembling lists of European institutions that have established policies, and many of these apply to all researchers, not only graduate students. Laurence Horton with the London School of Economics collaborated with the U.K.'s Digital Curation Centre to compile the 'Overview of UK Institution RDM Policies' web site (M. Donnelly, personal communication, June 28, 2019). As of June 21, 2022, it included 86 institutions plus one that was drafting a policy (although none are dated past 2016, this number has changed during the writing of this paper and some policies are undated). Kerstin Helbig at Humboldt-Universität zu Berlin noted their web site listing with links to dozens of German institutions' policies in a message to the RESEARCH-DATAMAN [listserv, a research data management] list (K. Helbig, personal communication, June 28, 2019). And FAIRsharing.org, a manually curated educational resource that describes and captures the relationships between standards, databases and data policies, indicated it would 'soon be inviting submissions to [its] data policy registry for these types of institutional research data policies' (P. McQuilton, personal communication, June 28, 2019)—it included 155 such policies as of June 21, 2022; however it does not appear to allow filtering by type of organization, e.g., IR, disciplinary repository, publisher, professional or research association, or other.

Few examples in the literature have evaluated social media as a marketing method to reach potential depositors, though. Boulton (2020) reviewed the literature around institutional repositories and engagement and found that, broadly, the repository community focused on improving systems and structures to make use of their repositories easier rather than on evaluating direct outreach to their audiences. Given that void, that paper turned to examining the social media practices of IRs but did not detail exactly how, nor how many IRs were examined, merely noting, 'Engagement through social media channels such as Facebook and Twitter appears common across institutional repositories …' (p. 1). This practitioner's aim was to go beyond using a single social media channel and describe instead a coordinated campaign using multiple media at Griffith University in Australia. The case study analyzed the traffic driven by two blog posts (covering the research backstories and results of 15 deposited articles) that were promoted in tweets from the Library's Twitter account and concluded that '…social media and blog posts could be used by the Library to increase engagement with an external audience and to drive traffic to the repository' (p. 4). In other words, the particular aim of this experiment was to increase research impact rather than to increase depositors but the same strategy could potentially achieve both ends. Certainly, the Griffith repository succeeded in highlighting the articles: during the two months of the campaign, '... the 15 featured articles were accessed close to 500 times, this being approximately 60% of their combined total for the previous six months …' (p 3).

Lafferty-Hess et al. (2018) discuss how their thought exercise of conceptually grouping the DCN's data curation activities helped them not only identify appropriate services for their respective institutions given different available resources, but also to consider communication strategies to make their services clear to researchers, and strategies for measuring their success. It is the authors' hope that the current study will also aid repositories in these important ways.

## Methodology

The population under examination in this study was North American data repositories, whether open or not, whether independent or based in an institution. (Note: respondents were not asked to identify the type of repository in which they worked. Also, the terms "researchers" and "faculty" are used interchangeably in this paper.) The focus was on those that accept datasets. While brief information was invited from repositories that do not, few non-data repositories participated. The appeal for survey participation (shown in Appendix A) was sent to repository professionals via three email lists heavily populated by data curation professionals: the membership list of the International Association for Social Science Information Services and Technology (IASSIST); that of the U.S.-based Research Data Access and Preservation (RDAP) Association; and datacure, a list established by geographically scattered participants in the Digital Curation Curriculum (DigCCurr, pronounced "dij-seeker") at the School of Information and Library Science (SILS) at UNC-Chapel Hill after returning to their home institutions, and whose membership has grown rapidly since. No organization is currently maintaining a directory of all repositories (OpenDOAR offers a directory only of open access repositories) so this approach was deemed best.

The initial appeal was sent to all three lists in August 2019, with two follow up reminders sent at the three- and five-week marks. To give an idea of the audience reached by the appeal, list membership at the time of the survey stood as follows: datacure – 231; IASSIST – 539; and RDAP – 557, for an estimated total of 1,327 invitations. However, there is some overlap among the three organizations' membership, and IASSIST includes members from outside of North America, so the actual number of eligible participants reached is unknown.

Data were collected by an online Qualtrics survey, which was pretested by several volunteers from the 2019 RDAP conference. The consent form was presented as the survey's first page (see Appendix B for the consent form and Appendix C for the full survey instrument). The survey solicitation and the survey instrument specified that one response per repository was requested, and the survey instrument was structured to allow response over multiple sessions to encourage input from multiple individuals as necessary to construct a full picture of each repository's characteristics. Response was requested within a six-week window, by September 30. The survey instrument was left available for some weeks after that deadline in case of late responses; the last response was recorded on October 10. The number of respondents was small, impacting what conclusions may be drawn from these results: 31 submissions overall, with two who did not accept data and one who accepted data but did not provide their number of depositors. Because of non-responses in number of depositors and other variables of interest, the sample size varied between 26 and 28 (see Appendix E for specific results by hypothesis). Still, the data were sufficient to identify multiple trends. Survey respondents were invited to indicate their willingness to participate in follow-up interviews but very few did so; consequently, plans for follow-up interviews were abandoned.

## Results

Eight hypotheses were tested for this paper:

1. Repositories with the highest staffing will have a significant positive correlation with larger numbers of depositors.

2. Larger numbers of depositors will be significantly correlated with offering advanced curation services.
3. Repositories with larger numbers of depositors will be those which have depositors referred by faculty versus referrals from other places.
4. Larger numbers of depositors will be significantly correlated with using social media to promote the repository.
5. Larger numbers of depositors will not be significantly positively correlated with infrastructure except where repository ingest is linked to electronic lab notebooks (ELNs).
6. Larger numbers of depositors will be significantly positively correlated with repositories that encrypt data.
7. Larger numbers of depositors will be significantly positively correlated with having more staff dedicated specifically to data deposits.
8. Larger numbers of depositors will be significantly positively correlated with older repositories.

All tests were performed at a 95% confidence level. Because of the small response rate, the authors focused on measuring relationships between just two variables at a time. This approach enabled rejection of a hypothesis but not assertion of causation nor indication of direction, i.e., which variable might have caused the other. The number of depositors as tested for each hypothesis does not follow a normal distribution and has large outliers, so a two-sided Asymptotic Wilcoxon ranked sum (aka Mann-Whitney) test was used for categorical data and a Spearman test for numeric data. Statistical testing was performed using R version 4.0.5 on a x86_64-w64-mingw32/x64 (64-bit) platform running Windows 10 x64 (build 19043). The R code and markdown files for the analyses are described and linked in Appendix D. Due to the uncertainty in how many the survey invitation actually reached, it was not possible to calculate response rates.

1. **Repositories with the highest staffing will have a significant positive correlation with larger numbers of depositors.**

   Analysis supported this hypothesis. Spearman correlation tests were performed (using the mid-ranks method to deal with ties) to examine relationships among numeric data and to control for large outliers. Two scatterplots are included in the markdown (.rmd) file to compare the lack of trend lines among data points when outliers were included, with the Spearman results which enabled exclusion of those extreme values and a more detailed view of the data. Having a larger staff was significantly correlated with having a larger number of depositors ($\rho$ =0.43, $Z$ = 2.25, *p-value* = 0.02). Because the correlation coefficient, rho ($\rho$), is positive, the relationship is positive: when one variable goes up, the other goes up.

2. **Larger numbers of depositors will be significantly correlated with offering advanced curation services.**

   Analysis supported this hypothesis. The survey asked respondents to indicate which of the following services each offers: minting digital object identifiers (DOIs); basic data curation tasks (e.g., checking data files, reading documentation, assignment of keywords/subject headings, running checksums); more staff-intensive data curation tasks (e.g., verification of file organization, file format normalization or migration over time, code checking, or replication); inserting an internal link between repository records for a deposited dataset and a deposited article based on those data; inserting a link from the repository record for a deposited dataset to a *different* repository's record for a deposited article; data encryption; scanning for personally identifiable information; and other services (a write-in category). Write-in answers for the Other category were excluded from analysis. The remaining services were grouped into basic versus advanced services offered as shown in Table 1 below. In most cases repositories offering advanced services provided

some if not all of the basic services as well, so testing focused on whether or not a repository offered advanced services.

A Wilcoxon-Mann-Whitney Test was employed to handle tie values and outliers in the data, since it is based on ranks instead of the raw values and p values were computed using mid-ranks to break ties. Offering advanced curation services had a significant association with having a larger number of depositors ($Z$ = -2.06, *p-value* = 0.04). Repositories offering advanced curation services had 12.5 more median depositors than those not offering such services.

*Table 1. Groupings of Curation Services*

| Basic | Advanced |
|---|---|
| · Minting DOIs<br>· Basic data curation tasks<br>· Inserting an internal link between repository records for a deposited dataset and a deposited article based on those data<br>· Inserting a link from the repository record for a deposited dataset to a *different* repository's record for a deposited article | · More staff-intensive data curation tasks<br>· Data encryption<br>· Scanning for personally identifiable information |

**3. Repositories with larger numbers of depositors will be those which have depositors referred by faculty vs referrals from other places.**

This hypothesis was undetermined. With only three of 28 repositories *without* referrals from faculty, conclusions drawn from statistical testing seemed unrepresentative, but the fact that so many repositories had referrals from faculty is notable.

**4. Larger numbers of depositors will be significantly correlated with using social media to promote the repository.**

Analysis supported this hypothesis. Using social media had a significant association with having a larger number of depositors ($Z$ = -2.40, *p-value* = 0.02). Those repositories using social media had on average 22 more median depositors than those that those not using it.

**5. Larger numbers of depositors will not be significantly positively correlated with infrastructure except where repository ingest is linked to electronic lab notebooks (ELNs).**

This hypothesis was undetermined. With only two of 28 repositories linking to ELNs, conclusions drawn from statistical testing seemed unrepresentative, but the fact that so few repositories offered such integration is notable.

6. **Larger numbers of depositors will be significantly positively correlated with repositories that encrypt data.**
   This hypothesis was undetermined. With only four of 27 repositories offering encryption, conclusions drawn from statistical testing seemed unrepresentative, but the fact that so few repositories offered this service is notable.

7. **Larger numbers of depositors will be significantly positively correlated with having more staff dedicated specifically to data deposits.**
   Analysis supported this hypothesis. An Asymptotic Spearman Correlation Test found a significant correlation between the number of depositors and the number of dedicated data staff: those repositories with more data staff also have larger numbers of depositors ($\rho$= 0.71, $Z$ = 3.64, *p-value* = [<0.01]).

8. **Larger numbers of depositors will be significantly positively correlated with older repositories.**
   This hypothesis was unsupported. An Asymptotic Spearman Correlation Test found repository age was not significantly correlated with numbers of depositors ($\rho$= 0.03, $Z$ = 0.16, *p-value* = 0.87).

## Limitations

The small sample size (*n* varied between 26 and 28) and non-response bias are the greatest limitations of these analyses. Stating the researchers' intention to deposit the data publicly may have discouraged more from participating, but perhaps many repositories chose not to respond which were similar to each other but different from those that did participate. Promoting the survey on several relevant email lists was intended to procure responses from a wide variety of repositories, and the outliers in the data seem to indicate this was successful. Possible analyses were limited to two-way tests, however; results of regressions would have been questionable with such a small dataset but in a larger study could control for the effect of multiple factors at once.

There is also a time-variant issue: the current study was conducted only at one point in time. If the study were administered year after year, changes in the number of depositors could be better related to changing repository characteristics.

The current study was also limited by what is generally known about the community of data repositories broadly. Since no complete directory of data repositories exists, targeted outreach and calculation of a response rate are impossible.

## Discussion

Data were so imbalanced for hypotheses three, five, and six as to make statistical analysis inappropriate—for a valid analysis, the smaller group should be big enough to be plausibly representative. These results still offer interesting insights, though. For hypothesis three regarding faculty referral, 25 of 28 responding repositories received referrals from faculty. The result is notable because so many repositories are doing well in this aspect of service. The importance of building trust with and offering good service to customers, as well as the importance of early adopters, make good service a critical factor in the success of repositories. Commercial retail research shows the importance of word-of-mouth advertising. BusinessWire (2011) quoted an American Express survey:

> Consumers will tell others about their customer service experiences, both good and bad, with the bad news reaching more ears. Americans say they tell an average of nine people about good experiences, and nearly twice as many (16 people) about poor ones – making every individual service interaction important for businesses. (p. 2)

Rogers' (2003) classic research into the spread of new ideas identified early adopters as key to persuading many more people to adopt a given technology. This more recent edition notes that, although the Internet has increased the speed of adoption of new technologies exponentially, early adopters are still key to the process. Taken together, these studies suggest that ensuring early repository users have a good experience to share with colleagues is important.

The insights generated by hypotheses five and six are similar: out of 26 respondents, only two linked ingest to electronic lab notebooks (hypothesis five), and of 28 respondents, only four offered encryption (hypothesis six). Both services require extensive resources but may be services for which demand will grow in the future. MacDonald and MacNeil (2015) point out the efficiencies to be gained by linking ELNs to repository ingest. Describing researchers' reaction to how this streamlined the (mandated) process of depositing data, they noted:

> When an initial, limited trial of [the new ELN] was rolled out to ten labs … researchers from no less than nine of the labs reported that it was the ability to use [the ELN] in conjunction with the … repository that was of most benefit. (p. 170)

Respondents to a survey by Lagzian, et al. (2015) ranked fourth most important out of 46 factors the statement, 'The IR is intuitive and easy to use.' Making data deposit easier and seamless with systems researchers already use (or that can otherwise save them time) may be key to boosting repositories' success. In addition, offering encryption could increase deposits by substantially broadening the types of data a repository could accept, particularly in the U.S. in light of the data management requirement the National Institutes of Health will be implementing in January 2023. This mandate may result in a substantial increase in the number of research studies to be archived which contain personally identifiable information, health data, and/or student data, all of which require a high degree of protection.

Other results of the current study are unsurprising, especially since this analysis offers no indication of directionality: it is perhaps predictable that repositories with the largest number of staff have the highest deposit rates as hypothesis 1 supposed. It may well be that having a larger staff means one or more of those employees is able to focus on promotion and outreach, so that the large staff directly results in more deposits. Likewise, offering higher levels of curation service (hypothesis 2) and having staff dedicated to data deposits (hypothesis 7) may indicate a larger staff that can, again, afford to devote time to those services and/or devote a person to promoting the repository, perhaps directly resulting in a larger number of deposits. On the other hand, this study assumed that a larger staff would be an indicator of a repository's success. However, some researchers suggest a small repository staff working either with a supportive data community or backed by a larger team of reference/subject librarians can also be successful (Akers & Green, 2014; Bailey, 2005; Bell, Foster et al., 2005; Springer & Cooper, 2020; Ruediger et al., 2022). Finally, the analysis not supporting hypothesis eight was unsurprising, that the age of a repository would correlate with more depositors, confirming Tillman's (2017) findings that older repositories do not automatically have more depositors, although that might seem counter-intuitive on the surface.

The result of the final hypothesis, number four, relating use of social media to more depositors, was perhaps not unexpected, but the fact that only half of respondents reported using social media for outreach was surprising given its direct access to patrons and economy over print marketing. Perhaps Boulton's (2020) observation of repositories' widespread use of social media to promote their contents rather than their services explains this. Repository staff may simply be more likely to promote their depositors than themselves. Nevertheless, although Chugh, Grose and Macht (2021) found that not all academics use social media, of those who do, a major purpose is for communication. They point out that O'Keeffe (2019) documented 'academics' perception that Twitter is a useful tool to assist with

informal academic development and learning, *in particular learning about academic knowledge and practices*' (p. 990 [emphasis added]). Referring to the previously mentioned importance of word-of-mouth advertising, it follows that repositories can benefit from having a presence in channels faculty are using for related purposes.

## Conclusion

The early twenty-first century has been a time of consciousness-raising with researchers about the importance of data management and data re-use, characterized by the proliferation of IRs in particular but also repositories more broadly. The 'if-you-build-it-they-will-come' approach was unsuccessful, and many repositories reconsidered their approach. Those that have survived and those that accept data deposits need to manage resources carefully and find the most efficient platforms, features and services to entice data deposit. This study extends the work of projects like Hudson-Vitale et al.'s Data Curation SPEC Kit (2017), to help repository staff understand which factors are associated with greater numbers of depositors.

Future researchers may want to explore further the circumstances in which repositories with smaller staff sizes are successful, and specifically the role of librarians as partners in the repository and the part they play in the success of repositories. They may also want to consider whether the small number of respondents in this study may have been a result of the authors' stated intention to deposit the data openly. Responses to the penultimate survey question about comfort with various models of sharing the data (i.e., with different audiences) indicated discomfort even among some who chose to participate, and the number willing to share the data did not vary much regardless of the audience with whom the data were to be shared. Table 2 below shows the distribution of answers. (Also, the last item in Appendix E, the detailed statistical test results, provides more detail about the pattern of respondents' answers.)

*Table 2. Answer Patterns for Opinions on Sharing Data*

| | No | Maybe | Yes |
|---|---|---|---|
| Open to All | 14 | 9 | 5 |
| Open Only to Researchers | 11 | 9 | 7 |
| Open Only to Data Curation Researchers | 12 | 10 | 5 |

The final question of the survey elicited reasons for the respondents' opinions on sharing data. Many indicated discomfort, either their own or that of their institutional leaders, with disclosing detailed information due to uncertainty about their own authority to release information publicly or (in particular) concern about making budget information public.

Overall recommendations from this study are:

- Further research should explore whether a stated intent to deposit (even de-identified) data deters participants; how repositories with small staff sizes achieve success; and what role reference/subject librarians play in making repositories successful.

- More repositories may want to connect with audiences through social media (while this study did not determine directionality, future research could also test this);
- Repositories may want to explore offering more advanced curation services such as checking code, linking to ELNs (or other university systems such as those that track grants), and/or offering encryption; and
- Repositories will want to continue to develop good relationships with researchers.

Finally, more repositories may want to evaluate whether their original structures and services are in fact meeting their audiences' needs and publish those evaluations. Only with more publicly available data will the repository community be able to benefit from past experience and more efficiently and effectively target services to their researchers.

## References


Abrams, S., Cruse, P., Strasser, C., Willet, P., Boushey, G., Kochi, J., Laurance, M., Rizk-Jackson, A. (2014). DataShare: Empowering Researcher Data Curation. International Journal of Digital Curation, 9(1), 110–118. https://doi.org/10.2218/ijdc.v9i1.305

Akers, K.G. and Green, J.A. (2014). Towards a Symbiotic Relationship Between Academic Libraries and Disciplinary Data Repositories: A Dryad and University of Michigan Case Study. International Journal of Digital Curation (9)1, 119–131. https://doi.org/10.2218/ijdc.v9i1.306

Association of American Universities-Association of Public & Land-Grant Universities Public Access Working Group. (2017, November 29). AAU-APLU Public Access Working Group Report and Recommendations. Washington, D.C. https://www.aau.edu/sites/default/files/AAU-Files/Key-Issues/Intellectual-Property/Public-Open-Access/AAU-APLU-Public-Access-Working-Group-Report.pdf

Bailey, C. W. (2005). The role of reference librarians in institutional repositories. Reference Services Review, 33(3), 259-267. https://doi.org/10.1108/00907320510611294

Bell, S., Foster, N.F., & Gibbons, S. (2005). Reference librarians and the success of institutional repositories. Reference Services Review, 33(3), 283-290. https://doi.org/10.1108/00907320510611311

Borgman, C. L., Golshan, M. S., Sands, A. E., Wallis, J. C., Cummings, R. L., Darch, P., & Randies, B. M. (2016). Data Management in the Long Tail: Science, Software, and Service. International Journal of Digital Curation, 11(1), 128–149. https://doi.org/10.2218/ijdc.v11i1.428

Boulton, S. (2020). Social engagement and institutional repositories: A case study. Insights, 33, 1-9. https://doi.org/10.1629/uksg.504

Bryant, R., Lavoie, B. and Malpas, C. (2017). Part 2: Scoping the University RDM Service Bundle. In The Realities of Research Data Management. Dublin, OH: OCLC Research. https://doi.org/10.25333/C3Z039

Bryant, R., Lavoie, B., & Malpas, C. (2018). Incentives for Building University RDM Services. The Realities of Research Data Management, Part 3. Dublin, OH: OCLC Research. https://doi.org/10.25333/C3S62F

Bryant, R. and Faniel, I.M. Works in Progress Webinar: Identifying and Acting on Incentives when Planning RDM Services. OCLC Research webinar, November 13, 2018, https://www.oclc.org/research/events/2018/111318-incentives-when-planning-rdm-services.html

Burris, B. (2009). Institutional Repositories and Faculty Participation: Encouraging Deposits by Advancing Personal Goals. Public Services Quarterly, 5(1), 69-79. https://doi.org/10.1080/15228950802634212

Burton, A., and Treloar, A. (2009). Designing for Discovery and Re-Use: the 'ANDS Data Sharing Verbs' Approach to Service Decomposition. International Journal of Digital Curation, 4(3), 44–56. https://doi.org/10.2218/ijdc.v4i3.124

BusinessWire. (May 3, 2011). Good Service is Good Business: American Consumers Willing to Spend More With Companies That Get Service Right, According to American Express Survey. https://www.businesswire.com/news/home/20110503005753/en/Good-Service-is-Good-Business-American-Consumers-Willing-to-Spend-More-With-Companies-That-Get-Service-Right-According-to-American-Express-Survey

CHORUS. (2019). How University of Denver Librarians used CHORUS Institution Dashboards in conjunction with their own internal data to help monitor public accessibility to the University's publicly funded research. [Brochure]. Author. https://www.chorusaccess.org/wp-content/uploads/UD-Success-Story-Final-011518-3.pdf

Chugh, R., Grose, R. & Macht, S.A. (2021). Social media usage by higher education academics: A scoping review of the literature. Education and Information Technologies 26(1) 983–999. https://doi.org/10.1007/s10639-020-10288-z

Data Curation Network (2018). Checklist of CURATED Steps. https://docs.google.com/document/d/1RWt2obXOOeJRRFmVo9VAkl4h41cL33Zm5YYny3hbPZ8/edit#heading=h.ir351ea2236s

Davis, P., and Connolly, M. (2007). Evaluating the Reasons for Non-use of Cornell University's Installation of DSpace. D-Lib Magazine, 13(3/4). http://www.dlib.org/dlib/march07/davis/03davis.html

Dunning, A. (2017, December 22). PhD Policies for Research Data in Netherlands and UK. [PowerPoint presentation]. Open Working web site. https://openworking.wordpress.com/2017/12/22/phd-policies-for-research-data-in-netherlands-and-uk/

Faniel, I.M., & Connaway, L.S. (2018). Librarians' Perspectives on the Factors Influencing Research Data Management Programs. College & Research Libraries, 79(1), 100-119. https://doi.org/10.5860/crl.79.1.100

Fecher, B., Friesike, S., & Hebing, M. (2015). What Drives Academic Data Sharing? PLOS ONE, 10(2), e0118053. https://doi.org/10.1371/journal.pone.0118053

Ferguson, L. (2014, November 3). How and Why Researchers Share Data (and Why They don't). https://www.wiley.com/network/researchers/licensing-and-open-access/how-and-why-researchers-share-data-and-why-they-dont

Gardner, D., W.Toga, A., Ascoli, G. A., Beatty, Jackson T., Brinkley, J. F., Dale, A. M., … Wong, S. T. C. (2003). Towards Effective and Rewarding Data Sharing. Neuroinformatics, 1(3), 289–296. https://doi.org/10.1385/NI:1:3:289

Hayslett, M. & Jansen, M. (2022). Factors contributing to repository success in recruiting data deposits (de-identified Repository Survey dataset in comma separated format), [Computer File]. https://doi.org/10.17615/a4hj-4w45

Hayslett, M. & Jansen, M. (2022). Factors contributing to repository success in recruiting data deposits (Repository Survey R script and markdown files), [Computer Files]. https://doi.org/10.17615/s6ps-qx27

Helbig, K. (2019). Data Policies: Institutionelle Policies. https://www.forschungsdaten.org/index.php/Data_Policies#Institutionelle_Policies

Hudson-Vitale, C., Imker, H., Johnston, L., Carlson, J., Kozlowski, W., Olendorf, R., & Stewart, C. Association of Research Libraries,. (2017). SPEC kit 354: Data curation.

Horton, Laurence and Data Curation Centre. (2016). Overview of UK Institution RDM Policies. http://www.dcc.ac.uk/resources/policy-and-legal/institutional-data-policies

Jaradeh, M.Y., Auer, S., Prinz, M., Kovtun, V., Kismihók, G., & Stocker, M. (2019, April 29). Open Research Knowledge Graph: Towards Machine Actionability in Scholarly Communication. https://arxiv.org/pdf/1901.10816.pdf

Jefferson, T. (1825, May 8). Letter to Henry Lee. http://www.nlnrac.org/american/declaration-of-independence/primary-source-documents/jefferson-to-lee

Lafferty-Hess, S., Rudder, J., Downey, M., Ivey, S., & Darragh, J. (2018, May 30). Conceptualizing Data Curation Activities Within Two Academic Libraries. LIS Scholarship Archive Works. https://doi.org/10.31229/osf.io/zj5pq

Lagzian, F., Abrizah, A., & Wee, M. C. (2015). Critical success factors for institutional repositories implementation. The Electronic Library, 33(2), 196–209. https://doi.org/10.1108/EL-04-2013-0058

Lowenberg, D. (2017, December 18). Where's the adoption? Shifting the Focus of Data Publishing in 2018 [Blog post]. https://medium.com/@UC3CDL/wheres-the-adoption-shifting-the-focus-of-data-publishing-in-2018-8506f80371cd

Lynch, C.A. (2003, February). Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. ARL: A Bimonthly Report. http://old.arl.org/resources/pubs/br/br226/br226ir.shtml

MacDonald, S. and MacNeil, R. (2015). Service Integration to Enhance Research Data Management: RSpace Electronic Laboratory Notebook Case Study. International Journal of Digital Curation, 10(1), 163-172. https://doi.org/10.2218/ijdc.v10i1.354

McHugh, A., Ross, S., Innocenti, P., Ruusalepp, R., and Hofman, H. Bringing Self-assessment Home: Repository Profiling and Key Lines of Enquiry within DRAMBORA. International Journal of Digital Curation, 3(2), 130-142. https://doi.org/10.2218/ijdc.v3i2.64

Mattern, E., Jeng, W., He, D., Lyon, L., & Brenner, A. (2015). Using participatory design and visual narrative inquiry to investigate researchers' data challenges and recommendations for library research data services. Program, 49(4), 408–423. https://doi.org/10.1108/PROG-01-2015-0012

Mayo, C., Vision, T. J., & Hull, E. A. (2016). The location of the citation: changing practices in how publications cite original data in the Dryad Digital Repository. International Journal of Digital Curation, 11(1), 150-155. https://doi.org/10.2218/ijdc.v11i1.400

Meishar-Tal, H., & Pieterse, E. (2017). Why do academics use academic social networking sites? The International Review of Research in Open and Distributed Learning, 18(1). https://doi.org/10.19173/irrodl.v18i1.2643

Minor, D., Critchlow, M., Hutt, A., Fleming, D., Bergstrom, M. L., & Sutton, D. (2014). Research Data Curation Pilots: Lessons Learned. International Journal of Digital Curation, 9(1), 220–230. https://doi.org/10.2218/ijdc.v9i1.313

National Research Council. (1985). Sharing Research Data. Washington, D.C.: National Academies Press.

O'Keeffe, M. (2019). Academic twitter and professional learning: Myths and realities. International Journal for Academic Development, 24(1), 35–46. https://doi.org/10.1080/1360144X.2018.1520109

Otto, J.J., (2016). A Resonant Message: Aligning Scholar Values and Open Access Objectives in OA Policy Outreach to Faculty and Graduate Students. *Journal of Librarianship and Scholarly Communication*. 4, p. eP2152. http://doi.org/10.7710/2162-3309.2152

Paul, S. (2012). Institutional Repositories: Benefits and incentives. International Information and Library Review, 44(4), 194–201. https://doi.org/10.1080/10572317.2012.10762932

Peer, L. and Green, A. (2012). Building an Open Data Repository for a Specialized Research Community: Process, Challenges and Lessons. International Journal of Digital Curation, 7(1), 151–162. http://dx.doi.org/10.2218/ijdc.v7i1.222

Pienta, A. M., Alter, G. C., & Lyle, J. A. (2010). The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data. In The Organisation, Economics and Policy of Scientific Research. Torino, Italy. https://doi.org/https://deepblue.lib.umich.edu/handle/2027.42/78307

Plale, B., McDonald, R. H., Chandrasekar, K., Kouper, I., Konkiel, S., Hedstrom, M. L., … Kumar, P. (2013). SEAD Virtual Archive: Building a Federation of Institutional Repositories for Long-Term Data Preservation in Sustainability Science. International Journal of Digital Curation, 8(2), 172–180. https://doi.org/10.2218/ijdc.v8i2.281

Rogers, E. M. (2003). Diffusion of Innovations. (5th ed.). New York: Free Press.

Ruediger, D., MacDougall, R., Cooper, D., Carlson, J., Herndon, J., & Johnston, L. (2022, August 9). Leveraging Data Communities to Advance Open Science: Findings from an Incubation Workshop Series. https://doi.org/10.18665/sr.317145

Salo, D. (2008). Innkeeper at the Roach Motel. Library Trends; Fall Library & Information Science Abstracts, 57(2), 98-123. https://muse.jhu.edu/article/262026/pdf

Savage, C. J., & Vickers, A. J. (2009). Empirical Study of Data Sharing by Authors Publishing in PLoS Journals. PLoS ONE, 4(9), e7078. https://doi.org/10.1371/journal.pone.0007078

Springer, R. and Cooper, D. (2020). Data Communities: Empowering Researcher-Driven Data Sharing in the Sciences. International Journal of Digital Curation, 15(1). http://dx.doi.org/10.2218/ijdc.v15i1.695

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., … Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. PLoS ONE, 6(6), e21101. https://doi.org/10.1371/journal.pone.0021101

Tillman, R.K. (2017). Where Are We Now? Survey on Rates of Faculty Self-Deposit in Institutional Repositories. Journal of Librarianship and Scholarly Communication, 5 (General Issue), eP2203. https://doi.org/10.7710/2162-3309.2203

TU Delft (2018, June 26). TU Delft Research Data Framework Policy. https://zenodo.org/record/2573160/files/TU%20Delft%20Research%20Data%20Framework%20Policy.pdf?download=1

University of Groningen (2013, September 1). University of Groningen PhD Regulations. Section 4.1.5. https://www.rug.nl/about-us/organization/rules-and-regulations/onderzoek/promotiereglement-14-en.pdf

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. PLoS ONE, 8(7), e67332. https://doi.org/10.1371/journal.pone.0067332

## Appendix A: Email Solicitation Message

A Word version of the invitation to participate is available in the Carolina Digital Repository at https://doi.org/10.17615/bj43-pw90.

## Appendix B: Consent Form

A Word version of the consent form is available in the Carolina Digital Repository at https://doi.org/10.17615/q0yy-8h57.

## Appendix C: Survey Instrument

A Word version of the survey instrument is available in the Carolina Digital Repository at https://doi.org/10.17615/7vn5-1g28.

## Appendix D: Code Used in Data Analysis

The code with which data analysis was performed is available in the Carolina Digital Repository at https://doi.org/10.17615/s6ps-qx27 and on GitHub at https://github.com/UNC-Libraries-data/repo-survey.

- The R script with which data analyses were performed is create-analysis-data.R, and
- The markdown file with which tables and diagrams were created is Analysis.rmd.

## Appendix E: Detailed Results of Statistical Tests by Hypothesis

*Note: Hypotheses marked with an asterisk (*) were not tested. Tables of these data are presented to show their imbalance.*

### Hypothesis 1: Repositories with the highest staffing will have a significant positive correlation with larger numbers of depositors. (*n*=28)

Asymptotic Spearman Correlation Test
*Z* = 2.2466, *p-value* = 0.02467
*alternative hypothesis: true rho is not equal to 0*
*ρ* = 0.4323595

| Staff | *n* | Median Depositors |
|---|---|---|
| 0.2 | 1 | 1.0 |
| 0.3 | 1 | 4.0 |
| 0.5 | 1 | 0.0 |
| 1.0 | 6 | 20.0 |
| 1.2 | 2 | 6.5 |
| 1.5 | 3 | 2.0 |
| 1.7 | 1 | 1.0 |
| 2.0 | 2 | 4.5 |
| 3.0 | 1 | 20.0 |
| 3.5 | 1 | 468.0 |
| 4.0 | 2 | 30.5 |
| 5.0 | 1 | 0.0 |
| 6.0 | 1 | 25.0 |
| 9.0 | 1 | 22.0 |
| 12.0 | 1 | 25.0 |
| 17.0 | 1 | 45.0 |
| 25.0 | 1 | 10.0 |
| 110.0 | 1 | 360.0 |

### Hypothesis 2. Larger numbers of depositors will be significantly correlated with offering advanced curation services. (*n*=26)

| Offer Advanced Curation Services | Count | Median Depositors |
|---|---|---|
| No | 10 | 4.0 |
| Yes | 16 | 16.5 |
| NA | 2 | 16.5 |

Asymptotic Wilcoxon-Mann-Whitney Test
*Z* = -2.6309, *p-value* = 0.008516
*alternative hypothesis: true mu is not equal to 0*

**Hypothesis 3.* Repositories with larger numbers of depositors will be those which have depositors referred by faculty vs referrals from many places. (*n*=28)**

| Receive Referrals from Faculty | Count |
|---|---|
| No | 3 |
| Yes | 25 |

Statistical testing is not supported by results with this level of category imbalance but the fact that so many repositories had referrals from faculty is notable.

**Hypothesis 4. Larger numbers of depositors will be significantly correlated with using social media to promote the repository. (*n*=28)**

| Use Social Media | Count |
|---|---|
| No | 15 |
| Yes | 13 |

Asymptotic Wilcoxon-Mann-Whitney Test
*Z* = -2.4026, *p-value* = 0.01628
*alternative hypothesis: true mu is not equal to 0*

**Hypothesis 5.* Larger numbers of depositors will not be significantly positively correlated with infrastructure except where repository ingest is linked to electronic lab notebooks (ELNs). (*n*=28)**

| ELN Linked | Count |
|---|---|
| No | 26 |
| Yes | 2 |

Statistical testing is not supported by results with this level of category imbalance but the fact that so few repositories linked with ELNs is notable.

**Hypothesis 6.* Larger numbers of depositors will be significantly positively correlated with repositories that encrypt data. (*n*=27)**

| Offer Encryption | Count |
|---|---|
| No | 23 |
| Yes | 4 |
| NA | 1 |

Statistical testing is not supported by results with this level of category imbalance but the fact that so few repositories offered encryption is notable.

**Hypothesis 7. Larger numbers of depositors will be significantly positively correlated with having dedicated staff for data deposits. (*n*=27)**

| Have Data Staff | Count | Median Depositors |
|---|---|---|
| False | 4 | 0 |
| True | 23 | 11 |
| NA | 1 | 468 |

Asymptotic Spearman Correlation Test
*Z* = 3.6407, *p-value* = 0.0002719
*alternative hypothesis: true rho is not equal to 0*
*ρ* = 0.71

**Hypothesis 8. Larger numbers of depositors will be significantly positively correlated with older repositories. (*n*=28)**

| Age in Yrs | Count | Median Depositors |
|---|---|---|
| 1 | 1 | 33 |
| 2 | 4 | 24.5 |
| 3 | 4 | 1 |
| 4 | 4 | 13 |
| 5 | 3 | 8 |
| 6 | 3 | 7 |
| 7 | 3 | 20 |
| 8 | 1 | 0 |
| 10 | 1 | 0 |
| 12 | 1 | 3 |
| 30 | 1 | 10 |
| 57 | 1 | 360 |

Asymptotic Wilcoxon-Mann-Whitney Test
*Z* = 0.16139, *p-value* = 0.8718
*alternative hypothesis: true rho is not equal to 0*
*ρ* = 0.03

### Opinions about Sharing Data, Detailed Answer Patterns

The table below displays how the 29 respondents answered on their feelings about sharing with different audiences.

| All who answered in each pattern | Open to Anyone | Open Only to Researchers | Open Only to Data Curation/Repository Researchers | Total Answers |
|---|---|---|---|---|
| | Maybe | Maybe | Maybe | 7 |
| | Maybe | Maybe | Yes | 2 |
| | No | Maybe | Maybe | 1 |
| | No | No | Maybe | 1 |
| | No | No | No | 11 |
| | No | Yes | Yes | 1 |
| | Yes | Yes | Yes | 4 |
| | Yes | (blank) | (blank) | 1 |
| | (blank) | (blank) | (blank) | 1 |

## Endnotes

[1] Michele Hayslett is the Librarian for Numeric Data Services and Data Management and Matthew Jansen is the Data Analysis Librarian in the University Libraries at the University of North Carolina at Chapel Hill. Questions may be sent by email to the lead author: michele_hayslett@unc.edu.