Open Science ETDs and Institutional Repositories: Making Research Data FAIRer

Abstract. Graduate students, as potential future full-time researchers, should show proficiency in data sharing because it provides credibility, increases impact and prepares students for grant writing. We compared the FAIRness of non-traditional research outputs (supplement materials) associated with theses and dissertations shared by individual students, with those shared through an institutional repository. Those shared in an institutional repository were significantly FAIRer and had higher views per month. We conclude that graduate students as a population are not yet proficient in applying the FAIR principles. And that they would measuarbly benefit from the review process that is part of most institutional repositories.

1 Introduction

Graduate students are at the beginning of their research careers and benefit from the practice of data sharing because it provides credibility, increases impact, and prepares students for grant writing [1]. While the concept of research data sharing, and increasingly FAIR data [2], are now widespread, putting the concepts into practice is still a difficult task for graduate students [3]. The last decade has seen initiatives to support graduate students in data sharing [4-6]. We wondered how well graduate students are adopting best practices when sharing data and non-traditional research outputs (NTROs) related to their electronic theses or dissertations (ETDs). We set out to do this by comparing the FAIRness of objects shared by graduate students directly and those likely shared with the help of research data management professionals or librarians through an institutional repository.

The Figshare repository platform uniquely provides the opportunity to study this comparison. Students can share outputs for free on https://figshare.com and research institutions can use the Figshare platform as an institutional or data repository and provide curation and review checks [7]. An open API enables the download of metadata across all Figshare repositories making it relatively easy to harvest information at scale.

2 Methods

We assessed metadata for two types of digital objects on the Figshare platform: records which hold zero to many files with accompanying metadata and Collections which aggregate records under unifying metadata. Data collection involved two metadata harvesting runs using the Figshare API [8] each searching all metadata fields for “thesis OR dissertation.” One run collected up to 1,000 metadata records each for datasets, figures, and media, and the other run collected metadata from as many Figshare Collections as possible. We also collected the number of views for each record using the statistics API endpoint.

We programmatically and manually checked the sample to include only records published from academic repositories or from individual researchers. We manually removed records and Collections that were not directly related to an ETD. We assumed that records and Collections from figshare.com are very likely published directly by graduate students (rather than mediated by a library professional) and that those from an institutional repository went through some level of curation or metadata enhancement.

We evaluated records against components in three of the FAIR principles [2]: Findable: Data are described with rich metadata; Interoperable: (Meta)data include qualified references to other (meta)data; Reusable: (Meta)data are richly described with a plurality of accurate and relevant attributes.

For each Collection we looked for the related ETD and documented what types of records were shared in the Collection. We also specifically looked for links from the records to the Collection.

3 Results

We collected a total of 2,606 records and 9,000 Collections through the Figshare API (Table 1). Cleaning the sample left 710 records and 46 Collections. 281 records represented 33 institutional repositories and 27 Collections represented 16 institutional repositories. The remainder are from figshare.com.

Table 1. Record and Collection sample information.

Object	Date Collected	Initial search results	Final sample set	Repositories represented
Record	9 Sept 2021	2,606	710	33
Collection	3 June 2021	9,000	46	16

3.1 Record results

Records from institutions are FAIRer with significantly longer titles (Mann–Whitney U = 28864.5, P < 0. 0.008), significantly longer descriptions (Mann–Whitney U = 36956.0, P < 0.008), significantly more references (Mann–Whitney U = 42776.0, P < 0. 0.008), and significantly more keywords (Mann–Whitney U = 43869.5, P < 0.008). The number of categories was not significantly different (Mann–Whitney U = 55393.5, P = 0.019) (Fig. 1). The number of views per month was significantly higher for records in institutional repositories (t(279)=-5.13, P < 0.008). For all tests the sample size of figshare.com records was 429 and institutional records was 281, tests were two tailed, and the significance level was Bonferroni corrected to 0.008 (0.05/6).

Sixty-five percent of records have no references and no hyperlinks in the description and 70% of those records are from figshare.com.

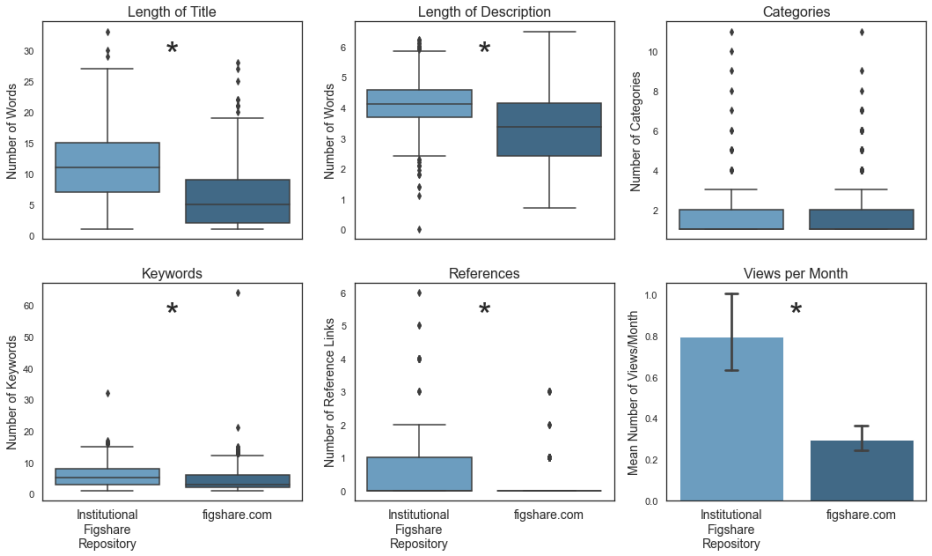

Fig. 1. Comparisons of records from figshare.com and institutional repositories various measures of metadata quality. Non-normal data was compared using Mann-Whitney tests and are displayed as box plots, normal data was compared with a t-test and are displayed in a bar plot. Asterisks indicate significant differences.

Over time, the difference in quality of record metadata between institutional repositories and individual users is widening. As two examples, in recent years records in institutional repositories show longer titles and more references than those shared by individuals on figshare.com (Fig. 2).

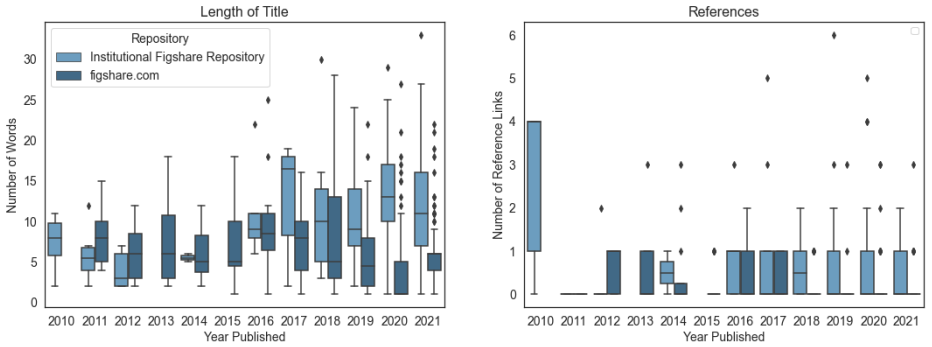

Fig. 2. Quality of metadata over time for figshare.com records and institutional repository records. Records from institutions are increasing in the number of words in the title and the number of references to other digital objects compared to figshare.com records. For references, there are many records with zero links, but starting in 2018 more institutional records have at least one reference link.

3.2 Collection results

Data, code, media, and figures make up most non-traditional research outputs in Collections. A slight majority of Collections (59%) do not contain the ETD and do not reference the ETD or a published paper. Collections shared in an institutional repository did not link to or contain an ETD more often than those shared by individuals (χ^2(df=1, N=46)=0.199, p=.655). Only 13% contained the ETD itself, but about 41% contained or

linked to a document with more information about the research. Nine percent link to a peer reviewed paper. Only nine percent of Collections contain records that link back to the Collection.

4 Discussion

ETD related NTROs shared in repositories with institutional oversight are more findable, interoperable, and reusable than those shared without institutional oversight. We interpret these results as indicating that graduate student mastery of FAIR sharing is limited but that they benefit from the services attached to institutional repositories. Many repositories are managed by librarians and they are likely the main reason for the increased FAIRness of ETD research outputs. Librarians can ensure metadata completion before publication and apply appropriate discovery terms, leading to a higher chance for reuse of a student's work. The higher number of views per month for institutional repository records indicate a measurable benefit to students.

Figshare Collections offer a natural way to group related records of many different file types, addressing some of difficulties [1] and [9] identify. However, linking between records and the Collection was rare. Collections in an institutional repository were just as likely to be missing a link to the ETD as those shared by individuals. We suggest that Figshare offer more help resources for users, especially around linking records and Collections, and institutions include specific metadata to link to the Collection in which a record is included.

References

1. Collie, W. A., & Witt, M. (2011). A Practice and Value Proposal for Doctoral Dissertation Data Curation. *International Journal of Digital Curation*, ***6*(2)**, 165–175. https://doi.org/10.2218/ijdc.v6i2.194
2. Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, ***3*(1)**, 160018. https://doi.org/10.1038/sdata.2016.18
3. Wiley, C. A. (2018). *Managing Research Data: Graduate Student and Postdoctoral Researcher Perspectives*. https://doi.org/10.5062/F4FN14FJ
4. Carlson, J., Fosmire, M., Miller, C. C., & Nelson, M. S. (2011). Determining Data Information Literacy Needs: A Study of Students and Research Faculty. *Portal: Libraries and the Academy*, ***11*(2)**, 629–657. https://doi.org/10.1353/pla.2011.0022
5. Adamick, J., Reznik-Zellen, R., & Sheridan, M. (2012). Data Management Training for Graduate Students at a Large Research University. *Journal of EScience Librarianship*, ***1*(3)**, 180–188. https://doi.org/10.7191/jeslib.2012.1022
6. Nelson, M. S., & Kong, N. N. (2020). Capturing their "first" dataset: A graduate course to walk PhD students through the curation of their dissertation data. *IASSIST Quarterly*, *44*(3), Article 3. https://doi.org/10.29173/iq971
7. *figshare for institutions—A repository for research data of all types*. (n.d.). Retrieved March 30, 2022, from https://knowledge.figshare.com/institutions
8. *Figshare documentation*. (n.d.). Retrieved March 30, 2022, from https://docs.figshare.com/
9. Van Tuyl, S. (2019). What's in the Box? Assessing the potential usability of four decades of thesis and dissertation supplementary files. *Journal of EScience Librarianship*, ***8*(1)**. https://doi.org/10.7191/jeslib.2019.1142

Data availability: To be filled after peer review