Journal of eScience Librarianship

Volume 8 | Issue 1 Article 2

2019-03-28

What’s in the Box? Assessing the potential usability of four decades of thesis and dissertation supplementary files

Steven Van Tuyl
Oregon State University

Recommended Citation

Van Tuyl S. What’s in the Box? Assessing the potential usability of four decades of thesis and dissertation supplementary files. *Journal of eScience Librarianship* 8(1): e1142. https://doi.org/10.7191/jeslib.2019.1142. Retrieved from https://escholarship.umassmed.edu/jeslib/vol8/iss1/2

ISSN 2161-3974

JeSLIB 2019; 8(1): e1142
doi:10.7191/jeslib.2019.1142

Journal of eScience Librarianship
putting the pieces together: theory and practice

Full-Length Paper

What’s in the Box? Assessing the potential usability of four decades of thesis and dissertation supplementary files

Steve Van Tuyl

Oregon State University, Corvallis, OR, USA

Abstract

Objectives: The objective of this study is to evaluate the quality and usability of supplementary data files deposited, between 1971 and 2015, to our university institutional repository. Understanding the extent to which content historically deposited in digital repositories is usable by today’s researchers can help inform digital preservation and documentation practices for researchers today.

Methods: I identified all graduate-level theses and dissertations (GTDs) in the institutional repository with multiple files as a first pass at identifying documents that included supplementary data files. These GTDs were then individually examined, removing supplementary files that were artifacts of either the upload or digitization process. The remaining “true” supplementary files were then individually opened and evaluated following elements of the DATA rubric of Van Tuyl and Whitmire (2016).

Results: Supplementary files were discovered in the repository dating back to 1971 in 116 GTD submissions totaling more than 25,000 files. Most GTD submissions included fewer than 30 files, though some submissions included thousands of individual data files. The most common file types submitted included imagery, tabular data, and databases, with a very large number of unknown file types. Overall, levels of documentation were poor while actionability of datasets was generally middling.

Conclusions: The results presented in this study suggest that legacy data submitted to our institutional repository with GTDs is generally in poor shape with respect to Transparency and somewhat less so for Actionability. It is clear from this study and others that researchers have a long road ahead when it comes to sharing data in a way that makes it potentially useable by other researchers.

Correspondence: Steve.Van Tuyl: steve.vantuyl@oregonstate.edu
Keywords: institutional repository, research data management, data sharing, metadata, data obsolescence

Disclosures: The authors report no conflict of interest.

Introduction

Over the past decade, researchers in academia and government have been increasingly called on to better manage and share the results of their research, including publications and datasets. The 2013 White House Office of Science and Technology Policy memorandum on increasing access to federally funded research (Holdren 2013) has put a fine point on the issue—requiring, at least to some extent, that research data be managed and shared in order to be reused. However, recent evaluations of the implementation of these policies have suggested that, at least in these early days, the quality of shared data is poor or that it is not clear to researchers what they should be doing to maximize the effectiveness of their data sharing (Van Tuyl and Whitmire 2016, Naudet et al. 2018).

A fair amount has been written on the need for best practices in data sharing, suggestions for best practices for data sharing, and how to solve the problem of quality in data sharing. The FAIR data principles (https://www.force11.org/fairprinciples) and the DATA rubric (Van Tuyl and Whitmire 2016) are two recent examples of broad scale best practices for high quality data sharing. There have also been a number of papers written recently discussing domain-specific (though sometimes cross-cutting) recommendations for data sharing including in Ecology (White et al. 2013), Life Sciences (Griffin et al. 2017), and Population Genetics (Leberg and Neigel 1999), among others. While not explicitly offering guidance for best practices, a number of researchers have explored the current behaviors and attitudes of researchers with respect to data management and sharing including a host of largely domain-agnostic institutional surveys (e.g. Van Tuyl and Michalek 2015, Akers and Doty 2013, Whitmire et al. 2016, and Rolando et al. 2013, among many others), and a number of domain-specific studies to the same effect in areas such as Neuroimaging (Borghi and Van Gulick 2018), Structural Engineering (Johnston and Jeffryes 2013), Atmospheric Science (Wiley and Mischo 2016), Biomedical Clinical Research (Federer et al. 2015), and Water Quality (Carlson and Stowell-Bracke 2013).

While some researchers have raised concerns about the quality of data (Merson et al. 2016) and its retrievability (Vines et al. 2014), relatively little has been written to evaluate data quality in data repositories or even what types of data are being deposited. Some notable examples of projects to evaluate quality data sharing have focused primarily on retrievability of the data (Savage and Vickers 2009) or repository contents (Wiley 2015), but relatively little has been written about the quality of the shared data. A notable exception to the dearth of investigation in this area is the work of Naudet et al. (2018), who examined the retrievability of datasets and the reproducibility of the results for two journals with strong data sharing policies, finding that only approximately45% of authors were able to produce their datasets with sufficient information to reproduce their results.

One of the challenges here is that “quality” can be interpreted to be subjective and domain specific while many digital repositories accepting datasets are generalist in nature. General best practices and domain-level experience (as mentioned above) have given data curators and those offering data sharing guidance to researchers fodder for offering best approaches to high quality data sharing, and those same best practices can be used to help evaluate the quality of shared data once it enters a digital repository. Using this guidance, Van Tuyl and Whitmire (2016) built a rubric for coarse evaluation of the quality of shared data, requiring that data be discoverable, accessible, transparent, and actionable (DATA), and found that almost none of the data produced by researchers at their institution, were compliant with this rubric.

In this paper, I look to the past to help understand what the future of data sharing might hold with respect to the transparency (i.e. documentation) and actionability (usability) of research data in the institutional repository of a research university. To do this, I evaluate the contents of supplementary files deposited to our institutional repository, ScholarsArchive@OSU, between 1971 and 2015 for transparency and actionability elements of the Van Tuyl and Whitmire (2016) rubric, characterize the types of problems encountered with these files, and offer guidance for future researchers in the areas of data sharing and data documentation.

At Oregon State University, Graduate Theses and Dissertations (GTDs) are submitted directly to our institutional repository, ScholarsArchive@OSU, where they are reviewed by the thesis coordinator of the Oregon State University (OSU) Graduate School and then by library staff before becoming available in the repository. This workflow has historically not included a separate review process for supplementary files, in part because the OSU Libraries and Press (OSULP) only recently (in the past five years) developed a Research Data Services program and because opportunities for interacting with graduate students depositing GTDs are limited due to the timing of GTD deposit and student graduation. Thus, datasets in our repository are, essentially, uncurated. This gives us a view of data sharing practices over time, across many domains, and can serve as a baseline for quality of data sharing against which to compare data curation activities at our institution, and potentially other institutions.

Methods

Source Data

Over the past five years, OSULP has digitized all of the GTDs produced by OSU graduate students back to 1902, and incorporated them into our institutional repository. During this process, all supplementary files submitted with the physical GTD, usually as contents of a computer disc, were also loaded to our institutional repository. Over the time period of this study (1975-2015), ScholarsArchive@OSU averaged about 400 GTD deposits per year. As a first pass at finding GTDs with supplementary datasets, I extracted all GTDs that included more than one file from the Electronic Thesis and Dissertation collection in our institutional repository—resulting in 2,792 candidate GTDs with supplementary files. I then manually screened these supplementary files to remove artifacts of the digitization and optical character recognition processes (including digitized plates and overlays from historic theses and dissertations), resulting in 116 GTDs with supplementary data files.

For each of the 116 GTDs with supplementary files, I reviewed the supplementary files(s) and recorded counts of file types, the presence or absence of documentation for the supplementary data, and any features of the dataset that created problems for reviewing the data. Compressed directories (e.g. zip, 7zip, gz) were opened and evaluated like other shared files. Where the number of files in a compressed directory exceeded 25 files, the directory contents were evaluated using a free tool for analyzing the contents of disk drives (JDiskReport—http://www.jgoodies.com/freeware/jdiskreport). I also manually reviewed these large file directories for documentation, unusual file types, or to further investigate anomalies or unexpected file types in directories.

For each supplementary dataset, I applied a score to each GTD indicating the level of Transparency and Actionability for the supplementary data using elements of the rubric from Van Tuyl and Whitmire (2016)—reproduced here in Table 1. I did not score these datasets on the other two elements presented by Van Tuyl and Whitmire (2016)—Discovery and Availability. Because all of these datasets are in an open repository (ScholarsArchive@OSU) that is indexed by major search engines and in other scholarly content aggregators (e.g. SHARE) all will have maximum scores for both of these elements.

Table 1: Rubric for Transparency and Actionability of identified supplementary files (from Van Tuyl and Whitmire 2016).

Transparent	
Score	***Journal Article & Project Scoring Criteria***
0	No documentation provided for the data
1	Some documentation provided for the data but lacks clear description of details such as how data were collected, analyzed or processed; description of units or headers; description of blanks; etc. Documentation may include a reference to the methods section of the paper
2	Readme file, data dictionary, or other metadata shared with the dataset that provide clear details about the nature and content of the data
Actionable	
Score	***Journal Article & Project Scoring Criteria***
0	Data are not in a format that is usable in an analysis application (e.g. shared in a PDF or as a figure)
1	Data are in a format usable in an analysis application but are formatted in a way that makes use difficult (e.g. spreadsheets not in regular row-column form). OR data are shared in a non-open format (e.g. .xls, .doc)
2	Data are in an open format (e.g. .csv, .xlsx, etc.) with usable formatting

Characterization of Transparency

Transparency was evaluated by searching each dataset for documentation either within individual files or as separate metadata either attached to the item record in the repository or as a separate documentation file. For large bodies of supplementary data contained in zip or other compressed file formats, I searched top level directories for documents that might contain metadata, data dictionaries, or documentation. This documentation screening was necessarily superficial, given that the author does not have domain-level expertise in every field of research for which data are deposited to our repository, nor is it feasible to open every file to determine the extent to which documentation may be embedded therein. That said, basic screening of documentation, at the very least, provides one with the ability to determine whether some attempt has been made by the data depositor to make re-use of the data easier through documentation, and providing documentation is a standard recommendation for data documentation that is not uncommon among researchers and data librarians (Tenopir et al. 2015, White et al. 2013).

Characterization of Actionability

To characterize the actionability of each dataset, whether a dataset could be readily examined in a modern analytical platform, I took three major evaluative steps. First, each individual file was coded into broad categories to evaluate patterns of supplementary file sharing (e.g. figures, datasets, etc.). Second, for objects for which I was not able to identify a native application or that I was unable to open in the native application, I attempted to open them using a text editor in order to assess whether access to the file was possible outside of the native application and, thus, whether the contents were at least recoverable. Last, for files that were unopenable in a text editor or in the native application, I determined the file format based on file extensions.

Results

As of 2015, ScholarsArchive@OSU held 25,339 graduate dissertations and theses. The earliest supplementary data that could be found came to ScholarsArchive@OSU in 1971 in the form of a physical box of learning materials that were subsequently photographed and included in the repository as part of our mass digitization project for GTDs. Supplementary data submissions came to ScholarsArchive@OSU from 40 academic departments, with only seven academic departments having five or more submissions with supplementary files. Geology had far more submissions at 21, followed by Geography (9), Civil Engineering (8), Water Resource Engineering (7), Botany and Plant Pathology (6), Industrial Engineering (5), and Wood Science (5).

From the time of the first deposit of supplementary materials until 2015, ScholarsArchive@OSU received about 400 GTD deposits per year. Deposit of supplementary data continued at a low rate until the mid-1990s, with only a few deposits per year, if any (Figure 2). Since the mid-1990s, ScholarsArchive@OSU has seen about 4-8 supplementary file deposits per year - a rate that remains to the end of the study period. Across all 116 GTDs with supplementary datasets, there were a total of 25,542 files, including all of the contents of compressed files. Summary statistics for all GTD supplementary datasets are presented in Table 2. The smallest number of supplementary files is 1 while the largest number of files included with one submission was 4641. The median number of supplementary files submitted with GTDs is 21.5 (mean 220.2, standard deviation 708.2), indicating a huge range in the number of files submitted.

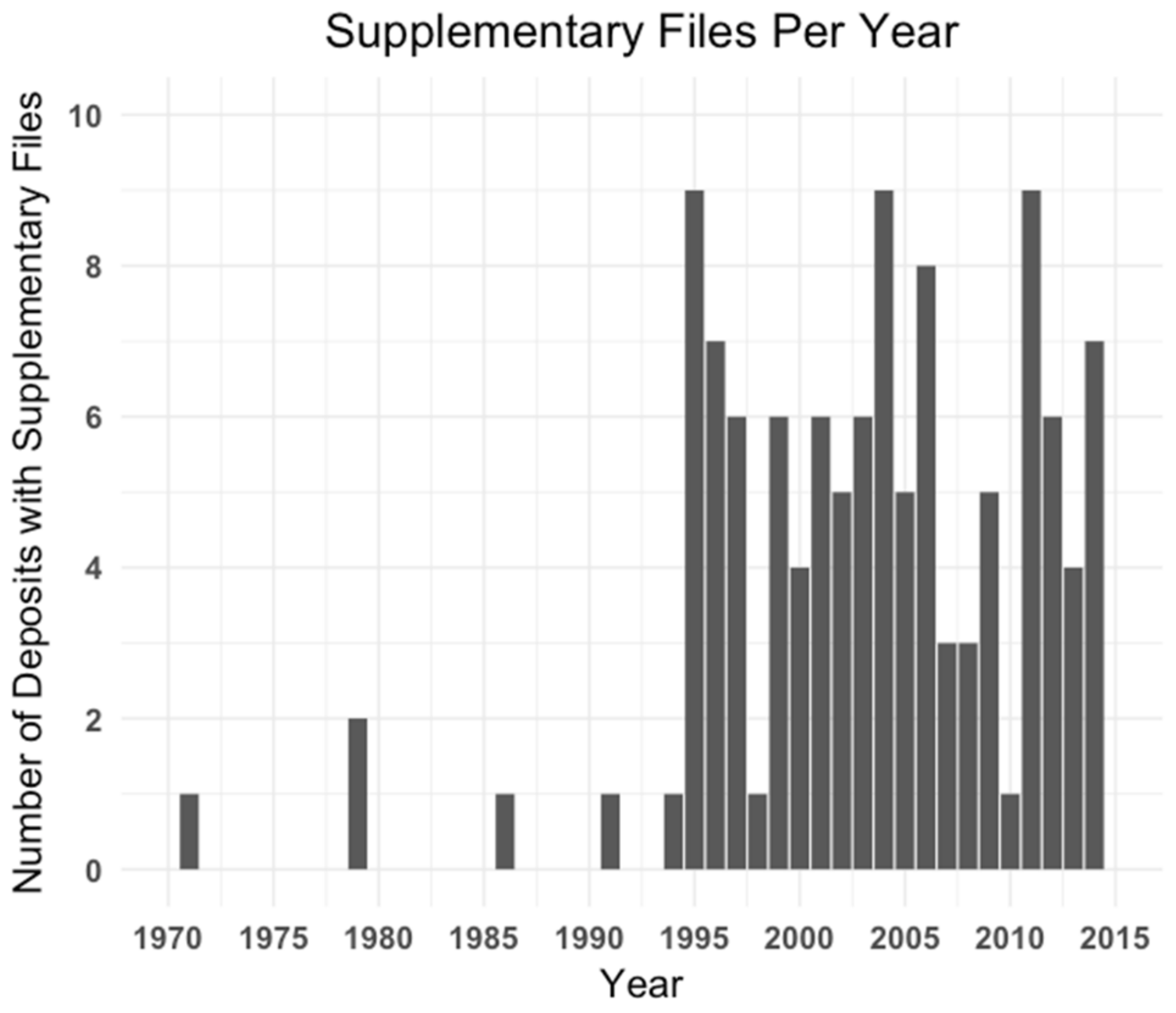


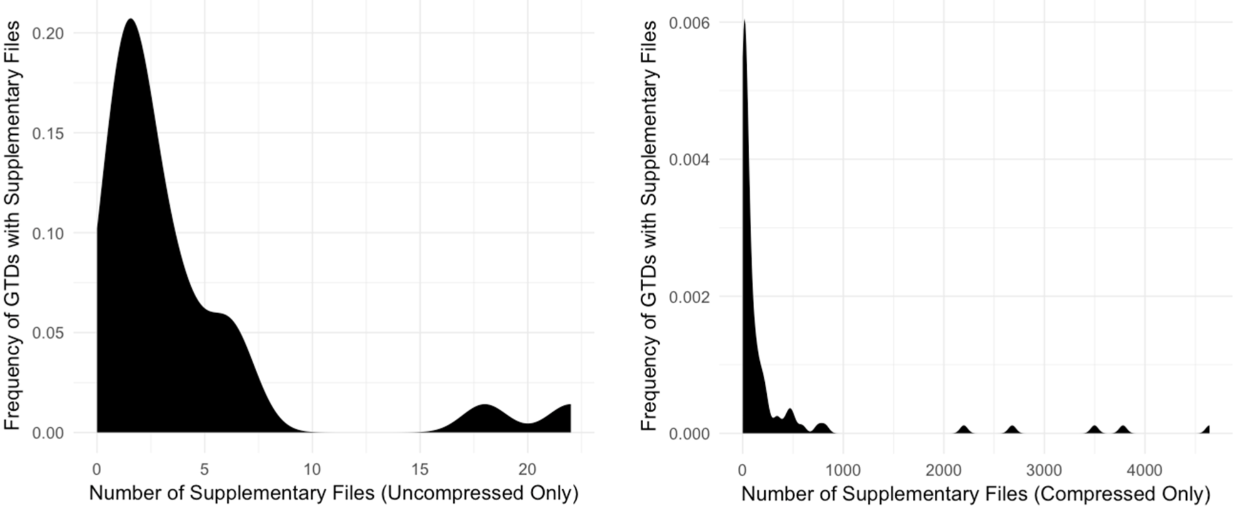
Figure 2: Number of supplementary file deposits with GTDs by year.

Table 2: Summary statistics for counts of supplementary files as well as for compressed and uncompressed filesets from GTDs.

	n	Mean	First Quartile	Median	Third Quartile	Max	Standard Deviation
All GTD Filesets	116	220.2	3.0	21.5	97.3	4641.0	708.2
Uncompressed Filesets Only	27	3.9	1.0	2.0	4.0	22.0	5.0
Compressed Filesets Only	89	285.8	8.0	31.0	147.0	4641.0	798.0

It is clear from the summary data that there are large differences in the number of supplementary datasets submitted with GTDs if there are compressed files included with the submission. 23.2% of submissions included no compressed files, with a median number of files is 2.0 files (mean 3.9, standard deviation 5.0), while 76.8% of submissions included one or more compressed files, with a median of 31.0 files (mean 285.8, standard deviation 798.0).

This difference in submissions including compressed versus those not including compressed files is further illustrated when examining the distributions of the number of files from compressed versus uncompressed supplementary datasets is presented in Figure 1, as is the total number of files in compressed supplementary files alone. For submissions without compressed files, the distribution of number of files peaks at a fairly low number (as indicated by the median number of files as 2.0) and drops to zero at about 10 files. There are a handful of deposits larger than 10 files represented in the dataset, though. For submissions with compressed files, the distribution is similar in shape, but the scale of values for the number of files is orders of magnitude larger. Most submissions with compressed files have relatively few files (as indicated by the median number of 31 files) and the distribution similarly drops to near-zero around 1000 files, with some trailing peaks of submissions larger than 1,000 files and up to the maximum of more than 4,500 files.



Figures 1a and 1b: Frequency distributions of number of supplementary files submitted with GTDs. 1a: distribution of supplementary files for submissions that only included uncompressed files. 1b: distribution of supplementary files for submissions that included compressed files.

The types of files represented in GTD submissions varies widely, when considering granular file types (e.g. file extensions or mime types), which is why grouping the file types into categories is helpful for summarization. Table 3 shows the number of files listed for each file type category.

Table 3: Summary of file types found in supplementary compressed files

File Type	Number of Files
audio/video	103
code	1085
database	1997
documentation	21
spreadsheet	928
executable	65
gis	1007
image	6340
text (data)	4239
text (document)	381
other file types	4659
unknown	2942

The most common known file type shared as supplements to GTDs is images, followed by text files containing data, files of unknown format, and database files of a variety of database formats. Computer code, Microsoft Excel files, varieties of Geographic Information Systems (GIS) data are approximately tied for the fourth most common file categories. The least common file types include audio/video, executables, and documentation. Over 6000 of the files submitted were either categorized as “Other” or “Unknown” filetypes. The “Unknown” file types are likely a mixture of obsolescence of the files, lack of experience on the part of the author, and a lack of information available about the variety of proprietary and custom file types used by researchers. The “Other” filetypes were a mixture of rarer types such as 3D modeling, archived websites, and identifiable proprietary software formats.

Transparency and Actionability

Of the 116 supplementary file submissions, only 10 scored the highest value (2) Transparency, 25 scored middling value (1), and the remaining 80 scored the lowest value (0). This indicates that most of the data submitted does not have sufficient documentation to allow reuse of the data. The distribution of Actionability scores was slightly different with 29 submissions scoring the highest value (2), 67 submissions scoring a middling (1) value, and 19 submissions scoring the lowest value (0; Figure 3). This indicates that it is relatively common for submissions to include data that is in a form that is readable in an analytical software application, though file formats and data organization may present challenges. One of the supplementary files was not included in the Transparency and Actionability analysis due to the unique nature of the contents—physical artifacts submitted with the GTD that were photographed as part of the GTD digitization process.

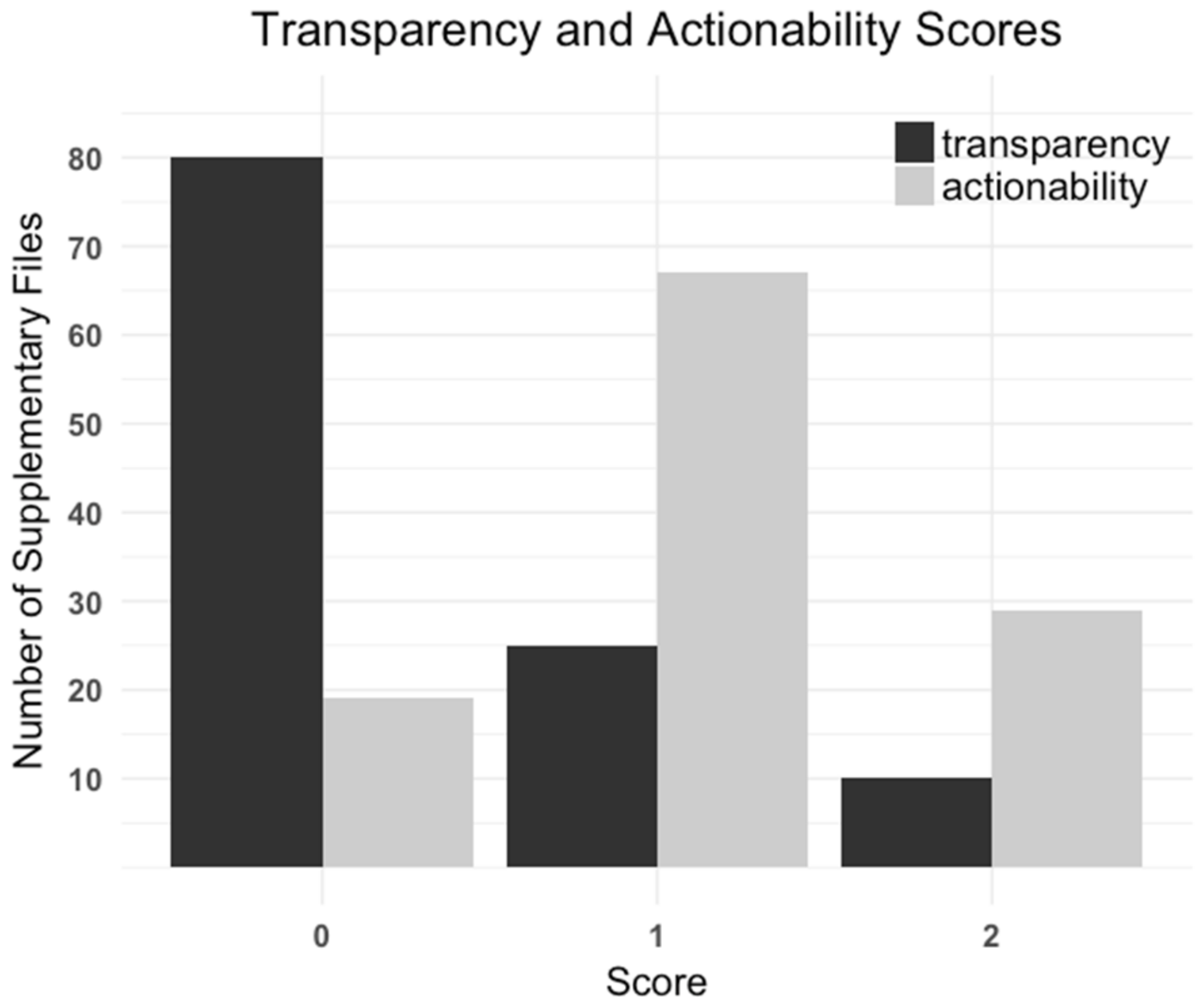
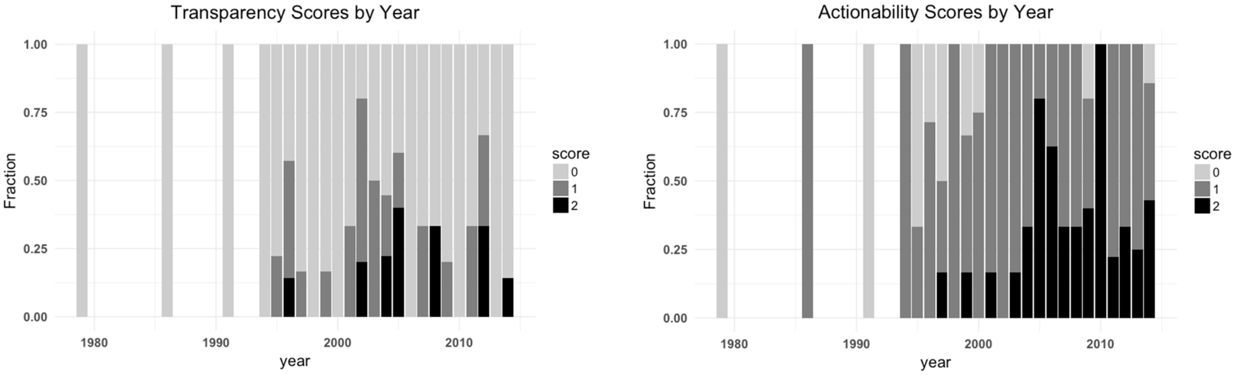


Figure 3: Summary of Transparency and Actionability scores.

Generally speaking, Transparency scores do not appear to degrade or improve over the time period of this study (Figure 4a). Actionability, on the other hand, appears to improve, slightly, through this time period (Figure 4b), though this improvement may have more to do with the modernity of the file types, formats, and software than with any decisions to make data actionable on the part of the depositors.



Figures 4a and 4b: Transparency (4a) and Actionability (4b) scores by year.

Discussion

Supplementary files associated with GTDs in ScholarsArchive@OSU are available as far back as four decades, and the current level of availability of the older datasets in this collection (approximately pre-2006 - the year online submissions for GTDs started; Boock and Kunda 2009) is thanks to the efforts of the digitization unit in the Oregon State University Libraries and Press. This highlights the importance of legacy digitization projects and their ability to make previously inaccessible content available more widely.

It should not be much of a surprise that much of the data associated with GTD submissions does not live up to emerging standards for data sharing in repositories, given the decades of difference between these submissions and our current thinking about data sharing. Mandated practices around research data sharing is only recently being formalized, and even discussed, in some fields of research, and making changes to researcher workflows should be expected to take time. Previous studies have suggested that a variety of domains of research are in the early stages of sorting out how best to manage data in a way that allows it to be useable to others (Akers and Doty 2013, Federer et al. 2015, Johnston and Jeffryes 2013). That said, the underlying principles and reasoning behind sharing data with a scholarly submission (be it to a journal or as a GTD) have persisted beyond recent policy mandates, as have the challenges of finding ways to help researchers comply with mandates and best practices (cf. Wolins 1962, Craig and Reese 1973, Leberg and Neigel 1999). There is value in exploring the quality of shared research materials to help us understand what types of content and challenges researchers face when accessing supplementary data files. Even more, the differences in quality of data submissions we see in the data from this project, over four decades, are similar to what we see in contemporary projects where data sharing expectation have been made more explicit - transparency and actionability of datasets is generally poor (cf. Van Tuyl and Whitmire 2016, Borghi and Van Gulick 2018).

Submission Contents

Admittedly, the number of files contained in these 116 submissions was much higher than we expected. It is clear that with many of the compressed file contents, very little curation of the deposited data was done given the lack of documentation and organization in the file structures. That said, some of the compressed files examined in this study were well organized (if not well documented) making it apparent that large bodies of complex and hierarchical supplementary files are sometimes required. The complexity of these supplementary datasets creates challenges for data depositors (organization and documentation), data curators (evaluation and feedback to depositors), and data users (comprehension and reuse of the data) that we are only beginning to understand how to overcome in many fields of research.

Many research data sharing best practices seem to focus on what is widely reported to be a common filetype—tabular data such as excel files or comma separated text files. While these are common data types in this and other studies (e.g. Van Tuyl and Michalek 2015, Whitmire et al. 2015, Wiley 2015, Parham et al. 2012) it is apparent that research data curators need to consider creating best practices for documenting and making actionable non spreadsheet/ tabular data. Certainly, the resources provided by national and international organizations (e.g. DataOne, FORCE11) as well as the resources provided at the “local” level by research data services workers (e.g. librarians) provide guidance for managing and sharing non-tabular data and large, hierarchical, and complex datasets. However, these efforts tend to focus at cross-domain best practices which are less likely to provide domain, filetype, or data organization specific needs of researchers sharing complex datasets. This is a complex puzzle to solve, but recent moves to aggregate research data services efforts across organizations and expertise are likely to provide some resolution to these problems. Some examples of these cross-cutting efforts include the Data Curation Network project (Johnston et al. 2017) project which seeks to offer distributed research data services expertise or the software citation efforts by Niemeyer et al. (2016).

Transparency

Generally speaking, it is not possible for this author or, really, any reviewer of this type of corpus of dataset to fully comprehend whether a dataset/datafile is documented well enough to be useable by another researcher. The level of documentation in the data presented here is simultaneously disappointing and unsurprising. Given the fact that much of the data presented here was submitted to ScholarsArchive@OSU before requirements of sharing data were far less common, it is inevitable that documentation would be lacking. However, one wonders at the reasoning behind providing research data with the submission of a GTD, other than for the purposes of another researcher being able to view, understand (at least in the context of the GTD), and possibly reuse the data.

Creating simple documentation, even in the form of a text file describing the contents of the dataset, may be enough to at least orient users of the data to its contents, how it was collected, and what concerns another data user might have. For most of the datasets in this study, the creation of this type of simple documentation could be relatively easy, given how few files were in a normal submission. One wonders if efforts to create tooling to facilitate the creation of simple documentation for research datasets might lower the overhead paid by researchers creating such documents.

Actionability

In some respects, scoring well in the category of Actionability for datasets is more challenging than Transparency. It is not uncommon for researchers to create custom data files and filetypes, especially when building novel systems or working in the areas requiring computer simulation. In these specific areas, it makes sense that some data files would not be actionable to an uninitiated user, as they may be meant to be actionable only to the system in question. An example of this is parameter or input files used as part of a larger computer simulation—these files are meant to be read by the computer model, not by a human. This issue does, however, speak to a need for clear documentation about the files in a shared dataset, what they are for, and what the expectations should be for actionability.

Similarly, in many fields of research, proprietary software is the norm, and it is challenging if not impossible to break away from using such tools. This make sharing highly actionable data difficult because of system requirements for reading the data (e.g. if the proprietary system used to generate the data can only read the proprietary file format) or because of restrictions on exporting or file conversion from the proprietary format to a more open format. Paradoxically, some proprietary systems have become such a norm in certain fields of research that they may, in some ways, be considered de facto “open.” While this is not ideal, because the software and data are still in proprietary formats, the fact that these software and data platforms have become the norm in certain fields of research helps ease the difficulty of data sharing - “everyone” is using proprietary system, so there is no need to create more open formats for sharing. Of course, “everyone” isn’t using the system, just those who have access to it, and this can be a common pain point when discussing data sharing and the need to provide open versions of files where possible.

For files that are meant to be human readable, and for which there is no need to use proprietary software, there are really no excuses for failing to provide some level of actionability for datasets. In more than one case in this study, submissions included such unactionable data as screen captures of spreadsheets, tables of data shared in PDF or image formats, and text files with no column headers or other descriptive information.

Drawbacks

The evaluation presented here is, admittedly, cursory—partially due to the coarse nature of the rubric developed by Van Tuyl and Whitmire (2016). However, this application of the rubric is appropriately coarse, given that the author (and others, such as data librarians and repository managers, in similar positions of evaluating submissions entering institutional or generalist repositories) is not a domain expert in all domains of work being submitted to the repository. Thus, it is not possible to tell, from a researcher/expert’s perspective, whether the data here would actually be useable by other researchers, but best practices suggest that, as a baseline, providing even basic documentation for shared datasets is a useful step towards reusability (Tenopir et al. 2015, White et al. 2013)

While a great deal of work could be done to identify and locate copies of native applications, and specifically the originating version of these native applications, the focus of this work is on identifying whether a dataset was readily accessible to a potential user. It is also clear that, given the size of some of the shared datasets, manual evaluation of every shared file is impossible. Computational approaches might be brought to bear in the future to evaluate corruption and obsolescence of deposited datasets at the time of deposit and over the lifespan of the dataset in the repository.

An additional issue with the DATA model, is that it looks at data sharing with a lot of assumptions about what is possible with respect to transparency and actionability (as mentioned above). With the help of domain experts, it may be possible to further refine or specify the DATA model and tailor it to specific types of data or domains of research.

Conclusions

The results presented in this study suggest that legacy data submitted to our institutional repository with GTDs is generally in poor shape with respect to Transparency and somewhat less so for Actionability. It is clear from this study and others that researchers have a long road ahead when it comes to sharing data in a way that makes it potentially useable by other researchers. On its own, this study suggests a lack of preparation of graduate students to share data in a meaningful way, and that there is more work to be done to prepare graduate students to enter the emerging world of expectations around data sharing. There already exist many programs to help graduate students in the area of data sharing (e.g. the ETD Plus Toolkit—https://educopia.org/etdplustoolkit), and many academic libraries in research institutions have hired scholarly communications and research data services librarians to help researchers and students understand data sharing best practices. Having these programs and personnel in place is a great start, but one wonders whether more could be done by the primary influencers on graduate students, the graduate advisor, to effect better data sharing practices.

The ability of this author and others with a background in data curation and research data services to effectively evaluate the quality of shared data is somewhat limited by lack of domain expertise in all areas being evaluated. In fact, that is a task that none could realistically live up to. This author is unaware of any other research investigating domain-level, peer-reviewed datasets in a way that would allow expert-based opinion as to the transparency and actionability of datasets deposited in institutional or other data repositories. It would be an extremely useful direction for future research into the usability of shared datasets, and I encourage others to pursue this route for enhancing our understanding of whether and how useable supplementary data are, and whether we are making improvements in this area, given recent policy initiatives.

Acknowledgements

I would like to thank Amanda Whitmire (Stanford University) and Chris Diaz (Northwestern University) for early and late (respectively) comments on this manuscript, and also the Research Data Access and Preservation (RDAP) community for supporting an early version of this work.

References

Akers, Katherine G., and Jennifer Doty. 2013. “Disciplinary differences in faculty research data management practices and perspectives.” *International Journal of Digital Curation* 8(2): 5-26. https://doi.org/10.2218/ijdc.v8i2.263

Boock, Michael, and Sue Kunda. 2009. “Electronic thesis and dissertation metadata workflow at Oregon State University Libraries.” *Cataloging & Classification Quarterly* 47(3-4): 297-308. https://doi.org/10.1080/01639370902737323

Carlson, Jake, and Marianne Stowell-Bracke. 2013. “Data management and sharing from the perspective of graduate students: An examination of the culture and practice at the water quality field station.” *portal: Libraries and the Academy* 13(4): 343-361. https://doi.org/10.1353/pla.2013.0034

Craig, James R., and Sandra C. Reese. 1973. “Retention of raw data: A problem revisited.” *American Psychologist* 28(8): 723. http://dx.doi.org/10.1037/h0035667

Federer, Lisa M., Ya-Ling Lu, Douglas J. Joubert, Judith Welsh, and Barbara Brandys. 2015. “Biomedical Data Sharing and Reuse: Attitudes and Practices of Clinical and Scientific Research Staff.” *PLoS ONE* 10(6): e0129506. https://doi.org/10.1371/journal.pone.0129506

Griffin Philippa C., Jyoti Khadake, Kate S. LeMay, et al. 2017. “Best practice data life cycle approaches for the life sciences [version 1; referees: 2 approved with reservations].” *F1000Research* 6:1618. https://doi.org/10.12688/f1000research.12344.1

Holdren, John P. 2013. *Increasing access to the results of federally funded scientific research.* Washington, D.C.: Office of Science and Technology Policy. Accessed April 26, 2018. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Johnston, L., and Jeffryes, J. 2013. “Data management skills needed by structural engineering students: Case study at the University of Minnesota.” *Journal of Professional Issues in Engineering Education and Practice* 140(2): 05013002. https://doi.org/10.1061/(ASCE)EI.1943-5541.0000154

Johnston, Lisa R., Jake R. Carlson, Patricia Hswe, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Robert K. Olendorf, and Claire Stewart. 2017. “Data Curation Network: How Do We Compare? A Snapshot of Six Academic Library Institutions’ Data Repository and Curation Services.” *Journal of eScience Librarianship* 6(1): e1102. https://doi.org/10.7191/jeslib.2017.1102

Leberg, Paul L., and Joseph E. Neigel. 1999. “Enhancing the retrievability of population genetic survey data? An assessment of animal mitochondrial DNA studies.” Evolution 53(6): 1961-1965. https://doi.org/10.1111/j.1558-5646.1999.tb04576.x

Merson, Laura, Oumar Gaye, and Philippe J. Guerin. 2016. “Avoiding data dumpsters—toward equitable and useful data sharing.” *New England Journal of Medicine* 374(25): 2414-2415. http://dx.doi.org/10.1056/NEJMp1605148

Naudet, Florian, Charlotte Sakarovitch, Perrine Janiaud, Ioana Cristea, Daniele Fanelli, David Moher, and John P. A. Ioannidis. 2018. “Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in *The BMJ* and *PLOS Medicine*.” *BMJ* 360: k400 https://doi.org/10.1136/bmj.k400

Niemeyer, Kyle E., Arfon M. Smith, and Daniel S. Katz. 2016. “The challenge and promise of software citation for credit, identification, discovery, and reuse.” *Journal of Data and Information Quality* 7(4): 16. https://doi.org/10.1145/2968452

Parham, Susan Wells, Jon Bodnar, and Sara Fuchs. 2012. “Supporting tomorrow’s research: Assessing faculty data curation needs at Georgia Tech.” *College & Research Libraries News* 73(1): 10-13. http://hdl.handle.net/1853/48706

Rolando, Lizzy, Chris Doty, Wendy Hagenmaier, Alison Valk, and Susan Wells Parham. 2013. “Institutional readiness for data stewardship: Findings and recommendations from the research data assessment.” *Georgia Institute of Technology*. http://hdl.handle.net/1853/48188

Savage, Caroline J., and Andrew J. Vickers. 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

Tenopir, Carol, Elizabeth D. Dalton, Suzie Allard, Mike Frame, Ivanka Pjesivac, Ben Birch, Danielle Pollock, and Kristina Dorsett. 2015. “Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide.” *PLoS ONE* 10(8): e0134826. https://doi.org/10.1371/journal.pone.0134826

Van Tuyl, Steve, and Gabrielle Michalek. 2015. “Assessing Research Data Management Practices of Faculty at Carnegie Mellon University.” *Journal of Librarianship and Scholarly Communication* 3(3): eP1258. https://doi.org/10.7710/2162-3309.1258

Van Tuyl, Steve, and Amanda L. Whitmire. 2016. “Water, Water, Everywhere: Defining and Assessing Data Sharing in Academia.” *PLoS ONE* 11(2): e0147942. https://doi.org/10.1371/journal.pone.0147942

Vines, Timothy H., Arianne YK Albert, Rose L. Andrew, Florence Débarre, Dan G. Bock, Michelle T. Franklin, Kimberly J. Gilbert, Jean-Sébastien Moore, Sébastien Renaut, and Diana J. Rennison. 2014. “The Availability of Research Data Declines Rapidly with Article Age.” *Current Biology* 24(1): 94-97. https://doi.org/10.1016/j.cub.2013.11.014

White, Ethan P., Elita Baldridge, Zachary T. Brym, Kenneth J. Locey, Daniel J. McGlinn, and Sarah R. Supp. 2013. “Nine simple ways to make it easier to (re) use your data.” *Ideas in Ecology and Evolution* 6(2): 1–10. http://dx.doi.org/10.4033/iee.2013.6b.6.f

Amanda L. Whitmire, Michael Boock, Shan C. Sutton. 2015. “Variability in academic research data management practices: Implications for data services development from a faculty survey.” *Program: electronic library and information systems* 49(4): 382-407. https://doi.org/10.1108/PROG-02-2015-0017

Wiley, Christie A.. 2015. “An Analysis of Datasets within Illinois Digital Environment for Access to Learning and Scholarship (IDEALS), the University of Illinois Urbana-Champaign Repository.” *Journal of eScience Librarianship* 4(2): e1081. http://dx.doi.org/10.7191/jeslib.2015.1081

Wiley, Christie, and William H. Mischo. 2016. “Data management practices and perspectives of atmospheric scientists and engineering faculty.” *Issues in Science and Technology Librarianship* 85(Fall 2016). http://dx.doi.org/10.5062/F43X84NJ

Wolins, Leroy. 1962. “Responsibility for Raw Data.” *American Psychologist* 17(9): 657-658. http://dx.doi.org/10.1037/h0038819