Data Management Plans: Implications for Automated Analyses

RESEARCH PAPER

NGOC-MINH PHAM
HEATHER MOULAISON-SANDY
BRADLEY WADE BISHOP
HANNAH GUNDERMAN

*Author affiliations can be found in the back matter of this article

ABSTRACT

Data management plans (DMPs) are an essential part of planning data-driven research projects and ensuring long-term access and use of research data and digital objects; however, as text-based documents, DMPs must be analyzed manually for conformance to funder requirements. This study presents a comparison of DMPs evaluations for 21 funded projects using 1) an automated means of analysis to identify elements that align with best practices in support of open research initiatives and 2) a manually-applied scorecard measuring these same elements. The automated analysis revealed that terms related to availability (90% of DMPs), metadata (86% of DMPs), and sharing (81% of DMPs) were reliably supplied. Manual analysis revealed 86% (n = 18) of funded DMPs were adequate, with strong discussions of data management personnel (average score: 2 out of 2), data sharing (average score 1.83 out of 2), and limitations to data sharing (average score: 1.65 out of 2). This study reveals that the automated approach to DMP assessment yields less granular yet similar results to manual assessments of the DMPs that are more efficiently produced. Additional observations and recommendations are also presented to make data management planning exercises and automated analysis even more useful going forward.

CORRESPONDING AUTHOR:
Ngoc-Minh Pham
University of Missouri, US
phamngocminhclc@gmail.com

KEYWORDS:
Data management plans (DMPs); Automated approaches; Text-mining; DMP evaluation

TO CITE THIS ARTICLE:
Pham, N-M, Moulaison-Sandy, H, Bishop, BW and Gunderman, H. 2023. Data Management Plans: Implications for Automated Analyses. *Data Science Journal*, 22: 2, pp. 1–14. DOI: https://doi.org/10.5334/dsj-2023-002

1. INTRODUCTION

To advance many frontiers of science, research data must be shared across the borders of both disciplines and countries, among the various organizations, institutions, and research teams that often stretch around the globe. In this internationally distributed and multidisciplinary environment, open science requires data that are discoverable and reusable in conjunction with the FAIR Principles (https://www.go-fair.org/fair-principles/). These principles serve to guide data producers, publishers, and other involved parties to share data in a way that could enhance the ability of machines to automatically find and use their scholarly data, besides the reuse by individuals. In other words, FAIR principles emphasize and support data sharing practices that could enable machine-actionability in finding and using data in the same manner that a human would but with different scope, scale, and speed (Wilkinson et al. 2016). Data management plans (DMPs) serve to address the what, how, who, and where of data management by formally outlining the roles, responsibilities and activities for managing data during and after research (Bishop & Hank 2020) in alignment with FAIR principles.

Recent calls to create machine-actionable DMPs (maDMPs) to allow for automatic exchange, integration, and validation (Miksa et al. 2019) could further move all of science toward standardization of research data management (RDM) practices and the formulation of DMPs. Making DMPs machine-actionable could facilitate data discovery and reuse and enable automated evaluation and monitoring (Cardoso et al. 2020; Simms et al. 2017). To facilitate the automated quality evaluation of proposed DMPs, Cardoso et al. (2020) offer the use of closed questions (e.g., name repositories, list metadata standards relevant to the discipline, identify file formats, etc.) which researchers could use during DMP creation, a process that is in alignment with Miksa et al.'s (2019) call for the implementation of 10 principles to promote maDMPs. When the results of the automated checks indicate the stated plans are not adequate, human evaluation will be of use.

At present, DMPs continue to be text-based, with the obligation to align with funder requirements and leave any considerations of the FAIR and other data principles up to the researcher. Tools have been created and evaluated to help with writing DMPs (e.g., Gajbe, Tiwari & Singh 2021), yet concerns remain that DMPs are unenthusiastically created, often as afterthought (e.g., Mannheimer 2018), with reviewers accepting them rotely (Berman 2017).

2. RESEARCH QUESTION

Despite efforts to make DMPs machine-actionable, maDMPs are not yet the norm. Further, evaluation work with the current text-based DMPs has been limited (e.g., Bishop et al. 2020) and the literature reveals that work has not extended to machine-based content analyses of DMPs. This project addresses this gap in the literature by addressing the following research questions:

> RQ1: How does an automated content analysis of a set of DMPs compare to a manual evaluation?
>
> RQ2: What does the analysis of successful DMPs reveal?

Using DMPs that accompanied successful proposals to an international and interdisciplinary funding competition, this research investigates whether automated means can serve to evaluate them effectively, and how automated evaluation methods may provide different insights from manual evaluations. The sample of this study is the entire population of DMPs from the 21 funded projects in two funding categories: ocean sustainability and arctic systems.

3. REVIEW OF THE LITERATURE

3.1. DMPS AND DATA SHARING

Top-down mandates for data sharing and reuse are compelling, yet researchers need to be convinced of the importance of managing data responsibly, in a way that will allow for future sharing and reuse. Resistance to this goal of DMPs, data sharing, has been observed (Bially Mattern & Moulaison Sandy 2018). Reasons cited for hesitations to publish data openly include the potential to have their work 'scooped' (Berman 2017; Kim & Adler 2015; Wiley Open Science Researcher Survey 2016), a low perceived return on investment, possible misuse or use against

them (Berman 2017), cyberinfrastructure issues, researcher foot dragging (Bially Mattern & Moulaison Sandy 2018), and problems with data interoperability and integration (Kim & Adler 2015; Wiley Open Science Researcher Survey 2016). Other studies quote scientists as lacking time and resources to make their data appropriate for sharing (Gewin 2016). If having a plan supports data sharing, not having a plan or not being ready to implement one seems to allow for plausible deniability. To make these new workflow tasks easier, tools have been created to assist researchers with DMPs such as the DMPtool.org; DMPonline.dcc.ac.uk; and Data Stewardship Wizard (https://ds-wizard.org/) to name only a few.

Another factor affecting attitudes toward DMPs and the end-goal of data sharing is the mixed messages that might be sent by funders. Dietrich et al. (2012) reviewed funding organizations and found varying disjointedness of coverage in data management policies concerning storage, licensing, metadata, and sharing. Therefore, while funders may require DMPs from researchers, the required elements within these DMPs vary significantly across organizations, and as a result, researchers following required RDM guidelines for one organization may render the data and digital outputs difficult to use (or even unusable) by the original research team or researchers from another organization or domain. In the early days of requiring DMPs, the U.S. National Science Foundation (NSF) reportedly felt the standard for content would emerge through the community of practice (Berman 2017). Although there are benefits to being descriptive versus prescriptive with the requirements for DMPs, to leave the content to the community of practice has the potential to be exclusionary for researchers outside of that immediate field. Leaving requirements fluid has the potential to penalize transdisciplinary teams seeking funding; it is likewise potentially uninviting to interdisciplinary or extra-disciplinary secondary users of the data or digital objects produced.

Still, funding agencies do have trainings and suggested parameters for the elements a DMP should have. The Belmont Forum's DDOMP provides specific questions about the data and other digital outputs for researchers to address in their proposals that directly map to a scorecard to be used in evaluation. The focus is on 16 criteria in 9 broad areas: (1) the data itself; (2) data storage and use; (3) data management personnel; (4) data security; (5) data preservation concerns; (6) restrictions (required only if necessary); (7) intellectual property; (8) supporting documentation; and (9) long-term costs. maDMPs as well as scoring of any DMP require measurable parameters. At these broad levels, each domain can address these key descriptive and contextual metadata.

3.2. AUTOMATED APPROACHES TO DMP ANALYSIS

The literature indicates that creating maDMPs could enable the automated assessment of DMPs. The format for maDMPs that can support automated assessment of DMPs includes having themes/data management components which represent the common topics addressed in DMPs and using controlled vocabulary associated with the themes (Miksa et al. 2019). Though there exists little to no literature on reported efforts to evaluate DMPs automatically, there are approaches DMP stakeholders can take to automate the evaluation process, some of which are the use of text mining techniques such as n-gram analysis. The n-gram analysis enables researchers to convert a text into a set of n-grams such as unigrams, bigrams, and trigrams. Those n-grams and their frequency distributions can be matched/compared with controlled words associated with each component of DMP to determine what key components of DMPs are present in the proposed plans. However, for some components such as ethical issues, the evaluation needs to go beyond assessing the presence of DMP components. In this case with DMPs, human input is inevitable (Miksa et al. 2019).

3.3. MANUAL DMP ASSESSMENT AND EVALUATION

The bulk of work to assess DMP has been done manually. One primary example is the Institute of Museum and Library Services (IMLS) Document Assessment and Review Tool (DART) project. The DART project developed a rubric to evaluate NSF DMPs in hopes to standardize their assessment and enable institutional comparisons on the absence or presence of certain elements. Ultimately, the rubric created was so exhaustive in the list of potential elements that any use of it would take considerable time (Rolando et al. 2015). Grant reviewers, for example, may have other elements such as scientific merit and impacts to prioritize in a review, but these assumptions are anecdotal.

Post-award assessment work, generally done through qualitative studies, has shown the quality of DMPs to be variable. An earlier version of the DART rubric was used to assess 29 DMPs, finding that overall, the quality of the DMPs varied greatly, with roles and responsibilities for data management, metadata standards for describing research data, and policies for protecting intellectual property rights the elements missing most (Samuel et al. 2015). Other work has been done in the form of case studies, examining DMPs for certain funding agencies (often, NSF) at specific universities (e.g., Berman 2017; Bishoff & Johnston 2015; Mischo, Schlembach & O'donnell 2014; Van Loon et al. 2017). Assessments of the contents of DMPs in the literature have been both positive and negative. A 2015 study reviewed 182 DMPs across many disciplines and found that 80% of plans adequately described how data would be archived (Bishoff & Johnston 2015), results that are construed as positive. Still, most research in this area highlights the shortcomings of DMPs, and generally from the perspective of information scientists or academic librarians. For example, while reviewing 119 DMPs, one team found that 51% did not identify the individual(s) responsible for data management, which was consistent with prior research findings (Van Loon et al. 2017). As indicated above, a number of these studies revealed less than enthusiastic approaches to DMP creation and, and potentially, implementation and use.

4. METHODOLOGY

Twenty-one DMPs (13 DMPs from the Transdisciplinary Research for Ocean Sustainability (https://www.belmontforum.org/cras/#oceans2018); abbreviated here 'Oc#') and 8 DMPs from the Resilience in Rapidly Changing Arctic Systems (https://www.belmontforum.org/cras/#arctic2019); abbreviated here 'Ar#') from projects funded by the Belmont Forum were analyzed. These DMPs were submitted to the Belmont Forum, which had published information about DMP expectations, including a scorecard (Belmont Forum 2018) for researchers to self-evaluate their DMPs (Bishop et al. 2019). The sample of this study is the entire population of DMPs from the 21 funded projects in these two categories. Figure 1 provides an overview of the research methodology.

4.1. AUTOMATED ANALYSIS

Automated analysis was conducted on MAXQDA Analytics Pro (VERBI Software 2022), using its MAXDictio functionalities. To determine terms associated with the scorecard (Belmont Forum 2018) criteria, we used text-mining features such as unigram frequencies and n-gram combinations offered in 'word frequencies' and 'word combinations' functions in MAXDictio. 'Word frequencies' and 'word combinations' functions in MAXDictio allow users to pre-process the corpus for noise reduction, using tidy text techniques like Lemmatization and Stop word removal. The lemmatization replaces words with their base form (Anandarajan et al. 2019). For example, the words 'plays,' 'played,' and 'playing' have 'play' as their lemma. Stop words are common and high frequency words like 'a,' 'the,' 'of,' 'and,' 'an' (Luhn 1958). Stop words can be customized. Any content words can be added to the list of stop words if they are viewed as insignificant to provide insights of the corpus. The stop word removal is used to reduce dimensionality of the datasets and therefore enhance performance of feature extraction algorithm in text mining (Aggarwal 2015). When using word frequencies and word combinations functions in MAXDictio, we applied the stop word removal using the built-in standard list of stop words, lemmatization, and search for word combinations with two to three words.

Results from word frequencies and word combinations show the frequency of words in the corpus, their unique occurrence across the documents, and their percentage of unique occurrence. The results were later reviewed, and terms associated with scorecard criteria were manually grouped together. Terms associated with the scorecard criteria (Belmont Forum 2018) used in phase two of the analysis were identified by the research team: (1) data: type / collection methods / size; (2) metadata: standards / formats; timeframe; data sharing; (3) personnel; (4) security; data security; (5) data retention; preservation personnel; (6) restrictions; sensitive data; limitations; (7) intellectual property; intellectual property rights; licensing; (8) supporting documentation; and (9) costs; long-term costs. The terms identified above, their lexical variants and synonyms and related bi- and trigrams were produced, providing an overarching view of the DMP content.

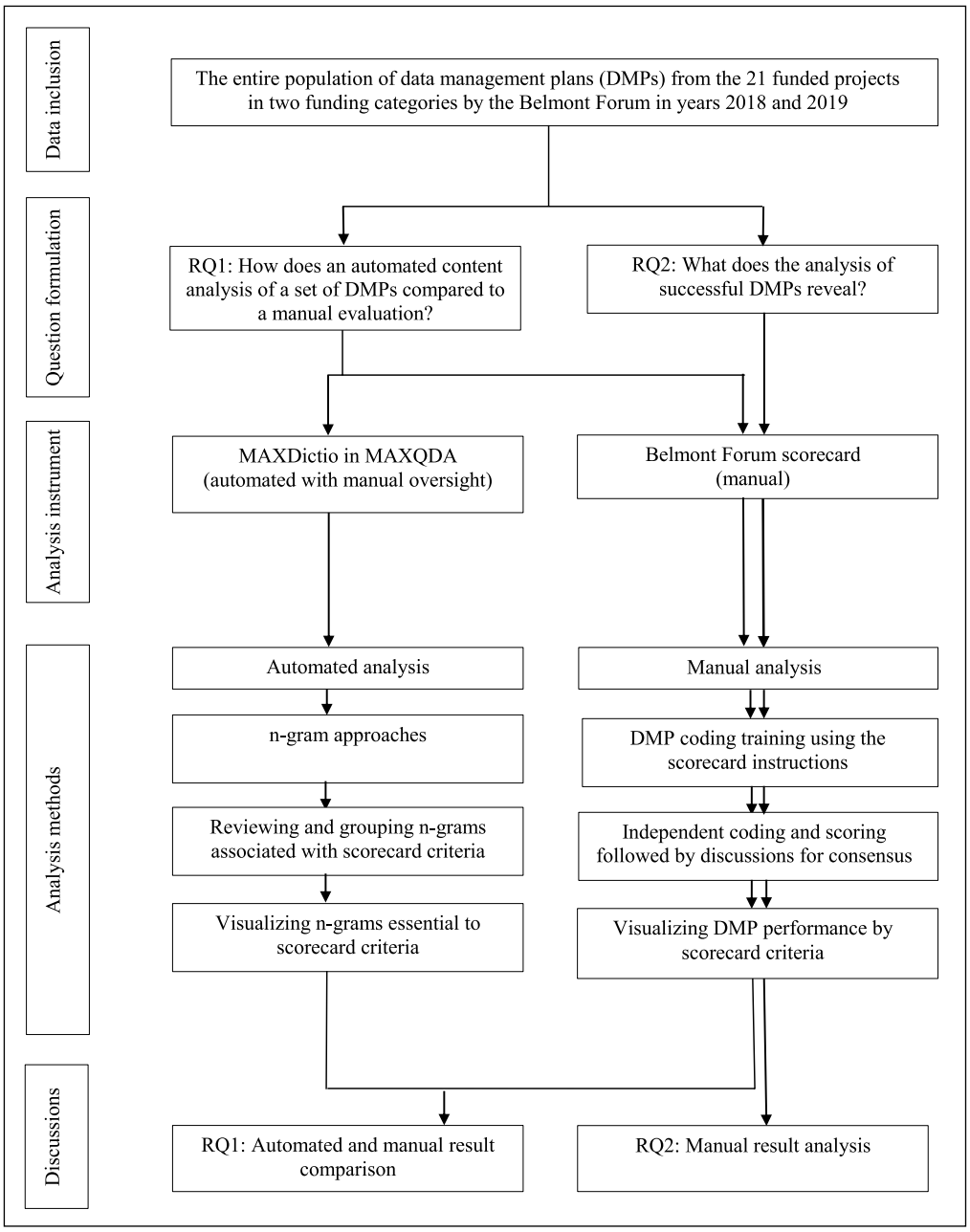

Figure 1 Overview of Research Methodology.

4.2. MANUAL SCORING

Manual scoring of the same set of DMPs was carried out for comparison using the Belmont Forum scorecard. Members of the research team were trained using the instructions given to Belmont Forum award applicants and reviewers; coding took place over several months in fall/winter 2021, requiring weekly meetings over the span of three months to finalize. The scorecard presents 16 criteria in 9 broad areas: (1) the data itself; (2) data storage and use; (3) data management personnel; (4) data security; (5) data preservation concerns; (6) restrictions (required only if necessary); (7) intellectual property; (8) supporting documentation; and (9) long-term costs. These high-level criteria and their sub-criteria are provided in Appendix A. Full conformance is worth 2 points, incomplete or partial conformance is worth 1 point, and no response or lack of conformance is worth 0 point. Partial conformance is deemed to be adequate, as it implies the criterion was addressed to an extent, but some aspect of the explanation was incomplete. For example, the DMP might not name a specific repository (e.g., Van Loon et al. 2017) or institution, or might otherwise address the criterion but not in an actionable way. When coding, team members met to discuss any differences in scores until

100% agreement was reached. An anonymized version of the dataset is available through Github at https://github.com/MinhphamMizzou/DMP_Belmont_Forum_analysis_.git.

Manually-assigned scores were then imported into R Studio (Version 1.1.463 – © 2009–2018 RStudio, Inc.) using the R language (R Core Team 2013) for analysis and visualization. The packages used during the process of importing, transforming, analyzing, visualizing, and formatting included readxl (Wickham & Bryan 2018), dplyr (Wickham et al. 2021), ggplot2 (Wickham 2016), scales (Wickham 2018), tidytext (Silge & Robinson 2016), and jtools (Long 2017). The Readxl package was used to import the scores in an .xlsx file into the R environment. Next, dplyr was used to transform the data, conduct an exploratory analysis, and obtain descriptive statistics. Ggplot2, scales, tidytext, and jtools were used to create and format figures. The scripts for the analysis can be found on GitHub https://github.com/MinhphamMizzou/DMP_Belmont_Forum_analysis_.git.

6. RESULTS

6.1. AUTOMATED ANALYSIS OF DMPS

An automated analysis of the corpus provides insight into the extent that required information is present. Terms (unigrams, bigrams and trigrams) included in the set that were essential to scorecard categories were organized according to the category they primarily supported. Table 1 presents these results by scorecard category, ordered by the percentage of documents in which the terms appear. In each category, the greater the number of documents, the better the coverage that can be inferred.

Table 1 Word frequencies of some of n-grams essential to scorecard categories.

CATEGORIES	TERM (LEMMATIZED)	FREQUENCY	NO. DOCUMENTS	DOCUMENT %
(1) data/size	type	19	11	52.38
	datum format	4	4	19.05
	size	2	2	9.52
(2) data storage/data use	available	91	19	90.48
	datum repository	15	11	52.38
	datum storage	10	7	33.33
(3) data security/data access	metadata	82	18	85.71
	security	14	9	42.86
(4) data management personnel	datum management	65	20	95.24
	responsible	32	16	76.19
(5) data preservation	project website	27	7	33.33
	preservation	12	7	33.33
	long-term use	3	3	14.29
	long-term preservation	2	2	9.52
(6) privacy (not required for all DMPs)	privacy	13	9	42.86
	sensitive	19	8	38.10
(7) intellectual property	intellectual property	17	12	57.14
(8) data sharing	share	65	17	80.95
	open access	26	15	71.43
	license	29	11	52.38
	make available	23	10	47.62
	open datum policy	6	6	28.57
(9) cost	cost	34	12	57.14

This automated analysis reveals that there is a considerable variation in the frequencies and the unique occurrence of the words associated with the scorecard criteria. Eight criteria had associated terms that appeared in at least half of the DMPs; information about (6) privacy was different and was only required if necessary. A bigram related to (4) data management appears in 95% of the DMPs; related unigram *responsible* is also seen to relate to data management personnel, and appears in 76% of the DMPs; these results are interpreted to be exceedingly promising for the adequate inclusion of information relating to personnel. Conversely, the bigram *intellectual property* relating to statements about ownership of (7) intellectual property was surprisingly inconsistent, appearing in only 12 documents total (57%); likewise, *cost* only appeared in 57% of the DMPs. Unigrams and bigrams supporting (5) digital preservation were also inconsistently supplied, with project website and preservation only appearing in 7 DMPs each (33%); long-term use in 14%; and long-term preservation in 10%. These variations seem to reflect the uneven attention of researchers to different components in the successful DMPs.

As noted, DMPs are free-text and do not use controlled vocabularies; the automated methods for analysis used here are predicted on the consistent use of terminology commonly associated with the required elements of a DMP. Table 1 shows that terms associated with each category should be evident and are the first steps to identifying the presence of key components of DMPs as well as evaluating their quality. This also reinforces previous assertions that to facilitate the creation of machine-actionable DMPs, a list of controlled keywords associated with each key component of DMPs is needed (Miksa et al. 2019). To be able to better identify the presence of key categories, and evaluate their quality, there is also a need for analysis mechanisms such as the Description Logics Queries system proposed by Cardoso et al. (2020) which can 'deduce implicit knowledge from the explicitly represented knowledge' (Lenzerini, Milano & Poggi 2004: 18).

6.2. MANUAL ANALYSIS

For comparison, a more granular score for the set of DMPs and for the individual categories was created through manual analysis. For each DMP scored manually, a composite score was calculated by adding the manual scores for each of the criteria and dividing by the total number of criteria. Plans with a composite score averaging 1 or more are deemed to be adequate to comply with the Belmont Forum's Data Policy and Principles (https://www.belmontforum.org/archives/resources/belmont-forum-data-policy-and-principles). These policies and principles were developed around the same time as the FAIR Data Principles and four sentiments on data to be discoverable, accessible, understandable, and future use in sustainable, trusted repositories map closely with to FAIR without a readable acronym. This study finds that 86% (n = 18) of the DMPs provided adequate information and had a composite score ≥1. Of those, 14 DMPs (77.8%) ranged from 1 to 1.5, with the mean of 1.27 (*SD* = 0.24). Four DMPs (22.2%) out of 18 adequate DMPs had composite scores greater than 1.5, or close to 2. Overall, the sample for the current study (*N* = 21) had composite scores ranging from 0.81 to 1.69 on a scale from 0 to 2; composite scores averaged 1.21 (*M* = 1.21, *SD* = 0.27). Detailed descriptive statistics of the sample are provided in Table 2.

Table 2 Descriptive statistics based on DMP composite scores.

	N (OUT OF 21)	*M*	*SD*	MEDIAN	MIN	MAX
Adequate DMPs (average composite score ≥1)	18 (86%)	1.27	0.24	1.22	1.00	1.69
Inadequate DMPs (average composite score <1)	3 (14%)	0.82	0.04	0.81	0.81	0.88
Total	21	1.21	0.27	1.19	0.81	1.69

Next, average scores for the 16 scorecard criteria were calculated; whereas in the automated analysis, the 16 criteria were analyzed according to the nine groups in which they are found, the manual analysis allows for increased granularity in the results. Average criteria scores ranged from 0.1 to 1.9 on a scale of 0 to 2 (see Figure 2). Three quarters of the criteria had scores averaging ≥1 (n = 12; 75%); with only one quarter having scores averaging <1 (n = 4; 25%). The average score for the criteria across all the plans was 1.21 (*M* = 1.21, *SD* = 0.44).

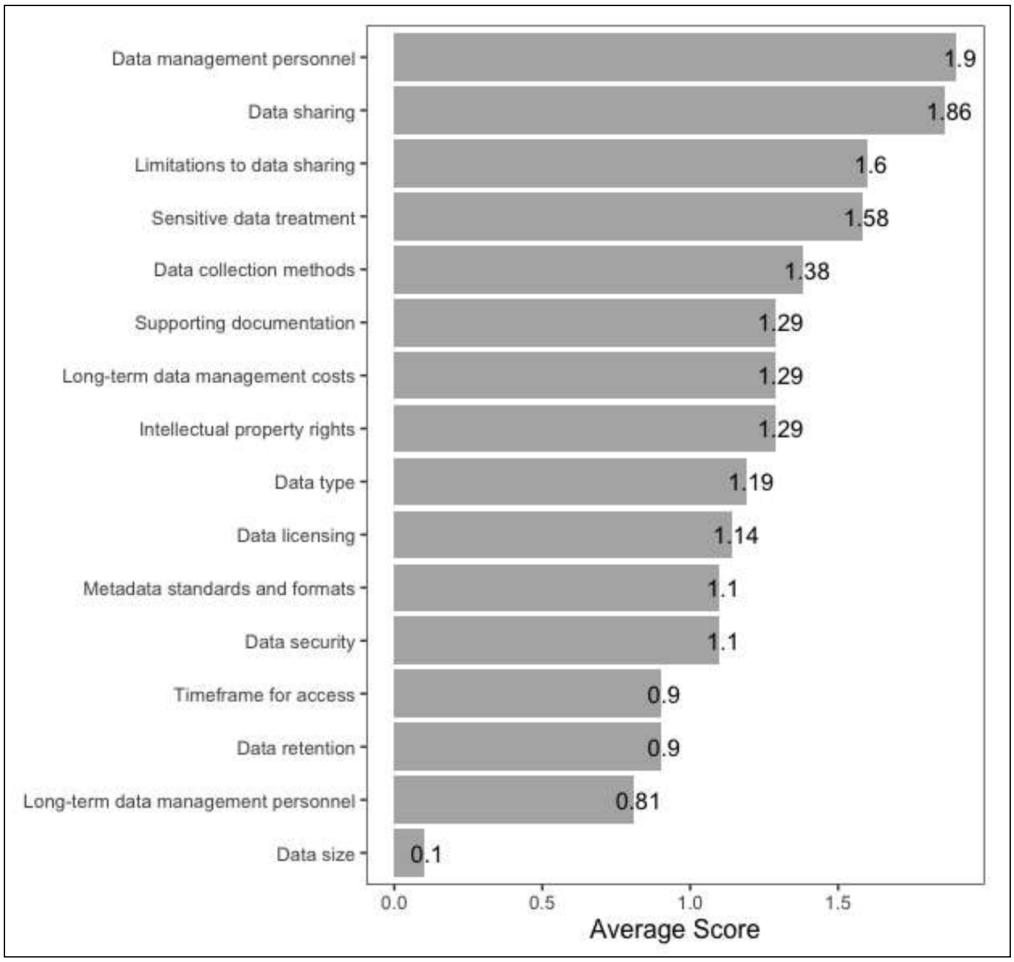

Figure 2 Average scores across all DMPs, by scorecard criterion.

The two criteria on which the plans performed best was a) Data management personnel (average score: 1.9) – full conformance required a named individual be responsible for data management; and b) Data sharing (average score: 1.86) – full conformance required a named repository be indicated. Although the level of conformance could not be confirmed using the automated means, these scores do approximate the results of the automated analysis. Conversely, the lowest average scores were associated with the following four criteria: a) Data size (average score: 0.1); b) Long-term data management personnel (average score: 0.81); c) Data retention (average score: 0.9); and d) Timeframe for access (average score: 0.9). Data size was by far the most overlooked criterion, having been addressed only in a single plan in the sample; this finding is also consistent with the results of the automated analysis.

Comparing criteria scores across adequate and inadequate DMPs presents a way to identify weak areas in order to seek improvements in inadequate DMPs (see Figure 3). The adequate set of DMPs (n = 18; 86%) had consistently strong scores across all the criteria. For the adequate DMPs, only three criteria averaged below 1: Data retention (average score: 0.94), Long-term data management personnel (average score: 0.94), and Data size (average score: 0.11). Of the three inadequate DMPs, seven criteria averaged below 1: the three mentioned for the adequate DMPs, plus Supporting documentation (average score: 0.67), Timeframe for access (average score: 0.33), Data security (average score: 0.33), and Long-term data management costs (average score: 0.00). In other words, the four criteria that differentiated DMPs averaging above or below a 1 were a) Supporting documentation; b) Data security; c) Timeframe for access, and d) Long-term data management costs.

To investigate possible effects of higher word counts on the adequacy of DMPs, the word counts of the DMPs were calculated. Overall, the average length of a DMP is about 1,000 words with great variability in word counts across the DMPs (M = 1,004.8, SD = 301.6). The DMP with the highest word count had 1,769 words, almost three times longer than the DMP with the lowest word count of 591 words. Table 3 shows that the set with composite scores averaging ≥1 tended to have higher word counts than the set with scores averaging <1. The average word count in the former set was 1,053.17 (M = 1,053.17, SD = 294.8), 47% higher than that in the

latter set (*M* = 714.3, *SD* = 151.2). The former set also had a greater variability in word counts (*SD* = 294.84) than the former set (*SD* = 151.19). This indicates that word counts in the former set are more dispersed than the former set and the higher mean of word counts in the former set may be biased by influential values. Any significant relationships between word counts and the adequacy of DMPs requires further studies with larger sample sizes.

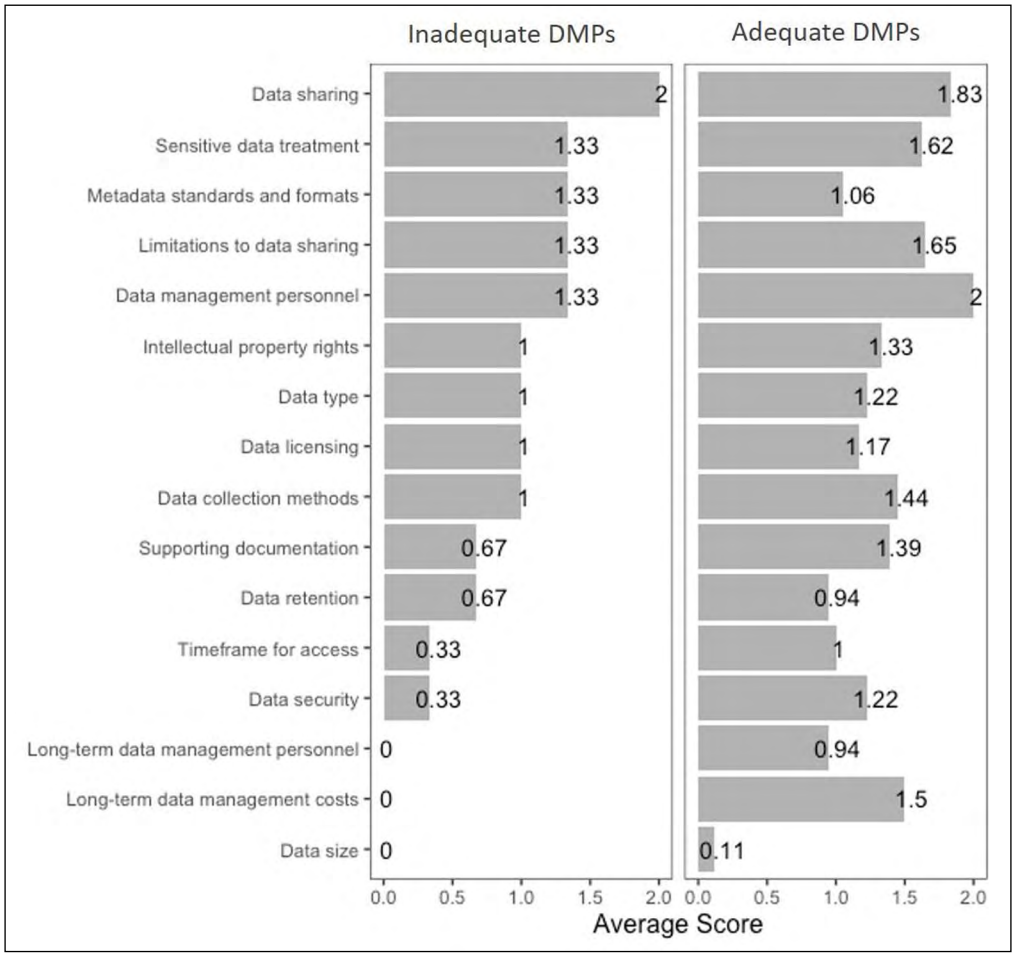

Figure 3 Average scores by scorecard criterion for adequate and inadequate DMPs.

	N (OUT OF 21)	M	SD	MEDIAN	MIN	MAX
Adequate DMPs (average composite score ≥1)	18 (86%)	1,053.2	294.8	1036	606	1,769
Inadequate DMPs (average composite score <1)	3 (14%)	714.3	151.2	714.3	591	883
Total	21	1,004.8	301.6	1025	591	1,769

Table 3 Word count for adequate and inadequate DMPs.

7. DISCUSSION AND IMPLICATIONS

This automated approach to DMP assessment yields a less granular (i.e., binary assessment) of the presence or absence of a term. In this study, the presence of a term as part of the automated analysis is equated with Level 1 conformance, and automated results are similar to manual assessments. Although a 1:1 comparison of results is not possible with the two differing methodologies, global results allow us to infer that machine-based approaches are a viable option for naïve or large-scale DMP assessment, at least in a preliminary way. Especially if a controlled vocabulary is developed, identification of Level 1 conformance or the binary presence of a term is easier to ascertain. Besides a controlled vocabulary, if funder requirements include standard parameters and elements, then DMPs could have required categorical responses to each element within a parameter; if this were to take place, it would make it easier for automated approaches to be implemented to assess Level 2 conformance. The problem here is buy-in on the part of funders and reviewers.

DMPs are reviewed by funders manually, even if DMPs will ideally support machine-actionability and the implementation of the FAIR data principles. This study shows that the manual analysis led to the assessment of Level 2 conformance at greater granularity than the automated approach, contributing to more accuracy in the evaluation of DMP quality. Although a one-by-one, manual approach to DMP assessment is a reasonable approach to adopt on the part of funders for their purposes, it does not scale and it does not ensure machine-actionability. We note that the extensive efforts required for manual evaluating will make manual coding difficult at scale. Further, the use of automated means aligns with recommendations for best practices (Miksa et al. 2019), implying that it is a more promising approach to pursue on the part of funders, especially. There is a disconnect between the two realities; this discussion will assess some of the strengths and weaknesses identified in each.

7.1. OBSERVATIONS

The top frequent unigrams, bigrams, and trigrams in the automated analysis reveal that the elements in the DMPs the researchers mentioned most are *data management* (n = 20; 95%) followed by *availability* (n = 19; 90%); *metadata* (n = 18; 86%), *sharing* (n = 17; 81%), and *responsibility* (n = 16; 76%). This suggests at least a Level 1 conformance in these areas with the binary inclusion of the concept inferred. *Size* and *long-term preservation* each were inconsistently included terms in the DMPs (n = 2; 10%), as were *long-term use* (n = 3; 14%), *data format* (n = 4; 19%) and *open data* policy (n = 6; 29%). These results were largely consistent with the results from the manual analysis.

Terms identified as supporting (4) data management personnel were two of the most frequently occurring: *data management* and *responsibility*. Manual assessments likewise performed best on this element, with the average score of 1.9 indicating high conformance. Full, Level 2 conformance requires that a named individual be included; such a name would not be captured using the automated methods employed, yet the results track together for both methodologies. The same phenomenon is observed in the case of the least frequently occurring term: *size*. Other terms identified as primarily supporting (1) data / size were *type* (n = 11, 52%) and *data format* (n = 4; 19%). Together, these were the weakest performing set of terms for an element according to the automated assessment. Data size was likewise the lowest performing criterion on the manual assessment, with an average score of 0.1. Although the nuances of the content were not captured, automated content analyses seem promising as an overarching proxy for the inclusion of DMP content that is most consistently revealed through manual analysis.

7.2. RECOMMENDATIONS FOR IMPROVEMENTS FOR DMPS

Close investigation of a number of elements provides insight in improvements for DMPs. First, we concur with the call to create and apply a controlled vocabulary of terms related to each element – this will provide for ease of assessment and confirmation of Level 1 conformance. Second, adequate DMPs consistently had higher word counts, implying that more content is covered through length. Researchers should be encouraged to use the needed number of words to explain their plan rather than providing minimal information or not adequately describing their intention. Increased word counts in DMPs have the potential to support context. Level 2 names should include Level 1 elements to clarify the kind of name they are.

Specifically, the mismatch between the mention of the term *metadata* in the DMPs found in Table 1 and the quality of performance in the manually encoded 'metadata standards and formats' criterion indicate that researchers grappled with the specifics of this element. Incomplete planning in regards to metadata may have negative impacts on the implementation of their DMP, which in return may hamper data sharing and data accessibility. The mismatch in the use of the terminology and the inclusion of consistent Level 1 or Level 2 information also may indicate that providing researchers with scorecard criteria may not suffice to support the creation of effective DMPs and to implement efficient DMP practices. Besides the criteria from funders, there is a need for guidelines on effective DMP practices from organizational funders and research institutions (Sallans & Donnelly 2012) and support on DMP writing by institutional authorities (Diekema, Wesolek & Walters 2014). In the case of 'metadata', researchers need help from information professionals in the development and implementation of their DMP.

8. LIMITATIONS AND FURTHER STUDY

This study analyzes successful DMPs from an international funder, the Belmont Forum; unfunded DMPs were not included in the analysis, nor were successful DMPs from different funding agencies. Additionally, the projects described were interdisciplinary, focusing on the sciences. Although social sciences methodologies were employed in data collection, this was not the case across the board, and no humanities data were collected or described. Further study should address these limitations based on the sample population analyzed. The methods of analysis used do not permit the one-to-one comparison of results in a particular DMP; rather, the automated analysis of a set of DMPs provides an overview of the set's conformance using the character strings in the documents; manual analysis was more granular on all counts. Future studies could look at other automated approaches to explore the possibility of generating results which can be compared one-to-one with results from manual work. Further, both analyses do not address whether there has been compliance with the DMPs as written; the projects for which these DMPs were written are still ongoing, and future research can look back to analyze the extent to which the DMPs were updated, to become living documents (Miksa et al. 2019), and to which of the parameters of the DMPs have been respected.

9. CONCLUSION

This project seeks to understand, through the analysis of 21 DMPs associated with funded research projects, both the extent to which successful DMPs can be considered useful at present, and how DMPs can be improved going forward. Requirements for DMPs have become increasingly visible in the landscape of funded research in an effort to promote more transparent, reproducible, and open scientific results. While the importance (and limitations) of DMPs have been extensively explored in the literature, less work has been dedicated to the evaluation of these documents using standardized metrics to determine their success in framing the data management considerations within a research endeavor. This project used the Belmont Forum scorecard to assess the completeness of 21 DMPs and how well they incorporated information on data and digital objects, including their storage, and sharing, and the data management personnel responsible for these activities, according to the metrics included in the scorecard. While many of the DMPs were found to be adequate, several areas of data management remained fairly overlooked, including data size, the timeframe for accessing the data and data retention policies, and information on long-term data management personnel.

As part of the process of analyzing DMPs and the potential for their evaluation, we consider the utility of any scoring initiative, as the DART project funded by the IMLS was widely perceived as being too *restrictive*, but the *scorecard* approach studied here had the limitations of not seeing the full proposal and of being evaluated by a team of data specialists, and not experts in the field. Yet, these scorecard evaluations provide a snapshot, allowing for comparisons across DMPs. Recommendations for improvement to support data sharing include funders' consideration as living documents, limiting options for responses to specific, actionable content, and continued evaluations by all relevant stakeholders, including data experts along with subject-area experts. Although DMP is the common term for most required plans, the Belmont Forum's DMPs emphases the other digital objects that make science reproducible and are not technically data. Future DMP literature should explore the curatorial considerations for the items, not just data, that support future reuse.

APPENDIX A: THE BELMONT FORUM (2018) DDOMP SCORECARD CRITERIA

1. **What types of datasets and other digital outputs of long-term value do you expect the project will produce or reuse?**
    - **Data type** [1.1 Plan lists the types of data and other digital outputs of long-term value.]
    - **Data collection methods** [1.2 Plan describes how the data and other digital outputs will be collected, captured, or created.]
    - **Data size** [1.3 Datasets and other digital outputs volume estimated.]

2. **How do you intend to ensure that the data and digital outputs from your project conform to the Belmont Forum Open Data Policy and Principles?**
    - **Metadata standards and formats** [2.1 Plan specifically addresses metadata standards or formats for the data and other digital objects.]
    - **Timeframe for access** [2.2 Plan describes when data and other digital outputs will be made available outside and within the project team.]
    - **Data sharing** [2.3 Plan describes how data and other digital outputs will be made available beyond the project team.]
3. **Which member(s) of your team will be responsible for developing, implementing, overseeing, and updating the Data and Digital Outputs Management Plan?**
    - **Data management personnel** [3 Plan describes which member(s) of the team will be responsible for developing, implementing, overseeing, and updating the DDOMP.]
4. **How do you intend to manage the data and digital outputs during the project to ensure their long-term value is protected?**
    - **Data security** [4.1 The plan describes the security measures to prevent unauthorized access to the data and other digital outputs.]
5. **How and by whom will the data and other digital outputs be managed after the project ends to ensure their long-term accessibility?**
    - **Data retention** [5.1 Plan indicates how long the data and other digital outputs will be retained.]
    - **Long-term data management personnel** [5.2 Plan indicates who will be responsible for managing data after the project ends.]
6. **What restrictions, if any, do you anticipate could be placed on how the data and digital outputs can be accessed, mined, or reused?**
    - **Sensitive data treatment** [6.1 (if applicable) Plan describes how sensitive data and other digital outputs will be made available beyond the project team.]
    - **Limitations to data sharing** [6.2 (if applicable) Plan describes any limitations on the ability to share data and other digital outputs.]
7. **How will you ensure that any data security, privacy, and intellectual property restrictions associated with datasets and digital outputs will be honored and preserved in derivative products?**
    - **Intellectual property rights** [7.1 Plan describes the intellectual property rights to the data and other digital outputs.]
    - **Data licensing** [7.2 Plan describes licensing of the data and other digital outputs.]
8. **What supporting documentation (i.e., metadata) do you plan to make publicly accessible to support the discovery and longer-term reuse of the data and digital outputs?**
    - **Supporting documentation** [8. Plan describes the supporting documentation and metadata that will be created to make data and digital outputs publicly accessible.]
9. **How have you accounted for the costs required to manage the data and digital outputs to ensure long-term accessibility?**
    - **Long-term data management costs** [9. Plan specifies the costs or estimated costs associated with long-term data management or an assigned data manager role.]

ACKNOWLEDGEMENTS

The authors would like to acknowledge the Belmont Forum for sharing the successful DMPs for analysis. We also would like to thank University of Missouri Graduate Professional Council for the Research Development Award and Dr. Rose Marra – Professor and Director of the School of Information Science and Learning Technologies, University of Missouri, Columbia in their support of this publication.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Ngoc-Minh Pham: Data curation, Formal Analysis, Investigation, Visualization, Writing- original draft. Writing – Reviewing & Editing. **Heather Moulaison-Sandy:** Conceptualization, Investigation, Methodology, Supervision, Writing – original draft, Writing – Reviewing & Editing, Validation. **Bradley Wade Bishop:** Conceptualization, Methodology, Writing – original draft, Writing – Reviewing & Editing. **Hannah Gunderman:** Data curation, Writing- Reviewing & Editing.

AUTHOR AFFILIATIONS

Ngoc-Minh Pham orcid.org/0000-0002-9776-3981
University of Missouri, US

Heather Moulaison-Sandy orcid.org/0000-0001-7783-7069
University of Missouri, US

Bradley Wade Bishop orcid.org/0000-0002-5022-2707
University of Tennessee, US

Hannah Gunderman orcid.org/0000-0002-7710-7055
Carnegie Mellon University, US

REFERENCES

Aggarwal, CC. 2015. *Data mining: the textbook* (Vol. 1). Springer. DOI: https://doi.org/10.1007/978-3-319-14142-8_1

Anandarajan, M, Hill, C and **Nolan, T.** 2019. Text preprocessing. In: *Practical Text Analytics*. Springer. pp. 45–59. DOI: https://doi.org/10.1007/978-3-319-95663-3_4

Belmont Forum. 2018. *Data and digital outputs management plan (DDOMP)*. Available at http://www.bfe-inf.org/resource/data-and-digital-outputs-management-plan-ddomp.

Berman, EA. 2017. An exploratory sequential mixed methods approach to understanding researchers' data management practices at UVM: Integrated findings to develop research data services. *Journal of eScience Librarianship*, 6(1): e1104. DOI: https://doi.org/10.7191/jeslib.2017.1104

Bially Mattern, J and **Moulaison Sandy, H.** 2018. Use, ethics, and governance of data repositories: A power-sensitive sociotechnical perspective [Paper presentation]. *The 14th Annual Social Informatics Research Symposium: Sociotechnical Perspective on Ethics and Governance of Emerging Information Technologies, 81st Annual Meeting of the Association for Information Science and Technology (ASIS&T)*, Vancouver, Canada.

Bishoff, C and **Johnston, L.** 2015. Approaches to data sharing: An analysis of NSF data management plans from a large research university. *Journal of Librarianship and Scholarly Communication*, 3(2): eP1231. DOI: https://doi.org/10.7710/2162-3309.1231

Bishop, BW and **Hank, C.** 2020. Digital curation. In Kobayashi, A (ed.), *International encyclopedia of human geography* (2nd ed.). Elsevier. DOI: https://doi.org/10.1016/B978-0-08-102295-5.10531-1

Bishop, BW, Ungvari, J, Davis, RI, Lee, T, Goudeseune, L, Virapongse, A and **Samors, RJ.** 2019. Belmont forum data management plan scorecard (Version v.20190819_final). *Zenodo*. DOI: https://doi.org/10.5281/zenodo.3530933

Bishop, W, Ungvári, J, Gunderman, HC and **Moulaison-Sandy, H.** 2020. Data management plan scorecard. *Proceedings of the Association for Information Science and Technology*, 57(1): e325. DOI: https://doi.org/10.1002/pra2.325

Cardoso, J, Proença, D and **Borbinha, J.** 2020. Machine-actionable data management plans: A knowledge retrieval approach to automate the assessment of funders' requirements. In: *European Conference on Information Retrieval*. Springer. pp. 118–125. DOI: https://doi.org/10.1007/978-3-030-45442-5_15

Diekema, AR, Wesolek, A and **Walters, CD.** 2014. The NSF/NIH effect: surveying the effect of data management requirements on faculty, sponsored programs, and institutional repositories. *The Journal of academic librarianship*, 40(3–4): 322–331. DOI: https://doi.org/10.1016/j.acalib.2014.04.010

Dietrich, D, Adamus, T, Miner, A and **Steinhart, G.** 2012. De-mystifying the data management requirements of research funders. *Issues in Science and Technology Librarianship*, 70(1). DOI: https://doi.org/10.29173/istl1556

Gajbe, SB, Tiwari, A and **Singh, RK.** 2021. Evaluation and analysis of data management plan tools: a parametric approach. *Information Processing & Management*, 58(3): 102480. DOI: https://doi.org/10.1016/j.ipm.2020.102480

Gewin, V. 2016. Data sharing: An open mind on open data. *Nature*, 529(7584): 117–119. DOI: https://doi.org/10.1038/nj7584-117a

Kim, Y and **Adler, M.** 2015. Social scientists' data sharing behaviors: Investigating the roles of individual motivations, institutional pressures, and data repositories. *International Journal of Information Management,* 35(4): 408–18. DOI: https://doi.org/10.1016/j.ijinfomgt.2015.04.007

Lenzerini, M, Milano, D and **Poggi, A.** 2004. Ontology representation & reasoning. *Proceedings of InterOp*.

Long, JA. 2017. jtools: Analysis and presentation of social scientific data (R package version 0.9.3) [Computer software]. *The Comprehensive R Archive Network*. Available at https://cran.r-project.org/package=jtools.

Luhn, HP. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2): 159–165. DOI: https://doi.org/10.1147/rd.22.0159

Mannheimer, S. 2018. Toward a better data management plan: The impact of DMPs on grant funded research practices. *Journal of eScience Librarianship*, 7(3): e115. DOI: https://doi.org/10.7191/jeslib.2018.1155

Miksa, T, Simms, S, Mietchen, D and **Jones, S.** 2019. Ten principles for machine-actionable data management plans. *PLOS Computational Biology*, 15(3): e1006750. DOI: https://doi.org/10.1371/journal.pcbi.1006750

Mischo, W, Schlembach, M and **O'donnell, M.** 2014. An analysis of data management plans in University of Illinois National Science Foundation grant proposals. *Journal of eScience Librarianship*, 3(1): 31–43. DOI: https://doi.org/10.7191/jeslib.2014.1060

R Core Team. 2013. *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Available at http://www.R-project.org/.

Rolando, L, Carlson, J, Hswe, P, Parham, SW, Westra, B and **Whitmire, AL.** 2015. Data management plans as a research tool. *Bulletin of the American Society for Information Science and Technology,* 41(5): 43–45. DOI: https://doi.org/10.1002/bult.2015.1720410510

Sallans, A and **Donnelly, M.** 2012. DMP online and DMPTool: Different strategies towards a shared goal. *International Journal of Digital Curation,* 7(2): 123–129. DOI: https://doi.org/10.2218/ijdc.v7i2.235

Samuel, SM, Grochowski, PF, Lalwani, LN and **Carlson, J.** 2015. Analyzing data management plans: Where librarians can make a difference. *Paper presented at the 122nd ASEE Annual Conference & Exposition*, Seattle, Washington. Available at https://www.asee.org/public/conferences/56/papers/12072/view.

Silge, J and **Robinson, D.** 2016. tidytext: Text mining and analysis using tidy data principles in R [Computer software]. *The Comprehensive R Archive Network*. DOI: https://doi.org/10.21105/joss.00037

Simms, S, Jones, S, Mietchen, D and **Miksa, T.** 2017. Machine-actionable data management plans (maDMPs). *Research Ideas and Outcomes*, 3: e13086. DOI: https://doi.org/10.3897/rio.3.e13086

Van Loon, JE, Akers, KG, Hudson, C and **Sarkozy, A.** 2017. Quality evaluation of data management plans at a research university. *IFLA Journal,* 43(1): 98–104. DOI: https://doi.org/10.1177/0340035216682041

VERBI Software. 2022. *MAXQDA Analytics Pro* [computer software]. Berlin, Germany: VERBI Software. Available at maxqda.com.

Wickham, H. 2016. ggplot2: Elegant graphics for data analysis [Computer software]. *The Comprehensive R Archive Network*. Available at https://ggplot2.tidyverse.org.

Wickham, H. 2018. scales: Scale functions for visualization (R package version 1.0.0) [Computer software]. *The Comprehensive R Archive Network*. Available at https://CRAN.R-project.org/package=scales.

Wickham, H and **Bryan, J.** 2018. readxl: Read excel files (R package version 1.1.0) [Computer software]. *The Comprehensive R Archive Network*. Available at https://CRAN.R-project.org/package=readxl.

Wickham, H, François, R, Henry, L and **Müller, K.** 2021. dplyr: A grammar of data manipulation (package version 1.0.4) [Computer software]. *The Comprehensive R Archive Network*. Available at https://CRAN.R-project.org/package=dplyr. DOI: https://doi.org/10.1007/978-1-4842-6876-6_1

Wiley Open Science Researcher Survey 2016. 2017. DOI: https://doi.org/10.6084/m9.figshare.4748332

Wilkinson, MD, Dumontier, M, Aalbersberg, IJ, Appleton, G, Axton, M, Baak, A, Blomberg, N, Boiten, J-W, Santos, LBDS, Bourne, PE, Bouwman, J, Brookes, AJ, Clark, T, Crosas, M, Dillo, I, Dumon, O, Edmunds, S, Evelo, C, Finkers, R and **Mons, B.** 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18

TO CITE THIS ARTICLE:
Pham, N-M, Moulaison-Sandy, H, Bishop, BW and Gunderman, H. 2023. Data Management Plans: Implications for Automated Analyses. *Data Science Journal*, 22: 2, pp. 1–14. DOI: https://doi.org/10.5334/dsj-2023-002

Submitted: 06 July 2022
Accepted: 15 December 2022
Published: 25 January 2023

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.