Journal of eScience Librarianship

Volume 10 | Issue 2 Article 1

2021-03-01

Use of Optional Data Curation Features by Users of Harvard Dataverse Repository

Ceilyn Boyd
Simmons University

ISSN 2161-3974

JeSLIB 2021; 10(2): e1191
https://doi.org/10.7191/jeslib.2021.1191

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

Full-Length Paper

Use of Optional Data Curation Features by Users of Harvard Dataverse Repository

Ceilyn Boyd

Simmons University, Boston, MA, USA

Abstract

Objective: Investigate how different groups of depositors vary in their use of optional data curation features that provide support for FAIR research data in the Harvard Dataverse repository.

Methods: A numerical score based upon the presence or absence of characteristics associated with the use of optional features was assigned to each of the 29,295 datasets deposited in Harvard Dataverse between 2007 and 2019. Statistical analyses were performed to investigate patterns of optional feature use amongst different groups of depositors and their relationship to other dataset characteristics.

Correspondence: Ceilyn Boyd: ceilyn.boyd@simmons.edu

Received: July 28, 2020 **Accepted**: September 27, 2020 **Published**: March 1, 2021

Data Availability: Data associated with this study is publicly available in the Harvard Dataverse repository (https://doi.org/10.7910/DVN/9STGWE).

Disclosures: The author reports no conflict of interest.

JeSLIB 2021; 10(2): e1191
https://doi.org/10.7191/jeslib.2021.1191

Abstract Continued

Results: Members of groups make greater use of Harvard Dataverse's optional features than individual researchers. Datasets that undergo a data curation review before submission to Harvard Dataverse, are associated with a publication, or contain restricted files also make greater use of optional features.

Conclusions: Individual researchers might benefit from increased outreach and improved documentation about the benefits and use of optional features to improve their datasets' level of curation beyond the FAIR-informed support that the Harvard Dataverse repository provides by default. Platform designers, developers, and managers may also use the numerical scoring approach to explore how different user groups use optional application features.

JeSLIB 2021; 10(2): e1191
https://doi.org/10.7191/jeslib.2021.1191

Introduction

Anyone tasked with electronically gathering user input has had to decide how to collect as much information as possible without overtaxing users' patience. User experience and survey designers often designate user-supplied data elements as either required or optional and provide application-appropriate defaults to reduce user input. Nevertheless, they must consider the downstream consequences of their choices. An optional setting, such as font size or color, may have little impact on most users. In contrast, low response rates for optional survey questions can limit researchers' data analysis or data reuse plans. These examples demonstrate that the specific use case, including the user community's immediate and long-term needs, should inform how designers categorize application features and data elements.

Within research data repositories, such as Harvard Dataverse (2020), the balance between optional-to-required application features shapes the characteristics of stored datasets contributing to their overall data quality, or fitness-for-use and reuse by different research communities (Tayi and Ballou 1998). Based on observations made by the Harvard Dataverse data curation and management team, journals, research projects, and other groups are more likely to use optional data curation features than individual researchers. This quantitative study investigates how individual data depositors differ from other user groups in their use of five optional Dataverse data curation features that provide additional support for the findability and reuse of datasets: 1) optional metadata blocks; 2) keywords; 3) dataset descriptions; 4) supplemental files, such as readmes and codebooks; and 5) additional terms of use. The study also examines how the use of optional features varies across other dataset characteristics, including the year of publication, the presence of restricted data files, and whether a curator reviewed the dataset before its publication in Harvard Dataverse.

The results of this analysis will inform future efforts to promote optional features to user groups who use them infrequently and to improve Dataverse training materials with the eventual goal of improving the findability, accessibility, and reuse of datasets housed within the Harvard Dataverse. Repository managers and depositors at other Dataverse installations may also benefit from improvements in documentation and instructional materials. Likewise, platform designers, developers, and managers may use the quantitative approach to explore how different user groups use optional application features.

Literature Review

Over the past decade, stakeholders across the research lifecycle, including funders, publishers, data curators, and researchers, have recognized the importance of research data repositories for supporting scientific reproducibility and furthering the secondary use of research data (Tenopir et al. 2011). During this period, the number of research data repositories has also increased. In 2013, the Registry of Research Data Repositories (re3data.org) identified 400 data

JeSLIB 2021; 10(2): e1191
https://doi.org/10.7191/jeslib.2021.1191

repositories (Pampel et al. 2013). Today, re3data.org lists over 2,000 disciplinary, institutional, and other repositories supported by non-profit and commercial organizations (Repository Types 2020 & Institution Types 2020). Likewise, the number of articles that explore dataset inventories, repository use, and the data curation services and workflows of research data repositories has also increased during this time (Thelwall and Kousha 2016; Llebot and Van Tuyl 2019; Wiley 2017; 2015; Jeng, He, and Chi 2017).

The term data curation describes the "encompassing work and actions taken by curators of a data repository in order to provide meaningful and enduring access to data" (Johnston et al. 2018a, 5). Data curation practices include uploading data to a repository, arranging data files and related materials for ease of access, ensuring that files conform to preservation best practices, describing datasets using discipline-specific metadata standards, and applying appropriate data use agreements and access controls (Johnston et al. 2018b, 132). These actions help to ensure that research datasets can be located and reused by future researchers. Research data repositories offer a variety of platform features and data curation services that support data sharing and preservation. These include metadata templates and smart defaults for optional fields, full-service support provided by trained curators, and self-service curation in which researchers and other data depositors perform curation tasks themselves.

Research data repositories also vary in their primary audiences, from government and institutional repositories; to disciplinary repositories; to general-purpose repositories that house research data from many disciplines (Austin et al. 2016; McNeill 2016). Harvard Dataverse is a multi-disciplinary, general-purpose repository that accepts data deposits, free of charge, from members of the worldwide research community. It allows self-deposit of datasets by individual researchers and allows groups, such as journals and research projects, to curate data deposits within their sub-repositories. Harvard Institute for Quantitative Social Sciences (IQSS) operates the repository with support from Harvard Library; it is one of 59 installations of the Open Source repository software developed by the Dataverse Project (Dataverse Project, 2020). On October 29, 2019, Harvard Dataverse housed 29,295 datasets with 383,685 files in more than eight subject areas that were deposited by, and are managed by individuals, research projects and organizations, and journals.

The Dataverse infrastructure is informed by the FAIR Guiding Principles that address the data sharing, interoperability, and reuse needs of humans and computational agents (Wilkinson et al. 2016). The Future of Research Communication and e-Scholarship (FORCE11) recommends that research objects, including research data, should be:

- *Findable*. Well-described using metadata standards and discoverable in data catalogs;
- *Accessible*. Citable, have a persistent identifier, reliable storage, and appropriate security and authentication;

JeSLIB 2021; 10(2): e1191
https://doi.org/10.7191/jeslib.2021.1191

- *Interoperable*. Use standard and open file formats and digital preservation best practices; and
- *Reusable*. Accompanied by unambiguous terms of use, clear provenance, meet the quality standards expected by the community of reuse.

These principles reflect best practices for data curation and serve as a framework within which different research communities can define and assess data fitness-for-use (Bishop and Hank 2018). Research data that meets these community quality expectations is FAIR data. Some data curation service providers, such as participants in the Data Curation Network (DCN) explicitly include an evaluation of research data FAIRness in their curation workflows (Johnston et al. 2018b, 132 and 134).

The Dataverse software platform supports FAIR data through a combination of default infrastructure functionality and required and optional user-facing features, including persistent identifiers, domain-specific metadata, support for Open and domain standard file formats, and support for supplementary documentation and code (Crosas 2019). These features capture datasets' research context, including the descriptive metadata, supplementary documentation, code, and other essential elements that support their interpretation and reuse by researchers (Faniel, Frank, and Yakel 2019). Members of the active Dataverse Open Source community may contribute new FAIR-supporting modules to the application codebase to share new functionality and features (Durand 2020). Repository owners may then customize their Dataverse installations with these modules to address their user communities' specific FAIR data needs and expectations.

The five optional data curation features examined in this study serve as research context for Harvard Dataverse datasets, beyond that supplied by Dataverse software by default. A previous study by Koshoffer et al. (2018) used a quantitative approach to examine differences in research context and data curation practices in research data repositories. They examined user-supplied metadata for 80 datasets in four institutional repositories. The repositories differed in the number and type of required and optional metadata fields they provided and the range of data curation services they provided, from no curation to pre-ingest curation, or post-ingest curation. Their study determined that datasets that were actively curated by repository staff, either pre- or post-ingest, more often included supplementary documentation. Also, across all repositories, data depositors contributed slightly more metadata, including keywords, than the minimum required. However, the level of data curation support did not appear to be a notable factor in the amount of contributed optional metadata.

Their results do not perfectly align with the observations made by the Harvard Dataverse data curation team. The team asserted that the level of dataset curation—signaled by the type of data depositor, individual researcher, or group—makes a substantive difference in the use of optional features, such as contributing supplementary documents and optional metadata elements. The

JeSLIB 2021; 10(2): e1191
https://doi.org/10.7191/jeslib.2021.1191

results of this research study may help to identify the source of this discrepancy.

Research Hypotheses

Harvard Dataverse allows self-curation by individual researchers and by organizations and groups, such as journals, teaching courses, and laboratories. Nonetheless, the data curation team periodically reviews data deposits to ensure that depositors use the Harvard Dataverse repository and its features as intended and follow best practices. Based on these reviews, the Harvard Dataverse curation team believed that datasets managed by groups were more likely to use optional data curation features than those deposited by individual researchers. If true, increased outreach to self-curators and improvements in Dataverse training materials could improve optional feature use and the fitness-for-use and FAIRness of self-curated datasets. To test the curation team's supposition, I chose a quantitative approach that uses a numerical optional feature use rubric to investigate the following three research questions:

RQ1. What is the overall distribution of the use of optional features for datasets in the Harvard Dataverse?

RQ2. What is the most frequently used optional feature?

RQ3. How does the use of optional features vary by key dataset characteristics?

In consultation with the Harvard Dataverse data curation team, I identified five optional features to examine that the team felt were instructional and outreach priorities. The optional features I selected fall within the findability, accessibility, and reuse categories of the FAIR Guiding Principles.

- *Optional metadata blocks*. Dataverse allows data depositors to add new, subject-specific metadata blocks to their datasets; custom metadata can improve the findability of datasets.
- *Keywords*. To support its findability by potential users, a dataset should have at least one keyword.
- *Description*. Descriptions also support dataset findability and reuse. Some depositors include the abstract for the research study or journal article associated with their dataset in this field.
- *Supplemental files*. Supplemental files, such as a codebook or a readme file, provide research context for the data itself and support data reuse; and
- *Additional terms of use*. By default, Harvard Dataverse assigned the Creative Commons License (2019), CC0 to datasets at their creation, but data owners can assign any CC by license category. Data owners may also choose to specify additional terms of use, if necessary, in a separate field.

JeSLIB 2021; 10(2): e1191
https://doi.org/10.7191/jeslib.2021.1191

I developed an Optional Feature Use Score (OFUS) rubric (Table 1) to indicate the presence or absence of an optional feature. For each optional feature, on a per dataset basis, I assigned one point if the use of the feature was present, and zero points if it was not; a perfect score for a dataset is 5 points.

Table 1: Optional feature use scoring.

Variable	**Description**	**Score**
has_OMB	Dataset uses of one or more optional metadata blocks	0 or 1
has_KW	Dataset uses one or more keywords	0 or 1
has_DESC	Dataset has a description	0 or 1
has_PSF	Dataset includes additional supplemental files	0 or 1
has_ATOU	Dataset has additional terms of use	0 or 1
	Perfect Score	5

The Dataverse database records granular information about datasets, including the type and affiliation of the depositor; the number and type of files associated with the dataset; the number of downloads associated with individual data files; and publications associated with the dataset. The use of optional features, indicated by the OFUS, may be related to these and other dataset characteristics. In this study, I examined the relationship of the OFUS score to the following seven characteristics: 1) publication year, 2) the number of keywords, 3) presence of related publications, 4) the category of Dataverse collection or dataverse, which indicates whether the dataset was deposited by an individual researcher or a group, 5) the number of file downloads for a dataset, 6) datasets with restricted files, and 7) datasets that require a formal review before publishing (Table 2).

JeSLIB 2021; 10(2): e1191
https://doi.org/10.7191/jeslib.2021.1191

Table 2: Dataset characteristics.

Variable	Description	Type
pub_year	The year in which the dataset was first published on Harvard Dataverse	Four-digit year
keyword_count	The number of keywords in the datasets keyword metadata fields	Integer
datasets_with_relpubs	True (1) means the dataset has a value in at least one of three related publication metadata fields. False (0) means the dataset has no value in any of the three metadata fields.	0 or 1
dataverse_cat	The affiliation type for the group or individual associated with the dataset: RESEARCH_PROJECT RESEARCHERS ORGANIZATION_OR_INSTITUTION RESEARCH_GROUP JOURNAL LABORATORY DEPARTMENT TEACHING_COURSE UNCATEGORIZED	String
submit_for_review	True (1) means that depositors must have datasets reviewed by an authorized data curator before publication. False (0) means that depositors can publish datasets without a review.	0 or 1
num_file_downloads	Total number of downloads for all files associated with the dataset_id	Integer
num_restricted_files	Datasets can contain multiple data files, some of which may have restrictions on their access or use. This variable is the total number of restricted files associated with the dataset	Integer

JeSLIB 2021; 10(2): e1191
https://doi.org/10.7191/jeslib.2021.1191

Also, in consultation with the Harvard Dataverse data curation team, I developed the following seven hypotheses, each shown with their associated research question. RQ1 is satisfied by the descriptive statistics generated for the datasets in the Harvard Dataverse repository.

RQ2. What is the most frequently used optional feature?

- H1. Additional terms of use will be the least frequently used option, followed by optional metadata blocks.

RQ3. How does the use of optional features vary by key dataset characteristics?

- H2. Datasets associated with groups, such as journals and laboratories, will have higher mean Optional Feature Use Scores than those associated with individuals.
- H3. Datasets associated with publications will have higher mean Optional Feature Use Scores than those not associated with publications.
- H4. Datasets associated with groups who require review before publishing will have higher Optional Feature Use Scores than those that allow self-publishing.
- H5. Datasets with higher Optional Feature Use Scores will have more file downloads.
- H6. Datasets with higher Optional Feature Use Scores will have at least one keyword.
- H7. There is a mean difference in Optional Feature Use Scores between datasets with restricted data files and those with no restricted data files.

Methods

Data Collection, Processing, and Analysis

The raw data for the study was gathered by the Harvard Dataverse database manager, who captured snapshots of the repository inventory of published datasets and metadata on 29 and 30 October 2019. He created two preliminary datasets and a codebook describing the dataset characteristics. Next, I cleaned his two raw data files using Excel and deposited the cleaned data into an SQLite v3.11.2 database for further complex processing and preparation. I used Python v3.6 and SQL queries to create the final project file that contained: 1) summarized metadata for 29,295 Harvard Dataverse datasets and their associated 383,685 files, 2) values associated with each of the five Optional Feature Use variables, and 3) the overall Optional Feature Use Score for each dataset (Boyd 2020).

JeSLIB 2021; 10(2): e1191
https://doi.org/10.7191/jeslib.2021.1191

The final tabular file contained 29,295 records, each associated with a published dataset housed in the Harvard Dataverse. In this context, published datasets refer to only those datasets that are no longer in draft form and whose existence has been made public to the worldwide research community; these datasets have a globally accessible document object identifier (DOI). Publication status is distinct from any restrictions placed upon the use of the dataset or its files. For instance, a dataset's metadata may be public, but its owner may require that the requester complete an application before they download and use it.

The record for each dataset includes 29 variables, including metadata fields and computed values, such as the Optional Feature Use Score. All variables are defined in the project codebook shown in Appendix A. Of note are the two variables: *dataverse_cat* and *ofus*. As shown in Table 2, *dataverse_cat* denotes one of nine types of user who created and manages the dataset. This study distinguished between three categories of *dataverse_cat*: individual researchers, groups, and uncategorized. The variable *ofus* corresponds to a dataset's computed optional feature use score, ranging in value from 0 to 5.

In a final step, I analyzed the 29,295 records using SPSS MacOS v26 to investigate each of the research hypotheses using the most appropriate test, either One-Way Analysis of Variance (ANOVA), Independent Samples t test, or the Pearson product-moment correlation.

Results

Characteristics of Datasets in the Harvard Dataverse Repository

The descriptive statistics in this section address RQ1, RQ2, and H1 and describe optional feature use across all datasets by year, by dataverse category, and for datasets with, and without associated publications. Across all 29,295 datasets, the mean Optional Feature Use Score was 1.89 (M = 1.89, SD = 1.09). The most frequent score was 2, with 40.4% (n = 11,833) datasets receiving this score. In descending order of frequency, 20.7% of datasets (n = 6,059) had a score of 3; 20.3% (n = 5,936) received a score of 1; 5.6% (n = 1,655) received a 4; and 12.3% (n = 3,600) had a score of 0. Only 0.7% of datasets (n = 212) used all five optional features.

Use of specific optional features across all Harvard Dataverse datasets

RQ2 H1 posits that additional terms of use will be the least frequently used option, followed by optional metadata blocks. However, the results show that optional metadata blocks were more commonly used than supplemental files. In descending order of frequency, 84.2% (n = 24,661) of datasets included a description; 49.8.1% (n = 14,593) included one or more keywords to facilitate discovery; 28.6% (n = 8,380) used optional metadata blocks; 16.4% (n = 4,796) included at least one possible supplemental file, such as a codebook or readme; and 10.3% (n = 3,029) included an additional terms of use statement. Therefore, H1 was not supported, as shown in Table 3.

JeSLIB 2021; 10(2): e1191
https://doi.org/10.7191/jeslib.2021.1191

Table 3: Summary of results by hypothesis.

Hypothesis	**Status**
H1. Additional terms of use will be the least frequently used option, followed by optional metadata blocks	Not supported
H2. Datasets associated with groups, such as journals and laboratories, will have higher mean Optional Feature Use Scores than those associated with individuals.	Supported
H3. Datasets associated with publications will have higher mean Optional Feature Use Scores than those not associated with publications.	Supported
H4. Datasets associated with groups who require review before publishing will have higher Optional Feature Use Scores than those that allow self-publishing.	Supported
H5. Datasets with higher Optional Feature Use Scores will have more file downloads.	Not supported
H6. Datasets with higher Optional Feature Use Scores will have at least one keyword.	Supported
H7. There is a mean difference in Optional Feature Use Scores between datasets with restricted data files and those with no restricted data files.	Supported

JeSLIB 2021; 10(2): e1191
https://doi.org/10.7191/jeslib.2021.1191

Optional feature use score across all datasets by publication year

Table 4: Optional feature use score by year, *N* = 29,295.

Publication Year	*M*	*SD*	*n*	% Total
2007	2.75	1.15	601	2.1%
2008	1.8	1.04	375	1.3%
2009	0.19	0.66	3,828	13.1%
2010	2.37	1.0	705	2.4%
2011	2.17	0.9	845	2.9%
2012	2.13	1.12	695	2.4%
2013	2.02	1.04	996	3.4%
2014	2.42	0.9	1,878	6.4%
2015	2.04	0.56	5,718	19.5%
2016	2.10	0.82	3,368	11.5%
2017	2.19	1.0	2,780	9.5%
2018	2.16	0.93	4,299	14.7%
2019	2.10	0.97	3,207	10.9%

On a yearly basis, for the majority of the 13 years during which data was published in Harvard Dataverse, datasets exceeded the mean Optional Feature Use Score of 1.89. The largest number of datasets (n = 5,728) was published in 2015 (19.5%, M = 2.04, SD = 0.56) and the fewest (n = 601) in 2007 (2.1%, M = 2.75, SD = 1.15). The year with the highest mean Optional Feature Use Score was 2007 (M = 2.75, SD = 1.15). The lowest mean OFUS occurred in 2009 (M = 0.19, SD = 0.66).

Optional feature use by dataverse category

When users create a new research space for their data, called a dataverse, they are required to choose one of nine categories that best describes its affiliation: 1) research project, 2) researchers, 3) organization or institution, 4) research group, 5) journal, 6) laboratory, 7) department, 8) teaching course, or 9) uncategorized. I designated all categories, except for researchers and uncategorized, as groups for the purposes of this study. Table 5 shows the optional feature use score by dataverse category.

JeSLIB 2021; 10(2): e1191
https://doi.org/10.7191/jeslib.2021.1191

Of the 29,295 datasets, 26.4% were uncategorized (*n* = 7,802, *M* = 2.09, *SD* = 0.71), followed by 21% associated with organizations or institutions (*n* = 6,156, *M* = 2.26, *SD* = 1.01), and 20.7% published by individual researchers (*n* = 6,125, *M* = 0.73, *SD* = 0.99). Journals accounted for 17.3% of datasets (*n* = 5,080, *M* = 2.34, *SD* = 0.92), closely followed by research projects at 11.3% (*n* = 3,320, *M* = 2.07, *SD* = 0.92), with research groups a distant sixth at 2.1% (*n* = 616, *M* = 2.41, *SD* = 0.9). The remaining three categories each accounted for a very small percentage of total datasets. The teaching course category was assigned to 0.4% of datasets (*n* = 111, *M* = 2.48, *SD* = 0.78), 0.2% of datasets were associated with laboratories (*n* = 61, *M* = 1.9, *SD* = 0.83), and 0.1% were associated with departments (*n* = 24, *M* = 2.63, *SD* = 1.41).

Table 5: Optional feature use score by dataverse category, *N* = 29,295.

Dataverse Category	***M***	***SD***	***n***	**% of Total**
Department	2.63	1.41	24	0.1%
Journal	2.34	0.91	5,080	17.3%
Laboratory	1.9	0.83	61	0.2%
Organizations and Institutions	2.26	1.01	6,156	21%
Research Group	2.41	0.9	616	2.1%
Research Projects	2.07	0.92	3,320	11.3%
Researchers	0.73	0.99	6,125	20.9%
Teaching Courses	2.48	0.78	111	0.4%
Uncategorized	2.09	0.71	7,802	26.6%

Optional feature use for datasets with related publications.

A dataset may be associated with a publication, such as a journal article or conference paper. The boolean variable datasets_with_relpubs indicates that the dataset may have a related publication. A value of 1 indicates that the dataset has a value in at least one of three related Dataverse publication metadata fields. In contrast, a 0 indicates that it does not have a known associated publication. Most datasets do not have a related publication (77.0%, *n* = 22,553) and have a slightly lower mean Optional Feature Use Score (*M* = 1.84, *SD* = 1.17) than the population mean of 1.89. Datasets with related publications (23%, *n* = 6,742) have a higher mean OFUS (*M* = 2.07, *SD* = 0.71) than the population.

JeSLIB 2021; 10(2): e1191
https://doi.org/10.7191/jeslib.2021.1191

Findings

The results in this section address the hypotheses associated with *RQ3: How does the use of optional features vary by key dataset characteristics?*

H2. Datasets associated with groups, such as journals and laboratories, will have higher mean Optional Feature Use Scores than those associated with individuals.

To compare the mean Optional Feature Use Score for the nine dataverse categories, I used a One-Way Analysis of Variance (ANOVA), with a *p-value* of 0.05. The results of the test indicated a significant difference across dataverse categories, $F(8, 29{,}286) = 1{,}650.92$, $p < 0.05$, $\eta 2 = 0.31$. I performed the Dunnett's C follow-up procedure to assess pairwise differences among the nine dataverse categories. The results indicated that the mean for researchers was significantly different ($p < 0.05$) from all other categories, including teaching courses and uncategorized, and therefore, the results support H2 (Table 3). Table 6 shows the mean OFUS for all dataverse categories. The mean OFUS was highest for departments ($M = 2.63$, $SD = 1.41$) and lowest for individual researchers ($M = 0.73$, $SD = 0.99$).

Table 6: Mean difference between dataverse categories. $N = 29{,}295$, $p < 0.05$.

	SS	***df***	***MS***	***F***	***p***
Within Groups	10,750.42	8	1,343.8	1,650.92	0.0
Between Groups	23,837.94	29,286	0.81		
Total	34,588.36	29,294			

H3. Datasets associated with publications will have higher mean Optional Feature Use Scores than those not associated with publications.

I performed an Independent Samples t test to investigate the OFUS for datasets without related publications. The test revealed that the equality of variances could not be assumed and that datasets without related publications ($n = 22{,}553$, $M = 1.84$, $SD = 1.17$) had significantly lower OFUS than datasets with related publications ($n = 6{,}742$, $M = 2.07$, $SD = 0.71$), $t(18{,}462.61) = -20.14$, $p < 0.05$, therefore H3 was supported.

H4. Datasets associated with groups who require review before publishing will have higher Optional Feature Use Scores than those that allow self-publishing.

An Independent Samples t test revealed that the equality of variances could not be assumed, and the Optional Feature Use Score for datasets that do not require a review before submission ($n = 3{,}357$, $M = 1.69$, $SD = 0.72$) is significantly lower than datasets that do require review ($n = 25{,}938$, $M = 1.92$, $SD = 1.12$), $t(5{,}724.58) = -15.94$, $p < 0.05$, therefore H4 was supported.

JeSLIB 2021; 10(2): e1191
https://doi.org/10.7191/jeslib.2021.1191

H5. Datasets with higher Optional Feature Use Scores will have more file downloads.

Hypothesis H5 posits that datasets with higher Optional Feature Use Scores will have more downloads. However, the results of a Pearson product-moment correlation analysis indicated no significant relationship between these factors, $r(29{,}293) = 0.007$, $p > 0.05$. Therefore, H5 was not supported.

H6. Datasets with higher Optional Feature Use Scores will have at least one keyword.

A Pearson product-moment correlation analysis indicated a significant correlation between Optional Feature Use Score and the presence of at least one keyword used to describe datasets, $r(29{,}293) = 0.62$ $p < 0.01$, therefore H6 was supported.

H7. There is a mean difference in Optional Feature Use Scores between datasets with restricted data files and those with no restricted data files.

The results of an Independent Samples t test, with equal variances not assumed, support H7. The Optional Feature Use Score for datasets that do not have restricted files (n = 26,798, M = 1.8, SD = 1.06) is significantly lower than datasets that have restricted files (n = 2,497, M = 2.85, SD = 0.89), $t(3{,}197) = -55.54$ $p < 0.05$.

Discussion

The results of this study indicate that five of the seven hypotheses were supported (Table 3). The general impressions of the Harvard Dataverse data curation team—that groups use optional features more often than individual researchers—were borne out by the statistical analyses. Similarly, their expectations that the datasets in Harvard Dataverse reviewed before submission, and those with an associated publication, such as a journal article, also make significantly higher use of optional features. The smaller sample size for Koshoffer et al. (2018) (N = 80) compared to this study (N = 29,295) could help to explain why their findings indicated that the level of data curation support did not strongly influence optional metadata contributions.

Hypotheses H1 and H5 were not supported. For H1, instead of additional terms of use, the least common optional feature was the inclusion of supplementary files. The results of H1 may be influenced by the approach that I used to identify supplementary materials such as codebooks and readme files. I used regular expressions in an SQLite query to search for occurrences of "code book," "codebook," and "readme" in the name of any file associated with a dataset. This approach will miss cases whose filenames do not reflect their contents, thereby undercounting the number of supplementary files and artificially lowering affected datasets' Optional Feature Use Scores.

JeSLIB 2021; 10(2): e1191
https://doi.org/10.7191/jeslib.2021.1191

In the case of H5, the results suggest that data seekers choose to download a dataset even if the research data context supplied by optional features is not present. Possibly the information provided by optional fields was extraneous for most data seekers. Alternatively, the behavior might be the result of data seekers using the Harvard Dataverse to locate known datasets (e.g., associated with an article they have read), rather than browsing for potential datasets to reuse. Additional research would be needed to investigate these or other possible explanations.

Finally, my analyses treated datasets associated with groups, individual researchers, and uncategorized datasets as distinct from one another. However, without manually inspecting each dataset it is not possible to determine how many datasets in the uncategorized category were deposited by either individuals or groups. The Harvard Dataverse User Guide neither provides formal definitions for the nine dataverse categories nor enforces their assignment. Therefore, users may interpret the meaning and scope of dataverse categories differently which affects the overall number of individual- or group-curated datasets and influencing the overall Optional Feature Use Score for each category. For instance, I performed a brief inspection of several uncategorized dataverses and noted a mix of Dataverses belonging to organizations as well as individual researchers. Additionally, a single researcher conducting a long-term research study might choose the research project category thereby reducing the number of dataverses associated with individual researchers.

The study's results indicate that members of groups make greater use of Harvard Dataverse's optional features than individual researchers. It is possible that groups, such as research institutes and publishers, are either more experienced or more frequent users of Harvard Dataverse and have a greater familiarity with its optional features. Alternately, their use of optional features may indicate the importance they place on the value of these specific data curation features.

Datasets that undergo a data curation review before publishing or are associated with a publication such as a journal article also make greater use of Harvard Dataverse optional features. Two possible explanations are: 1) groups are more likely to have workflows involving dataset review and 2) data depositors may populate optional fields, such as keywords or descriptions, with metadata from their publications.

In addition, datasets with restricted files use more optional features than those without restricted files. This may be the result of data curators attempting to mitigate the risk of inappropriate sharing by more thoroughly describing them or by applying additional terms of use.

Overall, the study results imply that there is a relationship between the attention that a dataset receives prior to its publication and the presence of characteristics that signal good research data curation and FAIRness. They also emphasize the value of experienced data curators to the repositories, journals, and other

JeSLIB 2021; 10(2): e1191
https://doi.org/10.7191/jeslib.2021.1191

organizations that employ them and to the current and future research community.

There are several key takeaways for managers of self-service research data repositories. First, the results show that individual depositors may benefit from increased outreach and training about data curation practices and repository features. Specific recommendations for the Harvard Dataverse repository include improving user documentation about optional metadata blocks, keywords, datasets descriptions, supplemental files, and supplying additional terms of use; more widely promoting these features; and offering feature-based training, such as brief videos, to individual users.

Next, the fact that optional data curation feature use was greater for datasets that have related publications implies that self-service repositories may encourage better data curation by more clearly communicating the value of adding bidirectional links between datasets and related articles to their users.

Finally, repository managers might apply the quantitative, rubric-based approach I have demonstrated here to analyze the characteristics of their own dataset inventories. These analyses may help repository staff to better understand how their user communities use optional repository features and generate ideas for repository improvement including new features, smarter defaults in optional fields, and better documentation and training.

Limitations

As noted in the Discussion section, the approach used to identify supplementary documentation may lead to artificially low Optional Feature Use Scores. The original SQL query used by the Harvard Dataverse database manager to identify publications related to the dataset may also have missed cases where data depositors used the dataset description field to mention a publication. If so, there are more than 6,742 datasets with related publications in the total population. Finally, in 2015, the dataset description field was made mandatory in the user interface. Therefore, post-2015 Optional Feature Use Scores may be influenced by the switch from optional to mandatory. However, the field is not mandatory for deposit through the Harvard Dataverse API, which may also confound the results.

Conclusion

This study used quantitative methods to investigate how different groups of depositors vary in their use of optional data curation features that provide support for FAIR research data in the Harvard Dataverse repository.

Its results indicate that members of groups make greater use of Harvard Dataverse's optional features than individual researchers. Additionally, datasets that undergo a data curation review before submission to Harvard Dataverse, are associated with a publication, or contain restricted files also make greater use of

optional features. Overall, the study contributes to the growing literature on the relationship of data curation practices to research data repository features and research context expressed by the characteristics of their datasets.

I conclude that individual researchers might benefit from increased outreach and improved documentation about the benefits and use of optional features to improve their datasets' level of curation beyond the FAIR-informed support that the Harvard Dataverse repository provides by default. Platform designers, developers, and managers may also use the numerical scoring approach to explore how different user groups use optional application features.

A follow-up study could refine the approaches for identifying supplementary documentation and related publications and compare the results. Other avenues of investigation include analyzing the presence and number of subjects associated with datasets and their relationship with the dataverse category and other dataset characteristics such as the number of keywords or related publications. These analyses would provide insight into how data cultures of practice might influence the use of data curation optional features.

Acknowledgments

I thank Mercè Crosas and members of the Harvard Dataverse data curation team, Sonia Barbosa and Julian Gautier for making the repository and its research datasets available for this quantitative investigation.

Supplemental Content

Appendix
An online supplement to this article can be found at http://dx.doi.org/10.7191/jeslib.2021.1191 under "Additional Files".

Data Availability

Data associated with this study is publicly available in the Harvard Dataverse repository (https://doi.org/10.7910/DVN/9STGWE).

References

Austin, Claire C, Susan Brown, Nancy Fong, Chuck Humphrey, Amber Leahey, and Peter Webster. 2016. "Research Data Repositories: Review of Current Features, Gap Analysis, and Recommendations for Minimum Requirements." *IASSIST Quarterly* 39 (4): 24–24. https://iassistquarterly.com/public/pdfs/vol_39_4_austin.pdf

Bishop, Bradley, and Carolyn Hank. 2018. "Measuring FAIR Principles to Inform Fitness for Use." *International Journal of Digital Curation* 13 (1): 35–46. https://doi.org/10.2218/ijdc.v13i1.630

Boyd, Ceilyn, 2020, "Harvard Dataverse Optional Feature Use Data." Harvard Dataverse, V1, UNF:6:wLA9qsuTuWjxTrPQOhAvXg== [fileUNF]. https://doi.org/10.7910/DVN/9STGWE

"Creative Commons Licenses." 2019. Creativecommons.Org. November 30, 2019. https://creativecommons.org/share-your-work/licensing-examples

Crosas, Mercè. 2019. "The FAIR Guiding Principles: Implementation in Dataverse." Massachusetts Institute of Technology (MIT), March 22. https://scholar.harvard.edu/files/mercecrosas/files/fairdata-dataverse-mercecrosas.pdf

Dataverse Project (version 4.20). 2020. Cambridge, MA, United States: Institute for Quantitative Social Sciences, Harvard University. https://dataverse.org

Durand, Gustavo. 2020. "Dataverse's Approach to Technical Community Engagement." *Septentrio Conference Series* no. 2 (March). https://doi.org/10.7557/5.5424

Faniel, Ixchel M., Rebecca D. Frank, and Elizabeth Yakel. 2019. "Context from the Data Reuser's Point of View." *Journal of Documentation* 75 (6): 1274–97. https://doi.org/10.1108/JD-08-2018-0133

Harvard Dataverse (version 4.20). 2020. Cambridge, MA, United States: Institute for Quantitative Social Sciences (IQSS), Harvard University. https://dataverse.harvard.edu

Institution Type. 2020. Registry of Research Data Repositories. https://doi.org/10.17616/R3D

Jeng, Wei, Daqing He, and Yu Chi. 2017. "Social Science Data Repositories in Data Deluge: A Case Study of ICPSR's Workflow and Practices." *The Electronic Library* 35 (4): 626–49. https://doi.org/10.1108/EL-11-2016-0243

Johnston, Lisa R, Jacob Carlson, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Robert Olendorf, and Claire Stewart. 2018. "How Important Is Data Curation? Gaps and Opportunities for Academic Libraries." *Journal of Librarianship and Scholarly Communication* 6 (1): 2198. https://doi.org/10.7710/2162-3309.2198

Johnston, Lisa R, Jake Carlson, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Robert Olendorf, Claire Stewart, et al. 2018. "Data Curation Network: A Cross-Institutional Staffing Model for Curating Research Data." *International Journal of Digital Curation* 13 (1): 125–40. https://doi.org/10.2218/ijdc.v13i1.616

Koshoffer, Amy, Amy E. Neeser, Linda Newman, and Lisa R Johnston. 2018. "Giving Datasets Context: A Comparison Study of Institutional Repositories That Apply Varying Degrees of Curation." *International Journal of Digital Curation* 13 (1): 15–34. https://doi.org/10.2218/ijdc.v13i1.632

Llebot, Clara, and Steven Van Tuyl. 2019. "Peer Review of Research Data Submissions to ScholarsArchive@OSU: How Can We Improve the Curation of Research Datasets to Enhance Reusability?" *Journal of eScience Librarianship* 8 (2): e1166. https://doi.org/10.7191/jeslib.2019.1166

McNeill, Katherine. 2016. "Repository Options for Research Data." In Making Institutional Repositories Work, edited by Burton B. Callicott, David Scherer, and Andrew Wesolek, 15–30. Purdue University Press. https://doi.org/10.2307/j.ctt1wf4drg.7

Pampel, Heinz, Paul Vierkant, Frank Scholze, Roland Bertelmann, Maxi Kindling, Jens Klump, Hans-Jürgen Goebelbecker, Jens Gundlach, Peter Schirmbacher, and Uwe Dierolf. 2013. "Making Research Data Repositories Visible: The Re3data.Org Registry." Edited by Hussein Suleman. *PLOS ONE* 8 (11): e78080. https://doi.org/10.1371/journal.pone.0078080

Pepe, Alberto, Alyssa Goodman, August Muench, Merce Crosas, and Christopher Erdmann. 2014. "How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in Aas Publications and a Qualitative Study of Data Practices among Us Astronomers." *PLOS ONE* 9 (8): 1–11. https://doi.org/10.1371/journal.pone.0104798

Repository Types. 2020. Registry of Research Data Repositories. https://doi.org/10.17616/R3D

Stall, Shelley, Martone, Maryann E., Chandramouliswaran, Ishwar, Crosas, Mercè, Federer, Lisa, Gautier, Julian, Hahnel, Mark, et al. 2020. "Generalist Repository Comparison Chart," July. https://doi.org/10.5281/ZENODO.3946720

Tayi, Giri Kumar, and Donald P Ballou. 1998. "Examining Data Quality." *Communications of the ACM* 41 (2): 54–57. https://doi.org/10.1145/269012.269021

Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. 2011. "Data Sharing by Scientists: Practices and Perceptions." Edited by Cameron Neylon. *PLOS ONE* 6 (6): e21101. https://doi.org/10.1371/journal.pone.0021101

Thelwall, Mike, and Kayvan Kousha. 2016. "Figshare: A Universal Repository for Academic Resource Sharing?" *Online Information Review* 40 (3): 333–46. https://doi.org/10.1108/OIR-06-2015-0190

Wiley, Christie. 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. https://doi.org/10.7191/jeslib.2015.1081

———. 2017. "Assessing Research Data Deposits and Usage Statistics within IDEALS." *Journal of eScience Librarianship* 6 (2): e1112. https://doi.org/10.7191/jeslib.2017.1112

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (March). http://dx.doi.org/10.1038/sdata.2016.18