# DATA CURATION IN THE TEXAS DATA REPOSITORY

## Spring 2020 Capstone Project


### Abstract

For my Capstone Project, I created a data curation workflow based on the Data Curation Network's CURATE(D) Model to improve the findability and reusability of datasets. The workflow is localized to the UT Austin Dataverse using needs identified in interviews with liaison librarians and an assessment of published datasets in the repository. My final product is a specialized Data Curation workflow and a list of recommendations that may be used by a team of liaison librarians to curate new datasets in the future.



By Brenna Wheeler

brennawheeler@utexas.edu

# Table of Contents

## Introduction to the Project

In 2017, the Texas Digital Library launched the Texas Data Repository (TDR) using an instance of Dataverse, an open source application developed by Harvard (Texas Digital Library, 2018a). Since then, researchers from 10 different Texas institutions have published over 600 datasets. Within TDR, the University of Texas at Austin Dataverse (here shortened to the UT Austin Dataverse) contains almost 300 datasets. As long as the data does not contain any sensitive or personally identifiable information, researchers can publish their datasets with little intervention to the UT Austin Dataverse. This allows for researchers to easily publish datasets to meet their deadlines. The consequence is that the metadata for these datasets is often minimal. According to the Metadata Dictionary (Texas Digital Library, 2018b), only the following metadata fields are mandatory:

- **Title**: Full title by which the Dataset is known.
- **Name**: The author's Family Name, Given Name, or the name of the organization responsible for this Dataset.
- **Affiliation**: The organization with which the author is affiliated.
- **Email**: The e-mail address(es) of the contact(s) for the Dataset.
- **Description**: A summary describing the purpose, nature, and scope of the Dataset.
- **Subject**: Domain-specific Subject Categories that are topically relevant to the Dataset.
- **Production Date**: Date when the data collection or other materials were produced (not distributed, published or archived).
- **Production Place**: The location where the data collection and any other related materials were produced.
- **Kind of Data**: Type of data included in the file.

When used correctly, these metadata fields have the potential to capture a lot of information about how and why they created the data, which would help other scholars check or reuse the data. This information also makes the dataset more discoverable by giving the search engines more metadata to work with. However, researchers do not often put this level of detail into their metadata before publishing their data to the UT Austin Dataverse.

To make the datasets within the UT Austin Dataverse more reusable and discoverable, Jessica Trelogan, Research Data Services Coordinator at UT Libraries, and I proposed a capstone project to create a localized Data Curation Team made of liaison librarians. These librarians would use their subject expertise to curate research data in their field and work with the researchers to enrich their dataset deposits. This method of data curation is based on the work of the Data Curation Network, which composes of librarians, repository managers, and subject experts assisting their fellow affiliated institutions with curating research data to meet FAIR standards (Data Curation Network, n.d.). According to Wilkinson, Dumontier, et al. (2016), FAIR refers to making data:

- **Findable** though unique identifiers in a searchable resource and rich metadata.
- **Accessible** by using an open, standardized communications protocol to retrieve data (with authorization procedures when needed). This can also include making the metadata accessible when the data is not available.

- **Interoperable** using formal, shared, and broadly applicable language for knowledge representation with vocabularies that follow FAIR principles and references to other metadata.
- **Reusable** by describing data with plenty of accurate and relevant attributes. This includes a detailed provenance, meeting domain-relevant community standards, and having a clear and accessible data usage license.

Currently, TDR fulfills many FAIR Principles by providing identifiers, a search engine, a communications protocol, an interoperable metadata schema, and associated data usage licenses. (For more information on this, readers can view the "Good Practice Principles for Scholarly Communication Services: A Self-Assessment by the Texas Digital Library" here.) This capstone project will focus on meeting FAIR Principles by enriching metadata.

The main goal for this project was to create a data curation workflow based on the Data Curation Network's CURATE(D) model (2018). A team of data curators, composed of the TDR Manager and a team of volunteer liaison librarians, would use this workflow to curate research data in their subject specialty. To investigate this approach, I worked with liaison librarians from the STEM and Social Sciences Engagement Team to understand the needs of our librarians and researchers. Together, we tested the CURATE(D) framework using checklists and tools recommended by the Data Curation Network and the Dataverse Developers on published datasets in the UT Austin Dataverse. Then, I used our findings to refine the generic framework for adoption by our local team and to train others in the curation activities.

We divided the project into three main tasks:

> The First Task focused on identifying and testing tools to help with assessing and identifying curation needs. For this project, I looked at existing APIs enabled in the Texas Data Repository (Dataverse Project, 2020) and Dataverse Data Curation Tool (Scholars Portal, 2020), which any institution could be able to use. I also investigated scripted approaches, but they pulled insufficient metadata for a complete assessment.
>
> The Second Task assessed published datasets in the Texas Data Repository to identify curation problems and reoccurring metadata issues. Curation problems could include unopenable files, file format conversion, or inappropriate data sets (e.g. a PowerPoint or a graph as opposed to the data underlying it). Metadata issues could include a lack of descriptive information about data provenance and process history, a lack of information about acceptable data use, or a lack of metadata to facilitate discovery.
>
> The Third Task involved hosting a training session on the new workflow to all interested librarians. Discussions and feedback from the training session would allow the workflow to be further developed in addressing the needs of librarians and researchers. The session would also ensure that the workflow made sense and could easily be followed by librarians, whose experience ranged from "receiving one or two questions about locating research data" to "supporting a large-scale research data management plan for a grant-funded project".

The planned outcomes and deliverables for the project included:

- A workflow for assessing current deposits and for curating future deposits in TDR.
- Resources to facilitate future curation.

- A training/sandboxing session with liaison librarians to test out the curation workflow.
- Recommendations for UT Libraries' deposit model and data curation activities.

## First Task – Tools & Utilities

The First Task focused on tools that would allow data curators to view and manipulate metadata in the UT Austin Dataverse. For this task, I looked at the Dataverse API (Dataverse Project, 2020) and the Data Curation Tool (Scholars Portal, 2020). In the end, I found these tools to be better suited for TDR Managers, who might be working with metadata in large batches. Liaison librarians are not likely to need these types of tools for their curation process since their data curation process would be working with each dataset individually.

### Dataverse API

I began working with the Dataverse API in October 2019, when Jessica Trelogan and I were first playing around with data curation workflows as an Ask-A-Librarian Graduate Research Assistant project. The goal was to find a way to view all the dataset metadata in the UT Austin Dataverse. I experimented with APIs by entering them into the address bar in Mozilla FireFox, which returned results in easy-to-read JSON.

The Dataverse API tool is includes five different APIs (Dataverse Project, 2020):

- **Search API**: most useful for searching through datasets and Dataverses; used in Command Line or FireFox.
- **Data Access API**: access and downloading individual files; more useful with Command Line.
- **Native API**: Dataverse and Dataset level access and manipulation; can use FireFox for viewing, but only Command Line can do manipulation.
- **Metrics API**: provides metrics of number of datasets or Dataverses in each category, how many downloads, and so on; can use FireFox or Command Line.
- **SWORD API**: allows users to remotely deposit files and metadata into a Dataverse; only uses Command Line.

After some experimentation, I got the following APIs to work:

**Metric API**
https://dataverse.tdl.org/api/info/metrics/datasets/bySubject
https://dataverse.tdl.org/api/info/metrics/dataverses/byCategory

**Search API**
https://dataverse.tdl.org/api/search?key=$apikey&q=*&type=dataset
https://dataverse.tdl.org/api/search?key=$apikey&q=Austin&type=dataverse&sort=description

**Native API**
https://dataverse.tdl.org/api/dataverses/root/contents
https://dataverse.tdl.org/api/dataverses/49/contents

In November 2019, the Dataverse Project added the “subtree” parameter to the Dataverse API, which allowed users to search in a specific Dataverse and its children Dataverses. However, the “subtree” parameter searched by Dataverse alias, rather than the unique Dataverse identifier, which is more common in other parameters. I was unable find the alias on the Dataverse front end, so I pulled the information from the API. I first used the following query from the Native API:

https://dataverse.tdl.org/api/dataverses/root/contents

This gave me the name, the type, and the ID for all the Dataverses at the University level (i.e. UT Austin's instance, Texas A&M University's instance, University of Houston's instance, etc.). So, I tried the Search API to see if that would pull more metadata. I ran one of the Search API queries I tried in October:

https://dataverse.tdl.org/api/search?key=$apikey&q=Austin&type=dataverse&sort=description

This kicked back 113 results, of which I could only see 10 at a time. I added "&per_page=113" onto the end to see all the results at once. This method gave me name, type, URL, identifier, description, and published_at elements for each result, and the alias I needed was under identifier. The successful API ended up being:

https://dataverse.tdl.org/api/search?key=$apikey&q=*&subtree=utexas&type=dataset

Here's a breakdown of the API query according to the Dataverse Project's API Guide (2020):

- **key**: an API Token needed for access (anyone who wants to use this query would need to replace $apikey with their own API Token).
- **q**: short for "query". This is a requirement to use the Dataverse Search API. If a user is not using it to narrow down search results, they can use an asterisk (*) to pull all results.
- **Subtree**: narrows it down by Dataverse (here, I'm using the UT Austin Dataverse).
- **Type**: can focus on dataset or Dataverse. I have it set to dataset to pull the dataset information.

This query allowed me to pull information about all the Datasets in the UT Austin Dataverse. However, the tool does not currently pull all metadata information needed to make an assessment. The JSON only allows the name, type (Dataverse or dataset), URL, global ID (DOI), description, publication date and time, citation, Dataverse identifier, Dataverse name, and the authors. The user would still need to look at the dataset in the front-end repository to curate the data.

## Dataverse Data Curation Tool

At the time of investigation, there were three options for using the Data Curation Tool:

1. In GitHub Pages (not recommended by the creators, but an option for testing; still requires downloading a .json file to the Dataverse server).
2. Inside Dataverse.
3. Inside a separate webserver.

Due to my inability to download items into the Dataverse server, I was unable to test the Dataverse Data Curation Tool. Instead, I relied upon the documentation provided on the software's GitHub (Scholars Portal, 2020). According to the ReadMe, the tool uses Command Line to edit and view "variable level" metadata. Users can also save changes to the metadata to the Dataverse and export metadata as XML. Looking at the documentation, the tool appears to be better suited for editing large batches of metadata (such as all the datasets in a specific Dataverse), so it may be helpful for standardizing specific metadata elements. However, the tool may be a bit much for evaluating one dataset at a time. It might be easier for curator to manually review each dataset to make a proper assessment.

## Second Task – The Assessment

The end-goal of the Data Curation Workflow is to improve the reusability and discoverability of datasets. To identify current curation needs, I assessed 295 published datasets in the UT Austin Dataverse, which was the total number of datasets in March 2020. As I assessed these datasets, I took note of patterns in the metadata and identified areas in need of improvement. These observations helped me refine the Data Curation Workflow.

### Creating the Assessment

In the first version of the assessment, I reviewed only a small sample of the datasets. There were only 290 datasets in January 2020, and I found that a good sample size was 166 by using a calculator from SurveyMonkey (2020). The original schema for the assessment included the following elements:

- Name
- DOI
- Dataverse
- Subject(s)
- Number of Files
- File type(s)
- Open [Yes/No]
- Need Update [Yes/No]
- Need Conversion [Yes/No]
- Rich Metadata
    - Dataset [Yes/No]
    - File [Yes/No]
- Documentation [Yes/No]
- Understandable [Yes/No]
- Keywords [Yes/No]
- Link to Publication [Yes/No]
- Link to Related Datasets [Yes/No]
- Link to Source Data [Yes/No]

This was a decent start. As indicated above, most of the elements were simple “Yes” or “No” answers, which would allow for easy quantitative analysis. Later, I added “Processing Details”, “Readme”, and “Software Used” to the elements list to better capture the richness of the metadata. During the assessment, I eliminated “Open”, “Update”, and “Convert” due to the large number of files. I added a column for “Texas ScholarWorks” to identify items suited for ingestion into Texas ScholarWorks. Using the Dataverse API, I pulled up all the datasets in UT Austin Dataverse using the successful Search API query on page 5. I selected each dataset by using a random number generator and pulling the corresponding result number.

After a month of using the above schema, I found it unable to accurately capture the richness of the metadata. For example, a dataset might describe how its creation and give no details about any manipulation or cleaning processes. Another dataset might explain the software used to manipulate the data and not describe how or why the data was collected. The old schema does not depict how rich each

data set is and in which aspect the dataset is rich in. I came up with a new schema for assessing the metadata richness:

- ID
- Name
- URL
- Date
- Dataverse
- Subject(s)
- # of Files
- File Type(s)
- Documentation [Yes/No]
- File Description [Yes/No/Some]
- Keywords [Yes/No]
- Controlled Vocab [Yes/No]
- Kind of Metadata
- Additional Schema
- Richness
    - Creation Process
    - Manipulation Process
    - Software Used
    - Purpose
    - What is Data?
- Publication Citation [Yes/No]
- Publication Status [Yes/No]
- Data Source
- Texas ScholarWorks? [Yes/No]
- Type for TSW

With this schema, I used RStudio to create a script that would pull the Name, URL, and Date for all datasets in the UT Austin Dataverse and arrange it into an CSV. (I included this code in Appendix A – R Code for converting API to JSON). Since the API pulled the DOI URL, I could click the URL and pull up the dataset. This made the process much simpler than randomizing, searching, and then pulling up the proper item. Using this CSV, I assessed all 295 datasets in Microsoft Excel.

The largest challenge I faced during the assessment was figuring out how to represent metadata richness. At first, I used a simple yes or no, due to a lack of knowledge of datasets in the UT Austin Dataverse. After a few days, I realized this would not depict the richness of the metadata, since each dataset has different information needs. For example, a dataset with images of single-celled organisms might need technical information about microscopes, imaging software, and photography to show the entire data creation process. In contrast, a dataset of interview responses might only need information about the software used for data analysis.

For this assessment, I created a “Metadata Richness Scale” using the following criteria:

1. Does the metadata mention the data creation process?

2. Does the metadata mention any manipulation or processing of the data?
3. Does the metadata include the software used for data creation, manipulation, or cleaning?
4. Does the metadata say why the data was collected?
5. Does the metadata explain what the data is or represents?

If the answer to a question was "Yes", the dataset gained 1 point on the Metadata Richness Scale, which ranged between 0 points and 5 points. Two major problems remained:

1. Complex technical jargon made it difficult to find answers to the above questions, especially since I was reviewing almost 300 datasets in fields where I had no previous experience or familiarity.
2. Many datasets were code used to manipulate data, but not the raw data itself.

For these situations, I tried to give points only to those that clearly answered the questions above, but the technical jargon still made some datasets difficult to assess. If a repository would like to adapt the richness scale for their own assessments, they will need create standards for the level of clarity and detail. For example, some datasets would explain why the creators were uploading data to the Dataverse, but they would not explain the reason why they collected the data in the first place. Would this be sufficient to answer Question 4? Some datasets would mention using a software, but they would not specify if the software generated, processed, or manipulated the data. Would this be enough for Question 3? Would that answer change depending on the research discipline? Only with established standards, the richness scale might useful in initial assessments of metadata and identifying areas in need of improvement.

## Assessment Results

Using this metadata schema and richness scale, I assessed a total of 295 datasets from the UT Austin Dataverse. I focused only on published datasets since unpublished datasets are usually in-progress or restricted. Due to technical and time restraints, I was unable to check, review, and assess the data files themselves, so I limited myself to the metadata. I created the following tables to identify what information is typically included or excluded.

Although time and technology restrictions made it difficult to open and examine every file, I still collected data on which file extensions were used in each dataset. Table 1 shows the most common file extensions used in the UT Austin Dataverse. The number represents the number of datasets that use that file type. Of these top 20 file types, 15 are open, and 7 are proprietary. Almost all of the proprietary software types are commonly used by a wide variety of other software (such as Adobe PDFs), with the only exception being m, which is used by MATLAB. This is great for digital preservation, since open or interoperable file types makes the files easier to access and preserve. This could be the result of the Data & Donuts workshop series by UT Libraries' Research Data Services, which discusses and revisits file types throughout the series.

| File Extensions | Number | Software |
|---|---|---|
| **pdf** | 121 | Proprietary |
| **zip** | 67 | Open |
| **txt** | 64 | Open |
| **jpg** | 63 | Open |
| **xls** | 43 | Proprietary |

| | | |
|---|---|---|
| **xml** | 40 | Open |
| **mp4** | 38 | Open |
| **csv** | 32 | Open |
| **tab** | 24 | Open |
| **docx** | 18 | Proprietary |
| **xlsx** | 18 | Proprietary |
| **R** | 8 | Open |
| **wav** | 8 | Open |
| **html** | 7 | Open |
| **pptx** | 7 | Proprietary |
| **gz** | 5 | Open |
| **m** | 5 | Proprietary |
| **mat** | 5 | Proprietary |
| **Rmd** | 4 | Open |

*Table 1: Top 20 File Extensions.*

Table 2 below looks at keywords and descriptions, which helps the search engine identify relevant content. The use of controlled vocabulary might make it easier for someone familiar with a specific field to narrow their search to a topic without knowing anything else about the dataset or Dataverse. File Level descriptions allow researchers to add details about very specific pieces of data, which would contribute to the metadata richness of the whole dataset. The Dataverse requires dataset descriptions, so those were not counted here. (Their use contributes to the metadata richness assessment below.)

| Used Keywords | Number | Percentage |
|---|---|---|
| **Yes** | 147 | 49.8% |
| **No** | 148 | 50.2% |
| Used Controlled Vocab | | |
| **Yes** | 6 | 2.0% |
| **No** | 289 | 98.0% |
| Used File Level Descriptions | | |
| **Yes** | 86 | 29.2% |
| **No** | 183 | 62.0% |
| **Some (Not All)** | 25 | 8.5% |
| **No Files** | 1 | 0.3% |

*Table 2: Controlled vocabulary and file-level description used in the UT Austin Dataverse.*

The number of datasets that use keywords is almost half of the total datasets, and the number using controlled vocabulary is very low. The low number of uses could be caused by the metadata element used to indicate the controlled vocabulary being a recent addition to the repository schema.

Table 3 lists the number of datasets using various types of documentation. The highest number of documentation type indicated was a document listing the contents for the whole dataset (listed here as "Contents Document"), but almost all instances of this file type belonged to the same research center. This represents a standardized practice, rather than a common trend among datasets. Outside of this research center, the most common documentation file was a readme file.

| Documentation Type | Number | Percentage |
|---|---|---|
| **Code Book** | 2 | 0.7% |
| **Contents Document** | 55 | 18.6% |
| **Data Dictionary** | 2 | 0.7% |
| **Manual** | 2 | 0.7% |
| **Readme** | 14 | 4.7% |
| **Report** | 5 | 1.7% |
| **Report, Documentation, & Diary** | 1 | 0.3% |
| **None** | 214 | 72.5% |

*Table 3: Documentation used in the UT Austin Dataverse.*

It is important to note here that 252 datasets were bundled files, such as gz or zip. Due to time constraints, not all bundled files were not opened and evaluated. I opened a couple that showed some documentation, usually readme files. These were not included in the table above for consistency. For their own assessments, repositories will need to decide the best practice here. The ideal situation is that the metadata would briefly summarize the data creation, manipulation, and use. Then, the documentation files would go into more detail. The repository will need to decide to include the file in the bundle or to list the file separately and alongside the bundled file. Storing the file in the bundle could make it easier for the documentation to be with the dataset and avoid getting lost. Storing the file separately would make it easier for browsing and quick access.

Another important note is that occasionally datasets used "reports" as documentation. The ones included here indicated in the metadata that the report focused on the data itself. Many other datasets included the final report of the entire experiment, not only the data gathering and manipulating. I left these out of the table above. I would recommend posting these reports to the Texas ScholarWorks repository. The repository is better suited for published works. As a former Ask-A-Librarian Graduate Research Assistant, I know from experience that patrons more often use Texas ScholarWorks to locate finalized documentation and publications. Posting (or cross-posting) would increase that document's findability, and the creator can link to the dataset in the metadata.

Table 4 indicates how many datasets use specialized schemas to describe their contents, in addition to the general, basic schema. For example, the GeoSpatial schema allows researchers to enrich their metadata with geographic location information, and the Journal schema gives more detailed publication information. Usually, datasets will only include 1-3 elements from these specialized schemas.

| Schemas | Number | Percentage |
|---|---|---|
| **GeoSpatial** | 16 | 5.4% |
| **GeoSpatial & Journal** | 1 | 0.3% |
| **Life Sciences** | 9 | 3.1% |
| **Life Sciences & Journal** | 1 | 0.3% |
| **Social Sciences & Humanities** | 1 | 0.3% |
| **No Additional Schemas** | 267 | 90.8% |

*Table 4: Additional schemas used in the UT Austin Dataverse.*

Finally, Appendix B – The Richness of Metadata in UT Austin Datasets contains a graph depicting metadata richness. (The R code used to make the graph is in Appendix C – R Code for Metadata Richness Chart.) According to my assessment 109 datasets had a Metadata Richness Scale of 1, and only three

datasets had a Metadata Richness Scale of 5. Table 5 shows the number of datasets that contain specific Metadata Richness Elements. Over half of the datasets seem to describe what the data is, but all other elements seem to be rarer in use.

Unfortunately, since I am not a specialist in most of the dataset subjects I looked at, the chart in Appendix A and Table 5 may not accurately reflect the true metadata richness due to knowledge lost in field-specific, technical jargon. The goal of the assessment was to identify data curation needs, and the need to describe datasets in a manner other people can understand is an important data curation need. The chart may look very different at another institution, especially if they have different description standards and if they decide that certain subjects or data types warrant their own richness scale.

| Richness Element | Number | Percentage |
|---|---|---|
| **Creation** | 83 | 28% |
| **Manipulation** | 101 | 34% |
| **Software Used** | 87 | 29% |
| **Purpose** | 70 | 24% |
| **What is Data** | 190 | 64% |

*Table 5: Metadata Richness in the UT Austin Dataverse.*

## Summary

The assessment identified the following areas in need of improvement:

- About half the datasets are using keywords, but only 6 datasets are using controlled vocabulary.
- Over half the datasets are not using file-level descriptions, which allow researchers to add details about specific pieces of data.
- A majority of datasets do not have a separate documentation file.
- A majority of datasets do not use a subject specific metadata schema.
- Over half the datasets explain what their data is, but the creation, manipulation, software, and purpose of the data is less frequently explained.

## Third Task – The Data Curation Workflow

During the assessment, I tested various aspects of the CURATE(D) framework and took notes about local implementation. I presented a rough draft of the workflow at a Sandbox Training Session to a group of about 20 liaison librarians. During the session, I divided the librarians up into five groups and gave each group an example dataset from the UT Austin Dataverse. For each step of the workflow, I gave them five minutes to work though it in a group, and then we discussed it together before moving onto the next step. The discussions and feedback from this group helped further develop the workflow.

In addition to the training, I also discussed this workflow with the STEM and Social Sciences Engagement Team. First, I presented the general project at one of their meetings, where I took note of any comments or concerns they had about the project. Then, I sent out a questionnaire to gauge their experiences with research data management and use. Finally, several librarians volunteered to meet with me in person to discuss the questionnaire and their experiences further. I used these discussions to inform the workflow, my final recommendations, and the list of helpful resources.

For each step below, I established the main goal of the step, and then created a check list to meet that goal. Ideally, this workflow would begin when a researcher deposits their data.

### Check

Goal: to establish the presence of content, metadata, and documentation.

- What file types are there?
- Do these files open as expected?
- Is there a description for the Dataverse?
- Is there a description for the dataset?
- Are there descriptions for each individual file in the dataset?
- What type of documentation is there?

### Understand

Goal: to ensure that the quality of the data, metadata, and documentation is high enough to facilitate reuse.

- Does the metadata explain what the data is?
- Does the metadata explain the purpose for creation and collection of the data?
- Does the documentation and/or metadata mention the software or hardware used?
- Does the documentation and/or metadata explain how the data was created?
- Does the documentation and/or metadata explain how the data was cleaned, processed, or manipulated?
- Does the documentation and/or metadata explain how the data is structured?

### Augment

Goal: to enhance and structure metadata for better discoverability and findability.

- Does the file names, description, or organization need development?
- Are there any keywords?
- Is there a controlled vocabulary the depositor could follow?

- Are there links to any related publications?
- Are there links to any related datasets?
- Are there links to the source data?
- Could any of these files be moved or copied to the Texas ScholarWorks Repository?

## Transform

Goal: to identify format, the restrictions of that format, and a non-proprietary equivalent for transformation.

- Can you identify the file formats?
- Can you identify the software used to create or access the files?
- Can the files be transformed into an open, non-proprietary format without losing data?

## Evaluate

Goal: to ensure that the dataset meets FAIR requirements.

- Does the metadata exceed author, title, and date?
- Is the data free to access in open, non-proprietary formats?
- Does the data use an interoperable, standardized schema?
- Does the metadata have a sufficient description?
- Are the creators, owners, and stewards of the data listed?

## Request

Goal: to work with the dataset owners to improve the dataset.

- Do you have 3-4 suggestions for improving the dataset?

## Document

Goal: to record necessary information for what happened to the dataset.

- Is this stored in the proper place?
- Can it be accessed by the dataset owner, curators, or users?
- Does it show detailed documentation of changes made to the dataset and why?

# Recommendations

At the conclusion of this project, I offer the following recommendations:

## Data Curation Implementation

- Add Dataverse Introduction videos created by Texas Digital Library to UT Libraries LibGuides.
- Host workshops for librarians on Research Data Management, Research Data Curation, and utilizing Texas Data Repository (especially for new features or updates).
- Establish limits of librarian involvement in the creation and publication of datasets and associated metadata. (Can researchers request that librarians describe and catalog their data for them? Can they request feedback on metadata for unpublished datasets? At what point should data owners be required to credit the librarian with data stewardship?)
    - Create procedures or guidelines for large research projects that may need additional librarian involvement beyond following the Data Curation Workflow and providing feedback. These procedures or guidelines should also ensure that librarians receive the proper amount of credit for their work (similar to being credited for their authorship or assistance with systematic reviews), as well as protect from being overwhelmed by student requests and additional projects. These guidelines should be made available either to the public or to relevant research groups.
- Create a statement for researchers to review that details the process and limits of librarian involvement. This statement could be part of the publishing process or available in another, easily accessible location.
- Solicit volunteers among liaison librarians to serve on the data curation team. Have the data curation team meet as a group for the first few months. After that, meetings should still be regular to discuss dataset problems, but frequency can depend on comfort levels.
- Project Management software (such as Trello, which is specifically made to organize checklists) should be used to manage the curation of new datasets.
- Create a general Research Data Services email for all Research Data related questions to be managed in shifts by Research Data Services and the Data Curation Team. General questions can be answered by the person on shift, while specialized questions can be forwarded to the subject librarian or to the data curator for those types of datasets.
    - If no project management software is identified for proper use, this email could be used for notifications about when datasets have been published. The notification email could then be sent to the proper librarian for curation.
- Work with a GRA to re-assess UT Austin Dataverse datasets each summer to identify progress, areas for improvement, and potential changes to the Data Curation Workflow to facilitate this improvement. Assessments and findings should be shared with Research Data Services and Data Curation Team for discussion, and solutions could be implemented before or at the start of the fall semester.

## Working with Researchers

- Create LibGuides or focus a Data & Donuts workshop on the proper method of describing and documenting datasets. LibGuides will allow librarians to link additional subject-specific tutorials

and show examples of good documentation. They can also provide links to recommended controlled vocabulary. Data & Donuts workshops are well advertised and could be used to reach a more general audience.

- Host subject-specific Research Data workshops. These should be arranged with the liaison librarian and a department (or researcher) to build relationships and attendance. *
- Meet and curate datasets with larger research groups. The datasets from large research centers appear to be very similar, so discussing datasets directly with them will improve many current and future datasets. *

* Try to involve the subject specialist librarian if they are not active in Research Data Services or the Data Curation Team.

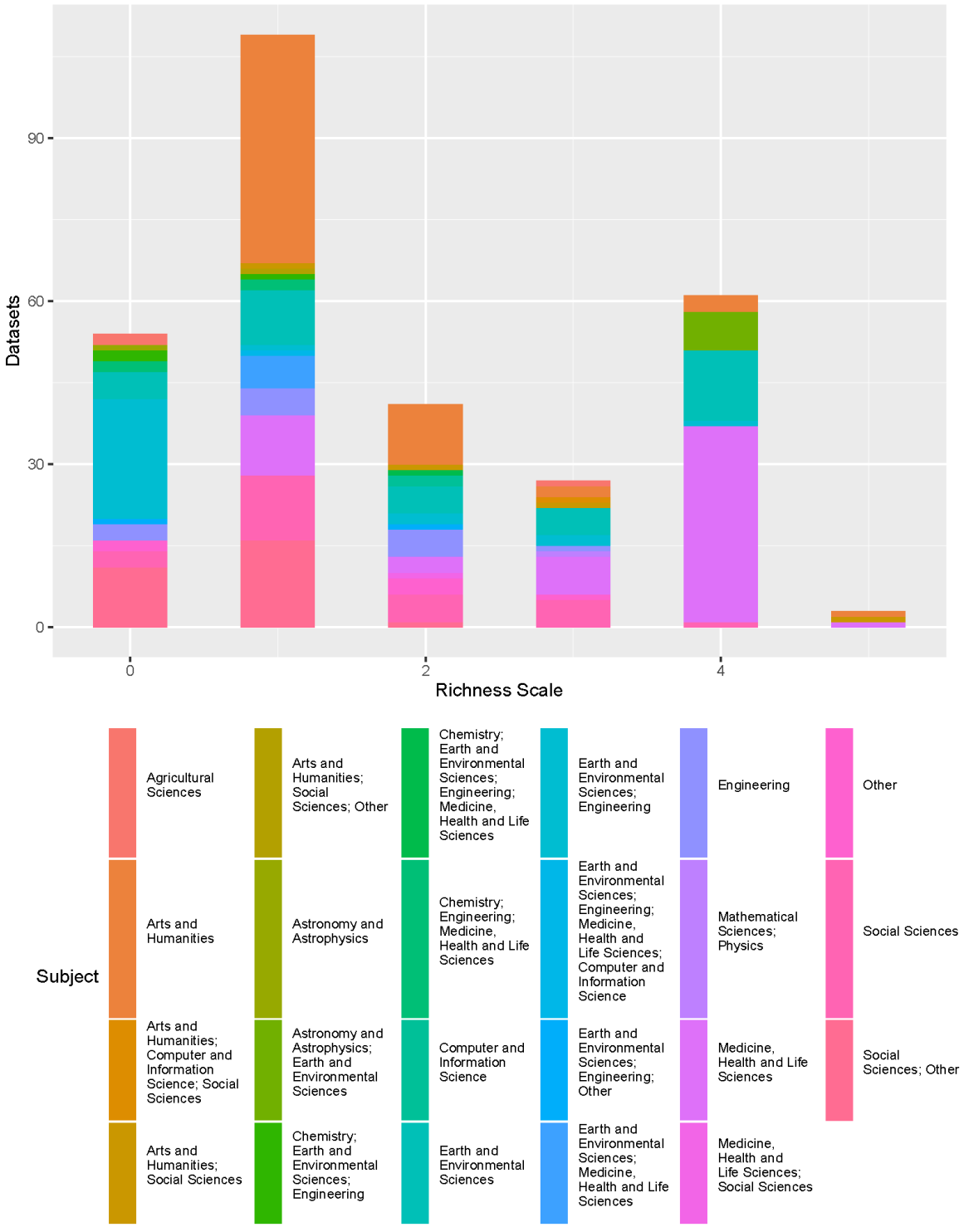
## Texas Data Repository

- Create a centralized location for TDR Members to share data curation workflows, strategies, and resources.
- Allow Publication Conformation emails for specific institutions to state that a librarian will be curating the data and will provide feedback by the institution's established turn-around time.
- Add general metadata guidelines to the "Uploading and Sharing Your Data" page in the Users Guide. These do not need to be required or structured a specific way. It can be as open as "We recommend that this type/piece of information should be somewhere in your metadata…". Although this is well-detailed in the Metadata Dictionary, the additional reminder may help the researcher while they are walking through the uploading and publication process.
- If there is enough interest, create a Data Curation Committee or Working Group to create and manage policies and practices across the consortium.
- Add "Data Curator" or "Data Steward" element to credit librarians involved in the data curation process.
- Give support to procedures and guidelines around crediting librarians for their work or protecting librarians from overwhelming workloads associated with data curation.

# Appendix A – R Code for converting API to JSON

```
#load library
library(httr)
library(lubridate)
library(plyr)

#pull data using Dataverse Search API replace $token with your TDR token replace $dataverseID with
#the ID of the dataverse you're pulling from
#if the number of datasets you are pulling exceed 10 add "&per_page=###" to the end of the string
#within the quotes and replace ### with the number of results
TDR_J <- GET("https://dataverse.tdl.org/", path =
"api/search?key=$token&q=*&subtree=$dataverseID&type=dataset")

#put data into dataframe
#replace "###" in "nrow = ###" with the number of datasets below
content(TDR_J)
test <- content(TDR_J)
dsets_J <- test[["data"]][["items"]]
test2 <- lapply(dsets_J, `[`, c('name', 'url', 'published_at'))
df <- data.frame(matrix(unlist(test2), nrow = ###, byrow = TRUE), stringsAsFactors=FALSE)

#clean data
df$date <- format(as.Date(df$X3), "%Y-%m")
df2 <- rename(df, c(X1 = "name", X2 = "url", X3 = "published_at"))

#export to My Documents
write.csv(df2, "datasets.csv")
```

# Appendix B – The Richness of Metadata in UT Austin Datasets



# Appendix C – R Code for Metadata Richness Chart

```
#load libraries
library(readxl)
library(ggplot2)
library(stringr)

#import Excel Data (insert your own filepath)
datasets <- read_excel("FilePath")

#clean up legend (subject names were really long)
datasets$Subject <- str_wrap(datasets$Subject, width = 15)

#subject bar graph
ggplot(datasets, aes(x = Richness, fill = Subject)) +
        geom_bar(width = 0.5) +
        theme(plot.margin = unit(c(1,1,1,1), "cm"),
                legend.position = "bottom",
                legend.text = element_text(size = 8, margin = margin(b = 10))) +
        guides(fill = guide_legend(ncol = 6)) +
        labs(x = "Richness Scale", y = "Datasets")
```

# Works Cited


Data Curation Network. (2018). *CURATE(D) Workflow*. Retrieved from Data Curation Network: https://datacurationnetwork.org/resources/resources-2/

Data Curation Network. (n.d.). *Mission*. Retrieved from Data Curation Network: https://datacurationnetwork.org/about/

Dataverse Project. (2020). *API Guide*. Retrieved December 28, 2019, from Dataverse: http://guides.dataverse.org/en/latest/api/index.html

Scholars Portal. (2020). *Dataverse Data Curation Tool*. Retrieved December 28, 2019, from GitHub: https://github.com/scholarsportal/Dataverse-Data-Curation-Tool

SurveyMonkey. (2020). *Sample Size Calculator*. Retrieved January 22, 2020, from SurveyMonkey.com: https://www.surveymonkey.com/mp/sample-size-calculator/

Texas Digital Library. (2017). *Announcing the official launch of the Texas Data Repository*. Retrieved April 16, 2020, from Texas Digital Library: https://www.tdl.org/2017/01/announcing-official-launch-texas-data-repository/

Texas Digital Library. (2018a). *About*. Retrieved April 16, 2020, from Texas Data Repository User Documentation: https://texasdigitallibrary.atlassian.net/wiki/spaces/TDRUD/pages/289144946/About

Texas Digital Library. (2018b). *Metadata Dictionary*. Retrieved April 16, 2020, from Texas Data Repository User Documentation: https://texasdigitallibrary.atlassian.net/wiki/spaces/TDRUD/pages/493551668/Metadata+Dictionary

Wilkinson, M. D., Dumontier, M., & et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*(1). doi:https://doi.org/10.1038/sdata.2016.18

# Additional Resources

## Data Repositories

Texas Data Repository Training Series.

Texas Data Repository User Documentation.

Examples of Other Repositories.

## Data Curation

### Readings


Inter-University Consortium for Political and Social Research (2012). Guide to Archiving Social Science Data for Institutional Respositories. 1st ed. Online here.

Johnston, Lisa R. (Ed.) (2017). Curating Research Data: Practical Strategies for Your Digital Repository. Vol. 1. Chicago: Association of College and Research Libraries. In UT Libraries Catalog here. Online here.

Johnston, Lisa R. (Ed.) (2017). Curating Research Data: A Handbook of Current Practice. Vol. 2. Chicago: Association of College and Research Libraries. In UT Libraries Catalog here. Online here.

Kowalczyk, Stacy T. (2018). Digital Curation for Libraries and Archives. Santa Barbara, CA: Libraries Unlimited. In UT Libraries Catalog here.

Texas Digital Library (2020). Good Practice Principles for Scholarly Communication Services: A Self-Assessment by the Texas Digital Library. Available for here.

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18


### Online Resources

Data Curation Network: CURATE(D), Data Primers

UK Data Archive

## Research Data Management

### Readings


DataOne (n.d.) "Primer on Data Management: What you always wanted to know (but were afraid to ask)." Retrieved from DataONE.org. Online here.

MozillaScience (2016). "Planning for Data Reuse Checklist". Retrieved from Mozillascience.github.io. Online here.

UK Data Archive (2011). “Managing and Sharing Data: Best Practices for Researchers.” Retrieved from UK Data Services. Online here.

## Online Resources

Australian National Data Service

Choose a License

DataONE: Education, Best Practices

Digital Management Short Course for Scientists