Protocol for Systematic Analysis Data Elements

Anushka Sheoran[1], Maryann Martone[1], Abel Torres-Espin[2,3,4]

[1]University of California, San Diego; [2]University of Waterloo;
[3]University of California San Francisco; [4]University of Alberta

A

Anushka Sheoran

JAN 15, 2024

OPEN ACCESS

DOI:
dx.doi.org/10.17504/protocols.io.5qpvo3wodv4o/v1

Protocol Citation: Anushka Sheoran, Maryann Martone, Abel Torres-Espin 2024. Protocol for Systematic Analysis Data Elements. **protocols.io** https://dx.doi.org/10.17504/protocols.io.5qpvo3wodv4o/v1

Protocol status: Working

Created: Dec 05, 2023

Last Modified: Jan 15, 2024

ABSTRACT

This protocol provides a detailed framework for the systematic annotation and analysis of the semantic interoperability of data elements in data shared through repositories. We applied this protocol for the analysis of spinal cord injury (SCI) research data shared through the Open Data Commons for Spinal Cord Injury (odc-sci.org), centering on the role and harmonization potential of Community-based Data Elements (CoDEs). It underscores the critical need for systematic analysis to achieve consistent, high-quality data standardization, integral to the reproducibility and evolution of SCI research. The protocol navigates researchers through the adoption and interpretation of data elements in publicly available datasets, focusing on their semantic interoperability and harmonization capabilities. By delineating a clear method for evaluating changes in data reporting practices and the efficacy of data elements, this protocol not only bolsters data transparency and reusability but also contributes significantly to the credibility and collaborative progress of the SCI research field.

MATERIALS

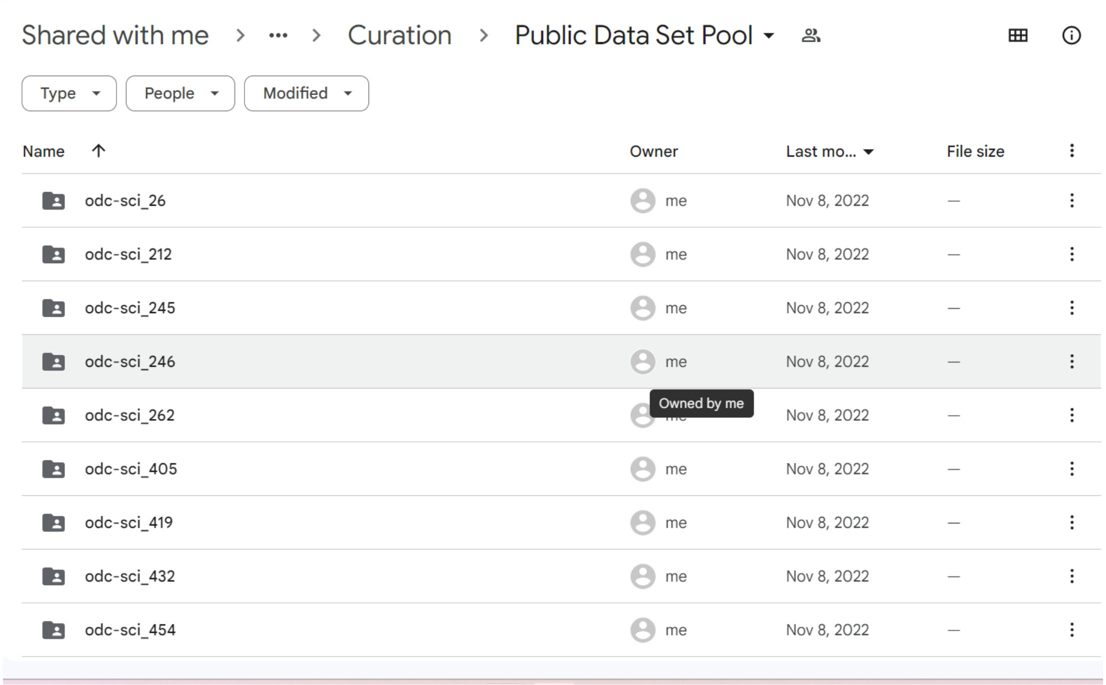
Templates for the different data annotation steps are provided. Please check the relevant steps to get the templates.

PROTOCOL integer ID:
91816

SET UP

1 **Establish the datasets and data elements to analyze with this protocol:** The initial step is to decide the goal of the analysis, the specific datasets, and data elements to include for systematic analysis. For example, the ODC-SCI team has developed and applied this protocol to the analysis of community-based data elements (CoDEs) required for the publishing release of data through odc-sci.org.

2 **Document Organization:** Download and systemically organize all required documents for curation, including datasets and data dictionaries (e.g., all data packages are stored in a folder called data set pool)



Shared with me › ··· › Curation › Public Data Set Pool ▾

Type ▾ People ▾ Modified ▾

Name ↑	Owner	Last mo... ▾	File size
odc-sci_26	me	Nov 8, 2022	—
odc-sci_212	me	Nov 8, 2022	—
odc-sci_245	me	Nov 8, 2022	—
odc-sci_246	me	Nov 8, 2022	—
odc-sci_262	me	Nov 8, 2022	—
odc-sci_405	me	Nov 8, 2022	—
odc-sci_419	me	Nov 8, 2022	—
odc-sci_432	me	Nov 8, 2022	—
odc-sci_454	me	Nov 8, 2022	—

Create one folder for all curation related materials, and one for all dataset-related content

TEMPLATE_ODC-SCI_Data_Dictionary.csv4KB

Note

This template can be found on ODC: Data Dictionary

3 **Curation Record Keeping:** Establish a DecisionLog/ READ-ME file to meticulously document all curation decisions.

TEMPLATE_DecisionsLog.docx15KB

4 Prepare an excel file for curation:

TEMPLATE_AnnotationSheet.xlsx60KB

4.1 Ensure essential columns, such as dataset number/DOI link and column headers, are frozen for easy reference.

4.2 Integrate DOI links directly within the curation spreadsheet.

4.3 Developed a set of standard curation questions across all data elements to be analyzed, with special consideration for edge cases (e.g., a column for if the data element value exists in other pieces of documentation such as an associated paper but not in the dataset)

4.4 Implemente data validation tools, including dropdown menus, for streamlined data entry.

5 **Qualitative Data Documentation:** Create a separate document to chronicle qualitative observations encountered during the curation process, noting the respective datasets for each observation and grouping them by data elements.

EXAMPLE_QualitativeObservations.docx15KB

CURATION

6 **Variable-by-Variable Approach:** Curate one data element after another and begin by formulating data element-specific questions (e.g., units may be curated in Age but not in Injury Device) in addition to the pre-established common queries.

7 **Data Source Reference:** Refer to the dataset or data dictionary for information. In cases where these sources are insufficient, consult relevant manuscripts and note them as a source of information.

8 **Tracking Deviations:** Once curation begins, note any relevant difference between the reported data as a new column in the curation spread

9 **Feature Extraction and Protocol Adaptation:** All potential queries and annotations for the data elements to be analyzed should be pre-specified. However, in our experience, it is often the case that there is a need to review the protocol after the analysis has started. To consider this need systematically, we suggest approaching the feature extraction in two waves, a preliminary wave with a pre-decided subset of datasets, followed by a revision of the protocol, and a second wave with the rest of the datasets to be annotated. That is, discuss the set of features to be extracted for each data element among members of the team from a preliminary annotation of the first *n* datasets to finalize the protocol of annotation for the whole set. Once a curation approach is adopted (as noted in the READ-ME file), keep it consistent.

10 **Quality Assurance:** After completing 50% of the curation, spot-check (by two other team members) to ensure the quality and consistency of the annotation process.

ANALYSIS

11 **Consolidation of Curation Data:** Following the completion of curation for all data elements, consolidate the individual curation sheets into a single comprehensive spreadsheet.

12 **Creation of Analytical Sheet:** Create a new sheet, listing data element names as row headers and maintaining the column headers from the curation sheets.

13 **Data Aggregation and Summary:** The analytical sheet can be summarized for analysis. For each column, tally the frequency of different responses (e.g., counting the instances of 'yes' or 'no' in response to the question, “Was the variable recorded?”) for each data element.

14 **Extended Data Analysis:** Conduct additional analyses, including calculating averages of various responses (for example determining the average data element reporting rate across all data elements),

and computing standard deviations, among other metrics.

Check out out paper to get examples of analysis