IRISS Work Package 6: Curation of Australian Research Data in the Social Sciences (CARDSS)
Technical Report 6.2

Data curation of social science data: Recommended curation practices and tools for IRISS

Author: Ryan Perry, Australian Data Archive
November 2023

1. Introduction

The Australian Data Archive has developed a proposed workflow for the harmonisation of social survey data, that takes account of the practical steps required to bring diverse content together in a machine-actionable way. This harmonisation workflow includes pre-processing of metadata consistent with established archival practises, and incorporates external registered, persistent content.

The workflow is oriented towards improving FAIR practices in the harmonisation process - through the use of reusable, accessible metadata structures that can both improve processing consistency for current projects, and be applied to future harmonisation projects.

There is a need for consistent pre-processing of data and metadata within repositories to reduce error handling in the harmonisation process. The ADA have developed an initial set of processing rules that will be implemented in CARAT.

2. Archival curation: Recommended practices

Combining survey data from different sources poses significant challenges in data harmonisation. Variables collected across different surveys often lack alignment in naming conventions, coding schemes, response categories, and measurement units.

This article presents curation undertaken by the Australian Data Archive to harmonize Australian survey data. The current project focuses on four major Australian survey programs: the Australian Census, Australian Social Survey, Australian Survey of Social Attitudes, and Australian Election Studies - comprising over 11,000 variables.

Harmonising these datasets is supported by systematic pre-processing of the metadata - the variable names, labels, and value codes - to improve compatibility across sources. We have developed a set of proposed curation standards to enable this harmonization process, structured around content cleaning, formatting, and conceptual processing.

2.1 Curation rules framework

Here we present a framework of curation rules and standards to prepare survey data for harmonisation, structured across three categories:

2.1.1 Content Pre-Processing

Rules related to text content, punctuation, spacing, spelling, and abbreviations. Content pre-processing rules are generally able to be easily implemented into the ADA archival workflow for data curation (using R Markdown and R-Shiny). Some rules in this category are already in place as recommendations made to data owners in preparation for dissemination.

- Punctuation: Removing punctuation like commas, parentheses, etc. from labels can be done with regular expressions in R. Label formats should be standardised by removing all punctuation. This rule also ensures interoperability across programming languages and statistical software packages that can include specific functions based on punctuation (e.g., the function of commas in comma-separated values text file format).
- Variable/Value Labels: Rules like removing spaces, correcting spelling, and expanding abbreviations. These rules may require lookup tables or dictionaries to map metadata to target content. Spelling correction is currently implemented at ADA and common abbreviations could be coded systematically.
- Capitalisation: Converting labels to sentence case can include capitalising first letter of each sentence. Implementing other appropriate capitalisation (e.g., proper nouns) may require manual input.

2.1.2 Encoding and Formats

Rules related to data types, encoding, and standard data formats including dates and measurement units.

- Date variables: Converting dates to ISO format YYYY-MM-DD is achievable using R date parsing and reformatting.
- Standard units: Converting measurement units like pounds to kg can be implemented using a lookup table of conversion factors.
- Values/Labels: Formatting ordinal scales with consistent punctuation.

2.1.3 Conceptual Processing

Rules related to semantic label content and recoding of scales to facilitate harmonisation.

- Likert Scales: Mapping different Likert scales to a consistent scale can be achieved by recomputing each scale value to the corresponding value in the target scale.
- Missing Data: A common scheme for labelling and coding missing values will be developed based on major national longitudinal projects in Australia.
- Binary Choices: Simplifying binary yes/no labels may be achieved by mapping synonyms with a suitable lookup table. The curation script will also flag value label content that overlaps with content in the corresponding variable label. This rule may require automated sematic matching.
- Vocabularies: Adopting standard vocabularies requires comprehensive taxonomies to map variable metadata to.

2.2 Summary

Many of the proposed rules related to text content, punctuation, date formatting, etc. can be fully implemented programmatically in R. Some conceptual rules require more complex logic or human intervention, such as mapping different Likert scales or interpreting abbreviations. Full implementation would require comprehensive lookup tables/dictionaries for abbreviations, units, vocabularies, etc. Some curation rules may therefore require the implementation or development of specialised data processing syntax or use of machine learning models.

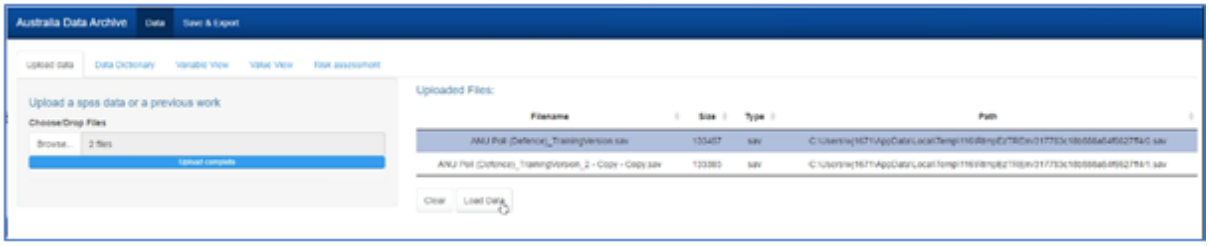
Rules should be prioritized based on impact and ease of implementation. Further development of these guidelines should focus on rules that balance optimal harmonisation while minimising loss of information, human readability, and project-specific metadata that should be retained for the purposes of dissemination and secondard data use – the core service functions at ADA.

3. CARAT: Curation and Risk Assessment Tool

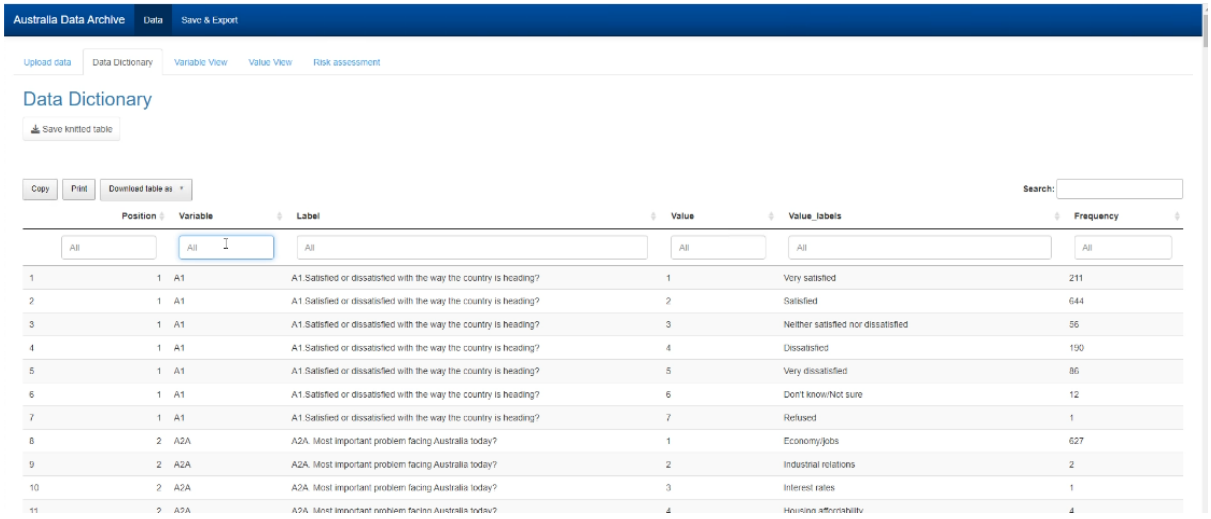
The Australian Data Archive has developed an R Shiny application for automating data curation and harmonization. CARAT implements a set of pre-processing workflows to improve metadata quality, enable comparability across datasets, and prepare data files for harmonisation where required.

3.1 Key Features

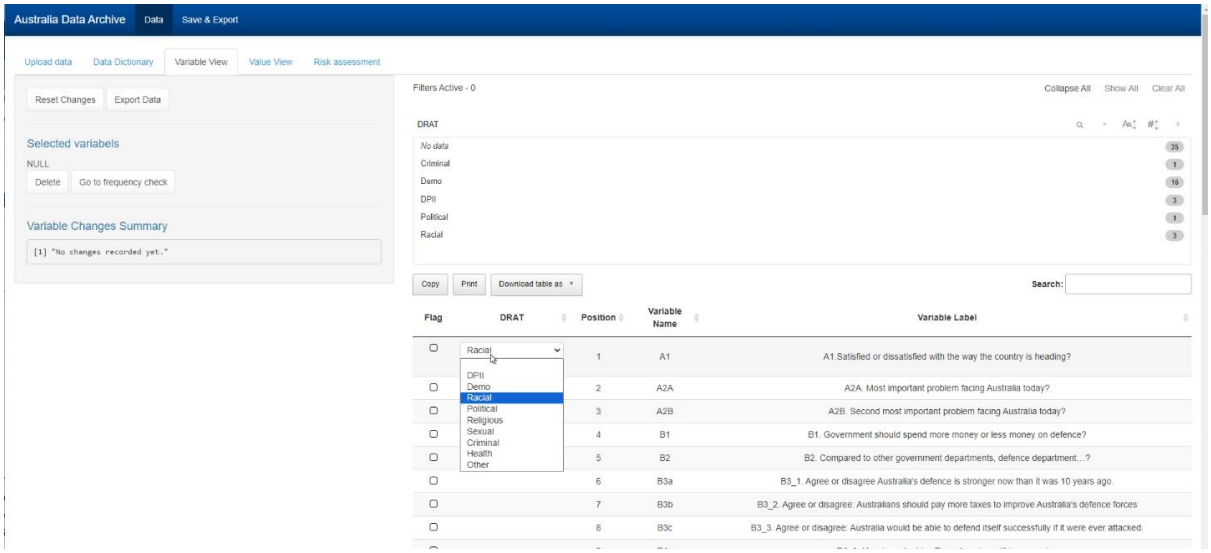
- File Upload: Upload multiple dataset files in SPSS .sav format
- Data Dictionary: Interactively explore variables, value labels, and coding schemes
- Risk Assessment Flags: Highlights variables with potential disclosure risk
- Curation Processing: Standardizes formats, fixes labels, handles missing values
- Save Progress: Bookmark feature reverts to previous saved versions
- Syntax Output: Exports code for replicating transformations and
- Data Output: Exports transformed data in multiple archival formats (SPSS, Stata, SAS, and .CSV) for publication.



File upload screen



Data dictionary screen



Risk assessment – Sensitive data categories

3.2 Future development

CARAT automates archival tasks involved in making diverse datasets compatible for analysis. Workflows such as standardizing date variables, mapping occupation codes, adapting Likert scales, and converting units will be added in forthcoming updates to the app.

The tool also flags variables requiring anonymization including potentially identifying information and sensitive attributes like religion, ethnicity, and health data. Deploying CARAT in the proposed hamonisation workflows improves FAIRness and leverages automation for more rigorous, transparent, and privacy-aware research.