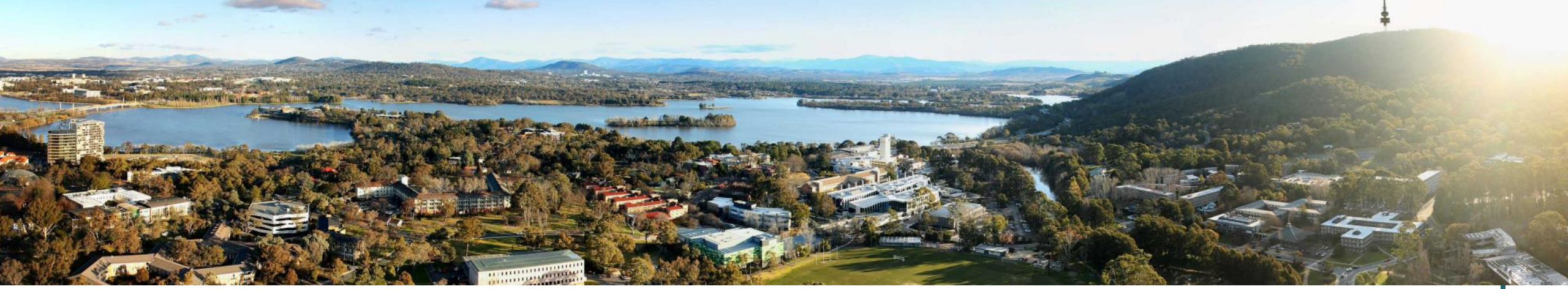
In the IRISS pipeline: A curation tool for integrated data risk assessment

Ryan Perry, Deputy Director
Weifan Jiang, Archivist

Australian Data Archive
The Australian National University



Australian Data Archive

- Est 1981, Research School of Social Sciences, Australian National University
- >5000 data sets from around 1500 studies
- >7500 registered users
- Wide range of data collections and data types

Introduction

- The Data Risk Assessment Tool (DRAT) is being developed in R-Shiny to support archivists and data owners
- Three main functions:
  1. Identifying categories of sensitive data
  2. Addressing privacy/identification risks
  3. ‘Quality’ assessment (usability and standardization of data)
- There is little consensus or guidance about standardization of data elements (Dubrow & Tomescu-Dubrow, 2015).
  - DRAT can support interoperability of data in the ADA collection

Data quality assessment

- Checks include:
  - spelling
  - label length, duplication, missing labels
- Metadata editing function records revisions in the syntax

How can we leverage these archiving procedures to support consistency, interoperability, and… harmonization?

DRAT in the IRISS pipeline

- Integrated Research Infrastructure for Social Science (IRISS)
  - data integration services – Vocabulary Access Service and GeoSocial – for conceptual, spatial and temporal data integration support
  - demonstrator projects testing the services and illustrating the implementation of the project infrastructure in applied social science research settings
  - data collection and curation services for the improved management of research data throughout the research process
- Curation work package
  - Aim to establish standardised curation practises, a program library, and training packages for the management of social science research data
  - Support data harmonisation through pre-processing source data

IRISS harmonisation crosswalk

Item Name	q53	g7comgov	h8fed
Missing		0	
NO RESPONSE	0		
No trust		1	
Just about always	1		1
Most of the time	2		2
Some trust		3	
Only some of time			3
SOME OF THE TIME	3		
Never			4
NOT AT ALL	4		
Great trust		5	

DRAT Demonstration

- Data file ingest
- Editing metadata (reproducible)
- Detection of non-standard content
  - Punctuation
  - Capitalisation
- Data and syntax file(s) export

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

server.R | save_page.R | data_page.R | DPtool.R | test.R | server.R | 0.ADA Data Quality Check V1.0.Rmd

Run App

reset | Next | Prev | All | Replace | Replace | All

In selection | Match case | Whole word | Regex | Wrap

```
1
2
3  # Library loading ------------------------------------------------------------
4  library(shiny)
5  library(shinyalert)
6  library(rhandsontable)
7  library(ggplot2)
8  library(tidyverse)
9  library(foreign)
10 library(haven)
11 library(DT)
12 # library(fs)
13 library (labelled)
14 library (dplyr)              # Managing the functions using pipes %>%
15 library (sjlabelled)         # Supports 'rename_at', 'var_labels' and 'add_value_labels' functions
16 library (Hmisc)
17 library(hash)
18 source('../Lib/DPtool.R')
19 # library(AMR)
20
21
22
23
24 # Function def ---------------------------------------------------------------
25
26
27 callback <- c(
28   "var tbl = $(table.table().node());",
29   "var id = tbl.closest('.datatables').attr('id');",
30   "function onUpdate(updatedCell, updatedRow, oldvalue) {",
31   "  var cellinfo = [{",
32   "    row: updatedCell.index().row + 1,",
33   "    col: updatedCell.index().column + 1,",
34   "    value: updatedCell.data()",
35   "  }];",
36   "  Shiny.setInputValue(id + '_cell_edit:DT.cellInfo', cellinfo);",
37   "}",
38
```

1002:52 | callbackR(value) | R Script

Console | Terminal | Background Jobs

R 4.2.2 · ~/

```
>
> runApp('C:/Users/wj1671/proj/ADA-R-Lib-test/ADA_DRAT_v2')

Listening on http://127.0.0.1:4225
```

Environment | History | Connections | Tutorial

Import Dataset | 628 MiB | List

R | Global Environment

var_view	List of 3
variable_view...	List of 3

Files | Plots | Packages | Help | Viewer | Presentation

Remaining challenges

Data types

- Code-category mismatches are fairly common due to different encoding and labelling practices, includinging between software packages. E.g.,
  - Variable and data types (e.g., string, numeric, ordinal, integer etc.)
  - Missing value encoding

Conceptual harmonising

- Potential for semantic similarity models to streamline conceptual mapping (requires considerable human input).
- Conceptual mappings not seen as equivalent
  - E.g., re-computing Likert scale points

Summary

- DRAT was initially conceived as an archiving tool that might assist the research community to prepare their own data for publishing.
- Increasingly supports more consistent/standardised archiving and more interoperable data.
- Standardised archiving practices can also support data harmonisation by *pre-processing* data for semantic mapping

ada.edu.au

ada@ada.edu.au

Integrated Research Infrastructure for Social Science