closer

The home of longitudinal research

Extract2DDI

Jon Johnson,

Technical Lead, CLOSER

September 2022

UKRI Economic and Social Research Council

What is Extract2DDI?

Open-source software

Extracts metadata from SPSS and Stata files

Supports DDI-Codebook, & Lifecycle

Configurable content

Development

Original code base funded by ESRC

Additional development at Cornell

CLOSER development of DDI Lifecycle writers

Added configuration

Running Extract2DDI

Command line tool

java -jar Extract2DDI.jar -f filename.sav --format 2.5
java -jar Extract2DDI.jar -f filename.sav --format 3.2
java -jar Extract2DDI.jar -f filename.sav --format 3.3Fragment

java -jar Extract2DDI.jar -f filename.sav --format 3.2 --config config.txt

DDI-Lifecycle Configuration

Add user defined content

--config

- agency={uk.closer}
- ddilang={en-GB}
- stats={max,min,mean,valid,invalid,freq,stdev}
- outputfile={example_file_name}

DDI-Lifecycle Configuration

Restrict output for individual variables and document why this is being done, for summary statistics

--exclude

- noofbedrooms={max}:{potentially disclosive}
- var2={freq}:{data holder policy}

Next steps

- Ironing out a range of bugs
- Add unit tests
- More testing
- Add support for DDI-Codebook 2.6
- Add configuration to DDI-Codebook output
- Beta version at:
  - https://github.com/CLOSER-Cohorts/Extract2DDI

closer

The home of longitudinal research

Thank you ☺

UKRI Economic and Social Research Council

closer

The home of longitudinal research

Questions and Comments

UKRI Economic and Social Research Council

Our Approach

- When we started in 2012, we did not intend to be spending a lot of time developing software
- We realised, that it was “easier” to develop software than it was to manage people and their outputs, so they could concentrate on content
- It has made it easier to collaborate within CLOSER and with studies to achieve higher and a more consistent quality of metadata.
- We also think it has provided a better quality of work for those work for and with us.

Summary

Managing questionnaire metadata in Archivist

Building a collaborative pipeline for questionnaire metadata ingest using Gitlab

Functional testing of Archivist using Selenium

Validation and management of ‘open text fields’ using a dashboard tool

Metadata extraction from statistical files using Extract2DDI