Pooler

Pool your study data and build your own datasets

Luke Stevens
Data Management Coordinator
Clinical Epidemiology and Biostatistics Unit (CEBU)
Murdoch Childrens Research Institute

Background

Any research project – and especially large and/or longitudinal projects – will have a large number of datasets containing different parts of the project data, e.g.

- Participant demographics
- Enrolment and consent
- Questionnaires
- Direct assessments
- Lab results
- Similar data capture from different study visits

Sometimes it is easy enough to manage project data by generating clean datasets that may be distributed – along with corresponding data dictionaries – to interested researchers in clearly demarcated "chunks": the "reference dataset" (demographics and other commonly used fields), "child questionnaire", "hearing test results" etc.

With a large and/or longitudinal project with many different datasets and large numbers of variables in disparate domains it is not always easy to do this.

Pooler

Pooler is a web-based application that is designed to facilitate researchers' access to a project's data by bringing together all of the project datasets' data dictionary information ("metadata" – variable names, labels, data types, coding etc.) in a searchable form. A researcher may then put together a customised dataset containing only the variables that they are interested in, drawn from whichever cleaned source data files have been made available.

Each request for data must go through an approval process before a researcher is able to download project data.

Pooler generates a raw data file in CSV format plus a Stata syntax file (do-file) that will read in the raw data and label it up ready for use.

Features

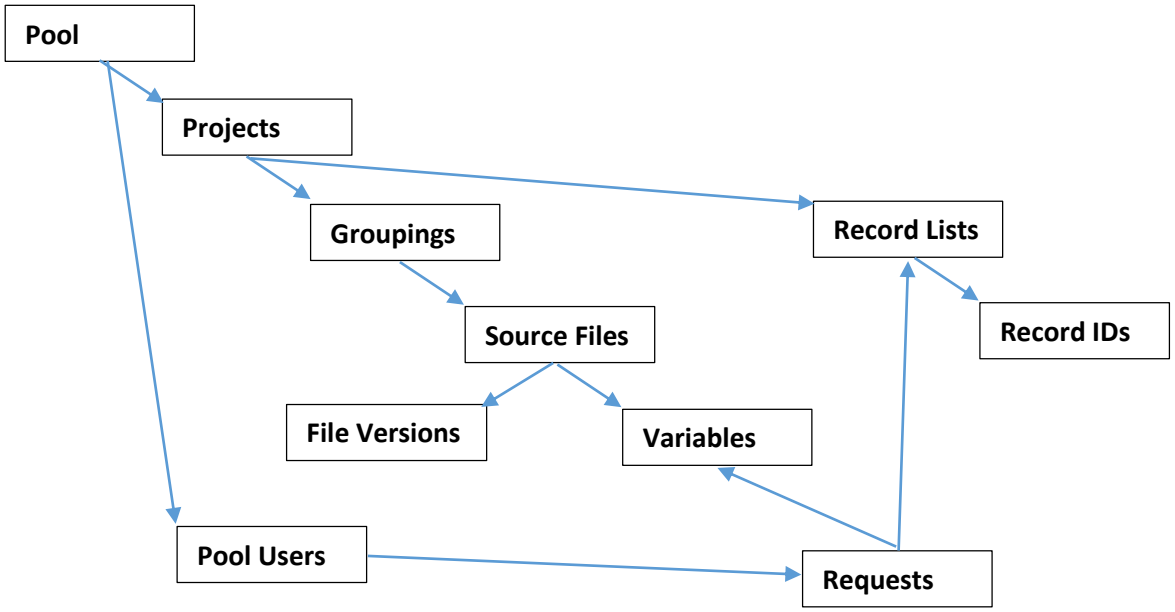
- Web-based
- Utilises MCRI's REDCap database for user authentication
- Request approval workflow
- Pool data from different projects and/or groupings of data within a project
- Control the version of a dataset that is available
- Pre-define convenient record filters
- Export raw CSV data plus a corresponding Stata syntax file containing full field metadata
- Permanent record of what data was provided to a researcher and when

Terminology and Architecture

- "Pool"
  The top level category: users are given access to browse the data contained in a pool.
- "Project"
  A pool has one or more projects from which data for the pool is sourced. Records from different projects appear as separate rows in a request data file.
- "Record List"

A project has at least one record list: a list of record id values from the project that meet some arbitrary criteria. "All eligible records", for example, or "Records with complete 12 month data".

- "Grouping"
  A project has at least one grouping, which is a mechanism that enables datasets to be categorised for more convenient searching. Groupings are determined by the person that configures the pool data and would typically relate to some feature of the project or data: study visits in a longitudinal project, for example, or perhaps child-level data / parent-level data.
- "Source File"
  A grouping has at least one source file: a clean dataset containing a set of "variables". One variable in each source file will contain record id values that match record ids in the source file project's record list(s).
- "File Version"
  A source file may have multiple versions of its data uploaded – as it gets progressively cleaner, perhaps. A specific version – or none – may be made available for use. The set of variables for a source file is the same across versions (changing the variables means it is a different "source file").
- "Request"
  A user browses the variables and record lists available in the pool and selects a set for inclusion in their request. The request is the combination of the variables selected (i.e. the columns) and the record lists (i.e. the rows).

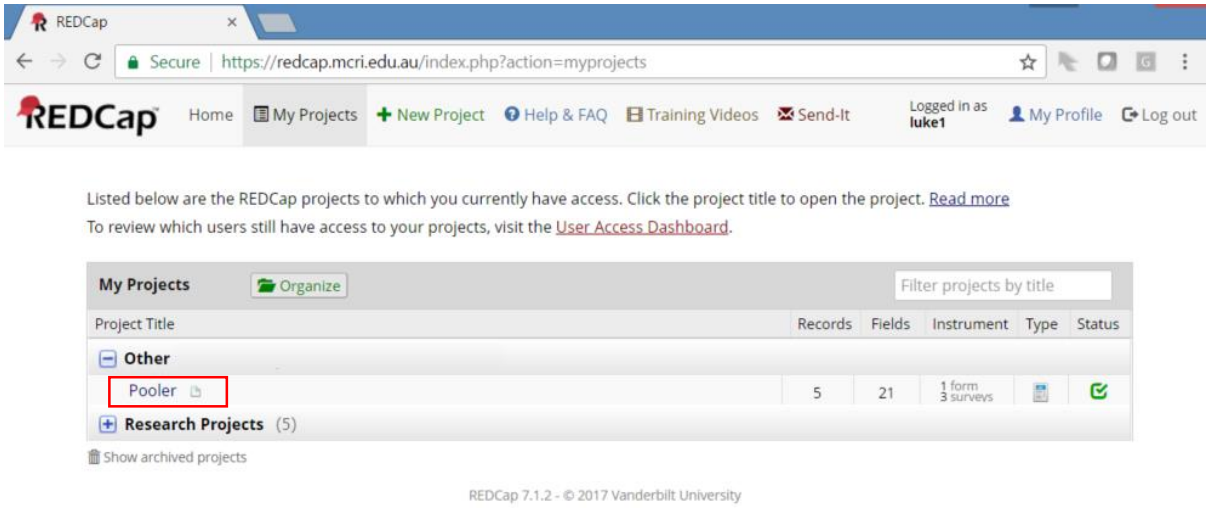


Usage

Login and Access

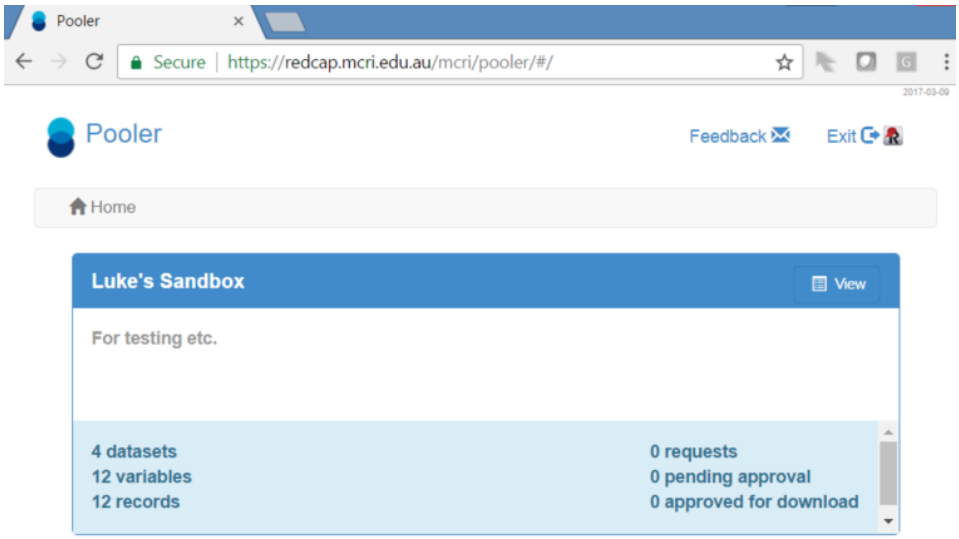
Pooler utilises MCRI's REDCap web-based database application for managing user accounts and authentication. You must be granted permission to access a specific REDCap project in order to access Pooler. Your Pool administrator (e.g. the project Data Manager) will organise this.

Log in to REDCap at https://redcap.mcri.edu.au and you will see "Pooler" listed on REDCap's "My Projects" page. Click the "Pooler" text to access the Pooler application.



Home Screen

The home screen shows the Pools that you have been granted permission to browse and make requests for. Here there is one: you may have more.

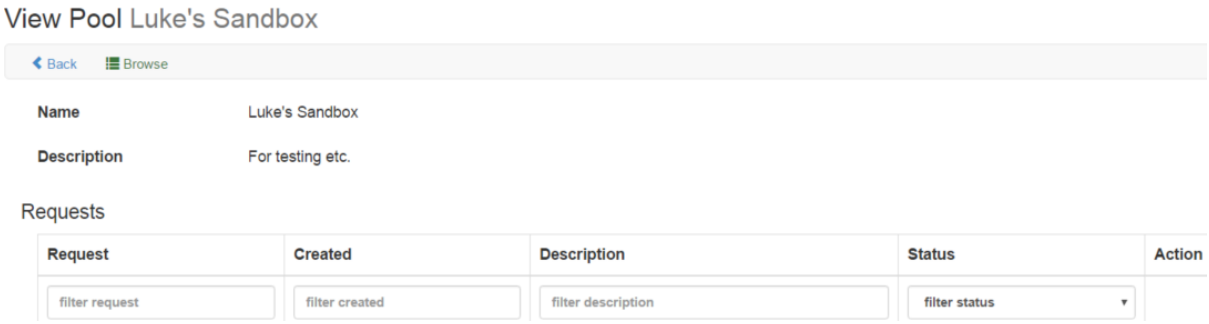


The following functions are available on the Home screen:

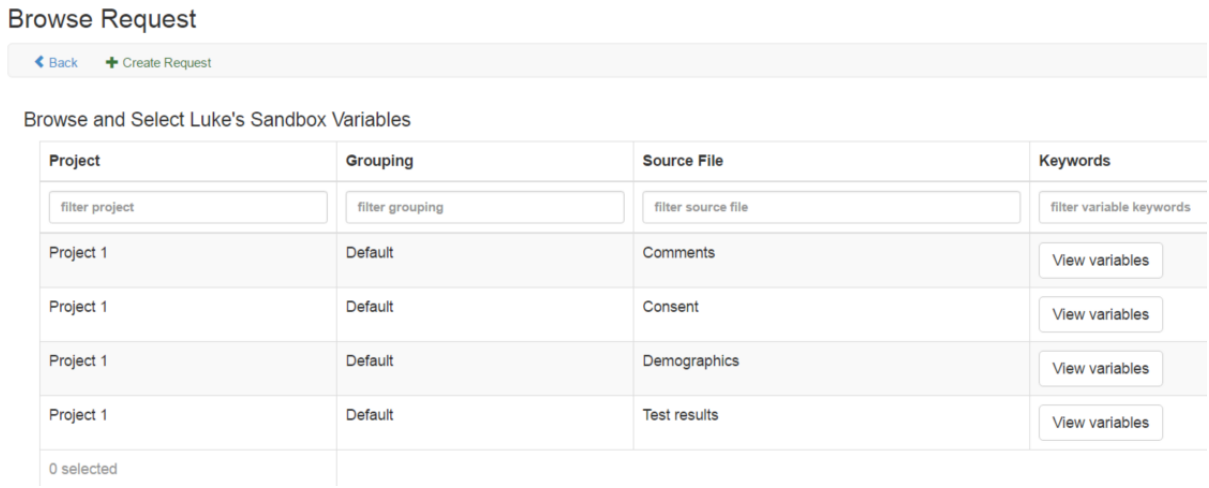
- **Pooler** Click the Pooler icon and title to return to this page
- **Feedback** Click the feedback link to report a bug or submit a feature request
- **Exit** Click Exit to return to the REDCap "My Projects" page
- **View** Browse the Pool variables and create requests for data

View Pool

The View Pool page shows you basic information about the Pool (title, description), plus a list of the requests that you have made.



Click Browse to view the variables that are available in the Pool. Clicking View Variables for a source file brings up a dialog containing a listing of the variables for that source file.

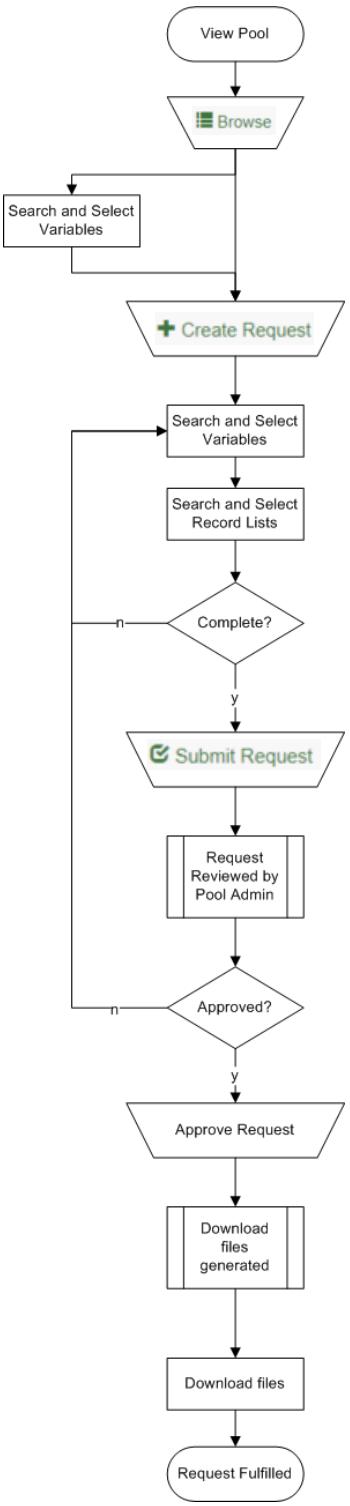


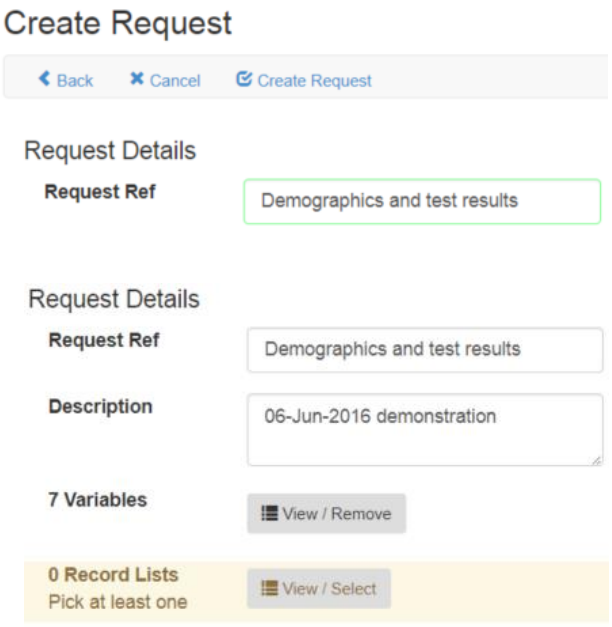
Build a Request

Workflow Diagram

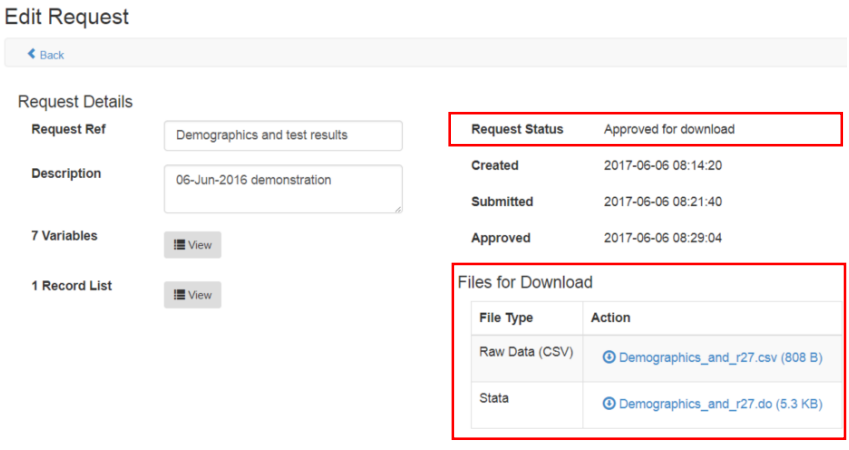
You must put together your request by selecting the variables and record lists that you require. Once this is complete you submit the request and then await approval of the request by a Pool Administrator (some users may have permission to approve their own requests).

This diagram illustrates the workflow:





(Nb. you need to select at least one variable and at least one record list)



Files for Download

Two files will be generated for you to download:

1. A file of raw data in CSV format (comma-separated values – plain text)
2. A Stata do-file that contains the variable metadata

Download both files to your Stata working directory and run the do-file. The do-file will:

1. Import the raw data
2. For each variable:
    a. Apply the variable label
    b. Apply a value label (for categorical variables) or format (for date/datetime variables)
    c. Add a note with information on the variable – source file, grouping, project etc.

CSV Data File

	A	B	C	D	E	F	G	H	I
1	projid	recid	recordid_26	dob	sex	recordid_28	starttm	endtm	result
2	8	1	1	31/07/2002	2	1	03/05/2016 08:37	03/05/2016 09:20	3
3	8	2	2	24/03/2003	1	2	04/05/2016 13:03	04/05/2016 14:37	4
4	8	3	3	30/07/2003	2	3	20/05/2016 11:09	20/05/2016 12:46	2
5	8	4	4	10/05/2003	1	4	01/06/2016 11:55	01/06/2016 13:38	4
6	8	5	5	20/08/2003	2	5	03/06/2016 18:35	03/06/2016 18:47	1
7	8	6	6	03/04/2002	1	6	05/06/2016 15:39	05/06/2016 15:57	5
8	8	7	7	20/07/2003	1	7	12/06/2016 17:32	12/06/2016 17:53	2
9	8	8	8	24/10/2002	2	8	12/06/2016 10:49	12/06/2016 12:34	2
10	8	9	9	24/12/2001	1	9	24/06/2016 10:54	24/06/2016 12:23	1
11	8	10	10	26/02/2003	2	10	29/06/2016 14:50	29/06/2016 15:25	1
12	8	11	11	15/08/2003	1	11	09/07/2016 16:06	09/07/2016 16:47	3
13	8	12	12	03/09/2003	2	12	09/07/2016 16:45		

Demographics_and_r27.csv

```
projid,recid,recordid_26,dob,sex,recordid_28,starttm,endtm,result
8,1,1,2002-07-31,2,1,2016-05-03 08:37:11,2016-05-03 09:20:07,3
8,2,2,2003-03-24,1,2,2016-05-04 13:03:00,2016-05-04 14:37:08,4
8,3,3,2003-07-30,2,3,2016-05-20 11:09:13,2016-05-20 12:46:56,2
8,4,4,2003-05-10,1,4,2016-06-01 11:55:22,2016-06-01 13:38:55,4
8,5,5,2003-08-20,2,5,2016-06-03 18:35:16,2016-06-03 18:47:42,1
8,6,6,2002-04-03,1,6,2016-06-05 15:39:01,2016-06-05 15:57:19,5
8,7,7,2003-07-20,1,7,2016-06-12 17:32:38,2016-06-12 17:53:16,2
8,8,8,2002-10-24,2,8,2016-06-12 10:49:12,2016-06-12 12:34:08,2
8,9,9,2001-12-24,1,9,2016-06-24 10:54:07,2016-06-24 12:23:05,1
8,10,10,2003-02-26,2,10,2016-06-29 14:50:22,2016-06-29 15:25:39,1
8,11,11,2003-08-15,1,11,2016-07-09 16:06:13,2016-07-09 16:47:59,3
8,12,12,2003-09-03,2,12,2016-07-09 16:45:12
```

Notes on the CSV data file:

- CSV files are plain text – they just open in Excel for convenience.
- Do not open and save the file using Excel! See how Excel applies its own formatting to certain columns? If you save the file using Excel you change the data and may break the do-file. Edit datasets using a do-file, not interactively.
- If your dataset contains multiple variables that have the same name then each is distinguished by an appended unique id number corresponding to the source file. In the example above, the variable *recordid* has been selected for inclusion in the dataset both from file 26 (as *recordid_26*) and from file 28 (as *recordid_28*).

Do-File

Notes on the Stata do-file:

- Check your working directory and the location of the raw data file before running
- The file works through each variable in turn and provides information about where each variable has come from
- After processing all the variables there is a *describe* statement
- Following the do-file run you have a fully labelled dataset in memory. It is not saved.

Demographics_and_r27.do

```
*************************************************************************
* Pooler Do File
* Generated at 2017-06-06 08:30:01
* Pool Luke's Sandbox
*************************************************************************
* View information about variables using the notes command:
* > notes <varlist>
*************************************************************************

clear
set more off

* SET YOUR WORKING DIRECTORY!
* cd ./path/to/location/of/raw/data/file
import delimited Demographics_and_r27.csv, varnames(1) case(preserve) asdouble

label define projid_label 8 "Project 1"
label values projid projid_label

label variable projid "Project source of record"
label variable recid "Record Id"

*************************************************************************
* Variable
* Label          Record id
* Name           recordid
* Data type      integer
* Status
* Domain
* Source file    Demographics (_26 resolves conflict)
* Data grouping Default
* Project        Project 1
* NOTE: This is a merge column. A missing value indicates the record does not e

label variable recordid_26 "Record id"
notes recordid_26: Source file="Demographics" (_26 resolves conflict) Data grou
*************************************************************************
```

Feedback

Please report any bugs or difficulties you experience via the Feedback link in the page header.

Thank you,
Luke

Luke Stevens
Data Management Coordinator
Clinical Epidemiology and Biostatistics Unit (CEBU)

Murdoch Childrens Research Institute
The Royal Children's Hospital
Flemington Rd Parkville, Victoria 3052 AUS
T: (03) 9345 6552
E: luke.stevens@mcri.edu.au
W: www.mcri.edu.au