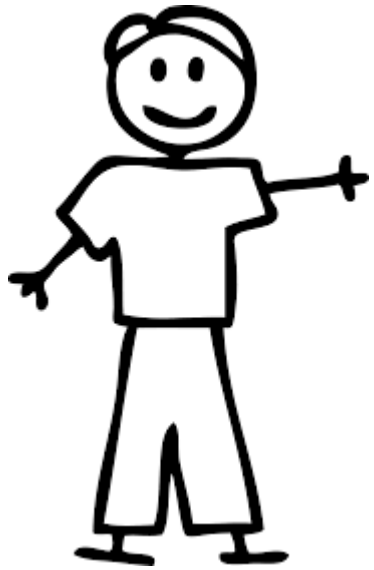
Extracting Metadata from Stata Datasets

Suzanna Vidmar and Luke Stevens

Clinical Epidemiology and Biostatistics Unit
Murdoch Children's Research Institute



Data sharing and storage

- To enable data sharing, the data should be stored in a format that does not required a particular version of a particular statistical package
- At the conclusion of a study, data should be stored in a retrievable format, and not one that may become obsolete
- The safest retrievable format is to have the data stored in CSV or text files
- Stata's `export delimited` command writes data from a Stata dataset to a text file

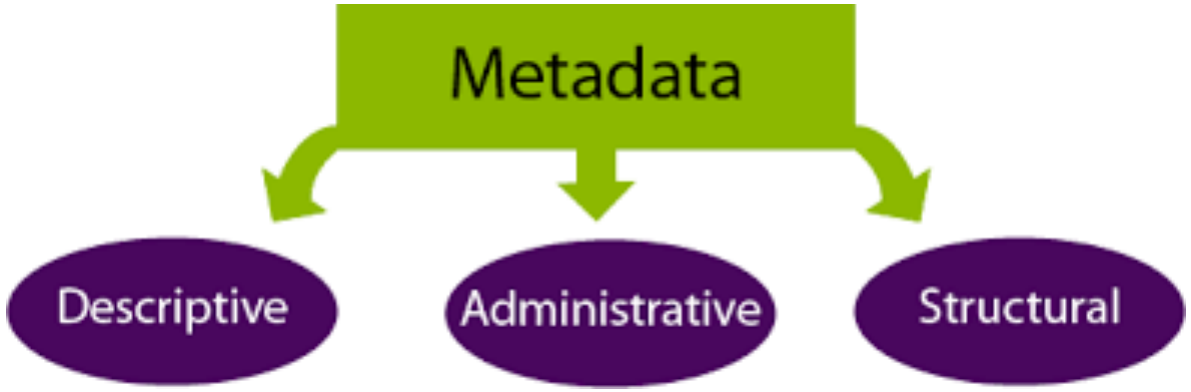
But what do the data mean?

partid	p1pt_nassda	ccpp_actical	p1sh_valudec
121	6	25461	.32
207	3	19483	.44
153	5	25618	.036
114	4	20159	.87
215	7	23876	.02

Without a description of the data, the data file is of limited use

Metadata

- **Metadata** is data that describes other data

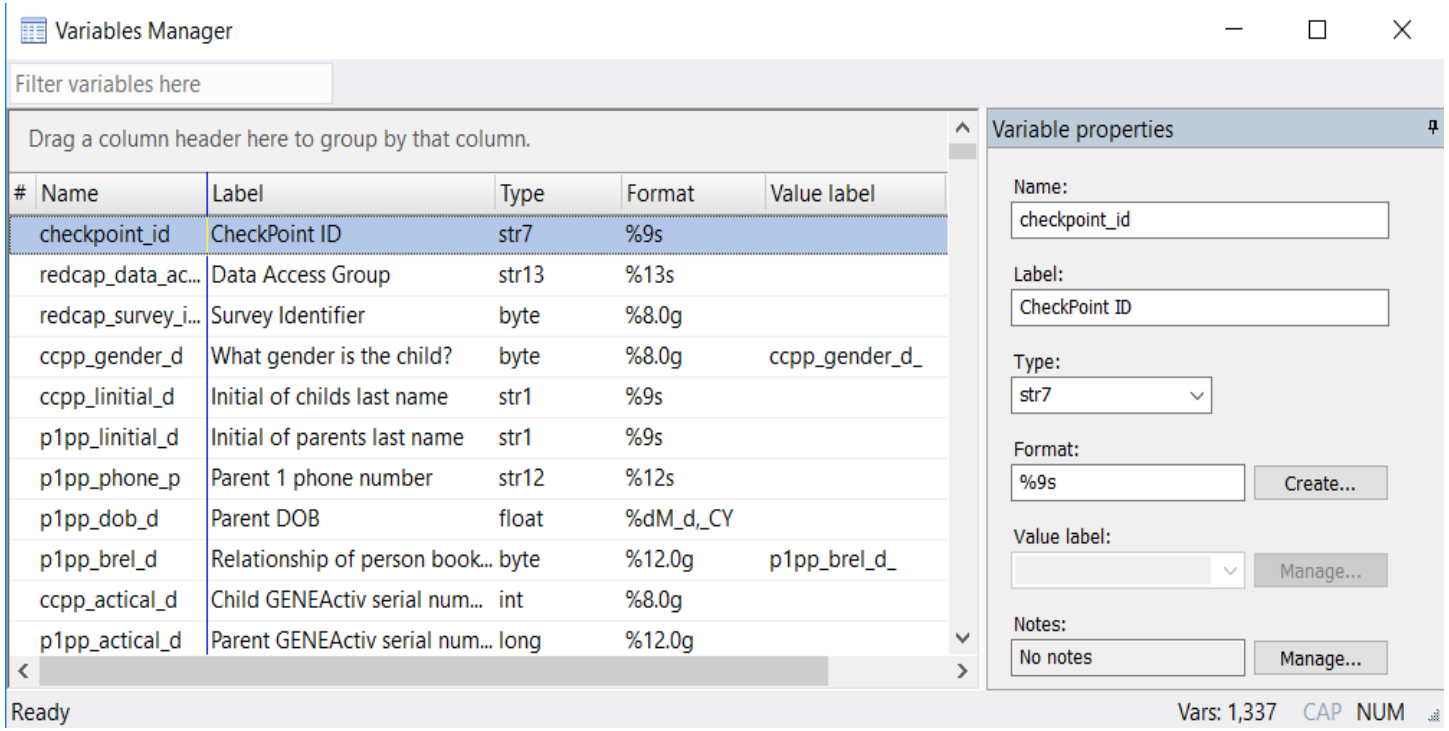


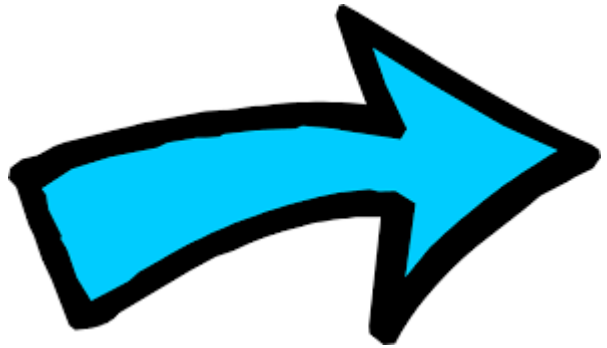
- My focus is on variable-level meta data, also known as a data dictionary
- Examples of variable-level metadata are data types, variable labels and value labels

Metadata is a love note to the future

Extracting the data dictionary from Stata

STATA release 15





filename.CSV

But wait, there's more!

Data and metadata can be imported into data capture software such as REDCap

Variable / Field N	Form Name	Section Header	Field Type	Field Label	Choices, Calculations, OR Slider Labels
enrolid	demographics		text	Enrolment ID	
screenid	demographics		text	Screening ID	
dob	demographics		text	DOB	
gender	demographics		dropdown	Gender	1, Male | 2, Female | -2, Query
dtenrol	demographics		text	Enrolment Date	
ageenrol_error	demographics		descriptive	Age is out of range	
constipation_yrs	demographics		text	Length of time constipated (years)	
constipation_mth	demographics		text	Length of time constipated (months)	
hprac_yn	demographics		dropdown	Has the child's constipation been managed by a medical or allied health practitioner?	0, No | 1, Yes | -2, Query
hprac_type	demographics		checkbox	What type of health practitioner has been treating your child's constipation?	1, GP | 2, Dietician | 3, Physiotherapist | 4, Psychologist
hprac_spec	demographics		text	Specify who managed your child's constipation.	
treated_yrs	demographics		text	Years treated for constipation	
treated_mths	demographics		text	Months treated for constipation	
meconium_24	demographics		dropdown	Meconium passed >24hrs	0, Absent | 1, Present | -1, Missing | -2, Query
meconium_48	demographics		dropdown	Meconium passed >48hrs	0, Absent | 1, Present | -1, Missing | -2, Query

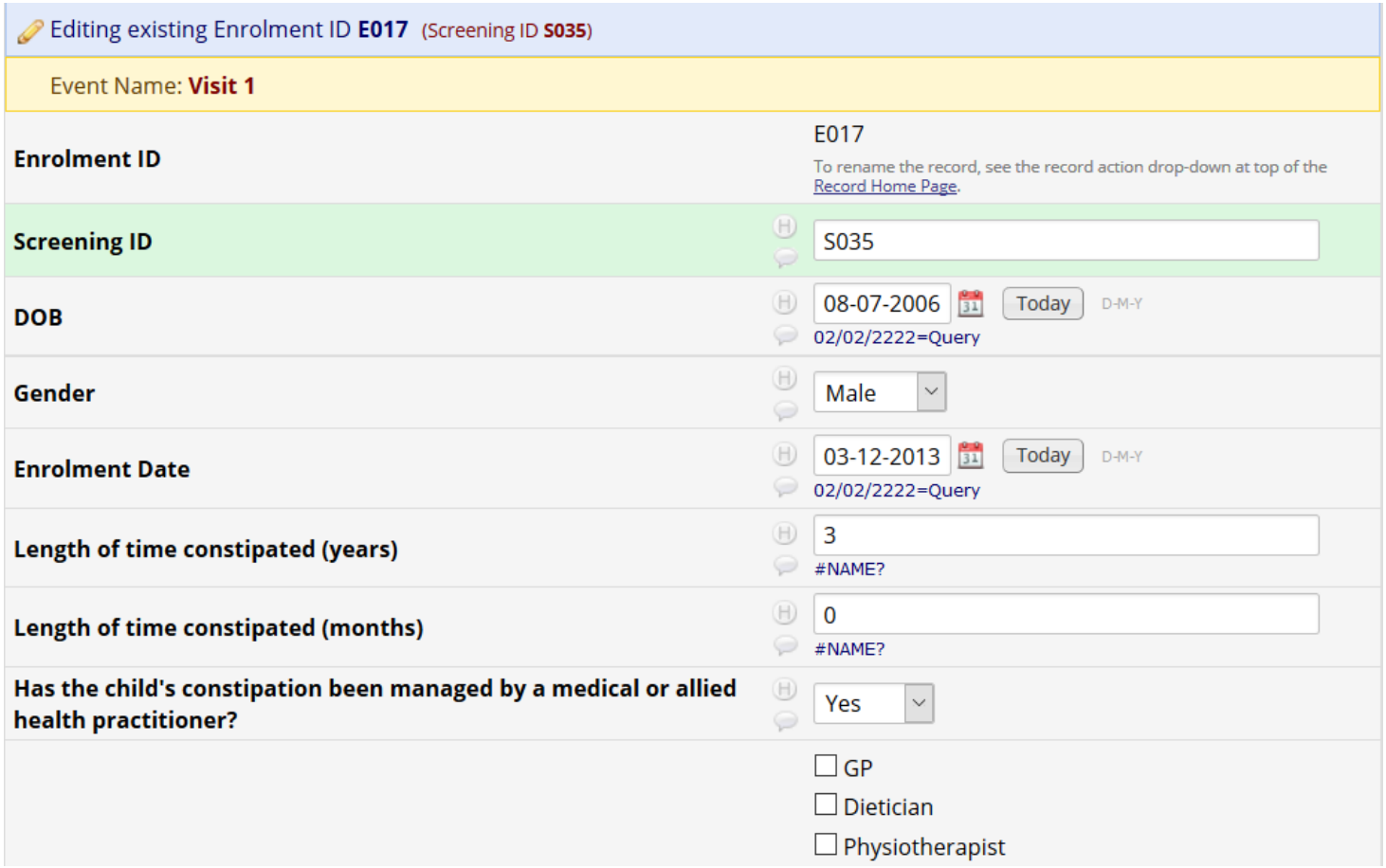
Features of REDCap

- Secure, web-based application for research databases and surveys
- Very easy to use
- Audit trail
- User permission controls
- Data quality measures
- Data export to statistical software
- Generate summary report and letters

https://projectredcap.org/

Building a REDCap database

- As with all data capture software, data entry forms can be developed within REDCap
- A REDCap database can also be built by uploading an external data dictionary



metadatacsv.ado

Example using metadatacsv.ado

example.dta

idno	hosp	cob	gender	qu1	qu2	qu3	age
1	RC	Australia	Male	Yes	Yes	Yes	7.442102
2	RC	Australia	Female	Yes	Yes	Yes	6.8682
3	WCH	Australia	Missing	No	No	No	5.956524
4	RC	Missing	Missing	Yes	No	Yes	5.199898
5	PCH	Australia	Missing	Yes	Yes	No	11.08253
6	WCH	Australia	Missing	No	No	No	7.455984

dict_example.csv

name	varlab	type	isnumeric	format	vallab	choices
idno	ID number	float	1	%9.0g		
hosp	Hospital	str3	0	%9s		
cob	Country of birth	float	1	%14.0g	coblab	-1, Missing | 1, Australia | 2, United Kingdom | 3, Vietnam | 4, China | 5, Singapore | 6, New Zealand
gender	Gender	float	1	%9.0g	genderlab	-1, Missing | 1, Male | 2, Female | 3, Intersex
qu1	Question 1	float	1	%9.0g	noyes	0, No | 1, Yes
qu2	Question 2	float	1	%9.0g	noyes	0, No | 1, Yes
qu3	Question 3	float	1	%9.0g	noyes	0, No | 1, Yes
age	Age (yrs)	float	1	%9.0g		

Directory and file name

```
describe, replace
local fullpath: char _dta[d_filename]


mata: st_local("fullname", pathbasename("`fullpath'"))


local length=strpos("`fullname'",".")-1


local filestub=substr("`fullname'",1,`length')
```

Directory and file name

```
describe, replace
local fullpath: char _dta[d_filename]
```

- di "`fullpath'"
- C:\Users\suzanna.vidmar\Documents\Suzanna\Metadata\example.dta

```
mata: st_local("fullname", pathbasename("`fullpath'"))



local length=strpos("`fullname'",".")-1



local filestub=substr("`fullname'",1,`length')
```

Directory and file name

```
describe, replace
local fullpath: char _dta[d_filename]
```

- di "`fullpath'"
- C:\Users\suzanna.vidmar\Documents\Suzanna\Metadata\example.dta

```
mata: st_local("fullname", pathbasename("`fullpath'"))
```

- di "`fullname'"
- example.dta

```
local length=strpos("`fullname'",".")-1
```

```
local filestub=substr("`fullname'",1,`length')
```

Directory and file name

```
describe, replace
local fullpath: char _dta[d_filename]
```

- di "`fullpath'"
- C:\Users\suzanna.vidmar\Documents\Suzanna\Metadata\example.dta

```
mata: st_local("fullname", pathbasename("`fullpath'"))
```

- di "`fullname'"
- example.dta

```
local length=strpos("`fullname'",".")-1
```

- di "`length'"
- 7

```
local filestub=substr("`fullname'",1,`length')
```

Directory and file name

```
describe, replace
local fullpath: char _dta[d_filename]
```

- di "`fullpath'"
- C:\Users\suzanna.vidmar\Documents\Suzanna\Metadata\example.dta

```
mata: st_local("fullname", pathbasename("`fullpath'"))
```

- di "`fullname'"
- example.dta

```
local length=strpos("`fullname'",".")-1
```

- di "`length'"
- 7

```
local filestub=substr("`fullname'",1,`length')
```

- di "`filestub'"
- example

Saving data dictionary

```
export delimited "dict_`filestub'.csv", replace
```

Saves the data file:

dict_example.csv

describe, replace

- `describe` usually produces a written report
- When the `replace` option is specified, instead of a report the data in memory are replaced with dataset containing the information that would have been presented in the report. The new dataset has an observation for each variable in the original data.

describe

variable name	storage type	display format	value label	variable label
idno	float	%9.0g		**ID number**
hosp	str3	%9s		**Hospital**
cob	float	%14.0g	coblab	**Country of birth**
gender	float	%9.0g	genderlab	**Gender**
qu1	float	%9.0g	noyes	**Question 1**
qu2	float	%9.0g	noyes	**Question 2**
qu3	float	%9.0g	noyes	**Question 3**
age	float	%9.0g		**Age (yrs)**

describe, replace

position	name	type	isnumeric	format	vallab	varlab
1	idno	float	1	%9.0g		ID number
2	hosp	str3	0	%9s		Hospital
3	cob	float	1	%14.0g	coblab	Country of birth
4	gender	float	1	%9.0g	genderlab	Gender
5	qu1	float	1	%9.0g	noyes	Question 1
6	qu2	float	1	%9.0g	noyes	Question 2
7	qu3	float	1	%9.0g	noyes	Question 3
8	age	float	1	%9.0g		Age (yrs)

uselabel

Creates a dataset containing value-label information

lname	value	label	trunc
coblab	-1	Missing	0
coblab	1	Australia	0
coblab	2	United Kingdom	0
coblab	3	Vietnam	0
coblab	4	China	0
coblab	5	Singapore	0
coblab	6	New Zealand	0
genderlab	-1	Missing	0
genderlab	1	Male	0
genderlab	2	Female	0
genderlab	3	Intersex	0
noyes	0	No	0
noyes	1	Yes	0

Extracting value label names

lname
coblab
coblab
coblab
coblab
coblab
coblab
coblab
coblab
genderlab
genderlab
genderlab
genderlab
genderlab
noyes
noyes

```
gen recnum=_n
```

- recnum contains the number of the current observation

```
levelsof lname, local(levels)
```

`"coblab"' `"genderlab"' `"noyes"'

- These are stored in the local macro `levels'

Creating the contents of each value label

	lname
1	coblab
2	coblab
3	coblab
4	coblab
5	coblab
6	coblab
7	coblab
8	genderlab
9	genderlab
10	genderlab
11	genderlab
12	noyes
13	noyes

```
foreach x of local levels {
     local fullab
     qui su recnum if lname=="`x'"
     local j=r(min)
     local k=r(max)
     forval i=`j'/`k' {
          local val=value[`i']
          local lab=label[`i']
          local fullab `fullab' `val', `lab' |
     }
     local lenlab=strlen("`fullab'")-2
     local fullab=substr("`fullab'",1,`lenlab')
}
```

Creating the contents of each value label

	lname
1	coblab
2	coblab
3	coblab
4	coblab
5	coblab
6	coblab
7	coblab
8	genderlab
9	genderlab
10	genderlab
11	genderlab
12	noyes
13	noyes

```
foreach x of local levels {
     local fullab
     qui su recnum if lname=="`x'"
     local j=r(min)
     local k=r(max)
     forval i=`j'/`k' {
          local val=value[`i']
          local lab=label[`i']
          local fullab `fullab' `val', `lab' |
     }
     local lenlab=strlen("`fullab'")-2
     local fullab=substr("`fullab'",1,`lenlab')
}
```

Example with coblab

	lname	value	label
1	coblab	-1	Missing
2	coblab	1	Australia
3	coblab	2	United Kingdom
4	coblab	3	Vietnam
5	coblab	4	China
6	coblab	5	Singapore
7	coblab	6	New Zealand

```
forval i=`j'/`k' {
        local val=value[`i']
        local lab=label[`i']
        local fullab `fullab' `val', `lab'
|
}
```

`i'=1

-1, Missing |

Example with coblab

	lname	value	label
1	coblab	-1	Missing
2	coblab	1	Australia
3	coblab	2	United Kingdom
4	coblab	3	Vietnam
5	coblab	4	China
6	coblab	5	Singapore
7	coblab	6	New Zealand

```
forval i=`j'/`k' {
        local val=value[`i']
        local lab=label[`i']
        local fullab `fullab' `val', `lab'
|
}
```

`i'=2

-1, Missing | 1, Australia |

Example with coblab

	lname	value	label
1	coblab	-1	Missing
2	coblab	1	Australia
3	coblab	2	United Kingdom
4	coblab	3	Vietnam
5	coblab	4	China
6	coblab	5	Singapore
7	coblab	6	New Zealand

```
forval i=`j'/`k' {
        local val=value[`i']
        local lab=label[`i']
        local fullab `fullab' `val', `lab'
|
}
```

`i'=3

-1, Missing | 1, Australia | 2, United Kingdom |

Example with coblab

	lname	value	label
1	coblab	-1	Missing
2	coblab	1	Australia
3	coblab	2	United Kingdom
4	coblab	3	Vietnam
5	coblab	4	China
6	coblab	5	Singapore
7	coblab	6	New Zealand

```
forval i=`j'/`k' {
        local val=value[`i']
        local lab=label[`i']
        local fullab `fullab' `val', `lab'
|
}
```

`i'=4

-1, Missing | 1, Australia | 2, United Kingdom | 3, Vietnam |

Example with coblab

	lname	value	label
1	coblab	-1	Missing
2	coblab	1	Australia
3	coblab	2	United Kingdom
4	coblab	3	Vietnam
5	coblab	4	China
6	coblab	5	Singapore
7	coblab	6	New Zealand

```
forval i=`j'/`k' {
        local val=value[`i']
        local lab=label[`i']
        local fullab `fullab' `val', `lab'
|
}
```

`i'=5

-1, Missing | 1, Australia | 2, United Kingdom | 3, Vietnam | 4, China |

Example with coblab

	lname	value	label
1	coblab	-1	Missing
2	coblab	1	Australia
3	coblab	2	United Kingdom
4	coblab	3	Vietnam
5	coblab	4	China
6	coblab	5	Singapore
7	coblab	6	New Zealand

```
forval i=`j'/`k' {
        local val=value[`i']
        local lab=label[`i']
        local fullab `fullab' `val', `lab'
|
}
```

`i'=6

-1, Missing | 1, Australia | 2, United Kingdom | 3, Vietnam | 4, China | 5, Singapore |

Example with coblab

	lname	value	label
1	coblab	-1	Missing
2	coblab	1	Australia
3	coblab	2	United Kingdom
4	coblab	3	Vietnam
5	coblab	4	China
6	coblab	5	Singapore
7	coblab	6	New Zealand

```
forval i=`j'/`k' {
        local val=value[`i']
        local lab=label[`i']
        local fullab `fullab' `val', `lab'
|
}
```

`i'=7

-1, Missing | 1, Australia | 2, United Kingdom | 3, Vietnam | 4, China | 5, Singapore | 6, New Zealand |

Example with coblab

```
foreach x of local levels {
        …
        forval i=`j'/`k' {
                local val=value[`i']
                local lab=label[`i']
                local fullab `fullab' `val', `lab' |
        }
        local lenlab=strlen("`fullab'")-2
        local fullab=substr("`fullab'",1,`lenlab')
}
```

-1, Missing | 1, Australia | 2, United Kingdom | 3, Vietnam | 4, China | 5, Singapore | 6, New Zealand |

-1, Missing | 1, Australia | 2, United Kingdom | 3, Vietnam | 4, China | 5, Singapore | 6, New Zealand

Allowing for extremely long strings

```
tempname mem

file write `mem' "`x'" _tab "`fullab'" _newline
```

- `file` allows for extremely long string values, up to 2-billion characters
- With `postfile` the limit is 2045 characters

One week after submitting my abstract for this meeting …

The Stata Journal (2017)
17, Number 1, pp. 130–138

Capturing a Stata dataset and releasing it into REDCap

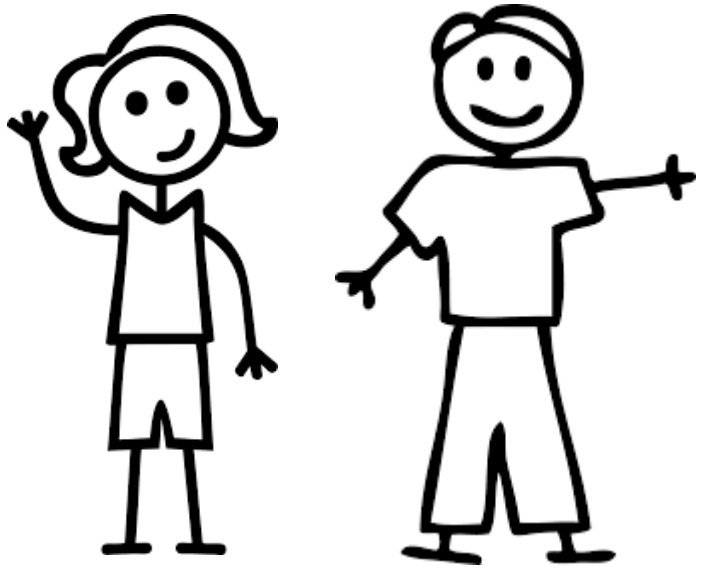
Seth T. Lirette
Department of Data Science
University of Mississippi Medical Center
Jackson, MS
and
Department of Biostatistics
University of Alabama at Birmingham
Birmingham, AL
slirette2@umc.edu

Samantha R. Seals
Department of Mathematics and Statistics
University of West Florida
Pensacola, FL

Chad Blackshear
Department of Data Science
University of Mississippi Medical Center
Jackson, MS

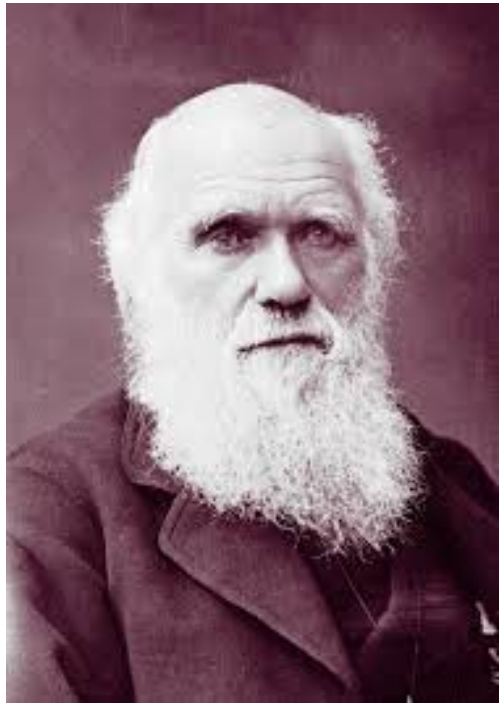
Warren May
School of Health Related Professions
University of Mississippi Medical Center
Jackson, MS

Beaten to the punch



Alfred Russel Wallace

Seth Lirette et al




R.I.P.
metadatacsv.a
do

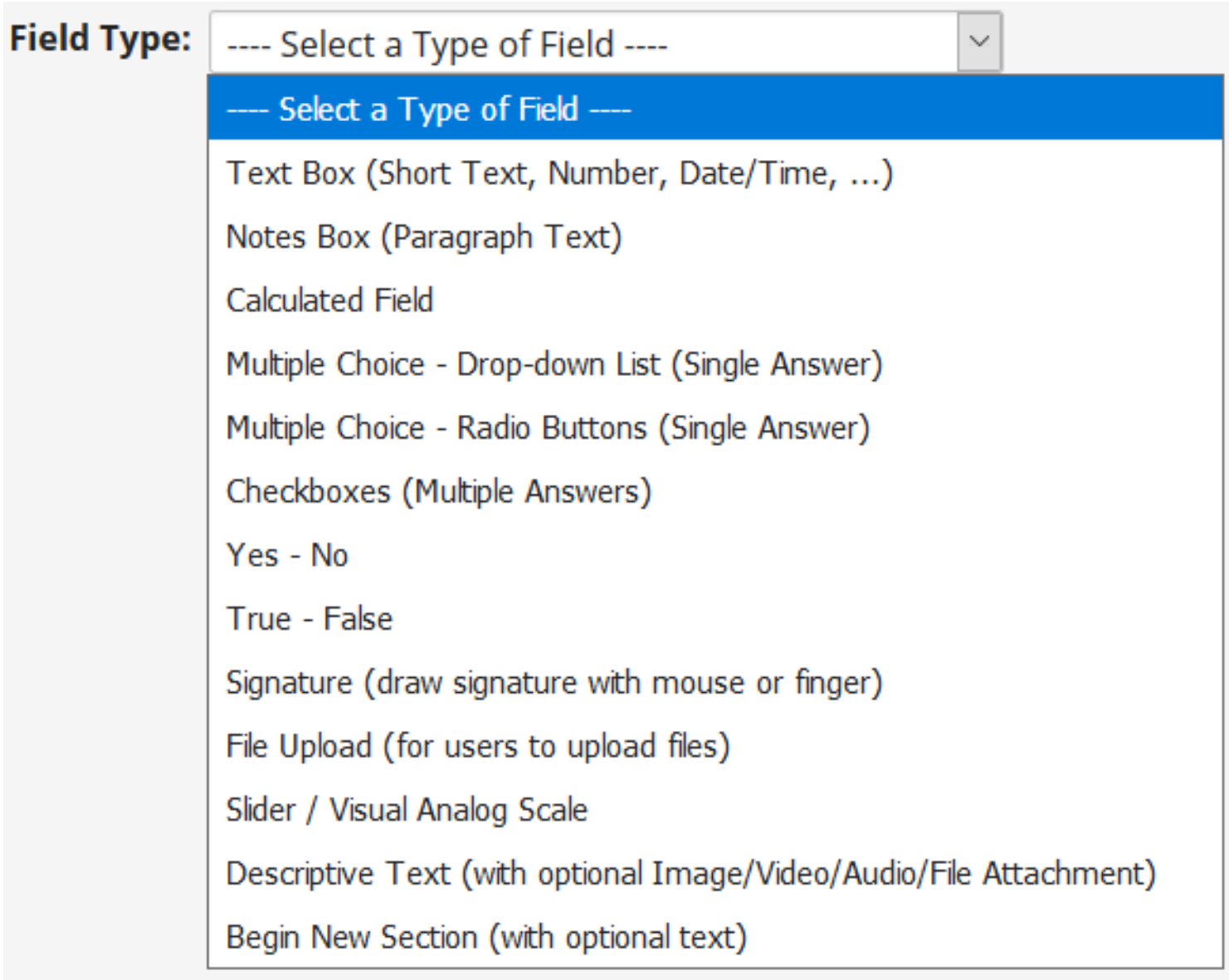
The `redcapture` command

redcapture syntax

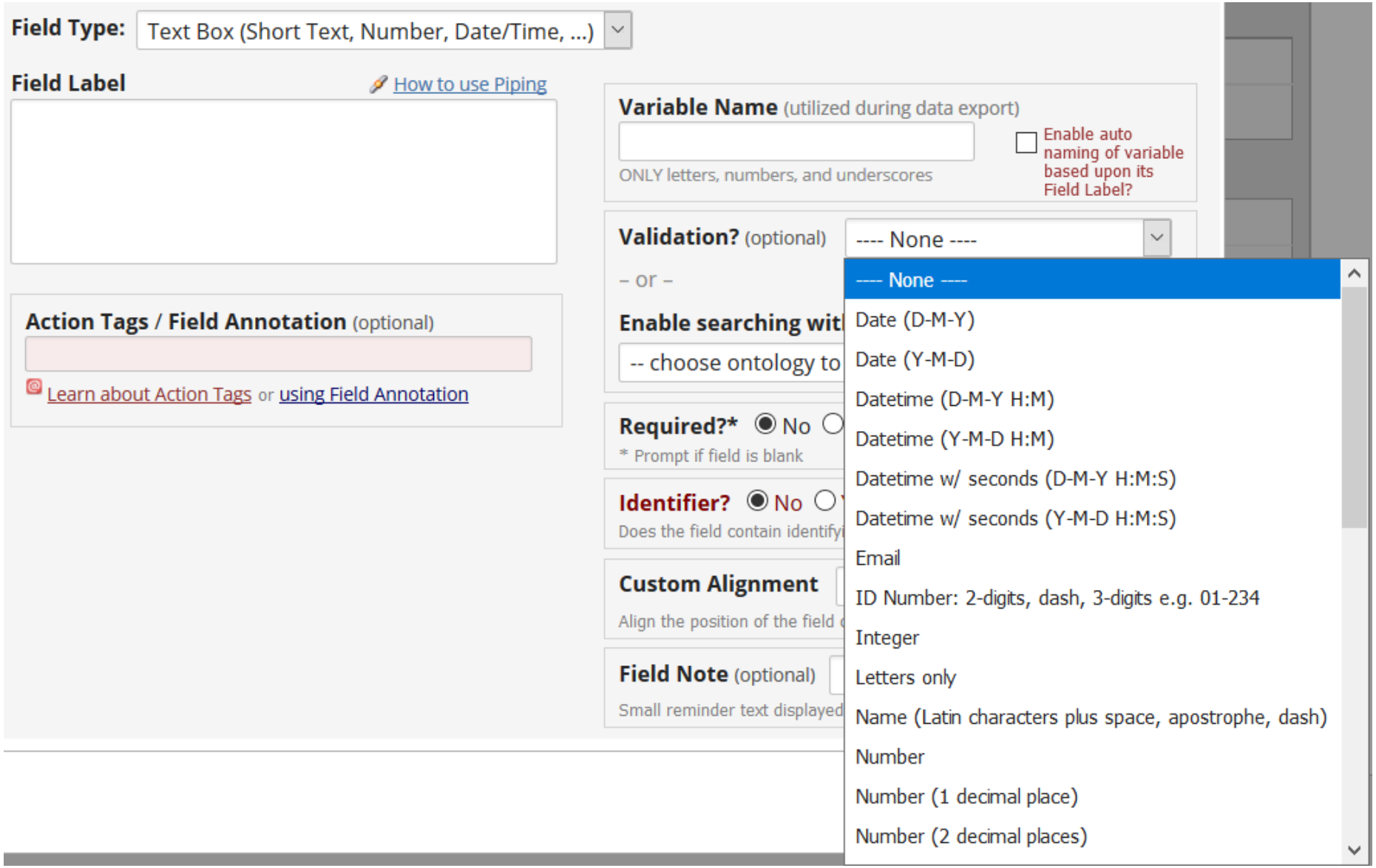
```
redcapture varlist, file(string) form(string)
[text(varlist) dropdown(varlist) radio(varlist)
header(string) validate(varlist)
validtype(validtypes)
validmin(minlist)validmax(maxlist)
matrix1(varlist) matrix2(varlist) matrix3(varlist)
matrix4(varlist) matrix5(varlist) matrix6(varlist)
matrix7(varlist) matrix8(varlist) matrix9(varlist)
matrix10(varlist)]
```

First, some background on

REDCap field types

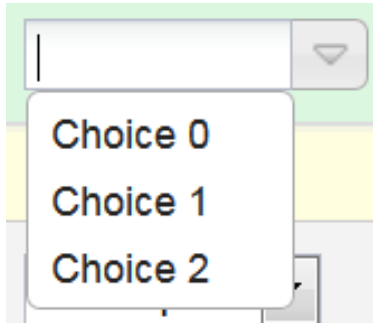


REDCap validations for text fields



Capturing categorical data in REDCap

Radio Group (Select One)	○ Choice 0 ○ Choice 1 ○ Choice 2
Drop-Down List (Select One)	
Checkboxes (Multi-select)	☐ Choice 0 ☐ Choice 1 ☐ Choice 2



	○ Choice 0 ◉ Choice 1 ○ Choice 2
	Choice 1
	☐ Choice 0 ☑ Choice 1 ☑ Choice 2

Checkbox groups are equivalent to a series of binary fields

	1 / Yes / Ticked	0 / No / Not ticked
Choice 0	○	◉
Choice 1	◉	○
Choice 2	◉	○

Checkboxes (Multi-select)	☐ Choice 0 ☑ Choice 1 ☑ Choice 2

X	Y	Z
checkboxquestion___0	checkboxquestion___1	checkboxquestion___2
0	1	1

Example Stata dataset

variable name	storage type	display format	value label	variable label
id	str9	%9s		**Participant ID**
consented	float	%9.0g	ynlab	**Is a consent document on file?**
age	float	%9.0g		**How old were you on your last birthday?**
race	float	%16.0g	rlab	**What is your race?**
sex	float	%9.0g		**What is your sex?**
bdate	float	%td		**What is your date of birth?**
sbp	float	%9.0g		**What was your last known systolic blood pressure?**
dbp	float	%9.0g		**What was your last known diastolic blood pressure?**
happy1	float	%26.0g	llab	**The staff greeted me in a professional and courteous manner.**
happy2	float	%26.0g	llab	**The waiting time to see a doctor was satisfactory.**
happy3	float	%26.0g	llab	**I would return to this hospital.**
comment	str4	%9s		**Comments**

Example script

```
redcapture *, file(example) form(example_form) header(Example) ///
     text(id age sex bdate sbp dbp comment)                      ///
     dropdown(consented race)                                    ///
     radio(happy1 happy2 happy3)                                 ///
     validate(id bdate dbp comment)                   ///
     validtype(ssn date_ymd integer alpha_only)             ///
     validmin(none 1/1/1900 20 none)                  ///
     validmax(none 12/31/2014 200 none)                  ///
     matrix1(happy1 happy2 happy3)
```

- **Metadata are saved in example.csv. This is the data dictionary that will be uploaded to REDCap.**
- **The form/instrument name in REDCap is example_form**
- **Its header is "Example"**

Example script

```
redcapture *, file(example) form(example_form) header(Example) ///
     text(id age sex bdate sbp dbp comment)                      ///
     dropdown(consented race)                                    ///
     radio(happy1 happy2 happy3)                                 ///
     validate(id bdate dbp comment)                   ///
     validtype(ssn date_ymd integer alpha_only)             ///
     validmin(none 1/1/1900 20 none)                  ///
     validmax(none 12/31/2014 200 none)                  ///
     matrix1(happy1 happy2 happy3)
```

For categorical variables. They must be numeric with value labels attached.

Example script

```
redcapture *, file(example) form(example_form) header(Example) ///
     text(id age sex bdate sbp dbp comment)                      ///
     dropdown(consented race)                                    ///
     radio(happy1 happy2 happy3)                                 ///
     validate(id bdate dbp comment)                   ///
     validtype(ssn date_ymd integer alpha_only)             ///
     validmin(none 1/1/1900 20 none)                  ///
     validmax(none 12/31/2014 200 none)                  ///
     matrix1(happy1 happy2 happy3)
```

- **These are text fields**
- **All variables in the validate() option must be declared as text fields**

Example script

```
redcapture *, file(example) form(example_form) header(Example) ///
     text(id age sex bdate sbp dbp comment)                      ///
     dropdown(consented race)                                    ///
     radio(happy1 happy2 happy3)                                 ///
     validate(id bdate dbp comment)                        ///
     validtype(ssn date_ymd integer alpha_only)                ///
     validmin(none 1/1/1900 20 none)                     ///
     validmax(none 12/31/2014 200 none)                    ///
     matrix1(happy1 happy2 happy3)
```

- **id is a social security number**
- **bdate is a date field in YMD format**
- **dbp is an integer**
- **comment is a string**

Example script

```
redcapture *, file(example) form(example_form) header(Example) ///
     text(id age sex bdate sbp dbp comment)                      ///
     dropdown(consented race)                                    ///
     radio(happy1 happy2 happy3)                                 ///
     validate(id bdate dbp comment)                   ///
     validtype(ssn date_ymd integer alpha_only)                ///
     validmin(none 1/1/1900 20 none)                  ///
     validmax(none 12/31/2014 200 none)                  ///
     matrix1(happy1 happy2 happy3)
```

- **To omit range checks for any or all of the validation variables, "none" should be entered into the corresponding location**
- **These are soft checks**

Example script

```
redcapture *, file(example) form(example_form) header(Example) ///
     text(id age sex bdate sbp dbp comment)                      ///
     dropdown(consented race)                                    ///
     radio(happy1 happy2 happy3)                                 ///
     validate(id bdate dbp comment)                   ///
     validtype(ssn date_ymd integer alpha_only)             ///
     validmin(none 1/1/1900 20 none)                  ///
     validmax(none 12/31/2014 200 none)                  ///
     matrix1(happy1 happy2 happy3)
```

- **Radio fields with a common set of response options can be grouped in a matrix**
- **See next slide**

Matrix of fields

Matrix group: matrix1

Variable: happy1

	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
The staff greeted me in a professional and courteous manner.	○	○	○	○	○

reset

Variable: happy2

The waiting time to see a doctor was satisfactory.	○	○	○	○	○

reset

Variable: happy3

I would return to this hospital.	○	○	○	○	○

reset

Data dictionary

The `redcapture` command created this data dictionary

A	B	C	D	E	F
VariableFieldName	FormName	SectionHe	FieldType	FieldLabel	ChoicesCalculationsSliderLabels
id	example_form	Example	text	Participant ID	
consented	example_form		dropdown	Is a consent document on file?	0,No | 1,Yes
age	example_form		text	How old were you on your last birthday?	
race	example_form		dropdown	What is your race?	1,Caucasian | 2,African American | 3,Other
sex	example_form		text	What is your sex?	
bdate	example_form		text	What is your date of birth?	
sbp	example_form		text	What was your last known systolic blood pressure?	
dbp	example_form		text	What was your last known diastolic blood pressure?	
happy1	example_form		radio	The staff greeted me in a professional and courteous manner.	1,Strongly Agree | 2,Agree | 3,Neither Agree nor Disagree | 4,Disagree | 5,Strongly Disagree
happy2	example_form		radio	The waiting time to see a doctor was satisfactory.	1,Strongly Agree | 2,Agree | 3,Neither Agree nor Disagree | 4,Disagree | 5,Strongly Disagree
happy3	example_form		radio	I would return to this hospital.	1,Strongly Agree | 2,Agree | 3,Neither Agree nor Disagree | 4,Disagree | 5,Strongly Disagree
comment	example_form		text	Comments	

…which can be uploaded into REDCap

In conclusion …

1. Ensure data will be retrievable 10 or 20 years from now
2. Ensure the next generation of researchers will be able to understand currently archived data

How?

By storing both data and metadata in text files

Stata's `export delimited` and `redcapture` commands facilitates this

Data and metadata can be uploaded to data capture software such as REDCap