COMMITTING TO DATA QUALITY REVIEW

Limor Peer, Yale University
Ann Green, Digital Lifecycle Research & Consulting
Elizabeth Stephenson, UCLA

@l_peer

IDCC14

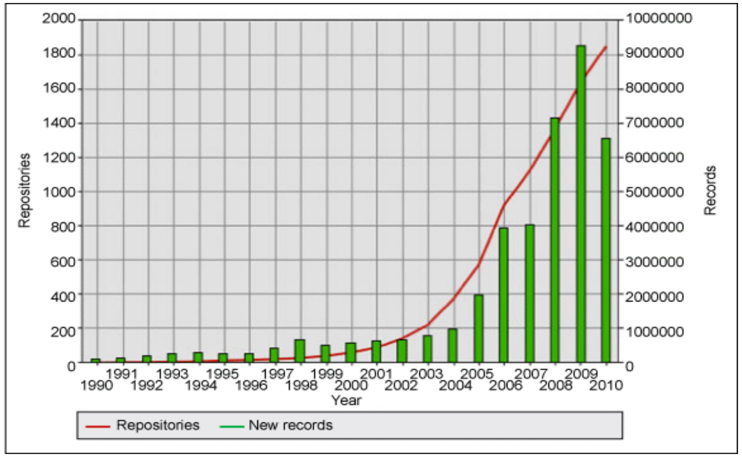
February 2014

More data are available in more ways

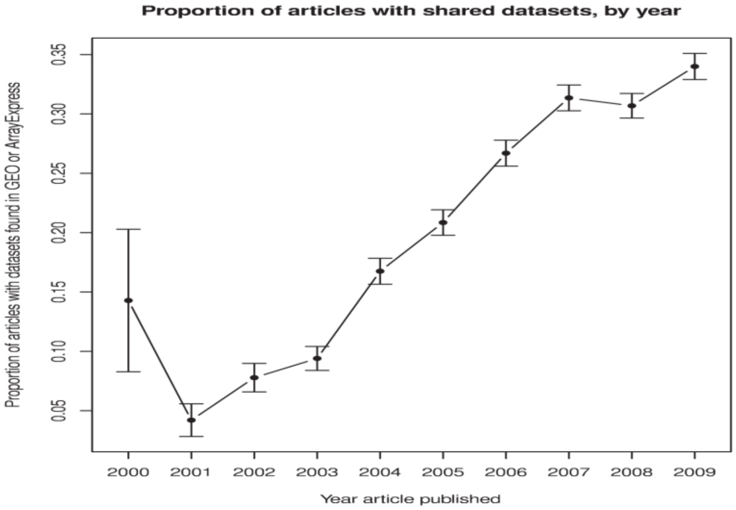
Databib
Find Repositories | Submit | Connect | About

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

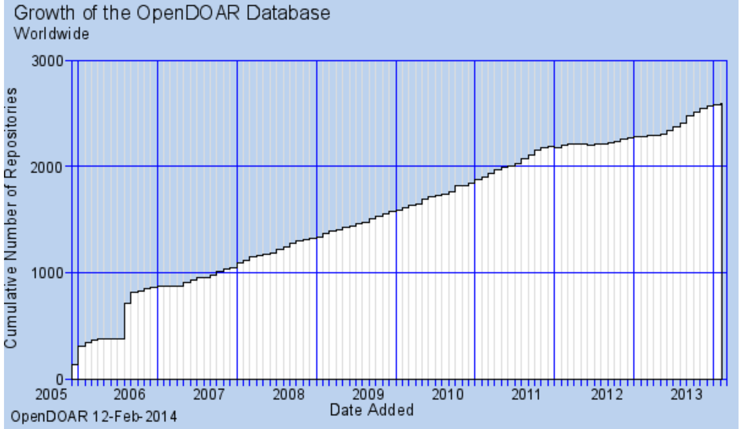
Over 600 data repositories in Databib and re3data (Feb 2014)



Source: ROAR



Source: Piwowar 2013



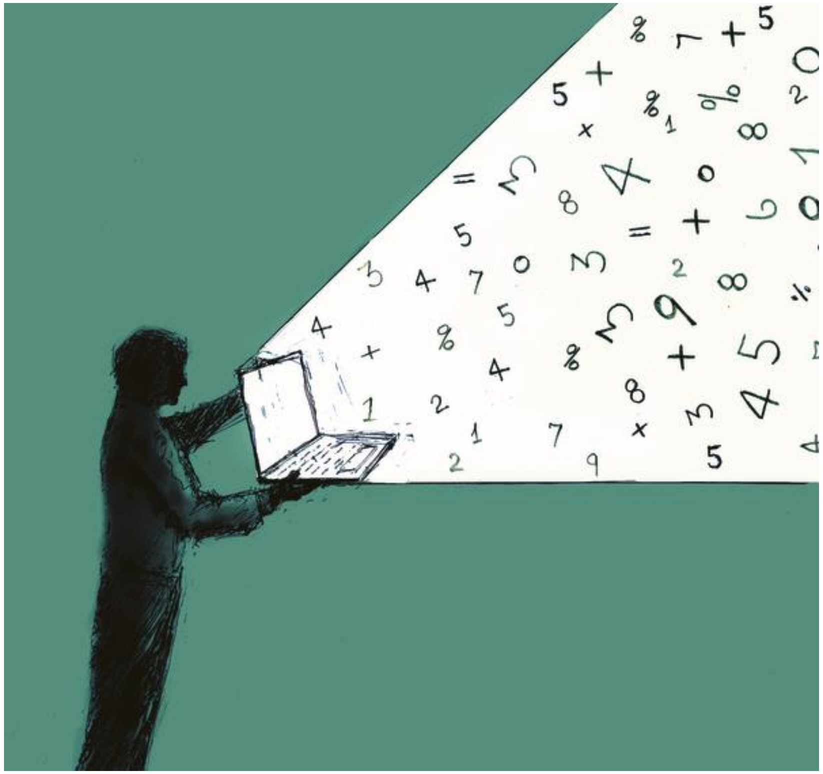
Source: OpenDOAR

Data are made available

…so they can be reused

“It's important to allow readers and reviewers to see exactly how you arrive at your results. Publishing data and code allows your science to be reproducible.”

-- Amy Zanne, biologist, George Washington University

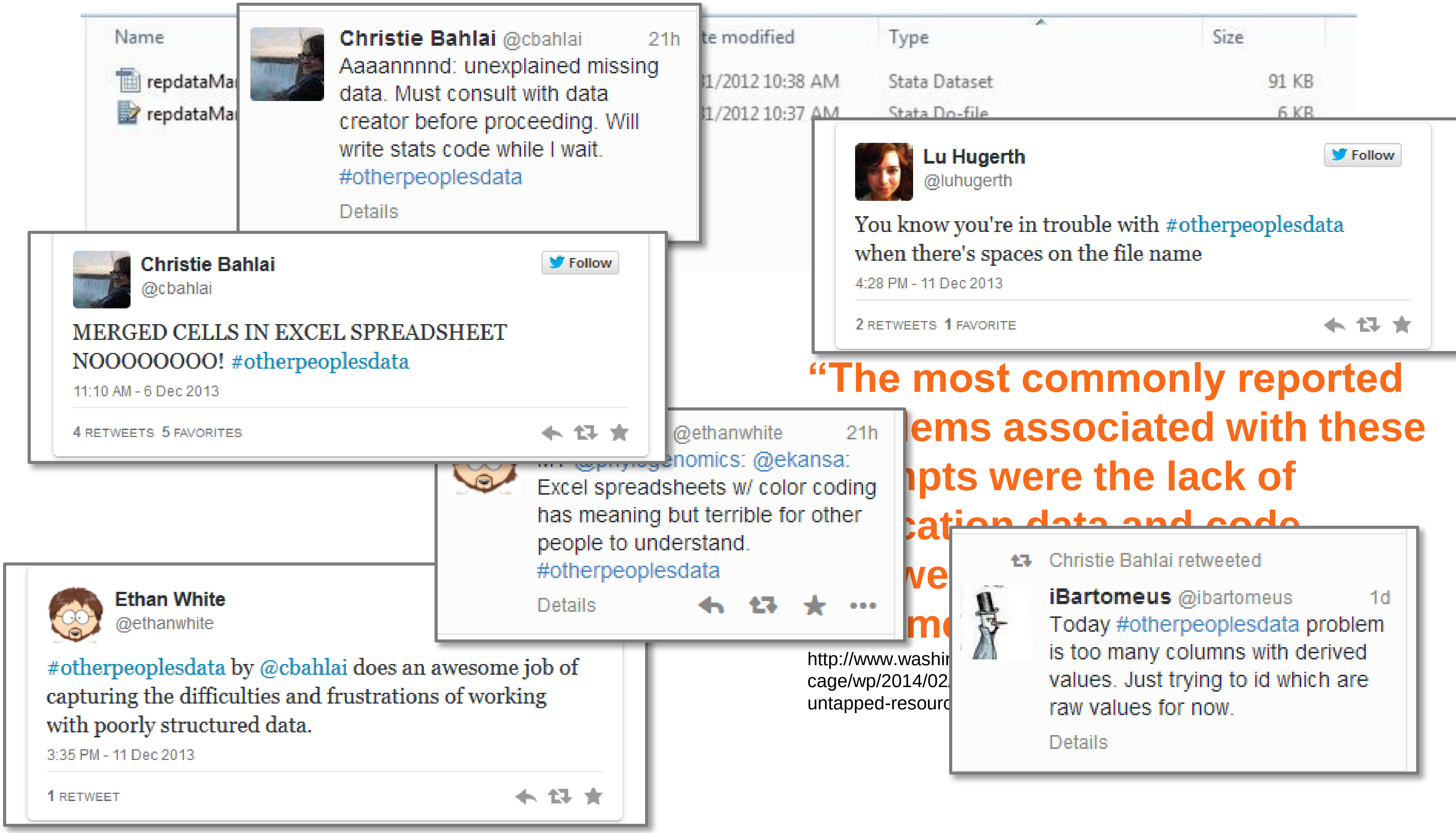


Source: Nature (GARY WATERS/IKON IMAGES/CORBIS)

“…we know from the Rheinhart Rogoff case that we simply need one student to reuse the data in order to achieve a huge impact.”

-- David Osimo, open-evidence.com

Using other people’s data…



Name	te modified	Type	Size
repdataMa	1/2012 10:38 AM	Stata Dataset	91 KB
repdataMa	1/2012 10:37 AM	Stata Do-file	6 KB

Christie Bahlai @cbahlai 21h
Aaaannnnd: unexplained missing data. Must consult with data creator before proceeding. Will write stats code while I wait. #otherpeoplesdata
Details

Christie Bahlai @cbahlai
MERGED CELLS IN EXCEL SPREADSHEET NOOOOOOOO! #otherpeoplesdata
11:10 AM - 6 Dec 2013
4 RETWEETS 5 FAVORITES

Lu Hugerth @luhugerth
You know you're in trouble with #otherpeoplesdata when there's spaces on the file name
4:28 PM - 11 Dec 2013
2 RETWEETS 1 FAVORITE

@ethanwhite 21h
RT @phylogenomics: @ekansa: Excel spreadsheets w/ color coding has meaning but terrible for other people to understand. #otherpeoplesdata
Details

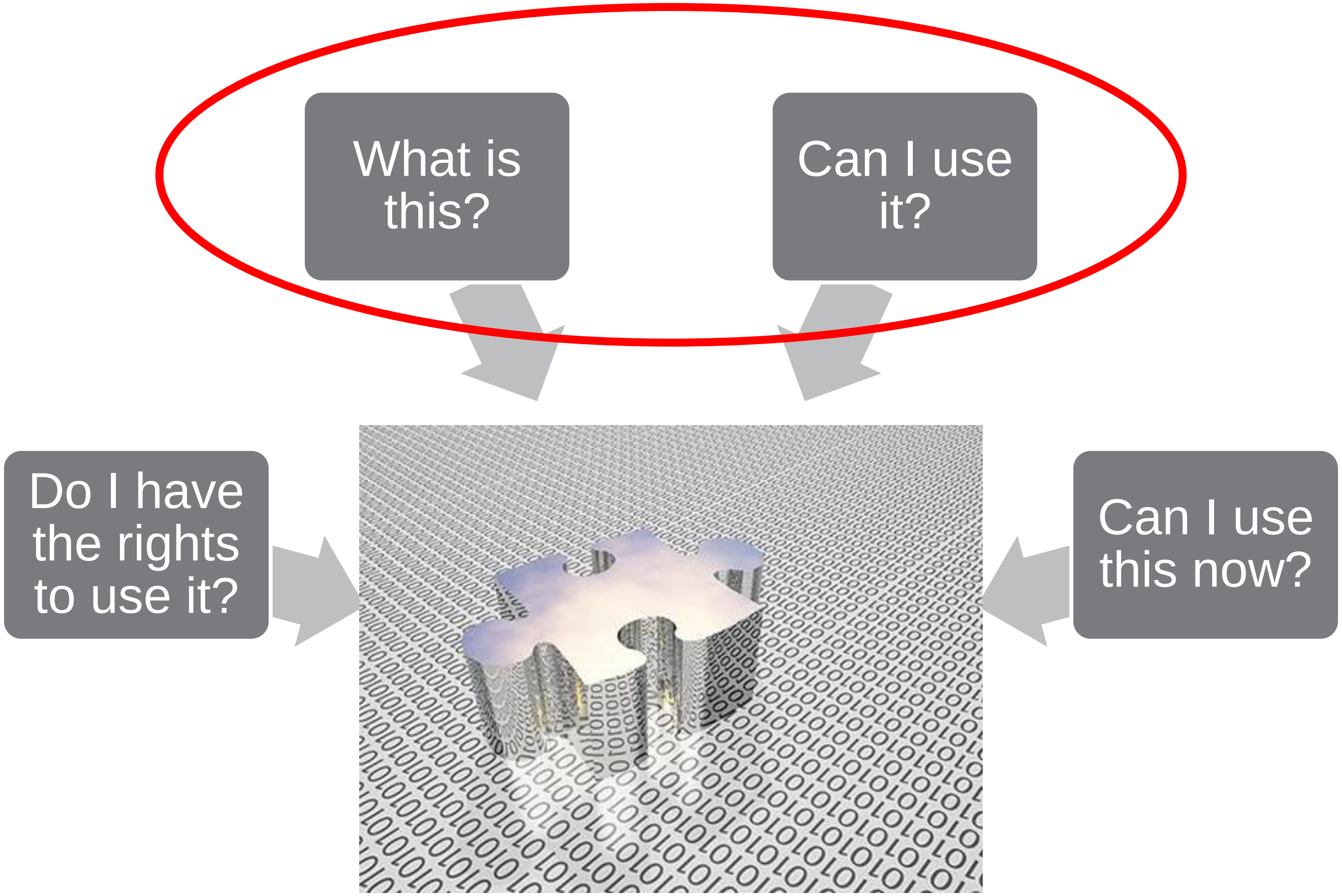
Ethan White @ethanwhite
#otherpeoplesdata by @cbahlai does an awesome job of capturing the difficulties and frustrations of working with poorly structured data.
3:35 PM - 11 Dec 2013
1 RETWEET

Christie Bahlai retweeted
iBartomeus @ibartomeus 1d
Today #otherpeoplesdata problem is too many columns with derived values. Just trying to id which are raw values for now.
Details

“The most commonly reported
lems associated with these
pts were the lack of
cation data and code
we
me

http://www.washi
cage/wp/2014/02/
untapped-resourc

Data reuse & the crisis of quality



What is this?

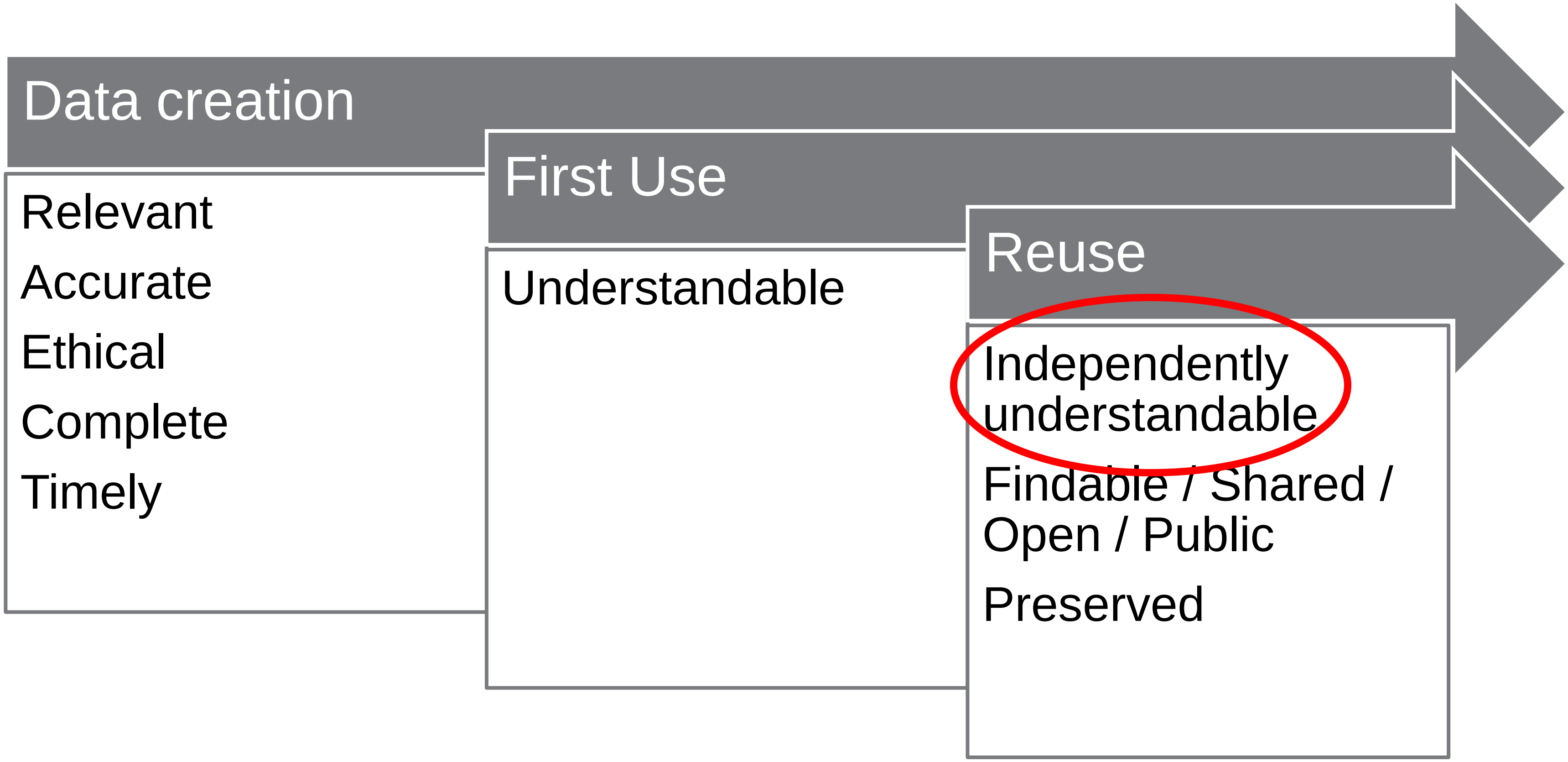
Can I use it?

Do I have the rights to use it?

Can I use this now?

Source: fehrandpeers.com

Aspects of data quality



Data creation

Relevant

Accurate

Ethical

Complete

Timely

First Use

Understandable

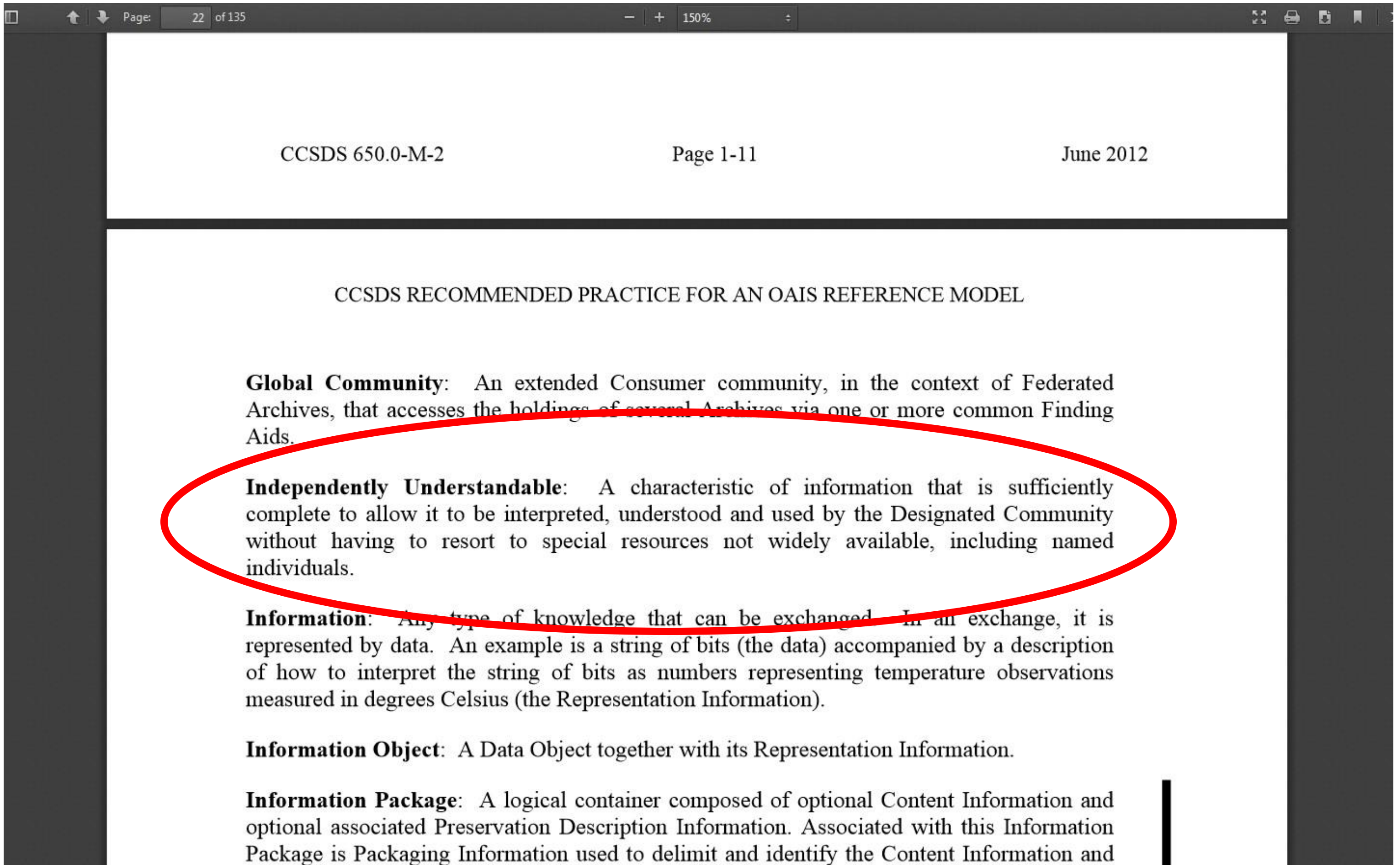
Reuse

Independently understandable

Findable / Shared / Open / Public

Preserved

Independently understandable data for informed reuse by a designated community



CCSDS 650.0-M-2 Page 1-11 June 2012

CCSDS RECOMMENDED PRACTICE FOR AN OAIS REFERENCE MODEL

Global Community: An extended Consumer community, in the context of Federated Archives, that accesses the holdings of several Archives via one or more common Finding Aids.

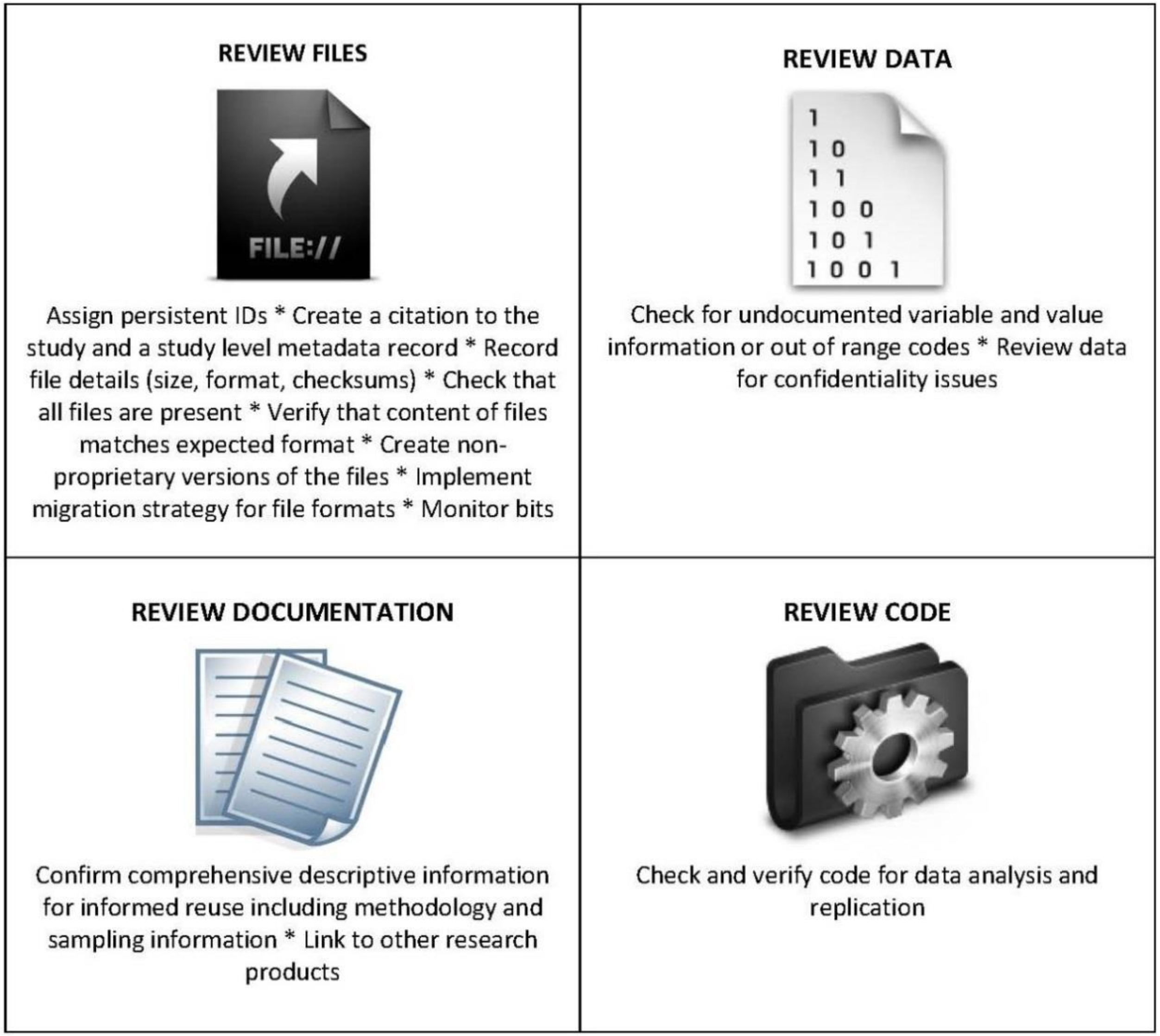
Independently Understandable: A characteristic of information that is sufficiently complete to allow it to be interpreted, understood and used by the Designated Community without having to resort to special resources not widely available, including named individuals.

Information: Any type of knowledge that can be exchanged. In an exchange, it is represented by data. An example is a string of bits (the data) accompanied by a description of how to interpret the string of bits as numbers representing temperature observations measured in degrees Celsius (the Representation Information).

Information Object: A Data Object together with its Representation Information.

Information Package: A logical container composed of optional Content Information and optional associated Preservation Description Information. Associated with this Information Package is Packaging Information used to delimit and identify the Content Information and

Data Quality Review



REVIEW FILES Assign persistent IDs * Create a citation to the study and a study level metadata record * Record file details (size, format, checksums) * Check that all files are present * Verify that content of files matches expected format * Create non-proprietary versions of the files * Implement migration strategy for file formats * Monitor bits	**REVIEW DATA** Check for undocumented variable and value information or out of range codes * Review data for confidentiality issues
REVIEW DOCUMENTATION Confirm comprehensive descriptive information for informed reuse including methodology and sampling information * Link to other research products	**REVIEW CODE** Check and verify code for data analysis and replication

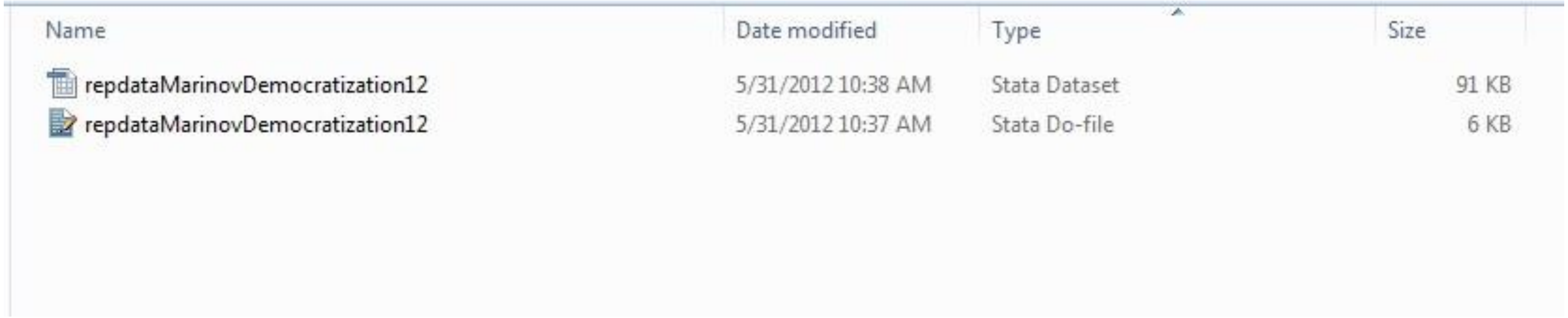
Data archives committed to DQR

Social Science Data Archive
University of California, Los Angeles

ICPSR
INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH

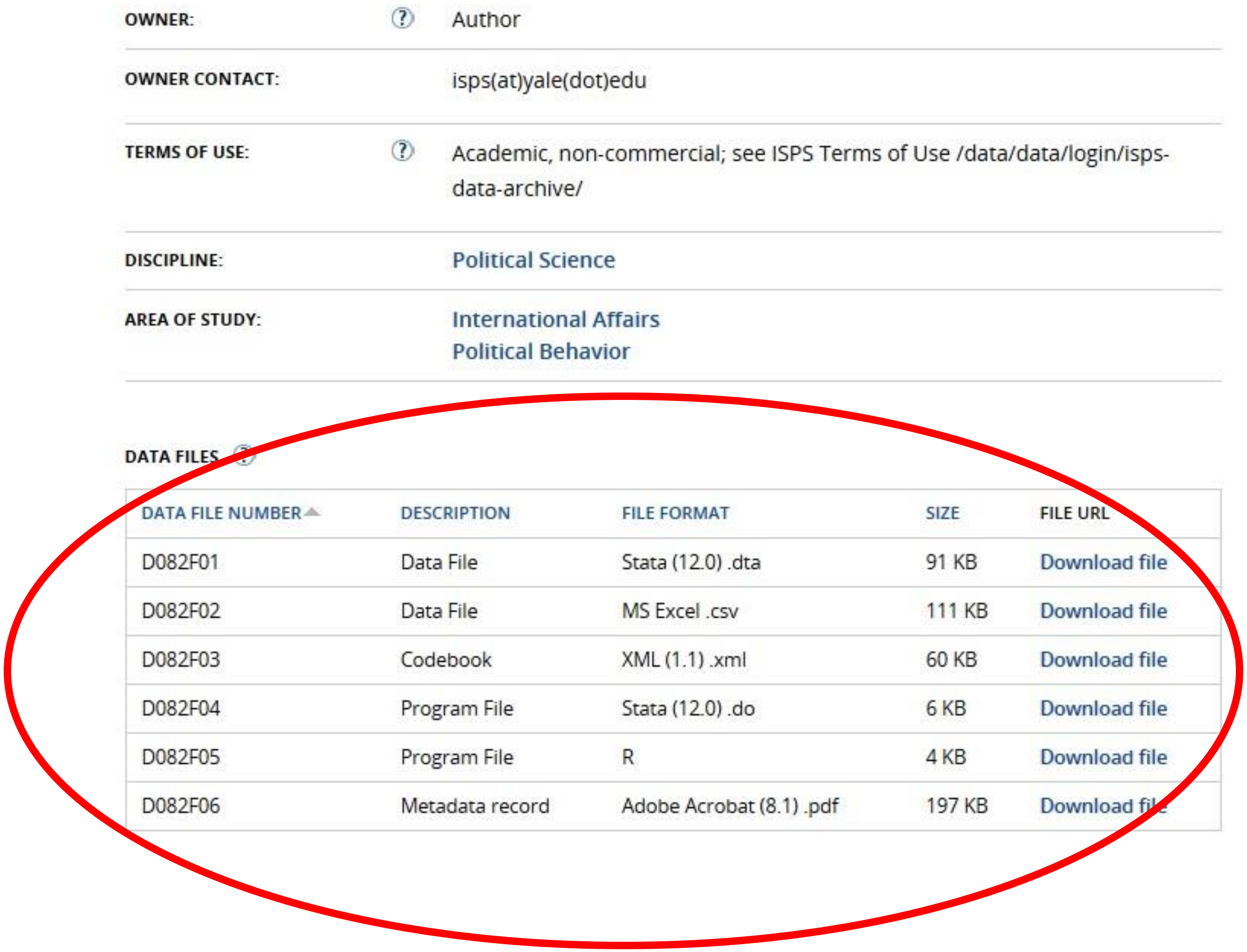
Yale
ISPS

File set before



Name	Date modified	Type	Size
repdataMarinovDemocratization12	5/31/2012 10:38 AM	Stata Dataset	91 KB
repdataMarinovDemocratization12	5/31/2012 10:37 AM	Stata Do-file	6 KB

File set after

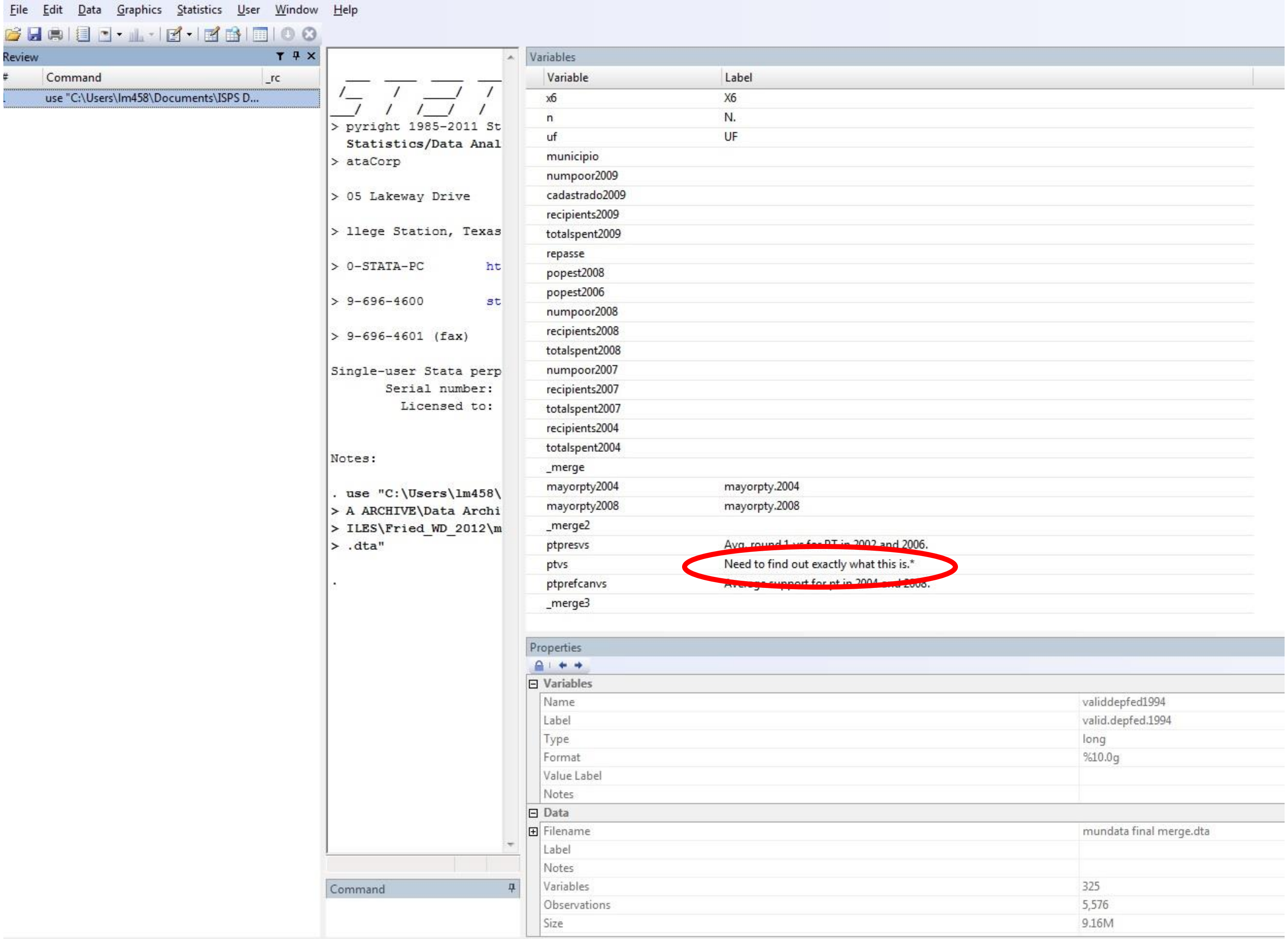


OWNER:	Author
OWNER CONTACT:	isps(at)yale(dot)edu
TERMS OF USE:	Academic, non-commercial; see ISPS Terms of Use /data/data/login/isps-data-archive/
DISCIPLINE:	Political Science
AREA OF STUDY:	International Affairs Political Behavior

DATA FILES

DATA FILE NUMBER	DESCRIPTION	FILE FORMAT	SIZE	FILE URL
D082F01	Data File	Stata (12.0) .dta	91 KB	Download file
D082F02	Data File	MS Excel .csv	111 KB	Download file
D082F03	Codebook	XML (1.1) .xml	60 KB	Download file
D082F04	Program File	Stata (12.0) .do	6 KB	Download file
D082F05	Program File	R	4 KB	Download file
D082F06	Metadata record	Adobe Acrobat (8.1) .pdf	197 KB	Download file

Missing labels



#	Command	_rc
	use "C:\Users\lm458\Documents\ISPS D...	

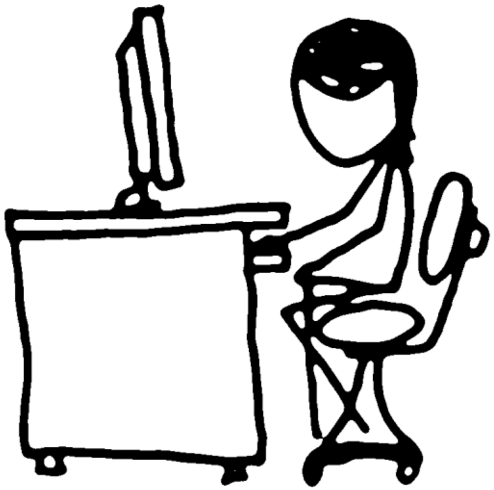
```
 ___  ____  ____  ____
/__    /   ____/   /
___/   /   /___/   /
> pyright 1985-2011 St
  Statistics/Data Anal
> ataCorp

> 05 Lakeway Drive

> llege Station, Texas

> 0-STATA-PC          ht

> 9-696-4600          st

> 9-696-4601 (fax)

Single-user Stata perp
       Serial number:
         Licensed to:


Notes:

. use "C:\Users\lm458\
> A ARCHIVE\Data Archi
> ILES\Fried_WD_2012\m
> .dta"

.
```

Variable	Label
x6	X6
n	N.
uf	UF
municipio	
numpoor2009	
cadastrado2009	
recipients2009	
totalspent2009	
repasse	
popest2008	
popest2006	
numpoor2008	
recipients2008	
totalspent2008	
numpoor2007	
recipients2007	
totalspent2007	
recipients2004	
totalspent2004	
_merge	
mayorpty2004	mayorpty.2004
mayorpty2008	mayorpty.2008
_merge2	
ptpresvs	Avg. round 1 vs for PT in 2002 and 2006.
ptvs	Need to find out exactly what this is.*
ptprefcanvs	Average support for pt in 2004 and 2008.
_merge3	

⊟ Variables	
Name	validdepfed1994
Label	valid.depfed.1994
Type	long
Format	%10.0g
Value Label	
Notes	
⊟ Data	
⊞ Filename	mundata final merge.dta
Label	
Notes	
Variables	325
Observations	5,576
Size	9.16M

Missing labels

RA: “We are missing labels for the following variables:
_n1, _n0, V1 and V0.”

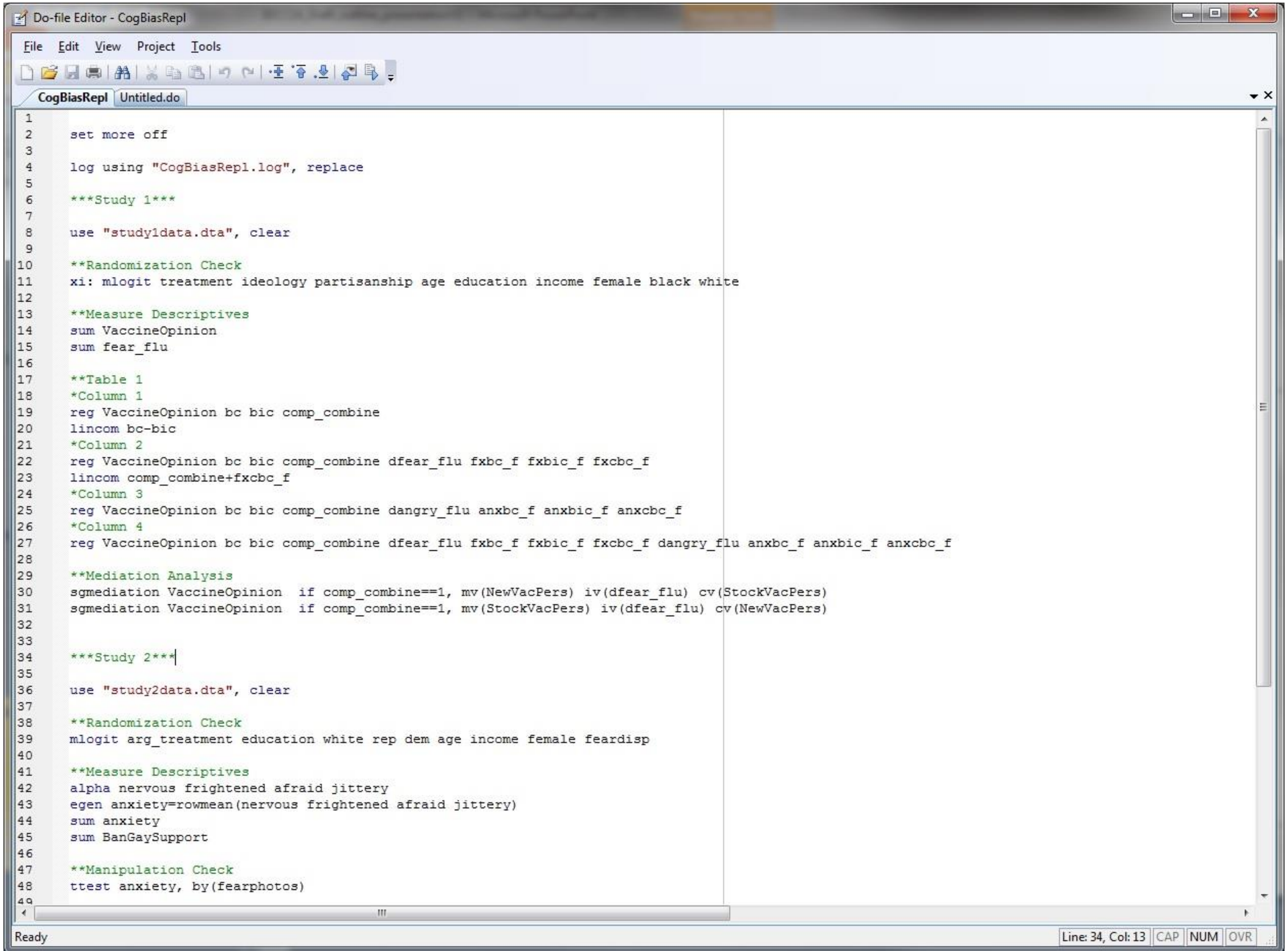


http://xkcd.com/662/ Creative Commons Attribution-Noncommercial

Researcher: “Here are the labels:
_n1 is the number of observations in the treated strata before matching
_n0 is the number of observations in the comparison strata before matching
v1 = turnout for treated observations
v0 = turnout for comparison observations

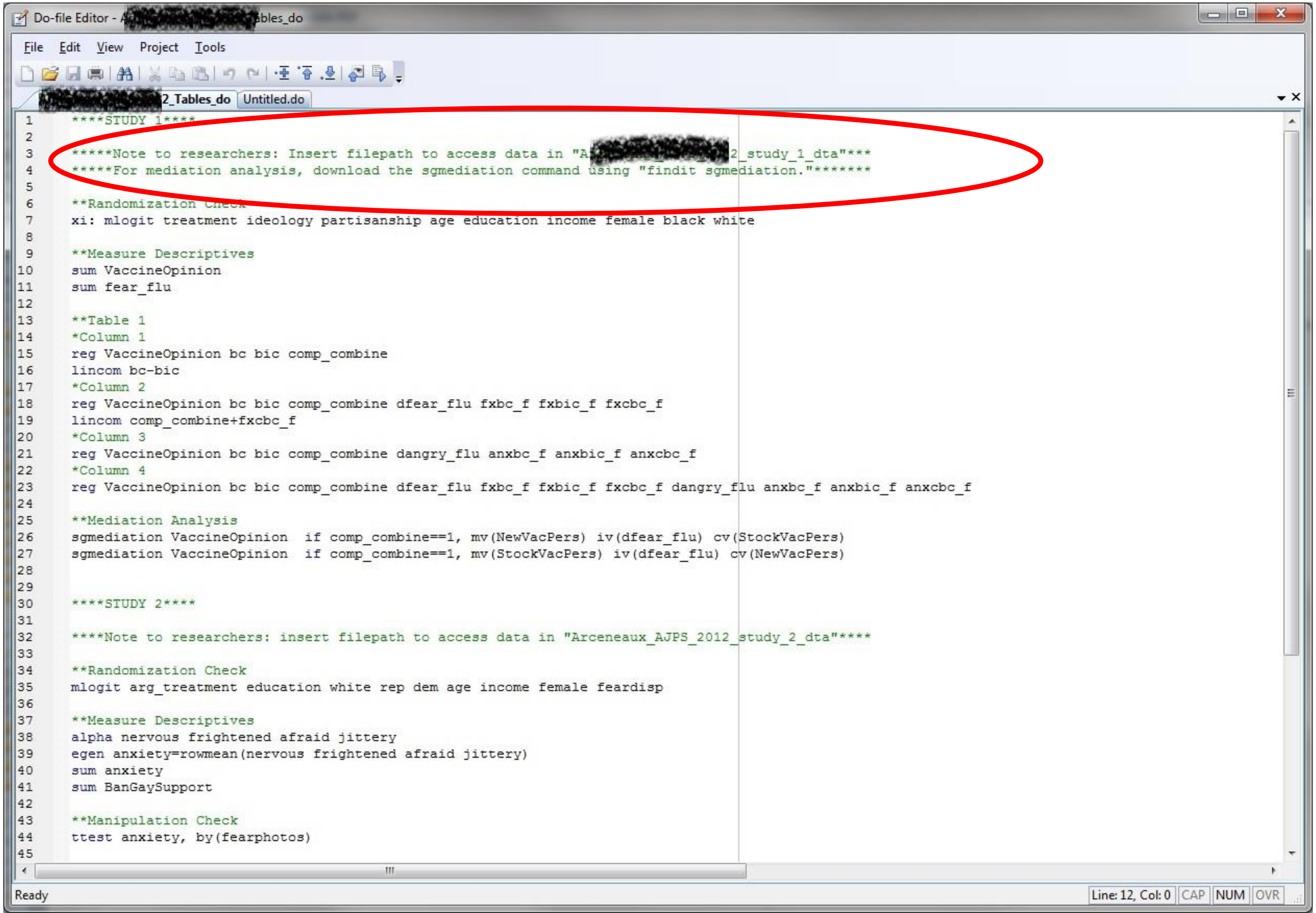
… this reminds me that I needed to include the .ado code in the Matching Code folder. I just did that and updated the readme file. Boy, the things your forget about after not thinking about something for two years!”

Code before



```

set more off

log using "CogBiasRepl.log", replace

***Study 1***

use "study1data.dta", clear

**Randomization Check
xi: mlogit treatment ideology partisanship age education income female black white

**Measure Descriptives
sum VaccineOpinion
sum fear_flu

**Table 1
*Column 1
reg VaccineOpinion bc bic comp_combine
lincom bc-bic
*Column 2
reg VaccineOpinion bc bic comp_combine dfear_flu fxbc_f fxbic_f fxcbc_f
lincom comp_combine+fxcbc_f
*Column 3
reg VaccineOpinion bc bic comp_combine dangry_flu anxbc_f anxbic_f anxcbc_f
*Column 4
reg VaccineOpinion bc bic comp_combine dfear_flu fxbc_f fxbic_f fxcbc_f dangry_flu anxbc_f anxbic_f anxcbc_f

**Mediation Analysis
sgmediation VaccineOpinion  if comp_combine==1, mv(NewVacPers) iv(dfear_flu) cv(StockVacPers)
sgmediation VaccineOpinion  if comp_combine==1, mv(StockVacPers) iv(dfear_flu) cv(NewVacPers)


***Study 2***

use "study2data.dta", clear

**Randomization Check
mlogit arg_treatment education white rep dem age income female feardisp

**Measure Descriptives
alpha nervous frightened afraid jittery
egen anxiety=rowmean(nervous frightened afraid jittery)
sum anxiety
sum BanGaySupport

**Manipulation Check
ttest anxiety, by(fearphotos)
```

Code after



```
****STUDY 1****

*****Note to researchers: Insert filepath to access data in "[redacted]2_study_1_dta"***
*****For mediation analysis, download the sgmediation command using "findit sgmediation."*******

**Randomization Check
xi: mlogit treatment ideology partisanship age education income female black white

**Measure Descriptives
sum VaccineOpinion
sum fear_flu

**Table 1
*Column 1
reg VaccineOpinion bc bic comp_combine
lincom bc-bic
*Column 2
reg VaccineOpinion bc bic comp_combine dfear_flu fxbc_f fxbic_f fxcbc_f
lincom comp_combine+fxcbc_f
*Column 3
reg VaccineOpinion bc bic comp_combine dangry_flu anxbc_f anxbic_f anxcbc_f
*Column 4
reg VaccineOpinion bc bic comp_combine dfear_flu fxbc_f fxbic_f fxcbc_f dangry_flu anxbc_f anxbic_f anxcbc_f

**Mediation Analysis
sgmediation VaccineOpinion  if comp_combine==1, mv(NewVacPers) iv(dfear_flu) cv(StockVacPers)
sgmediation VaccineOpinion  if comp_combine==1, mv(StockVacPers) iv(dfear_flu) cv(NewVacPers)


****STUDY 2****

****Note to researchers: insert filepath to access data in "Arceneaux_AJPS_2012_study_2_dta"****

**Randomization Check
mlogit arg_treatment education white rep dem age income female feardisp

**Measure Descriptives
alpha nervous frightened afraid jittery
egen anxiety=rowmean(nervous frightened afraid jittery)
sum anxiety
sum BanGaySupport

**Manipulation Check
ttest anxiety, by(fearphotos)

```

DQR process in other repositories

REVIEW FILES
Create persistent ID
Record file sizes and formats
Create checksums
Check for completeness, confirm all files are present (data, and required documentation and code if available)
Create study-level metadata record including file information
Create citation
Create non-proprietary file formats for preservation
REVIEW DOCUMENTATION
Confirm comprehensive descriptive information
Confirm methodology and sampling information
Create documentation compliant with community standards, e.g., DDI XML
REVIEW DATA
Run frequencies and check for undocumented or out of range codes
Standardize missing values; check for consistency and skip patterns
Check and edit variable and value labels
Check and add question wording (surveys)
Review data for confidentiality issues; Recode variables to address confidentiality concerns
Generate multiple data formats for dissemination
REVIEW CODE
Check and verify replication code
PUBLISH & LINK
Publish to access system
Link to other research products (e.g., publications, registries, grants)
PRESERVE
Migration strategy for file formats
Monitor bits

DQR by researchers

“No matter how invested in their own work, scientists appear to be “poor stewards” of their own work, the study concluded.”

– Kevin Fogarty, Slashdot

□ Data management plans

□ The research workflow

ROpenSci

CENTER FOR OPEN SCIENCE

Open Science Framework

Projects

□ Post-publication peer review

runmycode

DQR by journals

Uneven oversight of data deposits and no DQR

- Stricter policies & guides by journals
- Replication audits
- Data journals

SCIENTIFIC DATA

Journal of open psychology data

A peer-reviewed open-access journal
Biodiversity Data Journal
Making your data count! ISSN 1314-2828 (online)

F1000Research

DQR: A community commitment

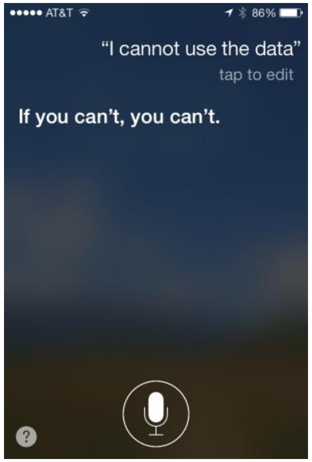
Reviewing the quality of the data is an obligation of *any* entity that assumes responsibility over the data.

It’s in everyone’s interest!

Unusable data = lost data

Image: Shutterstock.com/Lightspring

Thank you!



“I cannot use the data”

tap to edit

If you can’t, you can’t.

limor.peer@yale.edu; @l_peer

greenann@gmail.com; @annthegreen

libbie@ucla.com; @libbieatUCLA