ResBaz 2020 Pick n Mix

Intro to Tidy Data and OpenRefine

Yvette Wharton | Centre for eResearch
Te Whare Wānanga o Tāmaki Makaurau
The University of Auckland
@y_vettles
Pronouns: she, her

#resbaz2020 #resbazpicnmix

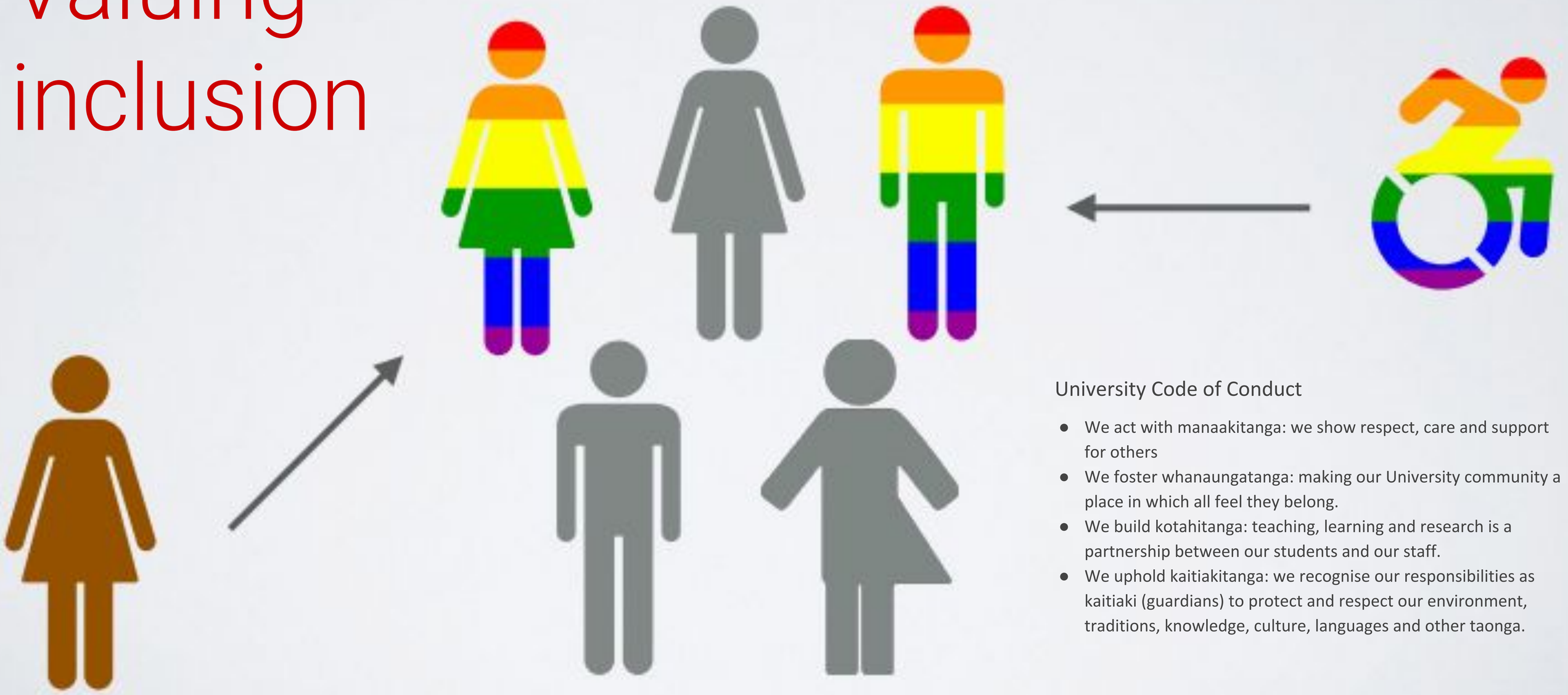
ResBaz 2020: Pick n Mix

Key Stories - join us to listen over lunch

Mon. 12-1pm	*Welcome to ResBaz 2020 : Pick n Mix* https://auckland.zoom.us/j/99720152410 Passcode: 640143
Tues. 12-1pm	*Harnessing the disruptive nature of portable sequencing for community empowerment* https://vuw.zoom.us/j/95143657235 Passcode: 718144
Wed. 12-1pm	*From Classics to Computer Science and back again* https://vuw.zoom.us/j/98866515882
Thur. 12-1pm	*Performance - Music performance and project showcase* https://vuw.zoom.us/j/98265221971
Fri. 12-1pm	*Exploring the cultural horizons of open science: your research and your life* https://auckland.zoom.us/j/98447731180 Passcode: 404029

#resbaz2020 #resbazpicnmix email: researchdata@auckland.ac.nz

Valuing inclusion



University Code of Conduct

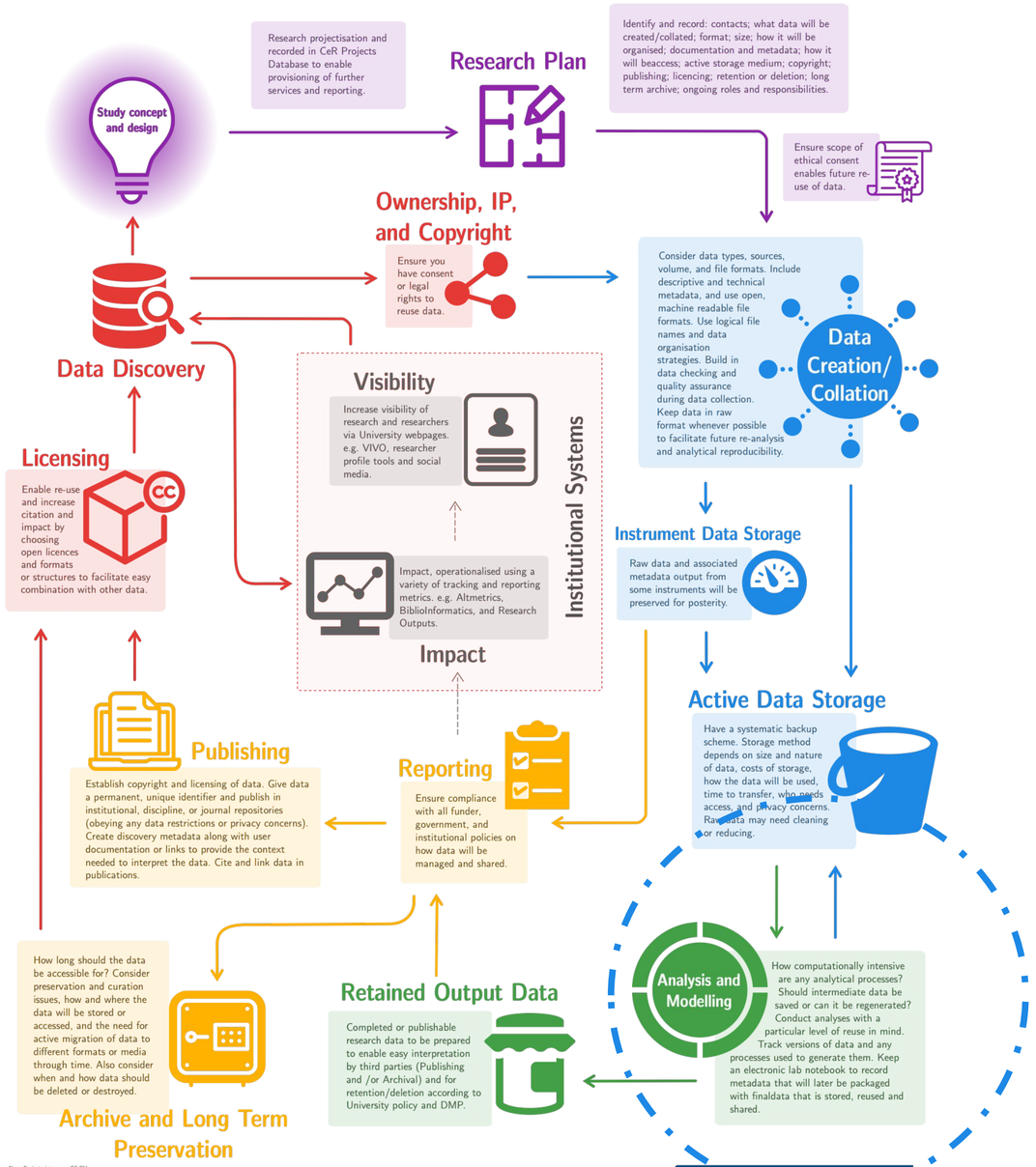
- We act with manaakitanga: we show respect, care and support for others
- We foster whanaungatanga: making our University community a place in which all feel they belong.
- We build kotahitanga: teaching, learning and research is a partnership between our students and our staff.
- We uphold kaitiakitanga: we recognise our responsibilities as kaitiaki (guardians) to protect and respect our environment, traditions, knowledge, culture, languages and other taonga.

Hi

Outline

1. Overview of Tidy Data
2. Dataset example
3. Tidying Data on OpenRefine

RESEARCH DATA MANAGEMENT



Data lifecycle model adapted from McLean, C. A. (2016): The Research Data Lifecycle. figshare. https://dx.doi.org/10.17608/k6.auckland.3102922.v1

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

Centre for eResearch

Typical Data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	CE	CF	CG	CH
1	Combined data from Rangitoto trip 2015																																		
2	16+ m island size																																		
3																																			
4	NORTH	Saturday															Sunday																		
5	Species	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	Mean	Std Dev	Sqrt n	Std Err
6	*Asplenium flaccidum*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0				
7	*Asplenium flabellifolium*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0				
8	*Asplenium oblongifolium*	1	3	3	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	1	2	0	0	0	0				
9	*Ctenopteris heterophylla*	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0				
10	*Hymenophyllum spp.*	0	3	3	0	0	1	0	2	1	0	0	1	1	0	0	0	0	0	2	0	0	0	0	0	0	0	2	0	1	0				
11	*Pellaea rotundifolia*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0				
12	*Microsorum pustulatum*	1	2	0	3	3	3	5	3	2	3	2	2	2	2	0	3	3	3	2	1	1	0	2	2	2	2	1	1	2	1				
13	*Pteridium esculentum*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0				
14	*Pyrrosia eleagnifoila*	3	1	3	0	0	1	1	2	2	0	2	1	1	0	0	0	0	3	1	0	3	0	0	2	2	3	2	3	1	0				
15	*Trichomanes reniforme*	1	3	4	0	0	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0				
16	*Acianthus sinclairii*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0				
17																																			
18	CENTRE	Saturday															Sunday																		
19	Species	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	Mean	Std Dev	Sqrt n	Std Err
20	*Asplenium flaccidum*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0				
21	*Asplenium flabellifolium*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	2	0	1	0				
22	*Asplenium oblongifolium*	0	0	1	0	0	2	3	1	2	0	2	1	1	0	0	0	0	0	0	0	1	1	0	1	1	0	3	3	1	0				
23	*Ctenopteris heterophylla*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	2	0	0	0				
24	*Hymenophyllum spp.*	0	2	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0				
25	*Pellaea rotundifolia*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0				
26	*Microsorum pustulatum*	2	2	3	0	0	1	2	1	1	1	1	2	1	0	1	2	2	1	2	3	0	0	0	2	0	3	2	3	1	1				
27	*Pteridium esculentum*	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0				
28	*Pyrrosia eleagnifoila*	0	0	0	0	0	0	0	1	1	0	1	0	1	0	0	1	0	0	2	0	0	0	0	0	3	0	1	0	1	0				
29	*Trichomanes reniforme*	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	4	3	0	1	0	0	0	0	0	2	0	0	0				
30	*Acianthus sinclairii*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	2	0	1	2				
31																																			
32	SOUTH	Saturday															Sunday																		
33	Species	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	Mean	Std Dev	Sqrt n	Std Err
34	*Asplenium flaccidum*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0				
35	*Asplenium flabellifolium*	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	2	0	0	1	0	1	0	1	1	1	0	0	1	0				
36	*Asplenium oblongifolium*	0	0	0	0	0	0	0	1	1	0	1	2	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	1	0				
37	*Ctenopteris heterophylla*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0				
38	*Hymenophyllum spp.*	2	2	3	4	4	3	6	0	2	1	2	1	0	0	0	0	0	0	3	0	3	1	4	3	3	3	2	0	1	2				
39	*Pellaea rotundifolia*	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0				
40	*Microsorum pustulatum*	1	3	2	0	0	1	2	1	2	1	2	2	0	2	1	0	1	2	1	2	3	1	3	2	2	2	2	0	1	2				


python


R


MATLAB
The Language of Technical Computing

The dream

```
# (ideal) data analysis process
raw_data = GET(data)
proc_data = PROCESS(raw_data)
SUMMARY(proc_data)
PLOT(proc_data)
model = FIT_MODEL(proc_data)
prediction = PREDICT(model)
PRINT(prediction)
> "Woo-hoo! validated model =)"
```

The reality

```
# (real) data analysis process
raw_data = GET(data)
clean_data = CLEAN(data)
proc_data = PROCESS(clean_data)
while (QUALITY(proc_data) != "good") {
   clean_data = CLEAN(proc_data)
   proc_data = PROCESS(clean_data)
   # while loop may run indefinitely
}
SUMMARY(proc_data)
PLOT(proc_data)
model = FIT_MODEL(proc_data)
prediction = PREDICT(model)
PRINT(prediction)
> "Ooops! model sucks =("
```

Images from http://www.gastonsanchez.com/visually-enforced/opinion/2012/09/19/Data-analysis-process/

Tidy Data Principles

- always keep a copy of the raw data
- have a separate copy which is your tidy dataset
- keep metadata record (codebook, readme.txt)
- keep a record of your ‘recipe’ (exact steps taken) to get from raw to tidy data

Reading:: Hadley Wickham, Tidy Data, Vol. 59, Issue 10, Sep 2014, Journal of Statistical Software. http://www.jstatsoft.org/v59/i10.

Tidy Data Tips

- 1 piece of information per cell
- 1 variable 1 column
- 1 observation 1 row

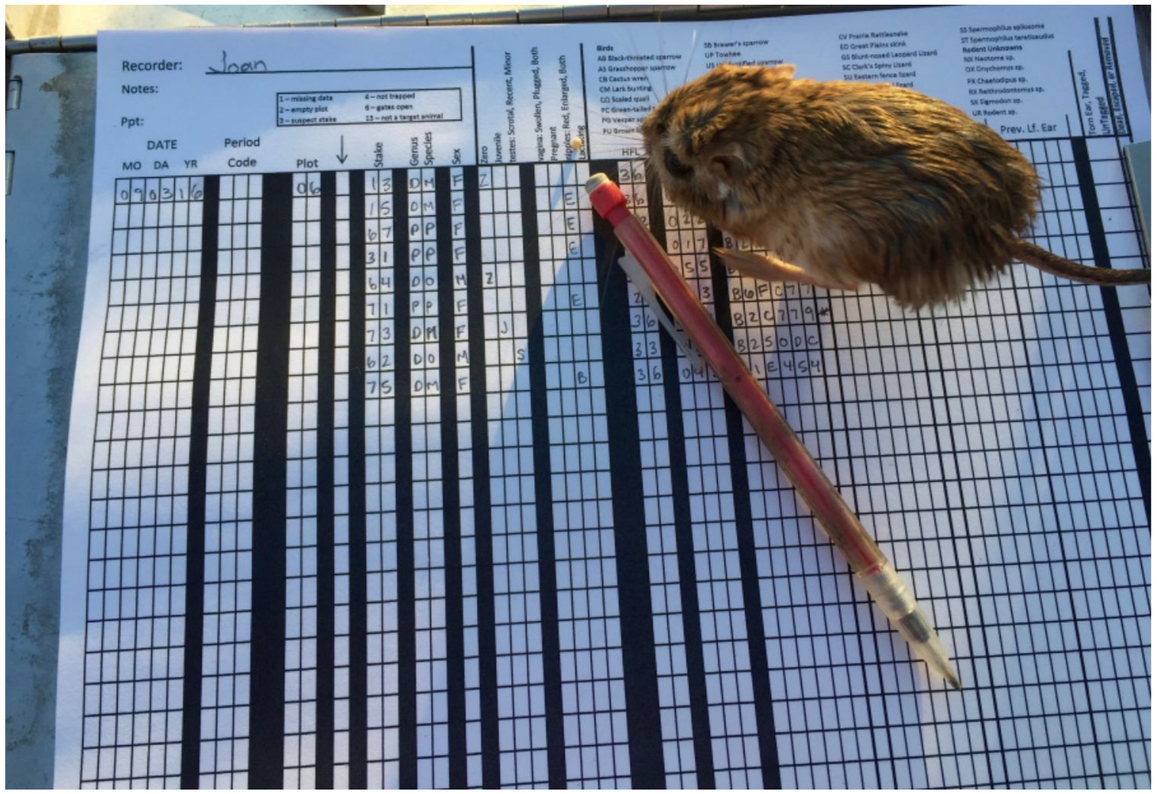
country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

table1

variables

observations

Exercise: Our dataset



Images from:
https://portalproject.wordpress.com/

Exercise: Tidy me up

Two field assistants conducted the surveys, one in 2013 and one in 2014, and they both kept track of the data in their own way. Let's have a look at the data (survey_data_spreadsheet_messy.xls) in a spreadsheet program.

In the chat note anything you think makes this dataset NOT TIDY

https://ndownloader.figshare.com/files/7823341

Tidy Data Tips

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

table1

variables

observations

- 1 piece of information per cell
- 1 variable 1 column
- 1 observation 1 row
- 1st row = variable names
- human readable names
- 1 table or file for each kind of variable
- use text-based format to save (e.g. .csv).
- multiple sheets use a README.txt file

Issues

- Using multiple tables
- Using multiple tabs
- Not filling in zeros
- Using problematic null values
- Using formatting to convey information
- Using formatting to make the data sheet look pretty
- Placing comments or units in cells
- Entering more than one piece of information in a cell
- Using problematic field names
- Using special characters in data
- Inclusion of metadata in data table
- Date formatting

Useful Resources

http://openrefine.org - great introductory videos

Google Group -good for beginner questions and problems.

OpenRefine Google Plus community - help

OpenRefine ecology data - Data Carpentry OpenRefine tutorials

OpenRefine for Social Science data

Cleaning Data with OpenRefine Van Hooland, Verborgh & De Wilde

Typical workflow

- Open OpenRefine (if it doesn't start automatically in your browser go to http://127.0.0.1:3333/)
- Import the dataset – CSV, tab – file, URL
- Explore data using 'facets' and 'filters'
- Use cluster analysis to make consistent
- Rearrange, split, sort
- Repeat
- Export dataset

Installing OpenRefine

1. OpenRefine website: http://openrefine.org/
2. Download section.
3. Choose the appropriate download for your operating system (Windows, Mac or Linux).
4. Follow the installation procedures

Tips

- I need more memory https://github.com/OpenRefine/OpenRefine/wiki/FAQ-Allocate-More-Memory
- Backup OpenRefine data https://github.com/OpenRefine/OpenRefine/wiki/Back-Up-OpenRefine-Data

Faceting

- Grouping a dataset based on 1+ parameters, properties, fields, columns
- Like tagging
- then explore just those records at the intersection of the facets
- A “species symbol” facet, for instance, would group all records that have the same species name
- Possible to facet on text, number ranges, pairs of numbers, etc.

More information: https://github.com/OpenRefine/OpenRefine/wiki/Faceting

Clustering

- Often faceting will reveal inconsistencies in the data
- Cluster analysis attempts to form clusters of data based on certain algorithms
- OpenRefine allows you try a variety of clustering methods
- These are quite good at revealing inconsistencies, e.g.:Kriesler vs Chrysler

More information:
https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth

Start the programme by double clicking the icon

Let’s try it

Regular expressions (regex) using GREL

- regex = codes used for matching patterns. Often there is more than one way to compose a regex pattern-match
- GREL General Refine Expression Language
- much of Refine's extensible and advanced power comes from regular expressions.
- Useful handout on regex and cheat-sheet.

Download data from a URL (API)

- fetch JSON from any web service (databases, registries, mapping services, etc.) based on values in a OpenRefine project
- Open Refine makes it relatively straightforward to call into an API, receive a response, and supplement your dataset with a portion of it.
- two step process - get data then parse data

More information:
https://github.com/OpenRefine/OpenRefine/wiki/Fetching-URLs-From-Web-Services

FAQs for problems

- Sometimes using a browser other than Firefox, OpenRefine does not automatically open. **Point your browser at http://127.0.0.1:3333/ or http://localhost:3333 to launch the program.**
- Issues getting OpenRefine to run on Windows. **Install Java (JDK + JRE) and add "JAVA_HOME" and "JDK_HOME" to the environment. This thread includes steps to diagnose possible issues, and links on how to set up environment variables.**
- Mac users with the newest operating system MAY have to allow this to run by "allowing everything" to run. They can change the setting back after the exercise OR right click and select open.

More resources

- OpenRefine web site
- OpenRefine Documentation for Users (Wiki site)
- Using OpenRefine book by Ruben Verborgh, Max De Wilde and Aniket Sawant
- Grateful Data is a fun site with many resources devoted to OpenRefine, including a nice tutorial.
- Margaret Heller shows how she uses OpenRefine for Measuring and Counting Impact in Repositories.
- Intersect Course Resources has Jared Berghold's Cleaning & Exploring your data with Open Refine.

ResBaz 2020: Pick n Mix

Key Stories - join us to listen over lunch

Mon. 12-1pm	*Welcome to ResBaz 2020 : Pick n Mix* https://auckland.zoom.us/j/99720152410 Passcode: 640143
Tues. 12-1pm	*Harnessing the disruptive nature of portable sequencing for community empowerment* https://vuw.zoom.us/j/95143657235 Passcode: 718144
Wed. 12-1pm	*From Classics to Computer Science and back again* https://vuw.zoom.us/j/98866515882
Thur. 12-1pm	*Performance - Music performance and project showcase* https://vuw.zoom.us/j/98265221971
Fri. 12-1pm	*Exploring the cultural horizons of open science: your research and your life* https://auckland.zoom.us/j/98447731180 Passcode: 404029

#resbaz2020 #resbazpicnmix email: researchdata@auckland.ac.nz