Cleaning & Exploring your Data

with Open Refine

Today

- A word about me
- Very short introduction
- Then we get down to putting Open Refine to use
- Working through a generally plausible example
- Goal: Find out more about historic police stations in NSW
- Start with basic set of data from OEH, then use Google Geolocation API, then State Records NSW API
- Not intended to be thorough research!

INTERSECT

Open Refine

- Sometimes described as Excel on steroids
- Kind of true
  - Exploring datasets
  - Cleaning up datasets
- Most datasets are messy
  - Variant spellings – Sydney, Sidney
  - Different number formats – "1000", "1,000", "1,000.00", "1.000,00"
  - Myriad of date formats
  - Compound fields – "Sydney, NSW"

INTERSECT

Open Refine

- Typical workflow:
    1. Import dataset – CSV, tab – file, URL
    2. Rearrange, split, sort
    3. Explore data using facets
    4. Use cluster analysis to make consistent
    5. Supplement with an API call
    6. Repeat 2-5
    7. Export dataset

INTERSECT

Open Refine

- Faceting:
    - Grouping the dataset based on one or more parameters, properties, fields, columns
    - Like tagging
    - You can then explore just those records at the intersection of the facets
    - A "suburb" facet, for instance, would group all records that have the same suburb
    - Possible to facet on text, number ranges, pairs of numbers, etc.

INTERSECT

Open Refine

- Clustering
  - Often faceting will reveal inconsistencies in the data
  - Cluster analysis attempts to form clusters of data based on certain algorithms
  - Open Refine allows you try a variety of clustering methods
  - These are quite good at revealing inconsistencies, e.g.:
    - **10-12 Church St**. vs **10,12 Church Street**

INTERSECT

Open Refine

- APIs

  – Increasingly, APIs are being used to expose services – databases, registries, mapping services, etc.

  – Open Refine makes it relatively straightforward to call into an API, receive a response, and supplement your dataset with a portion of it.

Open Refine

Let's have a go!

Get the slides, exercises and required dataset at:

http://www.intersect.org.au/course-resources

INTERSECT

Thanks for attending!

- Please complete our course survey at:
  - http://svy.mk/18c8dHa

- Any further questions, please contact us at training@intersect.org.au

INTERSECT