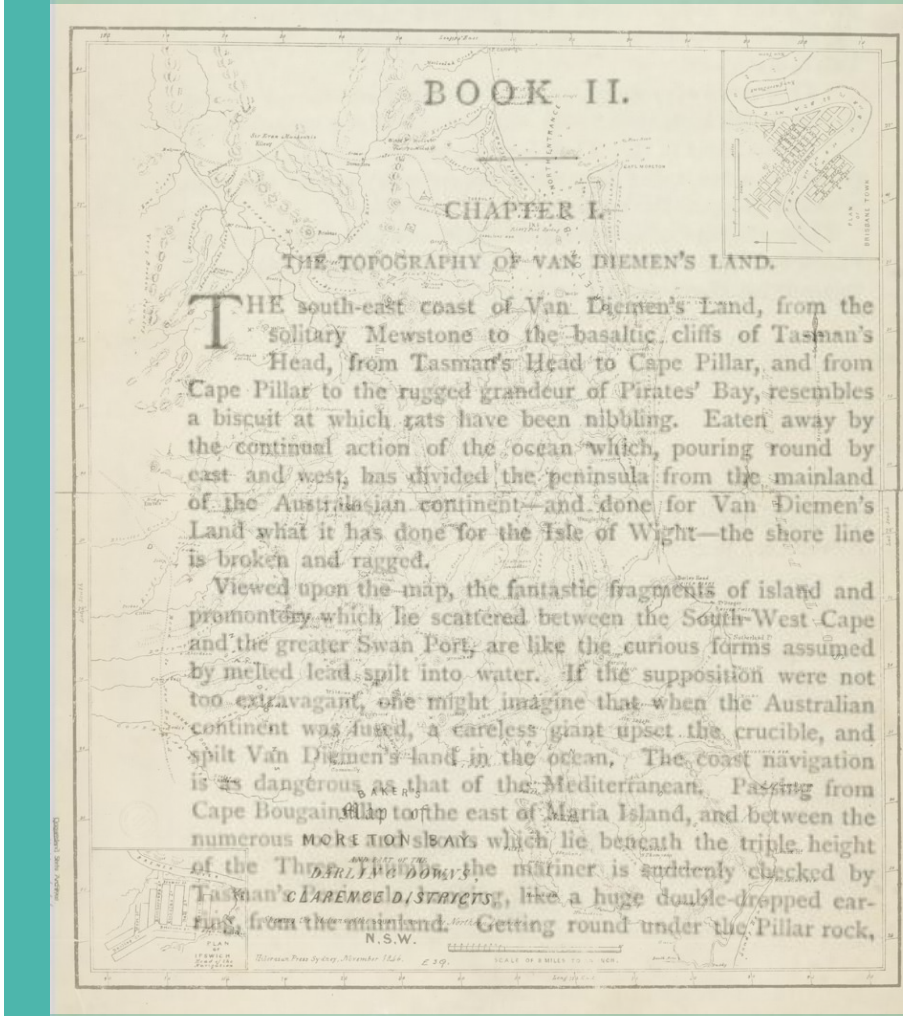
Geolocating Australian Historical Resources

Finding placenames and locations with gazetteers

Fiannuala Morgan (ANU, NLA)
Michael Niemann (UQ, Monash)
Simon Musgrave (UQ)

atap australian text analytics platform
atap.edu.au



BOOK II.

CHAPTER I.

THE TOPOGRAPHY OF VAN DIEMEN'S LAND.

THE south-east coast of Van Diemen's Land, from the solitary Mewstone to the basaltic cliffs of Tasman's Head, from Tasman's Head to Cape Pillar, and from Cape Pillar to the rugged grandeur of Pirates' Bay, resembles a biscuit at which rats have been nibbling. Eaten away by the continual action of the ocean which, pouring round by east and west, has divided the peninsula from the mainland of the Australasian continent—and done for Van Diemen's Land what it has done for the Isle of Wight—the shore line is broken and ragged.

Viewed upon the map, the fantastic fragments of island and promontory which lie scattered between the South-West Cape and the greater Swan Port, are like the curious forms assumed by melted lead spilt into water. If the supposition were not too extravagant, one might imagine that when the Australian continent was fused, a careless giant upset the crucible, and spilt Van Diemen's land in the ocean. The coast navigation is as dangerous as that of the Mediterranean. Passing from Cape Bougainville to the east of Maria Island, and between the numerous islands and bays which lie beneath the triple height of the Three Thumbs, the mariner is suddenly checked by Tasman's Peninsula, hanging, like a huge double-dropped earring, from the mainland. Getting round under the Pillar rock,

Outline

How the workshop is structured

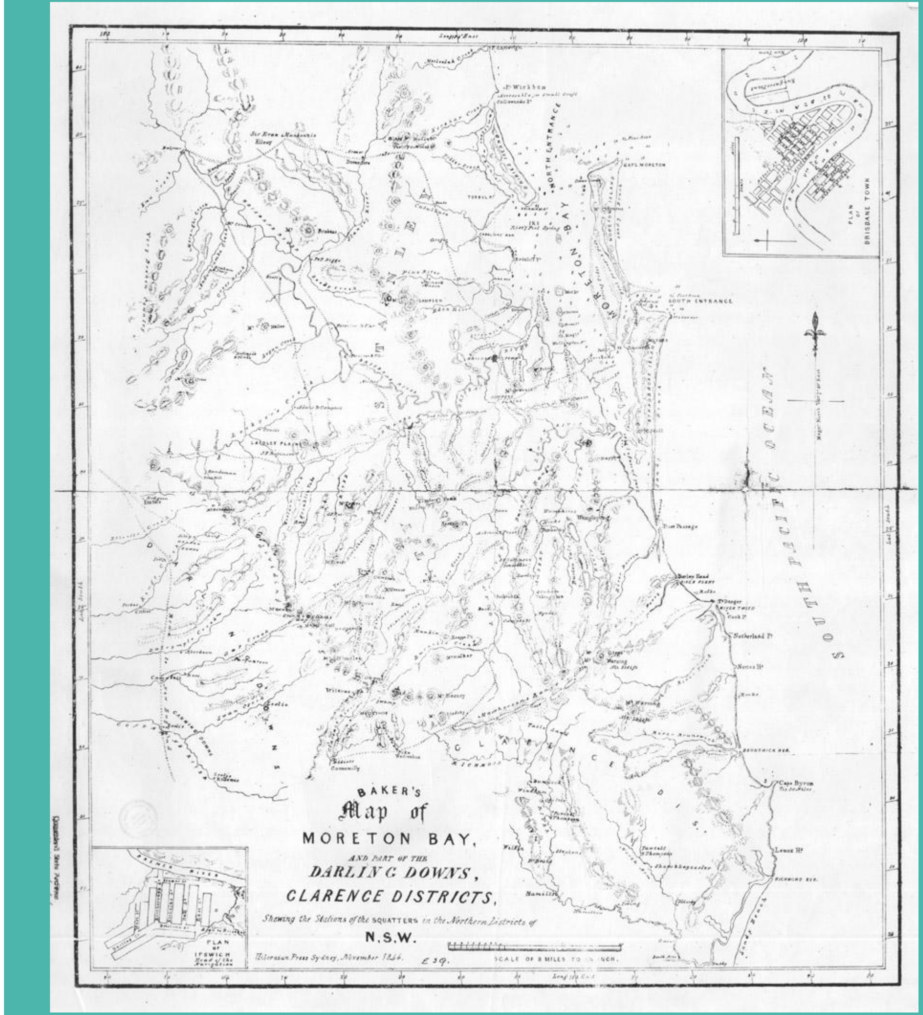
- Part 1 - Researching with geolocation (45 mins)
- Part 2 - Identifying placenames (45 mins)
- Part 3 - Finding locations for placenames (1 hour)

Slides which relate to live demonstrations are marked with ✦

We acknowledge and celebrate the First Australians on whose traditional lands we meet, especially those of the Ngambri and Ngunnawal, and pay our respect to the elders past, present and emergent..

Part 1 - Researching with Geolocation

Geolocating Australian Historical Resources



Geolocation - why bother?

Accurate coordinates for places open up two important possibilities:

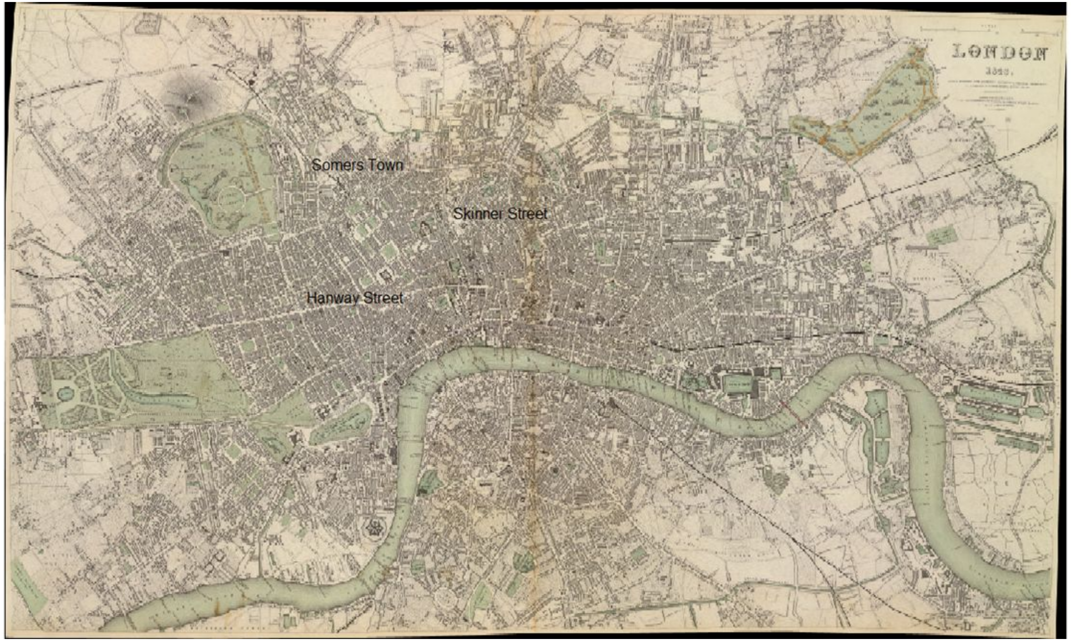
- **Analysis** - quantitative data can be analysed in spatial terms
  Example: identifying disease clusters
- **Visualisation** - data can be displayed on maps provided coordinate system is shared
  Examples to come

Shared coordinate system

- Locations are normally specified using latitude and longitude
    - But there are other possibilities, e.g., UK Ordnance Survey system, Cartesian coordinates
- Map images can have this information included also
- Plotting packages can therefore place locations on a map
- If coordinate systems are the same
    - NB - this includes projections being the same!
- Online map data services are the basis for such mapping mostly - but not always

William Godwin’s homes

Map from David Rumsey Collection (Stanford) has coordinate information added to the image

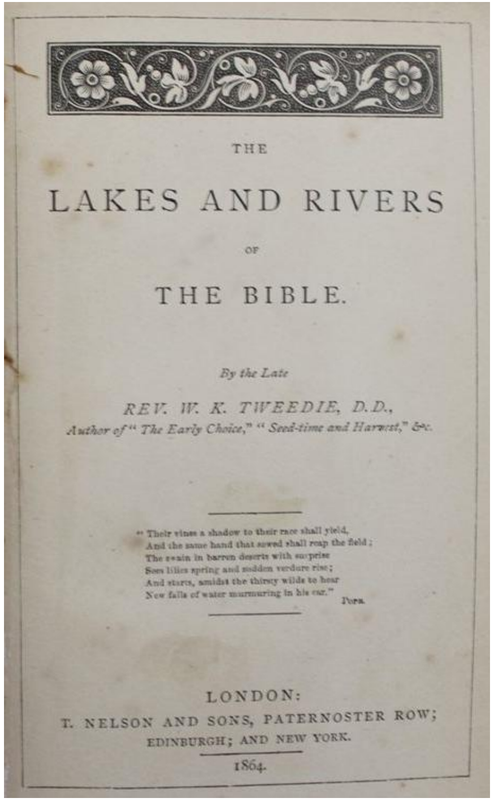


Colonial Frontier Massacres Map

- Created by an ARC-funded team based at the University of Newcastle
- Allows us to see distribution of massacres across space and across time
- Basic data for individual events is accessible by clicking on map points

Colonial Frontier Massacres, Australia, 1780 to 1930, v3 (newcastle.edu.au)

How is it done? An example in detail



THE

LAKES AND RIVERS

OF

THE BIBLE.

By the Late

REV. W. K. TWEEDIE, D.D.,

Author of "The Early Choice," "Seed-time and Harvest," &c.

"Their vines a shadow to their race shall yield,
And the same hand that sowed shall reap the field;
The swain in barren deserts with surprise
Sees lilies spring and sudden verdure rise;
And starts, amidst the thirsty wilds to hear
New falls of water murmuring in his ear."
Pope.

LONDON:
T. NELSON AND SONS, PATERNOSTER ROW;
EDINBURGH; AND NEW YORK.
1864.

- Belgum, Kirsten, Keith Handley, and Rachel Bott. 2018. "Mapping Travel Writing: A Digital Humanities Project to Visualise Change in Nineteenth-Century Published Travel Texts." *Studies in Travel Writing* 22 (3): 306–24. https://doi.org/10.1080/13645145.2019.1575765.
- Starting point is bibliographies of travel books published in C19
    - 3000 works published in Britain
    - Titles taken as basic data
- Each work was located geographically on the basis of places mentioned in the title
- Results can be accessed via interactive map Bibliography data show with heatmaps (keithhandley.com)

First attempt

- Used bibliography of French travel writing
- Place names identified by hand (student assistants)
- Relied on coder knowledge and reference works
- Processing 100 entries took on average one hour
    - Plus time to review and correct errors

Geoparsing

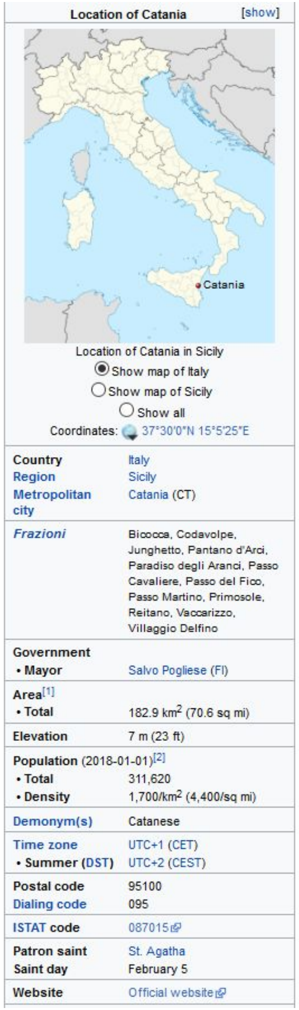
- Takes unstructured references to locations
- Tries to resolve them to unambiguous identifiers (e.g. coordinates)
- Steps:
  - Identifying references in text (cf. Named Entity Recognition)
  - Match text reference to entry or entries in geographic database
  - Resolve ambiguities if possible
- Problems:
  - Variant names (archaic names and spellings)
  - Ambiguities – Paris, France v. Paris, Texas

Geographic database

- Large digital gazetteers exist e.g. GeoNames
- GeoNames stores variant names
- But no information on frequency of use
- Belgum et al. used Wikipedia
  - Entries on places typically include geographic identifier (latitude and longitude)
  - Good coverage of higher level topography
    - Not many travel books have e.g. village name in title
  - Search history can be used for disambiguation
    - Paris, France is a common search term, Paris, Texas is not

Searching Wikipedia

- Procedure started with 5 word strings from title, then 4 word strings and so on down to individual words
    - Aim was to match e.g. New York City before New York
- Match of string to Wikipedia entry with geographic coordinates taken as identification of a place name
- Place name and coordinates added to bibliography database
- Manual checking as final step, with a map interface



Location of Catania [show]

Location of Catania in Sicily

● Show map of Italy

○ Show map of Sicily

○ Show all

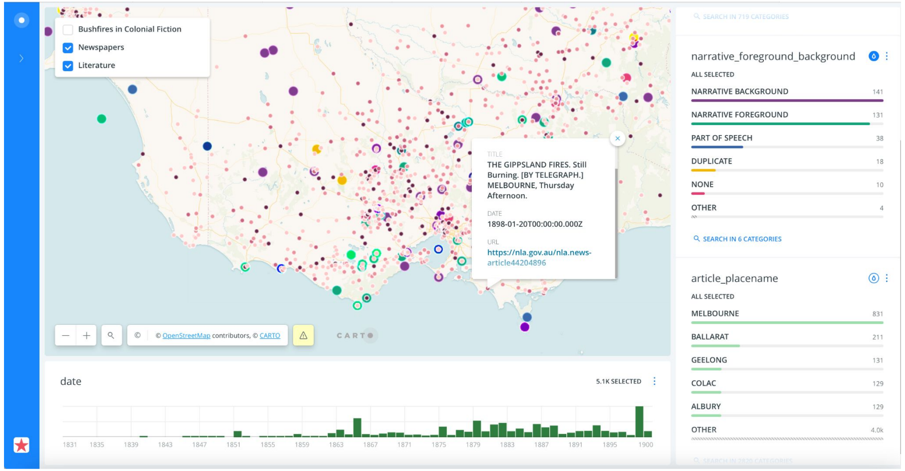
Coordinates: 37°30'0"N 15°5'25"E

Country	Italy
Region	Sicily
Metropolitan city	Catania (CT)
Frazioni	Bicocca, Codavolpe, Junghetto, Pantano d'Arci, Paradiso degli Aranci, Passo Cavaliere, Passo del Fico, Passo Martino, Primosole, Reitano, Vaccarizzo, Villaggio Delfino
Government	
• Mayor	Salvo Pogliese (FI)
Area[1]	
• Total	182.9 km^2 (70.6 sq mi)
Elevation	7 m (23 ft)
Population (2018-01-01)[2]	
• Total	311,620
• Density	1,700/km^2 (4,400/sq mi)
Demonym(s)	Catanese
Time zone	UTC+1 (CET)
• Summer (DST)	UTC+2 (CEST)
Postal code	95100
Dialing code	095
ISTAT code	087015
Patron saint	St. Agatha
Saint day	February 5
Website	Official website

Remaining problems

- Small number of cases where Wikipedia information did not align well with book
  - Manual input required
- Assigning point locations to regions is a general problem
  - Broad regions in titles (e.g. *The Islands of Greece)* were hard to identify

Historical Fires Near Me



☐ Bushfires in Colonial Fiction
☑ Newspapers
☑ Literature

TITLE
THE GIPPSLAND FIRES. Still Burning. [BY TELEGRAPH.] MELBOURNE, Thursday Afternoon.

DATE
1898-01-20T00:00:00.000Z

URL
https://nla.gov.au/nla.news-article44204896

© OpenStreetMap contributors, © CARTO

date — 5.1K SELECTED

SEARCH IN 719 CATEGORIES

narrative_foreground_background

ALL SELECTED

NARRATIVE BACKGROUND	141
NARRATIVE FOREGROUND	131
PART OF SPEECH	38
DUPLICATE	18
NONE	10
OTHER	4

SEARCH IN 6 CATEGORIES

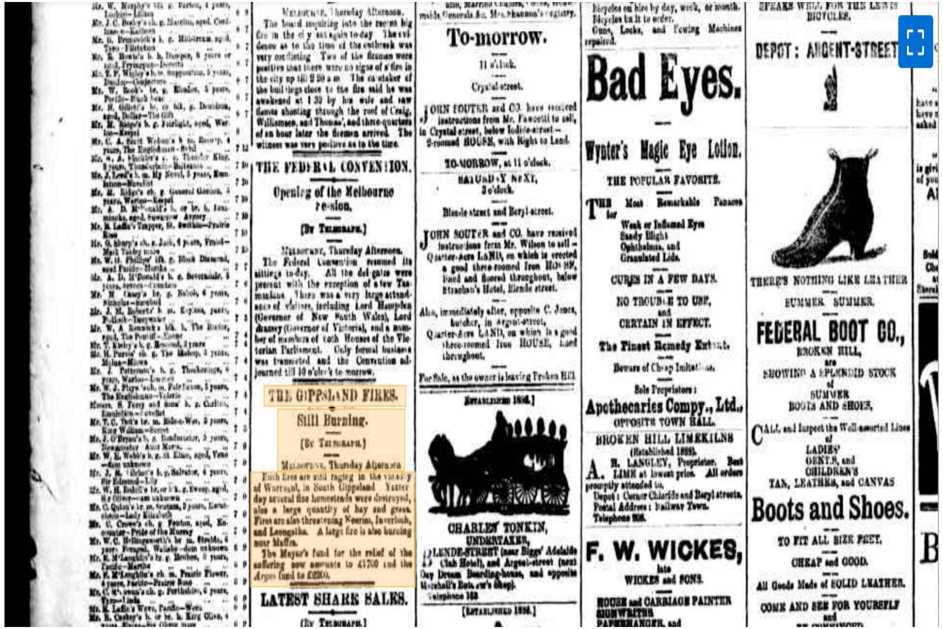
article_placename

ALL SELECTED

MELBOURNE	831
BALLARAT	211
GEELONG	131
COLAC	129
ALBURY	129
OTHER	4.0k

SEARCH IN 2820 CATEGORIES

Screenshot from Historical Fires Near Me



THE FEDERAL CONVENTION.

Opening of the Melbourne Session.

(By Telegraph.)

THE GIPPSLAND FIRES.

Still Burning.

(By Telegraph.)

Melbourne, Thursday Afternoon.

Bush fires are still raging in the vicinity of Warragul, in South Gippsland. Yesterday several fine homesteads were destroyed, also a large quantity of hay and grass. Fires are also threatening Neerim, Inverloch, and Leongatha. A large fire is also burning near Maffra.

The Mayor's fund for the relief of the suffering now amounts to £1700 and the *Argus* fund to £280.

LATEST SHARE SALES.

(By Telegraph.)

To-morrow.

Bad Eyes.

Wynter's Magic Eye Lotion.

FEDERAL BOOT CO.,

BROKEN HILL,

Boots and Shoes.

CHARLES TONKIN,

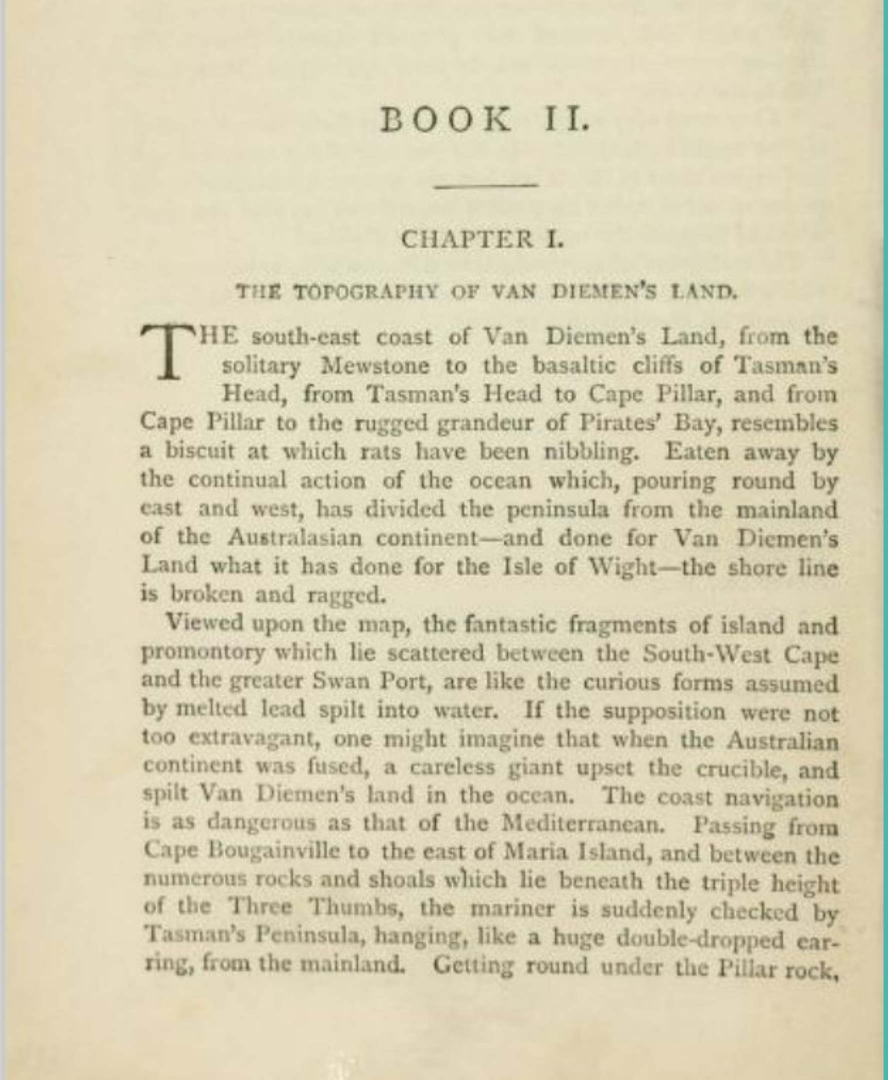
UNDERTAKER,

F. W. WICKES,

(1898, January 20). *Barrier Miner (Broken Hill, NSW : 1888 - 1954)*, p. 4 (SECOND EDITION).

Part 2 - Identifying placenames

Geolocating Australian Historical Resources



BOOK II.

CHAPTER I.

THE TOPOGRAPHY OF VAN DIEMEN'S LAND.

THE south-east coast of Van Diemen's Land, from the solitary Mewstone to the basaltic cliffs of Tasman's Head, from Tasman's Head to Cape Pillar, and from Cape Pillar to the rugged grandeur of Pirates' Bay, resembles a biscuit at which rats have been nibbling. Eaten away by the continual action of the ocean which, pouring round by east and west, has divided the peninsula from the mainland of the Australasian continent—and done for Van Diemen's Land what it has done for the Isle of Wight—the shore line is broken and ragged.

Viewed upon the map, the fantastic fragments of island and promontory which lie scattered between the South-West Cape and the greater Swan Port, are like the curious forms assumed by melted lead spilt into water. If the supposition were not too extravagant, one might imagine that when the Australian continent was fused, a careless giant upset the crucible, and spilt Van Diemen's land in the ocean. The coast navigation is as dangerous as that of the Mediterranean. Passing from Cape Bougainville to the east of Maria Island, and between the numerous rocks and shoals which lie beneath the triple height of the Three Thumbs, the mariner is suddenly checked by Tasman's Peninsula, hanging, like a huge double-dropped earring, from the mainland. Getting round under the Pillar rock,

Finding placenames in text

- Sounds easy right?
    - Read the text
    - Spot the placenames

This section will address how to do this using existing free software, including:

- An online Notebook
- Libraries of Language Technology software

This can be fully automated, but a bit of human interaction will improve the accuracy of the results.

Technical requirements

What do you need?

- A computer with a browser!

So you can use

- Binder!
  - GitHub
  - Jupyter Notebook
  - Python

atap
australian text
analytics platform
atap.edu.au

You are presumed to have **little or no** programming capability or familiarity with Python or GitHub.

Using a Jupyter Python Notebook

- What is a Jupyter Notebook
  - Free browser-based programming environment
  - Allows you to write, **run**, save and **load** (Python) programs
  - Uses *.ipynb* files
- How we are using it
  - Open source (so free to use, share and modify!)
  - Tool for researchers
  - Education
- Further information
  - Jupyter Notebooks
  - GitHub repositories

Running the notebook

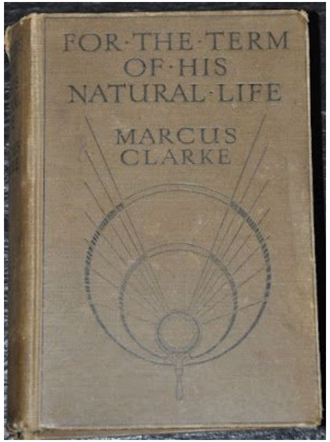
- Go to https://github.com/Australian-Text-Analytics-Platform/geolocation-tools-workshop
- Loading the notebook in Binder (or Jupytr)
- **Cells:** Python Code and Markdown
- **Kernel:** The memory and processing backbone
- Execution
    - Running Cells
    - Editting Cells
    - Rerunning cells
- Saving output from a notebook
    - Virtual environment vs local environments

Notebook: Installing & Importing Libraries

- Setting up the executable environment
- Installing libraries
- Importing libraries
- Virtual environment vs local environments

“For the Term of His Natural Life” by Marcus Clarke

Why use FtToHNL?



- **Pros:**
  - Freely available (Gutenberg Project Australia)
  - Complete and electronic (not partial or OCR of handwriting)
- **Cons:**
  - Only semi-structured (books, chapters and paragraphs)
  - No clues other than capitalisation as to what might be a placename
  - No indication about the semantic context of any placename,
    e.g., whether it is a building, city, region, country, or located within any of them

Notebook: Reading the historical text

- By default, it is as a single file
- For this Notebook, we have also broken it up per chapter - context!
    - No one way to do this, though we did write software to do this for FtToHNL
- As an example, we'll look at **Book 2, Chapter 3**
- Read the file expecting the UTF-8 encoding format, not unicode
    - Accented words and pound signs had already been converted to UTF-8 terms or characters

Named Entity Recognition

- Named Entities (NEs)
- Named Entity Recognition (NER)
- Natural Language Processing (NLP)
    - E.g., *spaCy*, *stanza*, *NLTK* for Python

Notebook: spaCy NER

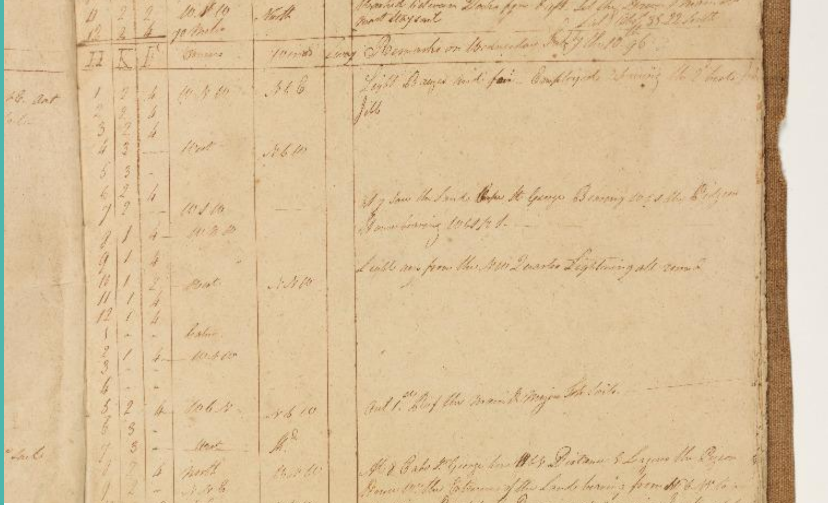
- Language Model
- spaCy Processing Pipeline
- spaCy Named Entity Recognition
- NEs as Multi-Word Expressions
- Semantic Categories
    - Complications
        - Single category per NE
        - No universal set of categories between systems
        - Corrections
        - Context

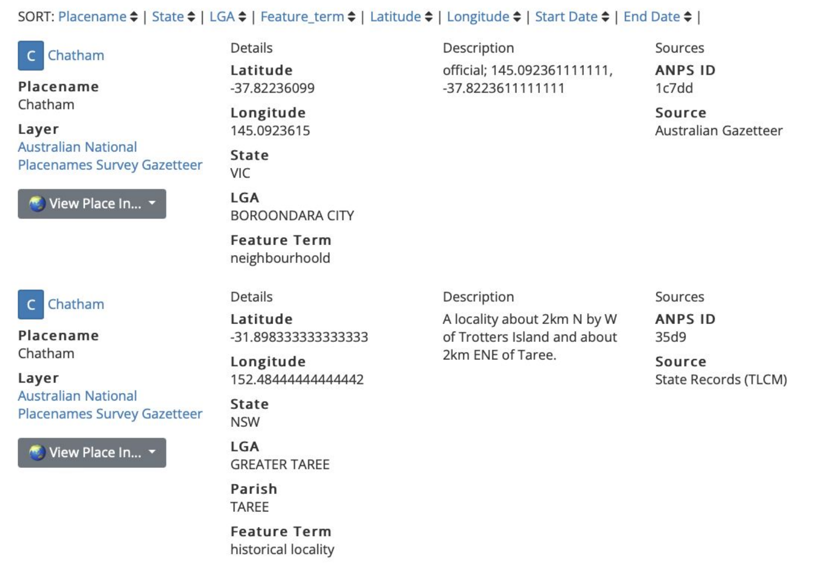
Notebook: Candidate Placenames

- Go to ATAP Geolocation Workshop notebook (via Binder)
- Reading multiple FtToHNL files
- Using spaCy NER
- Data about each placename
- Selecting a placename
- Saving the data
- Saving the selections

Part 3 - Finding locations for placenames

Geolocating Australian Historical Resources





SORT: Placename ⬍ | State ⬍ | LGA ⬍ | Feature_term ⬍ | Latitude ⬍ | Longitude ⬍ | Start Date ⬍ | End Date ⬍ |

	Details	Description	Sources
C Chatham **Placename** Chatham **Layer** Australian National Placenames Survey Gazetteer View Place In...	**Latitude** -37.82236099 **Longitude** 145.0923615 **State** VIC **LGA** BOROONDARA CITY **Feature Term** neighbourhoold	official; 145.092361111111, -37.8223611111111	**ANPS ID** 1c7dd **Source** Australian Gazetteer
C Chatham **Placename** Chatham **Layer** Australian National Placenames Survey Gazetteer View Place In...	**Latitude** -31.8983333333333 **Longitude** 152.484444444442 **State** NSW **LGA** GREATER TAREE **Parish** TAREE **Feature Term** historical locality	A locality about 2km N by W of Trotters Island and about 2km ENE of Taree.	**ANPS ID** 35d9 **Source** State Records (TLCM)

Finding Locations for Placenames

Issues

- Multiple names
- Multiple locations in different areas with the same names
- Multiple locations in the same area with the same names
- Who to trust?

Revisiting Gazetteers

- What they contain
  - No universal format
  - No universal interface - APIs
- Can they be trusted
  - Official vs Crowd-sourced “public” data
  - Using multiple gazetteers
- What rights do you have in using them
  - Ownership
  - API usage

Problems with using gazetteers

Considerations:

- Quality of data
  - Do you have place-names or full addresses?
  - Do you require point coordinates, or polygons?
  - What format is the data in?
- Open-source vs pay-as-you-go models
  - Google Maps (proprietary)
  - Geoscience Gazetteer of Australia (open-source)
- What is the scale of your analysis?
  - Regional, national or international?
  - Multiple locations in different areas with the same names
  - Multiple locations in the same area with the same names

Notebook: Unambiguous locations

- Loading the placename file
- Loading a reference file of unambiguous locations
  - Countries and continents
  - Major cities or region
  - Editting the file (or leave this to the final slides?)
- Matching against the placenames
  - Recording them as your best match
  - Degree of success for FtToHNL
  - No need to look elsewhere

Notebook: Open Street Map (OSM)

- What is the OSM?
- Nominatim as the OSM API tool
- Query
  - URL
  - Number of responses
- Response
  - Format vs reformatting
  - *PartOf*: postcode vs state for Australian locations, *Category*
- Recording the candidate locations
  - Not yet selecting a best candidate
  - Degree of success for FtToHNL

Notebook: TLCMap

- What is TLCMap?
  - Gazetteer of Historical Australian Places (GHAP)
- Query
  - Url
  - Public data
  - Search type: exact, contains. fuzzy
  - Number of responses
- Response
  - JSON format
  - Format vs reformatting
  - *PartOf*: state, *Category*
- Recording the candidate locations
  - Not yet selecting a best candidate
  - Degree of success for FtToHNL

Notebook: Ranking candidate locations

- Using two gazetteers
  - Degree of joint success
  - Context
    - Australian vs elsewhere
    - Multiple matches
    - Similar locations, based on coords
- Rank flags
  - Explanation
  - Priority ordering of flags
  - Ranking of candidates

Notebook: Manual selection of best locations

- Accordion user interface
- Outcome:

  Unambiguous vs
  Approved Best Matches vs
  Selected Best Matches vs
  Unknown (No location matched) vs
  Unlocated (No suitable location identified)
- Success of FtToHNL

Notebook: Modifying for your research

- NER
  - Category filter
  - Data storage
- Location selection
  - Unambiguous reference data
  - Gazetteers
    - New gazetteers
    - Number of candidates
    - Types of searches
    - Ranking heuristics
    - Ranking order
    - Localised contextual elements

Outcome: Provenance and workflow

- It is important to preserving data about the processing
  - The source: The text and the metadata about your rights and access
  - The result: the placenames
  - The process:
    - What technology you used (like this Notebook and spaCy)
    - What you did (workflow)
    - What choices you made (context)

Future: Stage 2 of the Geolocation notebook

- Incorporate recordkeeping of the provenance more in the process
    - Saving the UI choices
    - Reading the workflow records, to allow the textual context to be matched to the best match location
- Less Australia/Britain-centric
- More user interaction aspects, like settings for gazetteers
- Providing maps as visual aids

Acknowledgements and thanks

- Finn, for the collaboration with ATAP
- Alex Ip and others at AARnet for setting up and hosting the BinderHub and proxy cache for the notebooks
- The team at TLCMap for advising Finn on the options provided
- The greater team at ATAP and LDACA for making this workshop possible

atap
australian text analytics platform
atap.edu.au

ARDC
Australian Research Data Commons

This project is supported by the Australian Research Data Commons (ARDC). The ARDC is funded by NCRIS.