Mapping the Jouneys of Australian World War I Soldiers using Named Entity Recognition

Ashley Dennis-Henderson

December 5, 2025



THE UNIVERSITY of ADELAIDE

Acknowledgements

I would like to thank:

- The State Library of New South Wales for providing the transcriptions of the diaries and letters used in this project.
- The AIF Project for providing service records.
- My supervisors Matthew Roughan, Jonathan Tuke, and Gareth Pritchard for their guidance in this project.
- The aaDH for providing a bursary to help me attend this conference.

Talk Overview

1. Project Motivation
2. Data Acquisition and Standardisation
3. Named Entity Recognition
4. Geocoding
5. Shortest Paths through a Temporal Network
6. Mapping
7. Future Work

Project Motivation

PhD Aim

To analyse Australian World War I diaries using distant reading techniques in order to understand what the soldiers wrote about and how they felt over the course of the war.

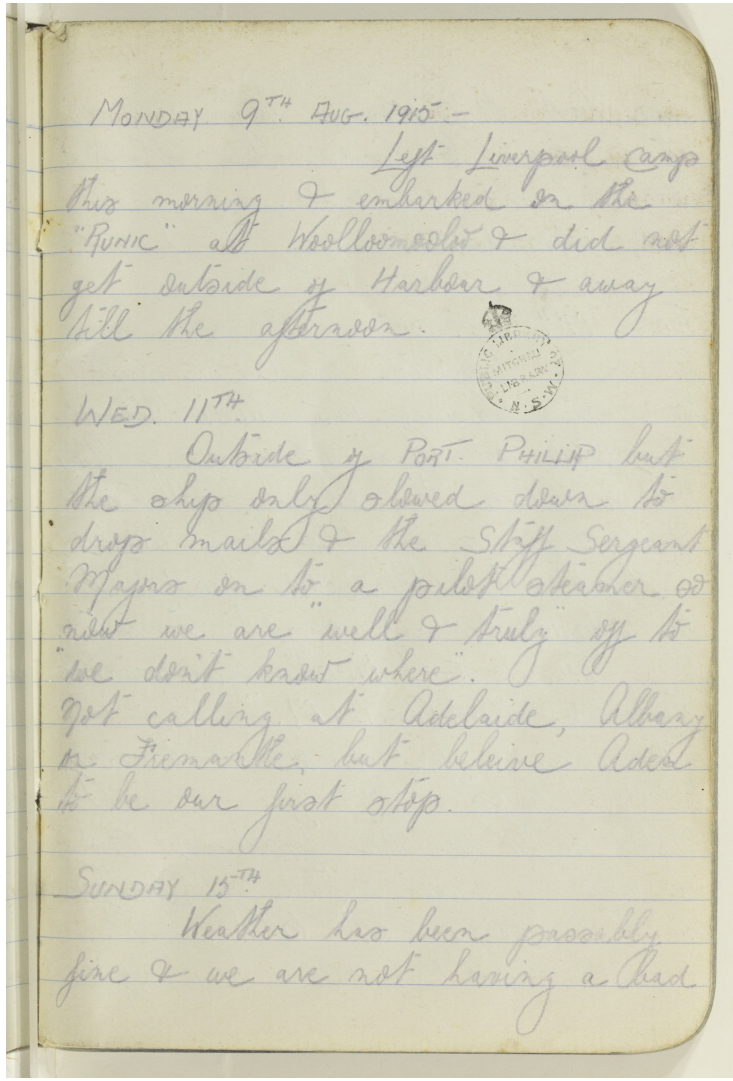
Talk Aim

To automatically extract locations from Australian World War I diaries and map them so we can view and explore each soldiers' personal journey through the war.

Data Acquisition and Standardisation

State Library of New South Wales Collection

- Began as the European War Collecting Project.
- Documents have been digitised and transcribed.
- 504 diaries formatted as date/entry pairs.
- Will use first 5 entries of Francis Addy's diary as a case study.



MONDAY 9TH AUG. 1915 –
Left Liverpool camp this morning & embarked on the "RUNIC" at Woolloomooloo & did not get outside of Harbour & away till the afternoon.

WED. 11TH
Outside of PORT PHILLIP but the ship only slowed down to drop mails & the Staff Sergeant Majors on to a pilot steamer & what we are well & truly off to we don't know where.
Not calling at Adelaide, Albany or Fremantle, but believe Adel to be our first stop.

SUNDAY 15TH
Weather has been possibly fine & we are not having a bad

Figure: Addy, Francis. *Item 01: Francis Addy diary, 9 August 1915-31 December 1917*. Mitchell Library, State Library of New South Wales.

Text Standardisation

Two main problems are abbreviations and spelling mistakes.

Abbreviations

Why?

- Likely to reduce time and materials required for writing.

What?

- Common military shorthand, e.g. “capt” for “captain”.
- Commonly used words, e.g. “lre” for “letter”.

Solution?

- Creating a custom mapping of known abbreviations in the corpus to their full form.

Spelling Mistakes

Why?

- Use of phonetics for unfamiliar words.

What?

- Mainly concerned about misspellings of locations, e.g. “Wooloomooloo” for “Wolloomooloo”.

- Many locations have different spellings or names today.

Solution?

- Create a custom mapping based on when code breaks in later steps.

Named Entity Recognition

Named Entity Recognition (NER)

- Currently using spaCy's pre-trained `en_core_web_trf` model.
- This uses the Roberta-base transformer.
- It is trained on various web-based documents (blogs, news, comments) from the OntoNotes5, ClearNLP, and WordNet 3.0 datasets.
- Focused on GPE and LOC entity types.

Issues with NER

Issues

- Some things that weren't locations were extracted.
  - Such as "Persia" (the ship), "Jerry", "Adolfs", and "Flank".
- Missed some actual locations.

Potential Solution

- Adding manually trained data to the model.

Geocoding

Geocoding

For each unique location:

- Used nominatum/Open Street Maps geocoder in R through the `tidygeocoder` package.
- As locations may exist in multiple places around the word, we found all possible matches.
- Used the `rnaturalearth` package to add country and continent to each possibility.

Case Study: Geocoding

Date	Location	Country	Continent	Lat	Long
597	Woolloomooloo	Australia	Oceania	-33.87	151.22
599	Port Phillip	Australia	Oceania	-38.10	144.78
599	Adelaide	Australia	Oceania	-34.93	138.60
599	Albany	Albania	Europe	41.00	20.00
599	Albany	USA	North America	42.65	-73.75
599	Albany	USA	North America	42.60	-73.98
599	Albany	USA	North America	31.58	-84.16
599	Albany	USA	North America	41.67	-105.68
599	Freemantle	UK	Europe	50.91	-1.43
599	Freemantle	Australia	Oceania	-33.27	149.33
599	Aden	USA	North America	38.65	-77.54
599	Aden	Yemen	Asia	12.83	44.92
599	Aden	USA	North America	38.24	-88.46
599	Aden	Kenya	Africa	0.11	36.28

Shortest Paths through a Temporal Network

Creating a Temporal Network

- Nodes on each date are potential locations.
- Directed edges exist between locations on two neighbouring dates if that journey is feasible.
- Edges are weighted by the Euclidean distance between the locations.
- An additional bypass node is added to each date (except the first) with edges to each neighbouring location weighted by a penalty term.

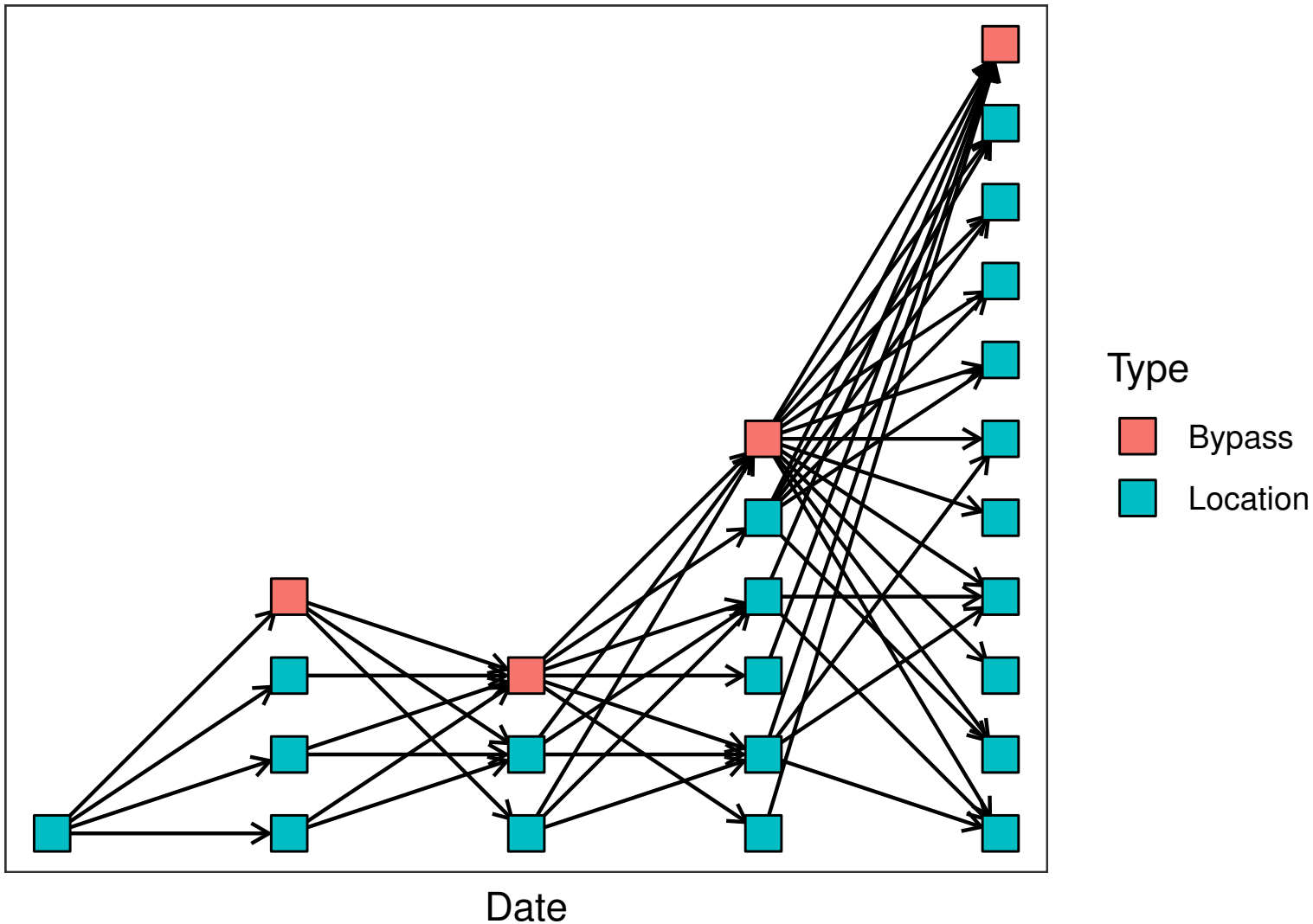
Feasible Journeys

- Travelling between continents (or within Australia):
  - Most likely via troopship.
  - Maximum 700km per day.
- Travelling between countries (same continent):
  - Most likely via train.
  - Maximum 500km per day.

Feasible Journeys

- Travelling within same country (not Australia):
  - Most likely walking/marching or via car.
  - Maximum 60km per day.
- Penalty:
  - 1000km per day

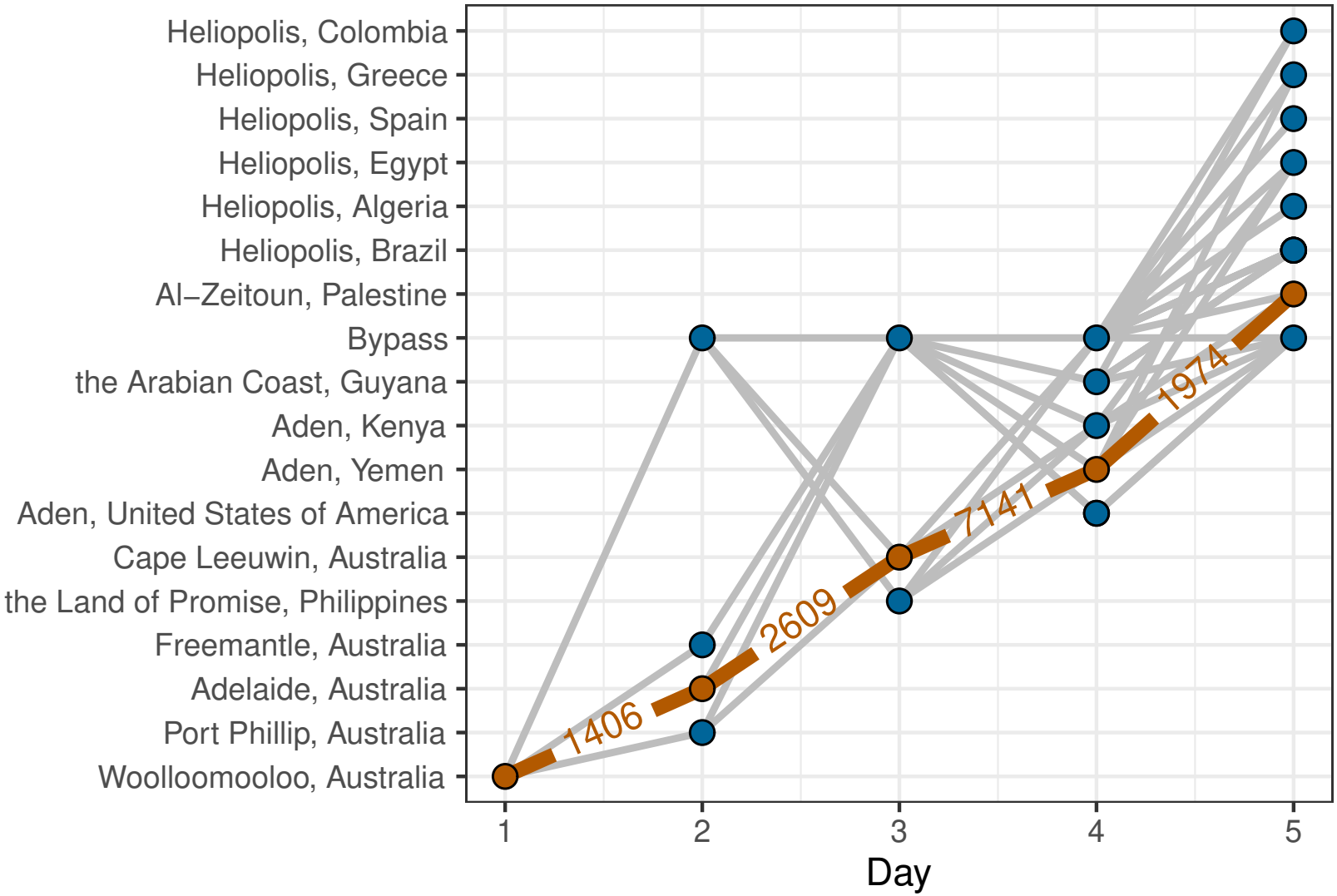
Case Study: Temporal Network



Shortest Paths

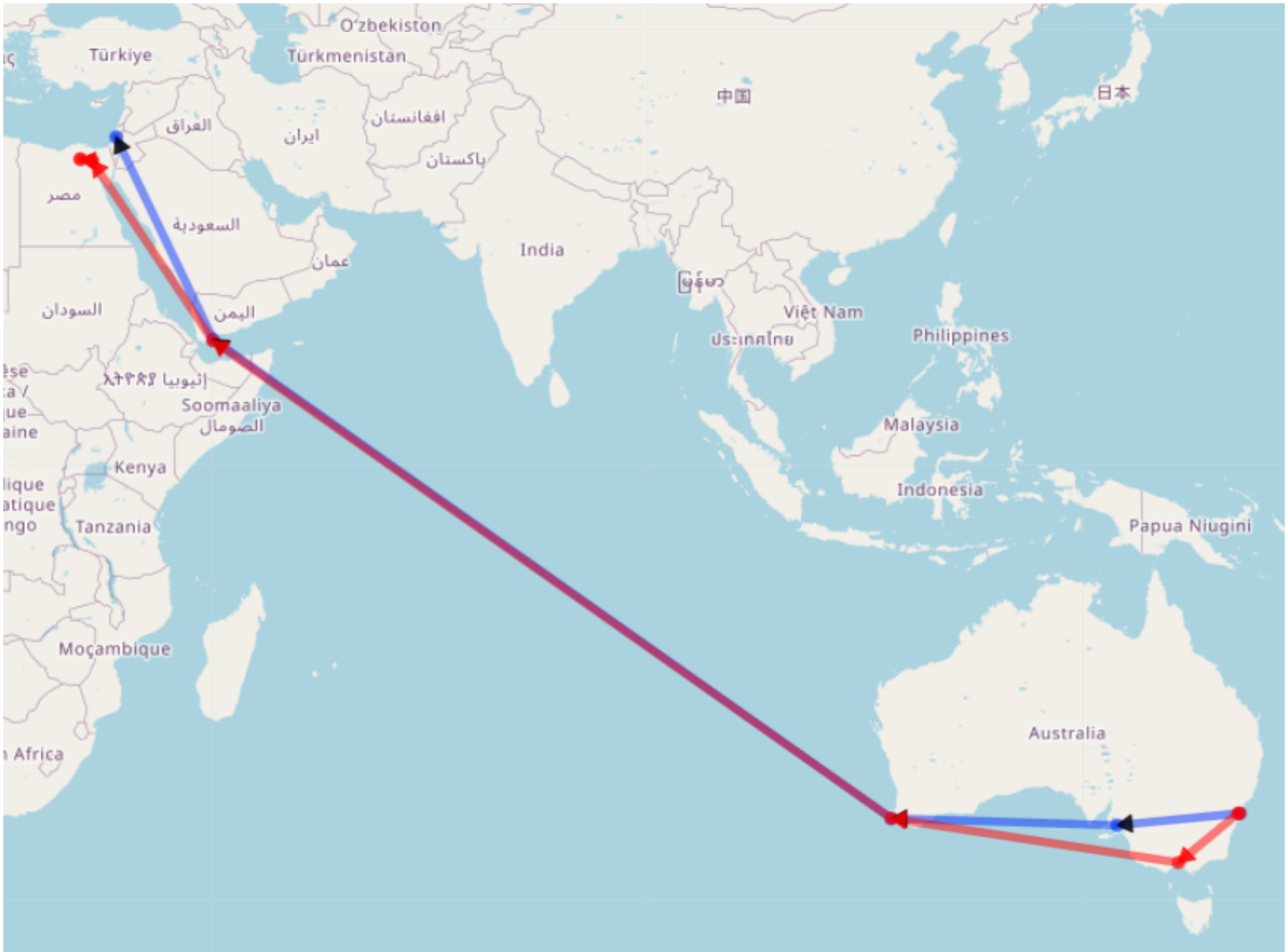
- Assume a diarist can only visit one location per date.
- We take the most likely journey as the shortest path through our temporal network.
- Use Dijkstra's algorithm to find the shortest path.

Case Study: Shortest Paths



Mapping

Case Study: Mapping



Future Work

Future Work

- Improve mappings for abbreviations and spelling mistakes.
- Add manually trained data to the NER model.
- Deal with locations where coordinates, country, and continent cannot be found.
- Tune parameters for feasible journeys.

Questions?

https://www.ashleydennishenderson.com/

ashley.dennis-henderson@adelaide.edu.au