

Using Regular Expressions and Optimisation to Create a Method for Automatically Extracting Dates from War Diaries

Ashley Dennis-Henderson

The University of Adelaide

August 5, 2024



THE UNIVERSITY
of ADELAIDE

Disclaimer

As this is Australian data, all dates are formatted as day - month - year unless otherwise specified.

Motivation: WWI Diary Collection

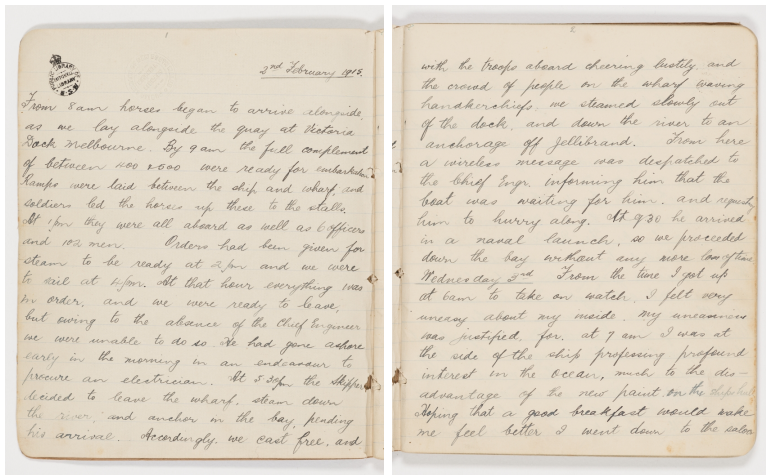


Figure: Gilroy, Norman. Norman Thomas Gilroy war diary, Volume 1: 2 February-29 April 1915. Mitchell Library, State Library of New South Wales.

Aim

Date	Entry
02/02/1915	from 8am horses began to arrive alongside, ...
08/02/1915	was so ill during this period that i was unable ...
28/03/1915	many large french and british transports ...
05/04/1915	at 5.30a we were towed out from the wharf ...
25/07/1915	arose about 6.30 and went down to church ...
27/07/1915	at 5am we were shifted to a berth further down ...

Challenges

- ▶ Variety of date formats, some with missing information
- ▶ Use of abbreviations or languages other than English in the dates
- ▶ Mistaken dates
- ▶ Non-entry dates
- ▶ Non-dates
- ▶ Out-of-order diaries

Challenges

“Tuesday 9th February 1915. At 6am woke to find we were ”

“Wed Jan 30th Left line tonight being relieved by the 54th BN.”

“Wednesday 24th Feby. Very disappointed when we received a”

“Tuesday May 8.5.17. Stand to 5.30 A.M. cool morning fed up”

“5th. Last night I had a party of 20 men repairing the blown in”

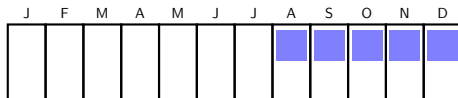
“17: Pte. Sellars fell down hold 50 ft., dying four hours later,”

“8th Field Ambulance. Breakfast had been previously arranged”

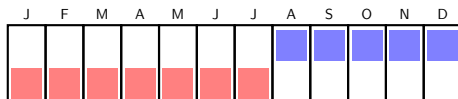
“Janvier 19 Dimanche Went to Orchestra Symphony Concert”

“One was written in April, the 24th & the other on June 23rd”

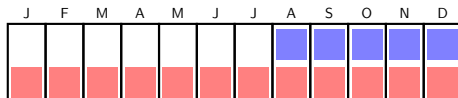
Out-of-Order Diaries



The diarist begins by writing the entries for all the 1916 dates after purchasing the diary.



The diarist then writes their 1917 entries from the beginning of the diary.



Some diarists then continued writing their 1917 entries underneath the 1916 entries until the year finished.

Legend:

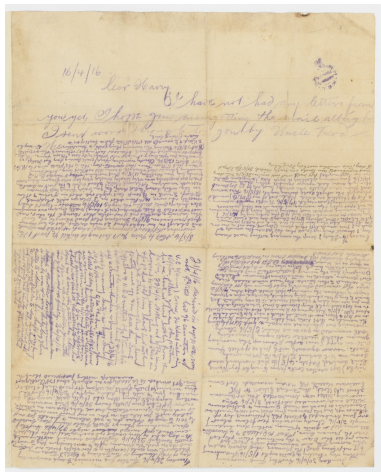


1916 entries,



1917 entries

Out-of-Order Diaries



Letter: 16/04/16			
		1 1	
08/05/16	2	14/05/16	3
to		to	
13/05/16	6	21/05/16	7
21/04/16	4	04/05/16	5
to		to	
24/04/16	2	07/05/16	5
25/04/16	6	30/04/16	7
to		to	
29/04/16	3	03/05/16	4

First Step: Regular Expressions

We extract all text matching a certain expression, e.g.

dd month yyyy

dow dd

dd-mm-yyyy

dd/mm/yyyy

Xmas, Christmas,
Good Friday

dd.mm.yyyy

dd-mm or dd/mm

month dd yyyy

dd

dow

month

Date Extraction

day	month	year
2	2	1915
	:	
2	4	
3	3	
4	3	
5	4	
	:	
25	7	
5	5	
26	7	

Gilroy, Norman. Norman Thomas Gilroy war diary, volume 1 - 4. Mitchell Library, State Library of New South Wales.

Date Extraction

day	month	year
2	2	1915
⋮		
2	4	0
3	3	0
4	3	0
5	4	0
⋮		
25	7	0
5	5	0
26	7	0

← mistaken dates

← non-entry date

Second Step: Optimisation

Goals:

- ▶ Fill in missing data
- ▶ Fix mistaken dates
- ▶ Change non-entry dates or non-dates to same value as previous or next date (to be removed afterwards)

Second Step: Optimisation

Assumptions:

- ▶ All months have 31 days
- ▶ We know the start and end date of each diary
- ▶ As we move through the diary we cannot go backwards in time

Notation

From the regular expressions we have:

- ▶ n dates extracted
- ▶ $\hat{d}_i, \hat{m}_i, \hat{y}_i$ are the extracted day, month and year values
- ▶ I_i^d, I_i^m, I_i^y are indicator variables with value 1 if the i th day, month, and year were extracted and 0 otherwise

Also,

- ▶ s_d and e_d are the known start and end dates of the diary
- ▶ s_y is the known start year of the corpus
- ▶ d_i, m_i, y_i are the outputted clean day, month, and year values
- ▶ $\alpha, \beta, \gamma, \delta, \omega$ are changeable parameters

Optimisation Program

$$\min_{d_i, m_i, y_i} \sum_{i=1}^n \left(\alpha l_i^d |d_i - \hat{d}_i| + 31 \beta l_i^m |m_i - \hat{m}_i| + 372 \gamma l_i^y |y_i - \hat{y}_i| \right) \\ + \delta \sum_{j=1}^{n-1} C(x_{j+1}, x_j) + \omega \sum_{k=1}^{n-1} (m_{k+1} - m_k)(1 - l_{k+1}^m)$$

$$\text{such that } x_i = 372(y_i - s_y) + 31(m_i - 1) + d_i - 1$$

$$x_i \leq x_{i+1}, \quad \text{for } i = 1, \dots, n-1$$

$$x_1 = s_d$$

$$x_n = e_d$$

$$C(x_{j+1}, x_j) = \begin{cases} 1, & \text{if } x_j = x_{j+1} \\ 0, & \text{if } x_j \neq x_{j+1} \end{cases}, \quad \text{for } i = 1, \dots, n-1$$

Epoch Dates

- ▶ Convert dates to number of days since 1st Jan 1914,

$$x_i = 372 * (y_i - 1914) + 31 * (m_i - 1) + d_i - 1$$

- ▶ Example, 25th April 1915

$$\begin{aligned}x_i &= 372 * (1915 - 1914) + 31 * (4 - 1) + 25 - 1 \\&= 489.\end{aligned}$$

Optimisation Program

$$\min_{d_i, m_i, y_i} \sum_{i=1}^n \left(\alpha l_i^d |d_i - \hat{d}_i| + 31\beta l_i^m |m_i - \hat{m}_i| + 372\gamma l_i^y |y_i - \hat{y}_i| \right) \\ + \delta \sum_{j=1}^{n-1} C(x_{j+1}, x_j) + \omega \sum_{k=1}^{n-1} (m_{k+1} - m_k)(1 - l_{k+1}^m)$$

$$\text{such that } x_i = 372(y_i - s_y) + 31(m_i - 1) + d_i - 1$$

$$x_i \leq x_{i+1}, \quad \text{for } i = 1, \dots, n-1$$

$$x_1 = s_d$$

$$x_n = e_d$$

$$C(x_{j+1}, x_j) = \begin{cases} 1, & \text{if } x_j = x_{j+1} \\ 0, & \text{if } x_j \neq x_{j+1} \end{cases}, \quad \text{for } i = 1, \dots, n-1$$

Optimisation Goal

day	month	year		day	month	year
2	2	1915		2	2	1915
	⋮				⋮	
2	4	0		2	4	1915
3	3	0		3	4	1915
4	3	0	⇒	4	4	1915
5	4	0		5	4	1915
	⋮				⋮	
25	7	0		25	7	1915
5	5	0		25	7	1915
26	7	0		26	7	1915

Gilroy, Norman. Norman Thomas Gilroy war diary, volume 1 - 4. Mitchell Library, State Library of New South Wales.

Optimisation Goal

day	month	year
2	2	1915
	⋮	
2	4	0
3	3	0
4	3	0
5	4	0
	⋮	
25	7	0
5	5	0
26	7	0



day	month	year
2	2	1915
	⋮	
2	4	1915
3	4	1915
4	4	1915
5	4	1915
	⋮	
25	7	1915
25	7	1915
26	7	1915

Gilroy, Norman. Norman Thomas Gilroy war diary, volume 1 - 4. Mitchell Library, State Library of New South Wales.

Determining Accuracy

- ▶ To determine the accuracy of the program dates will be manually extracted from two sets of diaries.
- ▶ The first set of diaries, referred to as the *training set*, is deliberately chosen to incorporate a variety of date formats and challenges. This set is used to determine the best parameters for the program.
- ▶ The second set of diaries, referred to as the *testing set*, is to get a final measure of the accuracy of the program. This set is partially randomly chosen.

Manually Extracting Dates using a GUI

Date Annotation GUI

Text File

Saturday 23/2/18 - Meeting of CO's at Bde. HQrs. Robinson went on leave - went for a ride.

Sunday 24/2/18 - Church parade, went for a long ride with Lloyd.

Monday 25/2/18 - Went to Boulogne to play football - match off - returned at 9.45 p.m.

Tuesday 26/2/18 - Route March 15 miles, walked all the way. Wrote to Doff.

Wednesday 27/2/18 - Div Boxing Competitions, stadium in this area, I judged.

Thursday 28/2/18 - Final of boxing - 26th won by two firsts & one second, rained in afternoon.

Friday 1/3/18 - Snowing & raining. Bn marched to Rifle Range but unable to shoot (10 miles).

Saturday 2/3/18 - Snowing & raining, had a game of bridge,

Metadata Entry

Raw Date

Date Type:
☐ Entry Date ☐ Non-entry Date ☐ Non-date

Extracted Values:
DD MM YYYY

Actual Values:
DD MM YYYY

Add

Progress

3%

Metadata CSV

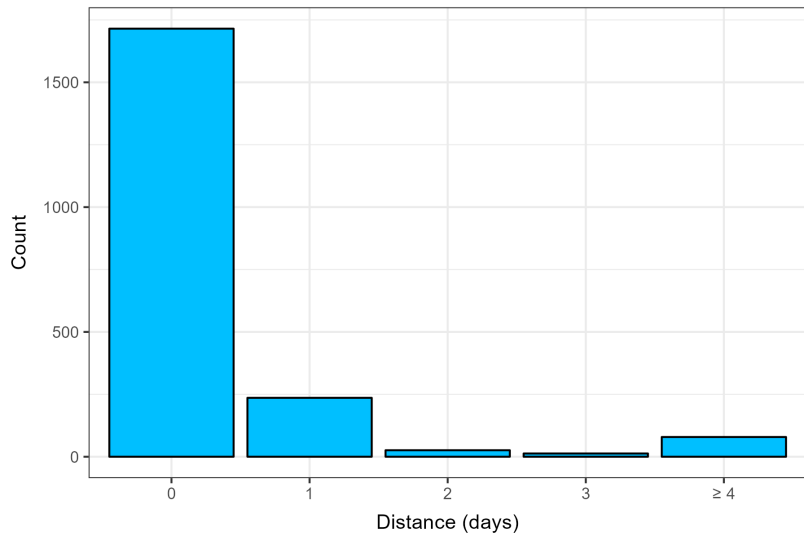
	L1	L2	Raw Date	ED	NED	ND	DD	E	MM	E	YYYY	E	DD	A	MM	A	YYYY	A
0	2	16	Friday 15/2/18	1	0	0	15	2	1918	15	2	1918						
1	111	127	Saturday 16/2/18	1	0	0	16	2	1918	16	2	1918						
2	240	254	Sunday 17/2/18	1	0	0	17	2	1918	17	2	1918						
3	336	350	Monday 18/2/18	1	0	0	18	2	1918	18	2	1918						
4	491	507	Saturday 23/2/18	1	0	0	23	2	1918	23	2	1918						

Save

Finish

Exit

Accuracy



Future Work

- ▶ Investigate the accuracy of the method further
- ▶ Include an indicator for out-of-order diaries
- ▶ Create a method for determining the optimal parameters of the program
- ▶ Research and test other methods of extracting dates

Questions?

www.ashleydennishenderson.com