HistoRadar

Alberto González Palomo
Uwe-Matthias Boltz
Johannes Braunias
Souhail Bouricha
Maria Jacob

Seminar “Unlocking the Secrets of the Past:
Text Mining for Historical Documents (WS 2009/10)”

2010-03-05

Historian's workflow

- Read documents in collection
- Collect interesting topics
- Snowball method:
  - Read again, collecting notes about selected topics
  - Add findings to “snowball”
  - Follow leads
  - Iterate

Maria

HistoRadar concept

Highlight places of potential interest in the historical document collection

- Extract information from text
- Radar shows points where information changes
  - Interesting places to start the “snowball”?
- Example:
  - Opinion change: A-supports-B → A-opposes-B

Alberto

HistoRadar concept

Highlight places of potential interest in the historical document collection

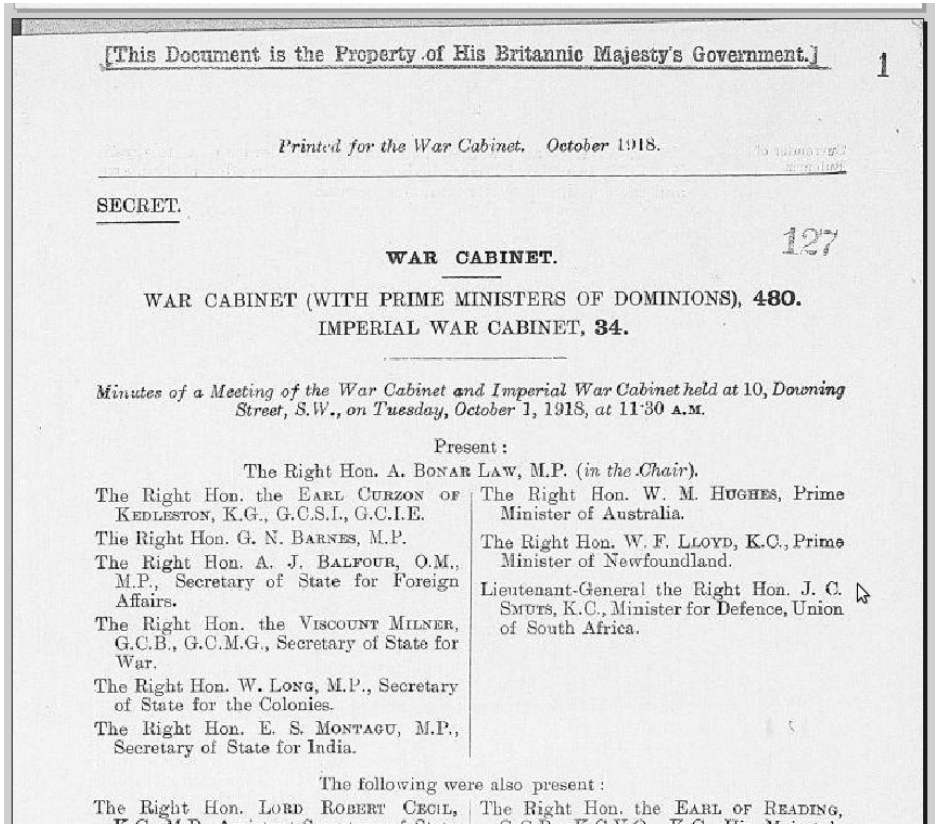
- Realistic first step
  - Track attendants to meetings of the British Cabinet
    - Who was suddenly absent?
    - Who re-appeared?
  - Named entities
    - Which countries start/stop being mentioned?
    - Which persons?

Alberto

Source text acquisition

Source text acquisition

British Cabinet Papers, http://www.nationalarchives.gov.uk/cabinetpapers/



[This Document is the Property of His Britannic Majesty's Government.] 1

Printed for the War Cabinet. October 1918.

SECRET.

127

WAR CABINET.

WAR CABINET (WITH PRIME MINISTERS OF DOMINIONS), **480.**
IMPERIAL WAR CABINET, **34.**

Minutes of a Meeting of the War Cabinet and Imperial War Cabinet held at 10, *Downing Street, S.W., on Tuesday, October* 1, 1918, *at* 11·30 A.M.

Present:

The Right Hon. A. BONAR LAW, M.P. (*in the Chair*).

The Right Hon. the EARL CURZON OF KEDLESTON, K.G., G.C.S.I., G.C.I.E.
The Right Hon. G. N. BARNES, M.P.
The Right Hon. A. J. BALFOUR, O.M., M.P., Secretary of State for Foreign Affairs.
The Right Hon. the VISCOUNT MILNER, G.C.B., G.C.M.G., Secretary of State for War.
The Right Hon. W. LONG, M.P., Secretary of State for the Colonies.
The Right Hon. E. S. MONTAGU, M.P., Secretary of State for India.
The Right Hon. W. M. HUGHES, Prime Minister of Australia.
The Right Hon. W. F. LLOYD, K.C., Prime Minister of Newfoundland.
Lieutenant-General the Right Hon. J. C. SMUTS, K.C., Minister for Defence, Union of South Africa.

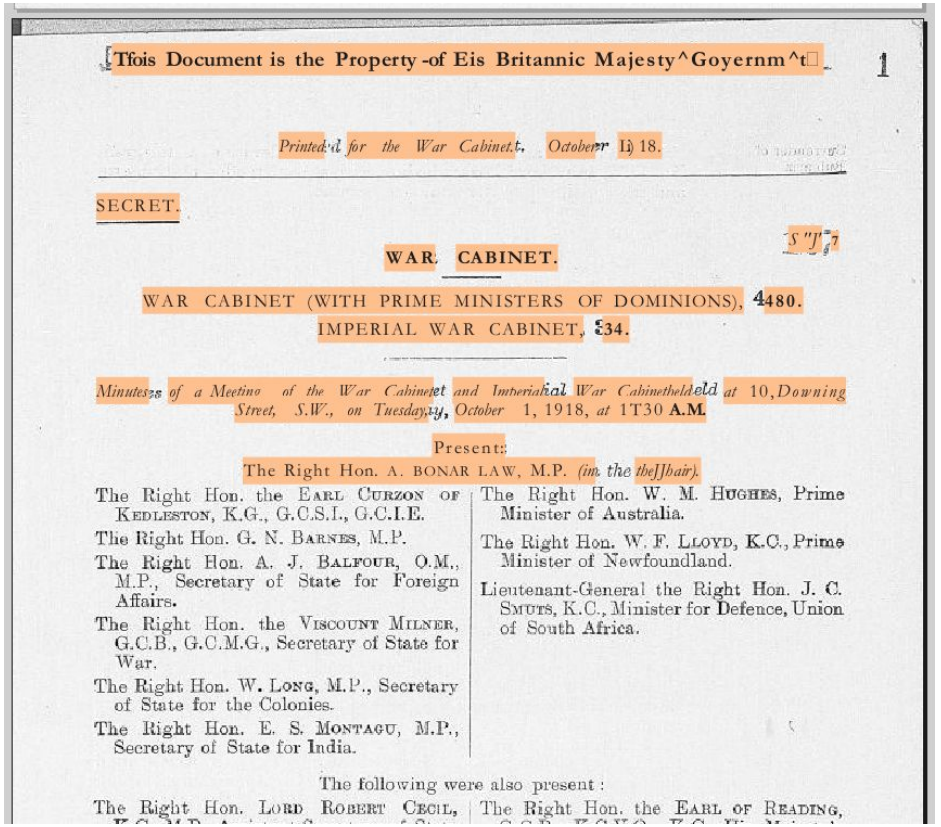
The following were also present:

The Right Hon. LORD ROBERT CECIL,
The Right Hon. the EARL OF READING,

- PDF with OCR text
  - Extraction of text
  - Document splitting

Alberto

Source text acquisition



[Tfois Document is the Property -of Eis Britannic Majesty^Goyernm^t□ 1

Printed for the War Cabinet. October Li) 18.

SECRET.

S "J" 7

WAR CABINET.

WAR CABINET (WITH PRIME MINISTERS OF DOMINIONS), 4480.
IMPERIAL WAR CABINET, 34.

Minutes of a Meeting of the War Cabinet and Imperial War Cabinet held at 10, Downing Street, S.W., on Tuesday, October 1, 1918, *at* 1T30 A.M.

Present:

The Right Hon. A. BONAR LAW, M.P. *(in the Chair).*

The Right Hon. the EARL CURZON OF KEDLESTON, K.G., G.C.S.I., G.C.I.E.
The Right Hon. G. N. BARNES, M.P.
The Right Hon. A. J. BALFOUR, O.M., M.P., Secretary of State for Foreign Affairs.
The Right Hon. the VISCOUNT MILNER, G.C.B., G.C.M.G., Secretary of State for War.
The Right Hon. W. LONG, M.P., Secretary of State for the Colonies.
The Right Hon. E. S. MONTAGU, M.P., Secretary of State for India.
The Right Hon. W. M. HUGHES, Prime Minister of Australia.
The Right Hon. W. F. LLOYD, K.C., Prime Minister of Newfoundland.
Lieutenant-General the Right Hon. J. C. SMUTS, K.C., Minister for Defence, Union of South Africa.

The following were also present:

The Right Hon. LORD ROBERT CECIL,
The Right Hon. the EARL OF READING,

*iTfois Document is the Property -of Eis Britannic Majesty^Goyernm^tX

Printed SECRET.

for the War Cabinet.

October

Li) 18.

WAR

CABINET.

S "J' 480.

7

W A R CABINET (WITH PRIME MINISTERS OF DOMINIONS), IMPERIAL W A R CABINET, ++Minutes of a Meeting of the War Cabinet Street, S.W., on Tuesday, 34.

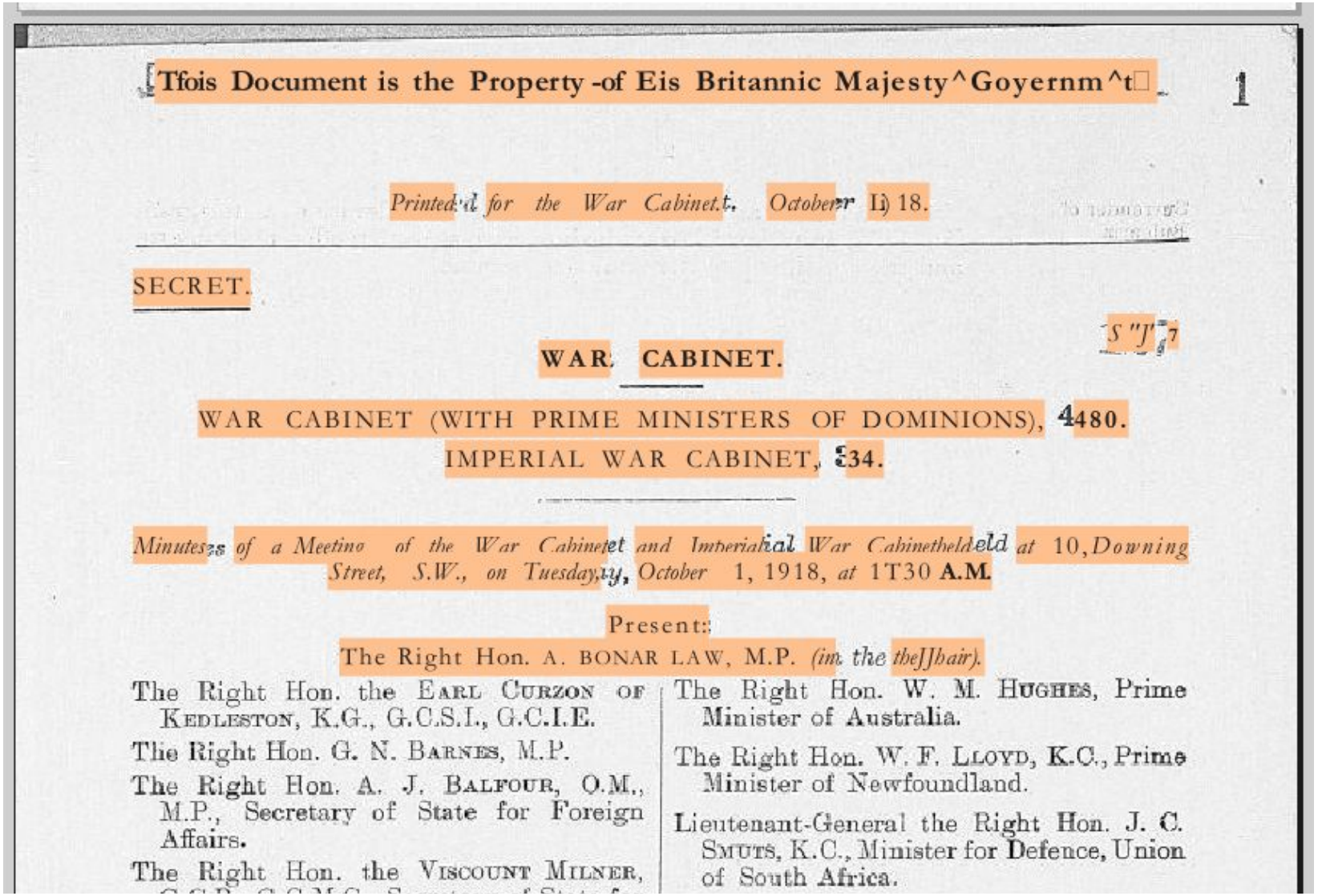
and Imperial War Cabinetheld at 1 0 , D o w n i n g October 1, 1918, at 1T30 A.M.
Present:

theJJhair). The Right Hon. A . BONAR L A W , M.P. (in The Right Hon. the E A R L CURZON OP The Right Hon. W. M. HUGHES, Prime Minister of Australia. KEDLESTON, K . G . , G-.C.S.L, G.C.I.E. The Right Hon. G. N . BARNES, M . P . The Right Hon. W. F. LLOYD, K G , Prime Minister of Newfoundland. The Right Hon. A . J . BALFOUR, O.M., M . P . , Secretary of State for Foreign Lieutenant-General the Right Hon. J . C. Affairs. SMUTS, K G , Minister for Defence, Union The Right Hon. the VISCOUNT MILNER, of South Africa. G.C.B., G.C.M.G., Secretary of State for War. The Right Hon. W . LONG, M . P . , Secretary of State for the Colonies. The Right Hon. E . S. MONTAGU, M . P . , Secretary of State for India.

T

Alberto

Source text acquisition



Tfois Document is the Property -of Eis Britannic Majesty^Goyernm^t 1

Printed for the War Cabinet. October 1918.

SECRET.

WAR CABINET.

WAR CABINET (WITH PRIME MINISTERS OF DOMINIONS), 480.
IMPERIAL WAR CABINET, 34.

Minutes of a Meeting of the War Cabinet and Imperial War Cabinet held at 10, Downing Street, S.W., on Tuesday, October 1, 1918, *at* 11 30 **A.M.**

Present:

The Right Hon. A. BONAR LAW, M.P. *(in the Chair).*

The Right Hon. the EARL CURZON OF KEDLESTON, K.G., G.C.S.I., G.C.I.E.

The Right Hon. G. N. BARNES, M.P.

The Right Hon. A. J. BALFOUR, O.M., M.P., Secretary of State for Foreign Affairs.

The Right Hon. the VISCOUNT MILNER,

The Right Hon. W. M. HUGHES, Prime Minister of Australia.

The Right Hon. W. F. LLOYD, K.C., Prime Minister of Newfoundland.

Lieutenant-General the Right Hon. J. C. SMUTS, K.C., Minister for Defence, Union of South Africa.

Alberto

Source text acquisition

```
*iTfois Document is the Property -of Eis Britannic Majesty^Goyernm^tX

Printed SECRET.

for the War Cabinet.

October

Li) 18.

WAR

CABINET.

S "J' 480.

7

W A R CABINET (WITH PRIME MINISTERS OF DOMINIONS), IMPERIAL W A R CABINET, ++Minutes
of a Meeting of the War Cabinet Street, S.W., on Tuesday, 34.

and Imperial War Cabinetheld at 1 0 , D o w n i n g October 1, 1918, at 1T30 A.M.
Present:
```

Text extracted with “pdftotext” from poppler.freedesktop.org

Alberto

Source text acquisition

```
*iTfois Document is the Property -of Eis Britannic Majesty^Goyernm^tX
Printed SECRET.
for the War Cabinet.
October
Li) 18.
WAR
CABINET.
S "J' 480.
7

W A R CABINET (WITH PRIME MINISTERS OF DOMINIONS), IMPERIAL W A R CABINET, ++Minutes
of a Meeting of the War Cabinet Street, S.W., on Tuesday, 34.

and Imperial War Cabinetheld at 1 0 , D o w n i n g October 1, 1918, at 1T30 A.M.
Present:

theJJhair). The Right Hon. A . BONAR L A W , M.P. (in The Right Hon. the E A R L
CURZON OP The Right Hon. W. M. HUGHES, Prime Minister of Australia. KEDLESTON, K . G .
, G-.C.S.L, G.C.I.E. The Right Hon. G. N . BARNES, M . P . The Right Hon. W. F. LLOYD,
K G , Prime Minister of Newfoundland. The Right Hon. A . J . BALFOUR, O.M., M . P . ,
Secretary of State for Foreign Lieutenant-General the Right Hon. J . C. Affairs.
SMUTS, K G , Minister for Defence, Union The Right Hon. the VISCOUNT MILNER, of South
Africa. G.C.B., G.C.M.G., Secretary of State for War. The Right Hon. W . LONG, M . P .
, Secretary of State for the Colonies. The Right Hon. E . S. MONTAGU, M . P . ,
Secretary of State for India.

T
```

Alberto

Source text acquisition

*iTfois Document is the Property -of Eis Britannic Majesty^Goyernm^tX
Printed SECRET.
for the War Cabinet.
October
Li) 18.
WAR
CABINET.
S "J' 480.
7

W A R utes
of a M

and Im .
Presen

theJJh
CURZON OF The Right Hon. W. M. HUGHES, Prime Minister of Australia. KEDLESTON, K . G .
, G-.C.S.L, G.C.I.E. The Right Hon. G. N . BARNES, M . P . The Right Hon. W. F. LLOYD,
K G , Prime Minister of Newfoundland. The Right Hon. A . J . BALFOUR, O.M., M . P . ,
Secretary of State for Foreign Lieutenant-General the Right Hon. J . C. Affairs.
SMUTS, K G , Minister for Defence, Union The Right Hon. the VISCOUNT MILNER, of South
Africa. G.C.B., G.C.M.G., Secretary of State for War. The Right Hon. W . LONG, M . P .
, Secretary of State for the Colonies. The Right Hon. E . S. MONTAGU, M . P . ,
Secretary of State for India.

T

Problem: several documents per file

Approach: find document start line, split there

Alberto

Source text acquisition

```
*iTfois Document is the Property -of Eis Britannic Majesty^Goyernm^tX
Printed SECRET.
for the War Cabinet.
October
Li) 18.
WAR
CABINET.
S "J' 480.
7
```

```
patterns = [
    re.compile(r"\bthis\b.*\bdocument\b.*\bproperty\b", re.I),
    re.compile(r"\bdocument\b.*\bproperty\b.*\bhis\b +\bbritannic\b", re.I),
    re.compile(r"\bproperty\b.*\bbritannic\b +\bmajesty\b", re.I),
    re.compile(r"\bdocument\b.*\bproperty\b.*\bmajesty\b", re.I),
    re.compile(r"\bthis\b +\bdocument\b.*\bgovernment\b", re.I),
    re.compile(r"\bproperty\b +\bof\b.*\bgovernment\b", re.I),
]
```

Split if more than one pattern matches the line.

Alberto

Source text acquisition

```
*iTfois Document is the Property -of Eis Britannic Majesty^Goyernm^tX
Printed SECRET.
for the War Cabinet.
October
Li) 18.
WAR
CABINET.
S "J' 480.
7
```

Header repeats in first pages of some documents

→ split only if document length > 60 lines

Alberto

Document clean-up and date extraction

Document clean-up

- Biggest problem: words with spaces in them
- Regexp replacement: `(\b\S)\b\s\b`
- Unwanted side effect: single characters (like the article "a") concatenated to next word
- Use a word splitting library

Date extraction

- Which method?
  - Browse through provided NLP links:
    DANTE:

DANTE web demo

Reference date and time (assumed document creation date):

2010-03-05 18:49:48

Document:

```
(THIS DOCUMENT' IS TET PROPERTY OF HIS BRITANNIC MAJE3TY S
G-OVxRNMSN'
1

S E C

R E T . C A 3 I N E T SO (36),

COPY NOj

Meeting of the Cabinet to be held at No. 10, Downing Street,
3.W.I., on WEDNESDAY, 23th OCTOBER, 1936, at 11.0 a.m.

AGENDA. 1. FOREIGN AEF/-IR3. (9-) Preparations for the Five
Power Conference. (Reference Cabinet 58 (36) Conclusion 3 )
. Memoranda by the Secretary of State for Foreign Affairs. C
P . 268 (36) - already circulated. C P . 278 (36) - already
circulated. C P . 284 (36) - to be circulated. ("b) The
Situation- in Spain (if required).

(Reference Cabinet 58 (36) Conclusion 6 ) .
2


The Cabinet had before them the following
```

Submit

Johannes

Date extraction

- Which method?
  - Browse through provided NLP links:
    DANTE:
  - Problems: doesn't deal with
    "24 hours after 3 October"
    "between 5 and 7 October"

Date extraction

- What do we need all **dates** for?
  - most important is the date
    of cabinet meeting
    → extract "held on" date
  - we can do that with regular expressions

Date extraction

- Dates we have to deal with:

```
Tuesday, 7th August, 1 9 4 5 , at 5 - 0 p.m. Street, S.W. 1,
Thursday, 1st November, 1 9 4 5 , at 1 1 a.m.
Tuesday, 1st January, 1 9 4 6 , at 1 1 a.m.
December 9, 1916, at 11-30 A.M.
March 1 , 1 9 1 7 , at 1 1 - 3 0 A . M .
Tuesday, June 5, 1917, at 11*30 A.M.
Tuesday, January 1 , 1 9 1 S , at 1 1 * 3 0 A.M.
Monday, April .1, 1918, at 1T30 A.M.
Monday, July 1, 1018, at 12 noon.
October 1, 1918, at 1T30 A.M.
Thursday,3 , 1 9 1 9 , at 12
Tuesday, July 1 , 1 9 1 9 , at 1 1 * 3 0 A.M.
Friday, June 8, 1917, at IE.30 a.m.
Friday, August 15, 1919, at 1 1 3 0 A.M
Friday, June 8, 1917, at IE.30 a.m.
Tuesday, January 2, 1940, at 11 A . M
WEDNESDAY, 21st JUNE, 1939, at 10o30 a,m
MONDAY, 24th APRIL, 1939 at 5.6 p.m
WEDNESDAY, 15th MARCH, 1939, at 11.0 a.m
WEDNESDAY, 22nd MARCH, 1959, at 10.0 a.m
```

Johannes

Date extraction

- Dates we have to deal with:

- Preprocessing (post-OCR):

```
pattern = Pattern.compile("(\\b\\S)\\b\\s\\b");
matcher = pattern.matcher(text);
correctedText = matcher.replaceAll("$1");
```

www.myregexp.com for Java

- Slot extraction:
  - year, month, day, time
  - day of week?
  - order of elements

- Remaining problems:

```
1T30 A.M.
IE.30 a.m.
10o30
```

```
Tuesday, 7th August, 1 9 4 5 , at 5 - 0 p.m. Street, S.W. 1,
Thursday, 1st November, 1 9 4 5 , at 1 1 a.m.
Tuesday, 1st January, 1 9 4 6 , at 1 1 a.m.
December 9, 1916, at 11-30 A.M.
March 1 , 1 9 1 7 , at 1 1 - 3 0 A . M .
Tuesday, June 5, 1917, at 11*30 A.M.
Tuesday, January 1 , 1 9 1 S , at 1 1 * 3 0 A.M.
Monday, April .1, 1918, at 1T30 A.M.
Monday, July 1, 1018, at 12 noon.
October 1, 1918, at 1T30 A.M.
Thursday,3 , 1 9 1 9 , at 12
Tuesday, July 1 , 1 9 1 9 , at 1 1 * 3 0 A.M.
Friday, June 8, 1917, at IE.30 a.m.
Friday, August 15, 1919, at 1 1 3 0 A.M
Friday, June 8, 1917, at IE.30 a.m.
Tuesday, January 2, 1940, at 11 A . M
WEDNESDAY, 21st JUNE, 1939, at 10o30 a,m
MONDAY, 24th APRIL, 1939 at 5.6 p.m
WEDNESDAY, 15th MARCH, 1939, at 11.0 a.m
WEDNESDAY, 22nd MARCH, 1959, at 10.0 a.m
```

Johannes

Named Entity Recognition

Named Entity Recognition

- Need to extract named-entities to derive facts about them
- At the very least:
  - whether they are present
  - how many times in a document

Named Entity Recognition

- Three approaches:
  - Own regexp-based tagger
  - Stanford NER
  - OpenNLP NER
- Technical difficulties for compilation
  - Solved finally for OpenNLP
  - Likely similar for Stanford NER

Named Entity Recognition

- Adaptation to our Document.SegmentList:
  - OpenNLP tokenizer removes spaces
    - Span offsets do not match source text
  - Otherwise fine
  - Possible to use different libraries and compare

Cabinet meetings attendant list extraction

Attendant list extraction

- List of attendants to meetings of the Cabinet
- Regular structure in documents
- List of attendants separated at beginning
  - labeled with words like "Present:"
  - finished with "1." or "]."(as OCR-error)
  - allows us to extract it with good recall even with relatively simple techniques

Attendant list extraction

- Approaches
  - Regular expressions
  - OpenNLP NER
- Structure of block elements not easy to parse
- Names variably denoted
  - Titles of honor
  - Position in the office
  - First name(s) and last name

Attendant list extraction

Example:

```
Major-General F. B. MAURICE, C.B.,

AdmiralSr.RJ . R . JELLICOE , GOB . , O.M.,

Directorof Militarv Office.

The Hon. SIRJ. S.

Operations, WarMESTON, K.C.S.L,

G.O.V.O., FirstSeaLord . TheHon . R . ROGERS ,
Ministerof PublicWorks , Canada. TheHon . J . L).
HAZEN , Ministerof Marine, andFisheries , andof
theNavalService, Canada.

Mr. H. 0. M. LAMBERT, C . B . , Colonial

Lieutenant-Governor Provinces, India.
```


Implementation

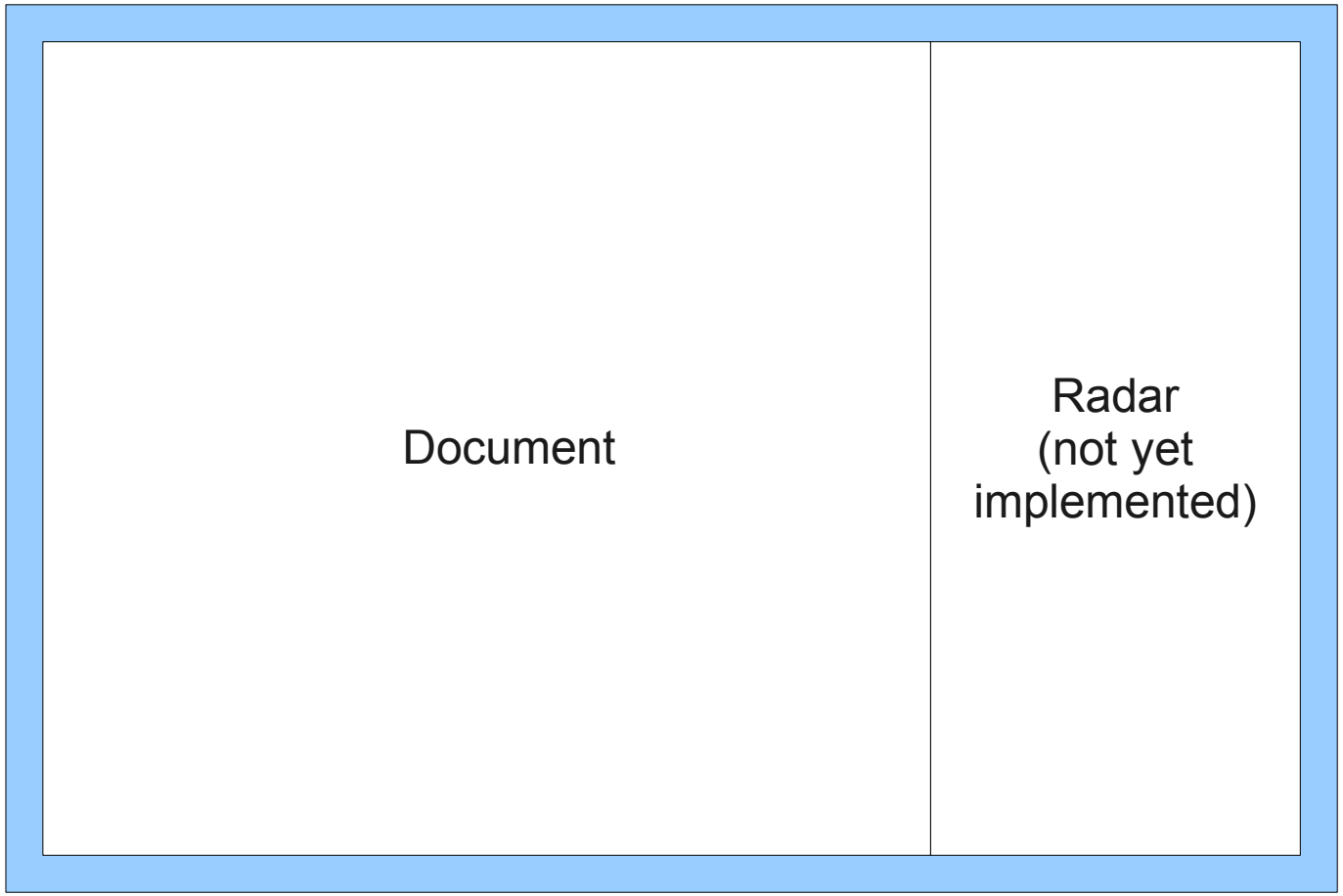
Implementation

- Complete GUI application in Java
- Full source code available
- Free Software / Open Source: GPL v3
- Completed after this presentation
- Details in the final report

Implementation

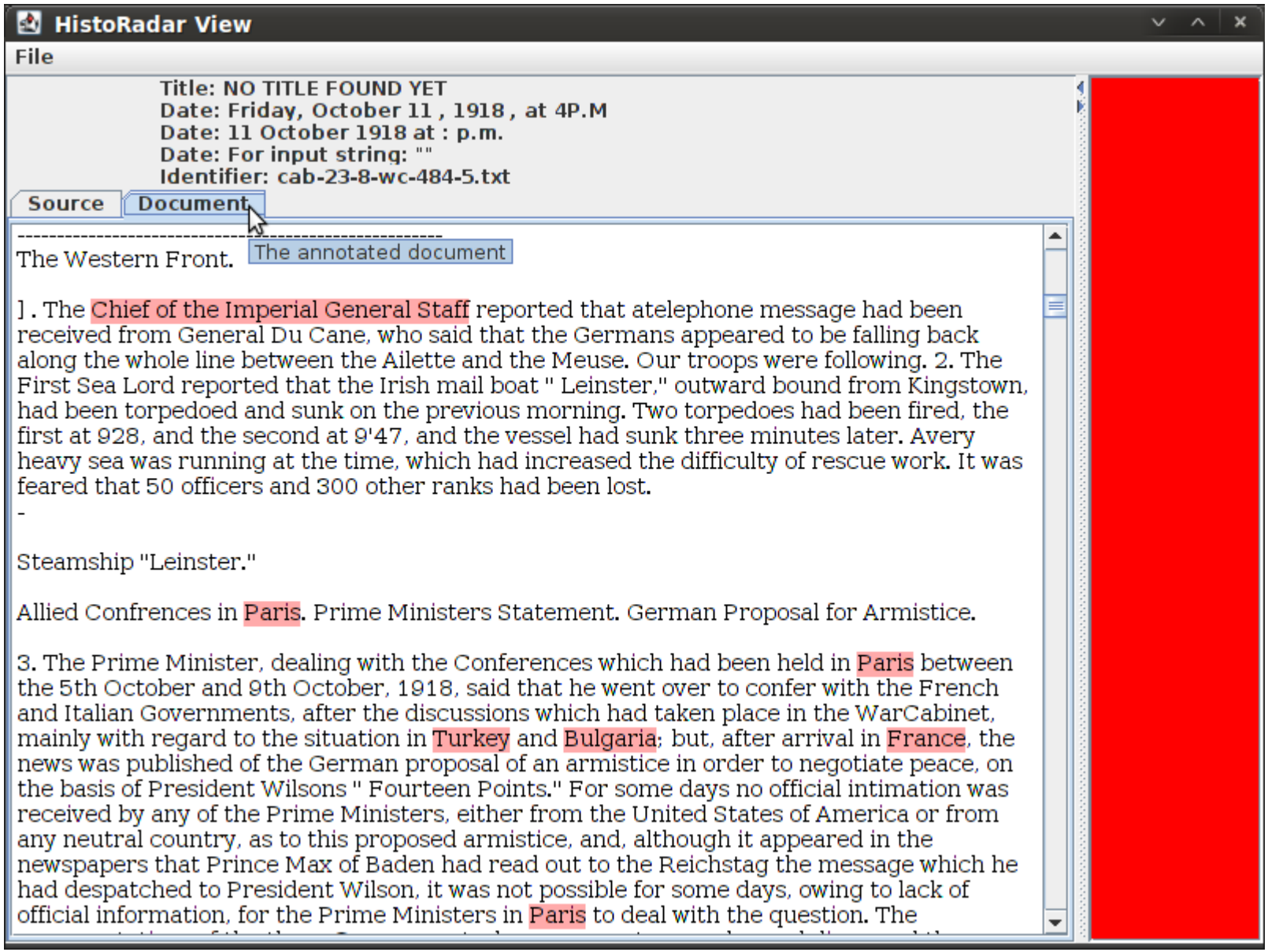
- Attributed document segments
  - Begin and end character offsets
  - Arbitrary string attributes
- Analogous to other implementations
  - OpenNLP opennlp.tools.util.Span
  - GATE gate.SimpleAnnotation
- XML export of document with annotation

User Interface



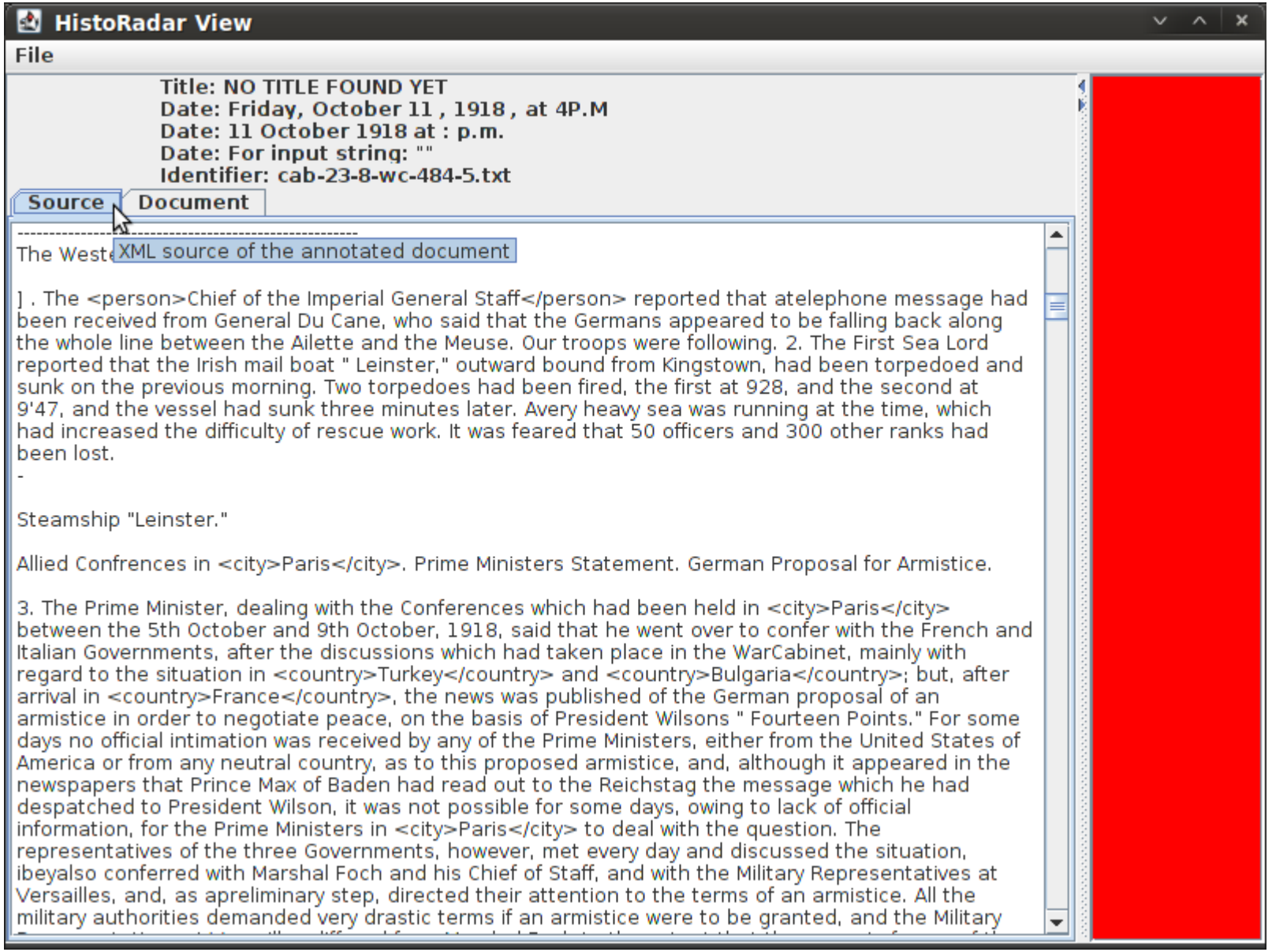
Alberto

User Interface



Alberto

User Interface



Alberto

User Interface

- Ideas for additional features
  - Support “snowball” method
    - One-click bookmarking
      - Select text in document if desired
      - Click bookmark button
      - Bookmark added with optional citation text and notes
      - Export to HTML, BibTeX, Zotero, …
    - Display area for “snowball” bookmarks
  - Integrated search with query expansion

Questions?

http://historadar.googlecode.com/

Image sources

- Snowballing

  - http://www.flickr.com/photos/artsmonkey/3250602627/

Alberto