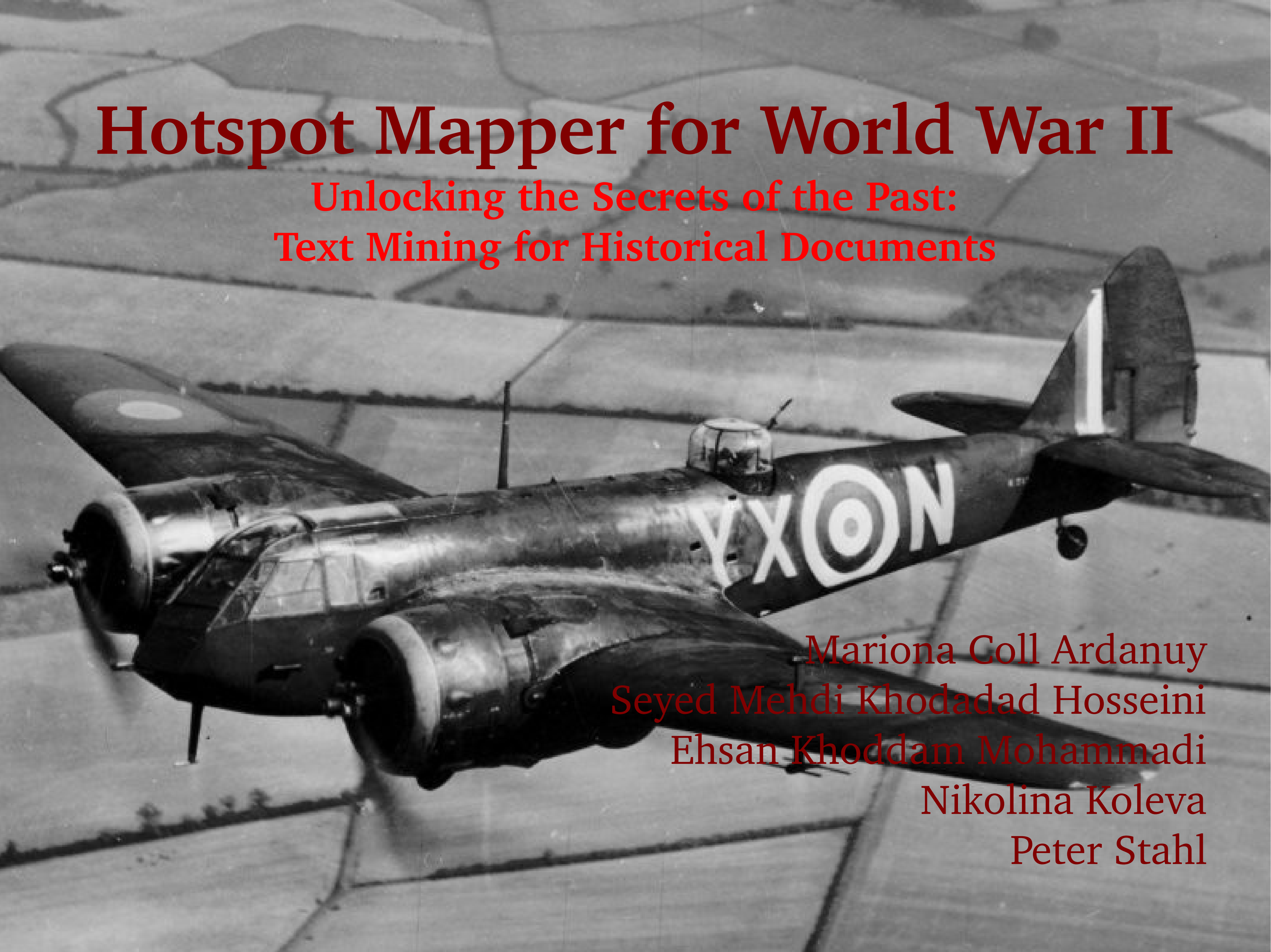


Hotspot Mapper for World War II

Unlocking the Secrets of the Past:
Text Mining for Historical Documents

Mariona Coll Ardanuy
Seyed Mehdi Khodadad Hosseini
Ehsan Khoddam Mohammadi
Nikolina Koleva
Peter Stahl

Demo

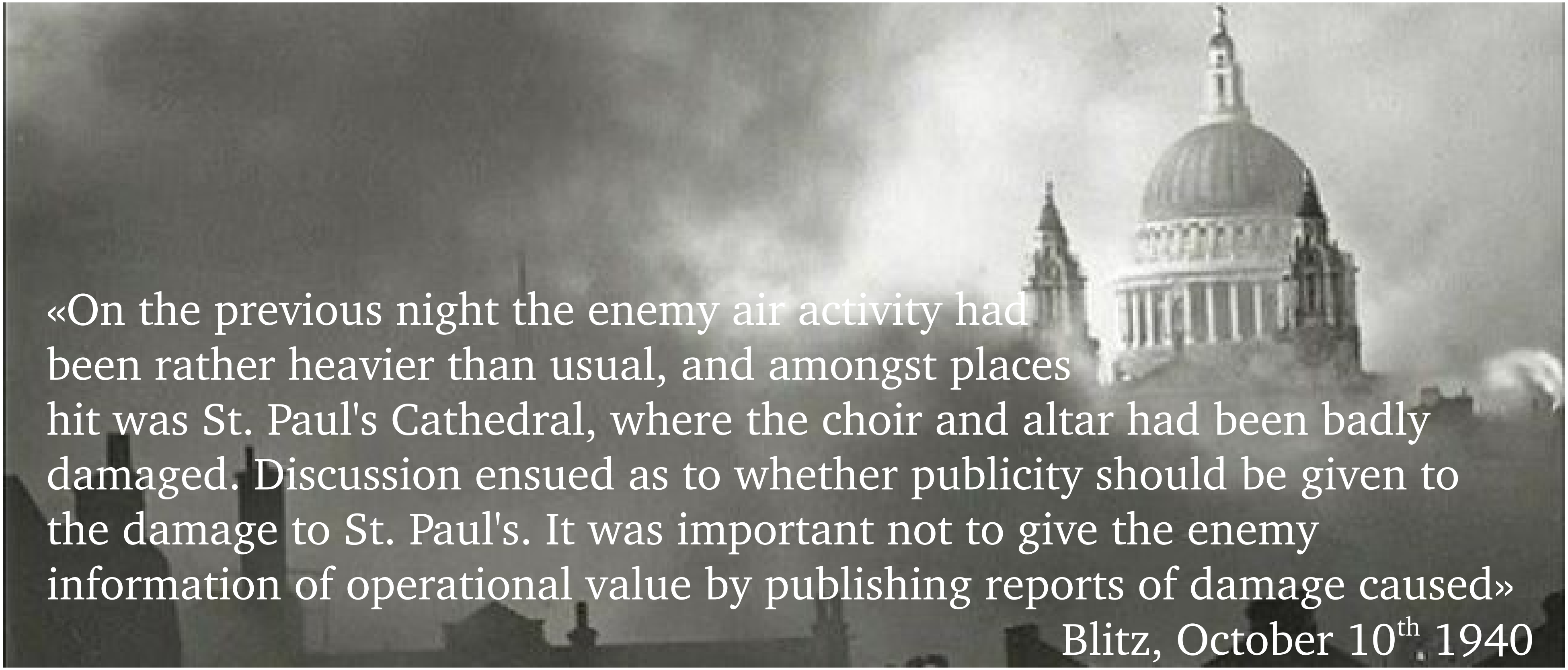
Historical Motivation

- August 27th 1939: Imminence of war, underground War Rooms in London became fully operational
- September 3rd 1939: Britain declared war to Germany
- Cabinet Room: Prime Minister, military strategists and Government ministers plotted the war there: 115 cabinet meetings, 226 documents issued

«This is the room from which I will direct the war»

The Collection

- British Cabinet Papers, part of The National Archives
- 216 texts from the period 1939-1945, total of 842,496 words
- Written in contemporary and descriptive style
- Development and magnitude of events, fears and reliefs, war strategy



«On the previous night the enemy air activity had been rather heavier than usual, and amongst places hit was St. Paul's Cathedral, where the choir and altar had been badly damaged. Discussion ensued as to whether publicity should be given to the damage to St. Paul's. It was important not to give the enemy information of operational value by publishing reports of damage caused»

Blitz, October 10th, 1940

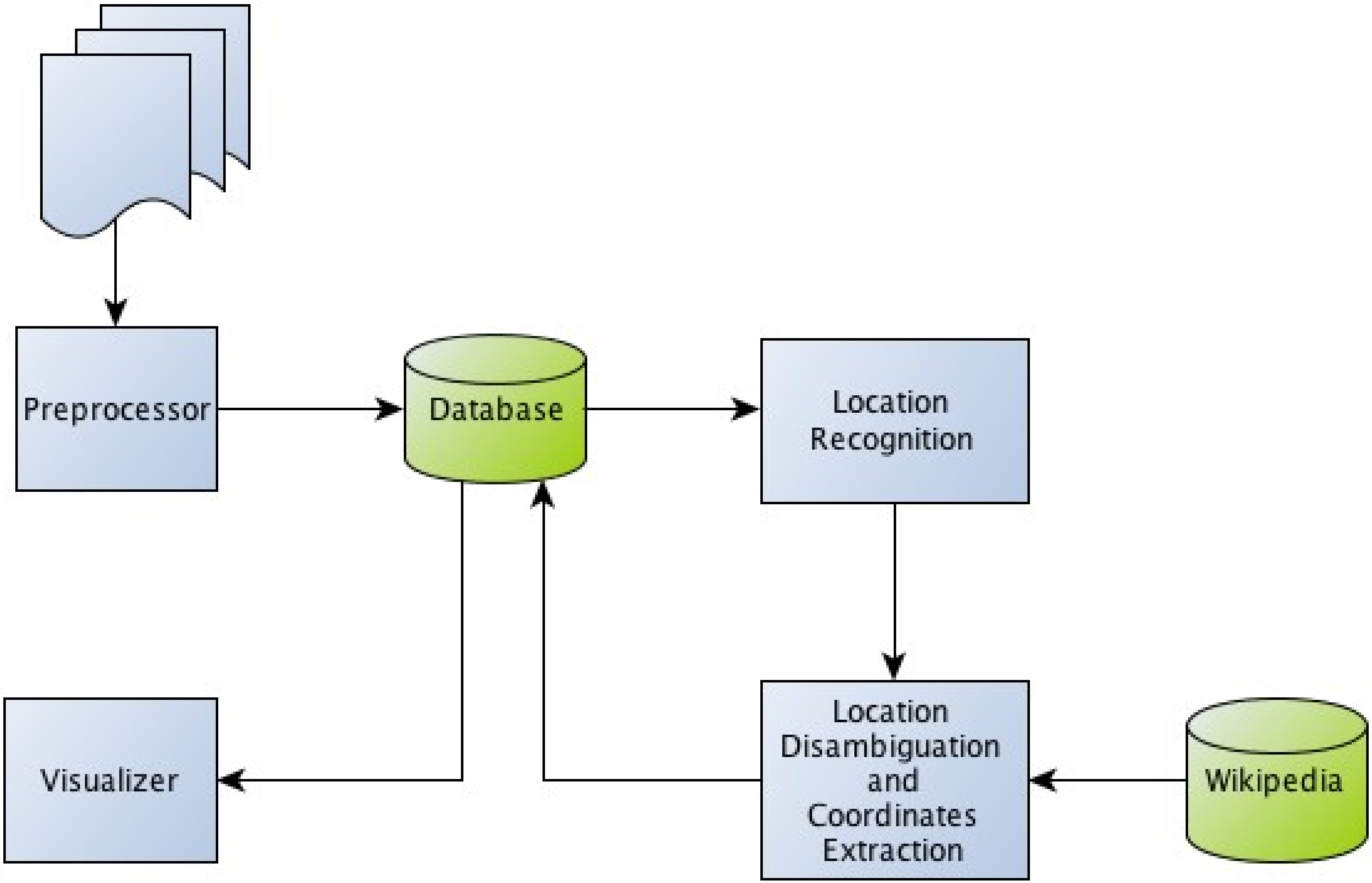
Our Project: an Approach to History

- WWII: Probably the most-studied conflict ever
- User-friendly access to primary sources on the development of the war
- Interesting both for historians and non-experts
  - Historians: different perspective of the conflict, easy access to the primary sources
  - Non-experts: overview of the conflict, what countries played into it, etc.

Overview

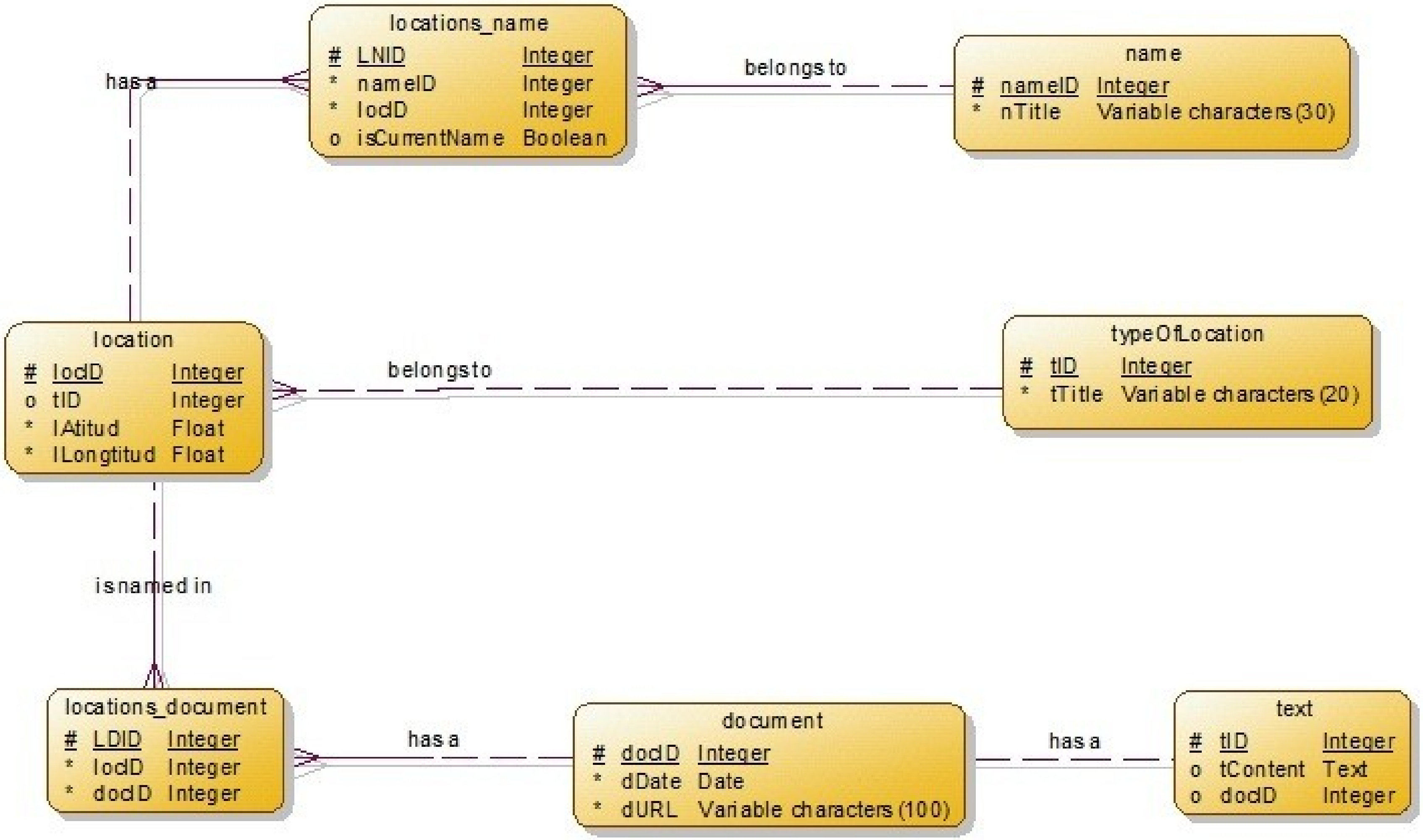
- **Motivation**
- **System architecture**
- **Components**
  - Preprocessor
  - Location Recognition
  - Coordinates Extraction
  - Location Disambiguation
  - Visualizer
- **Evaluation**
- **Future work**

System architecture

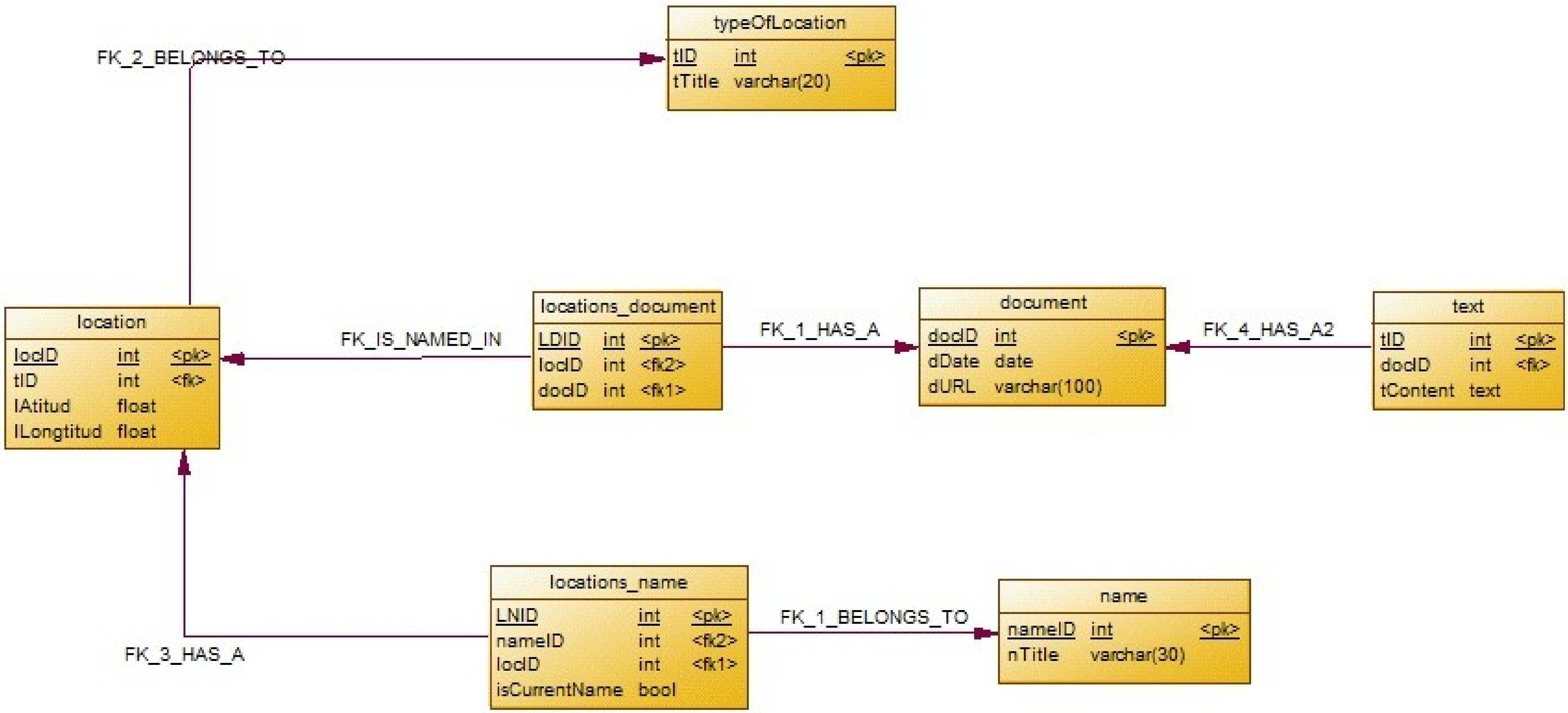


Database and MySQL

Logical Model Diagram



Physical Model Diagram



typeOfLocation		
tID	int	<pk>
tTitle	varchar(20)	

location		
locID	int	<pk>
tID	int	<fk>
lAtitud	float	
lLongtitud	float	

locations_document		
LDID	int	<pk>
locID	int	<fk2>
docID	int	<fk1>

document		
docID	int	<pk>
dDate	date	
dURL	varchar(100)	

text		
tID	int	<pk>
docID	int	<fk>
tContent	text	

locations_name		
LNID	int	<pk>
nameID	int	<fk2>
locID	int	<fk1>
isCurrentName	bool	

name		
nameID	int	<pk>
nTitle	varchar(30)	

FK_2_BELONGS_TO

FK_IS_NAMED_IN

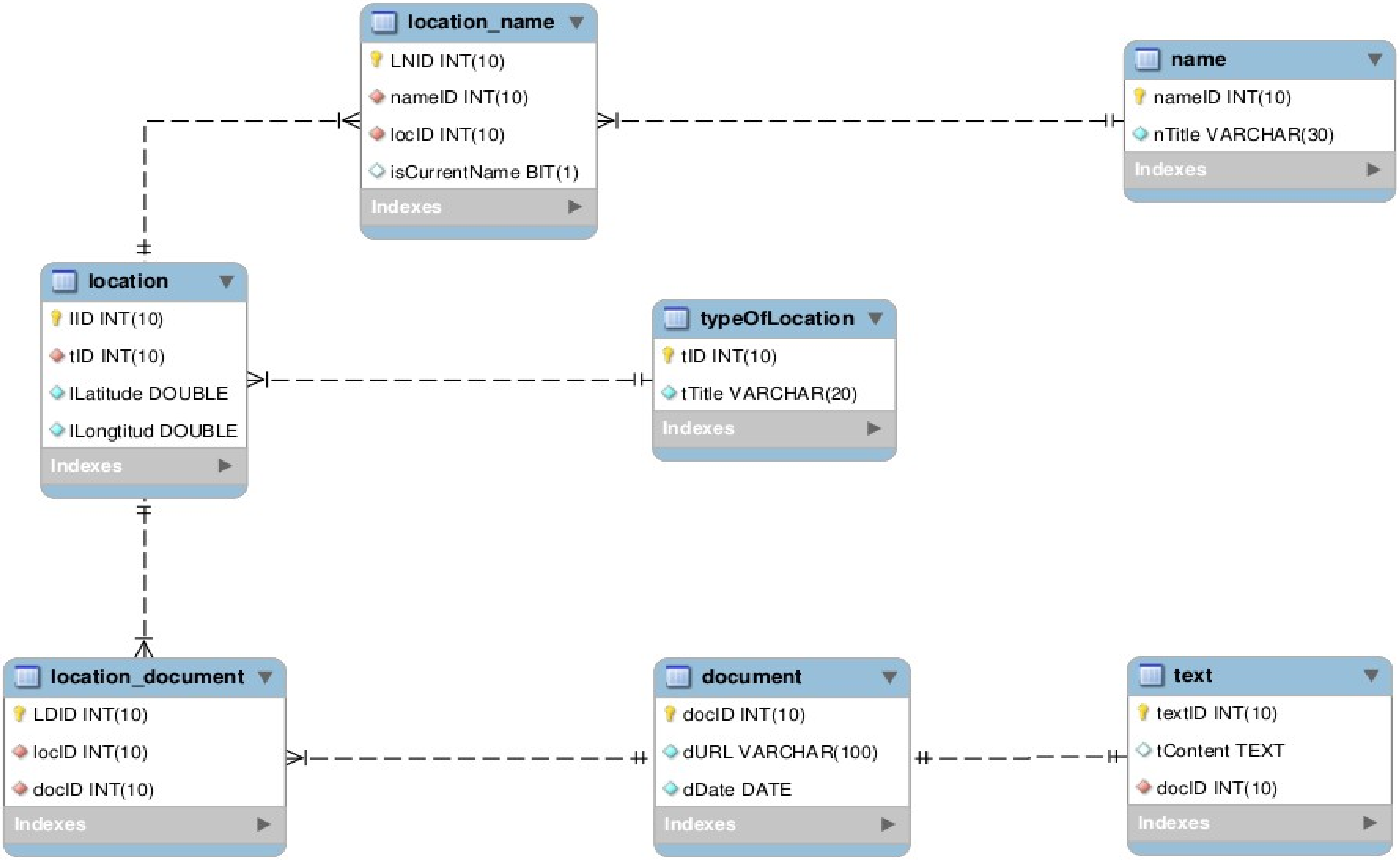
FK_1_HAS_A

FK_4_HAS_A2

FK_3_HAS_A

FK_1_BELONGS_TO

Database Diagram



location_name
LNID INT(10)
nameID INT(10)
locID INT(10)
isCurrentName BIT(1)
Indexes

name
nameID INT(10)
nTitle VARCHAR(30)
Indexes

location
lID INT(10)
tID INT(10)
lLatitude DOUBLE
lLongtitud DOUBLE
Indexes

typeOfLocation
tID INT(10)
tTitle VARCHAR(20)
Indexes

location_document
LDID INT(10)
locID INT(10)
docID INT(10)
Indexes

document
docID INT(10)
dURL VARCHAR(100)
dDate DATE
Indexes

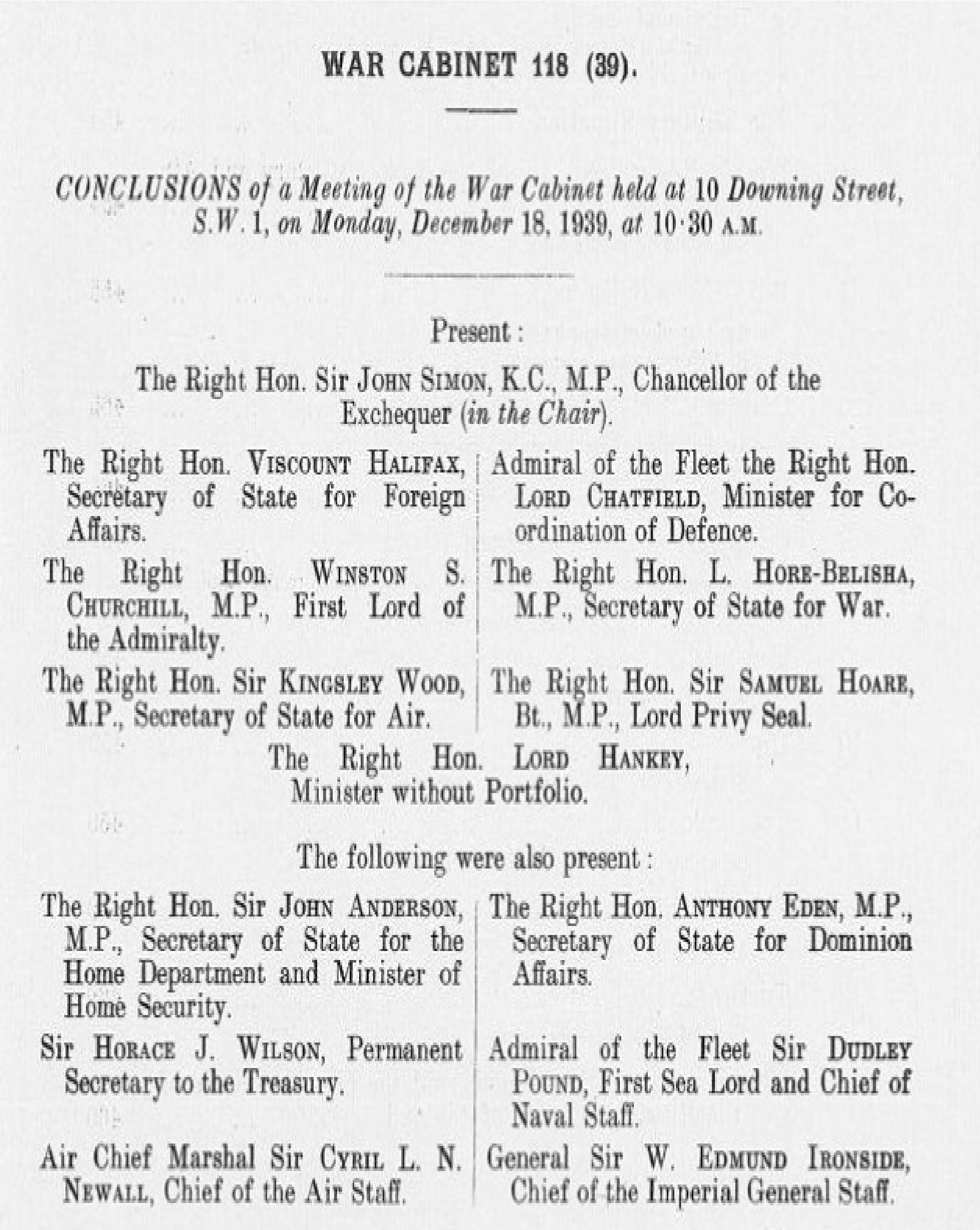
text
textID INT(10)
tContent TEXT
docID INT(10)
Indexes

Preprocessor

- Input: 216 text files converted from OCRed pdfs
- Stored following attributes for each document in the data base:
    - url of the original pdf document
    - month and year (when was the document written)
    - preprocessed text

Preprocessor

Filtering tables with names of present people



WAR CABINET 118 (39).

CONCLUSIONS of a Meeting of the War Cabinet held at 10 *Downing Street, S.W.* 1, *on Monday, December* 18, 1939, *at* 10·30 A.M.

Present :

The Right Hon. Sir JOHN SIMON, K.C., M.P., Chancellor of the Exchequer (*in the Chair*).

The Right Hon. VISCOUNT HALIFAX, Secretary of State for Foreign Affairs.	Admiral of the Fleet the Right Hon. LORD CHATFIELD, Minister for Co-ordination of Defence.
The Right Hon. WINSTON S. CHURCHILL, M.P., First Lord of the Admiralty.	The Right Hon. L. HORE-BELISHA, M.P., Secretary of State for War.
The Right Hon. Sir KINGSLEY WOOD, M.P., Secretary of State for Air.	The Right Hon. Sir SAMUEL HOARE, Bt., M.P., Lord Privy Seal.

The Right Hon. LORD HANKEY, Minister without Portfolio.

The following were also present :

The Right Hon. Sir JOHN ANDERSON, M.P., Secretary of State for the Home Department and Minister of Home Security.	The Right Hon. ANTHONY EDEN, M.P., Secretary of State for Dominion Affairs.
Sir HORACE J. WILSON, Permanent Secretary to the Treasury.	Admiral of the Fleet Sir DUDLEY POUND, First Sea Lord and Chief of Naval Staff.
Air Chief Marshal Sir CYRIL L. N. NEWALL, Chief of the Air Staff.	General Sir W. EDMUND IRONSIDE, Chief of the Imperial General Staff.

The Right Hon. Sir <PERSON>JOHN SIMON</PERSON>, <LOCATION>**K.C.**</LOCATION>, <LOCATION>**M.P.**</LOCATION>, Chancellor of the
Exchequer (in the Chair).

The Right Hon. VISCOUNT <LOCATION>**HALIFAX**</LOCATION>, Admiral of the Fleet the Right Hon.
Secretary of State for Foreign LORD <LOCATION>**CHATFIELD**</LOCATION>, Minister for <ORGANIZATION>Co</ORGANIZATION>-
Affairs. ordination of <ORGANIZATION>Defence</ORGANIZATION>.

The Right Hem, WINSTON S. The Right Hon. <PERSON>L. HORE-BELISHA</PERSON>,
<LOCATION>**CHURCHILL**</LOCATION>, <LOCATION>**M.P.**</LOCATION>, <LOCATION>**First Lord of M.P.**</LOCATION>, Secretary of State for War.
the Admiralty.

The Right Hon. Sir <PERSON>KINGSLEY</PERSON> WOOD, I The Right Hon. Sir <LOCATION>**SAMUEL HOARE**</LOCATION>,
<LOCATION>**M.P.**</LOCATION>, Secretary of <ORGANIZATION>State for Air</ORGANIZATION>. I Bt.,
<LOCATION>**M.P.**</LOCATION>, Lord Privy Seal.
The Right Hon. LORD <PERSON>HANKEY</PERSON>,
Minister without Portfolio.

The following were also present:

The Right Hon. Sir <PERSON>JOHN ANDERSON</PERSON>, The Right Hon. <LOCATION>**ANTHONY EDEN**</LOCATION>, <LOCATION>**M.P.**</LOCATION>,
<LOCATION>**M.P.**</LOCATION>, Secretary of State for the Secretary of <ORGANIZATION>State for Dominion
Home Department</ORGANIZATION> and Minister of Affairs.
Home Security.

Sir <PERSON>HORACE J. WILSON</PERSON>, Permanent Admiral of the Fleet Sir DUDLEY
Secretary to the <ORGANIZATION>Treasury</ORGANIZATION>. POUND, <ORGANIZATION>First Sea Lord</ORGANIZATION> and Chief of
Naval Staff.
Air Chief Marshal Sir <PERSON>CYRIL L. N. General Sir W. EDMUND IRONSIDE</PERSON>,
<LOCATION>**NEWALL**</LOCATION>, Chief of the Air Staff. Chief of the <ORGANIZATION>Imperial General Staff</ORGANIZATION>.

Location Recognizer

- Integrate Stanford NER
- Use CoNLL model
  - Precision: 93.40
  - Recall: 83.33

for a random document

1. extracts the tagged locations of the output
2. filters acronyms (PVS, PX, S.C, etc.)
3. filters relative locations (East, West, South, etc.)
4. lists the found locations for each document

Location Disambiguation

Problems

- Ambiguities:
  - Different names for same location (temporal ambiguity, political ambiguity,...)

    Petrograd vs. *St. Petersburg*
  - Same name for different locations (local ambiguity)

    Frankfurt (Am Main) vs. *Frankfurt (Oder)*

Location Disambiguation

Solutions

- Temporal ambiguity: use Wikipedia and redirection links
- Local ambiguity:

  1) Document-Wikipedia similarity by measuring similarity of feature vectors where dimensions are words

  2) Document-Wikipedia similarity by measuring similarity of feature vectors where dimensions are locations

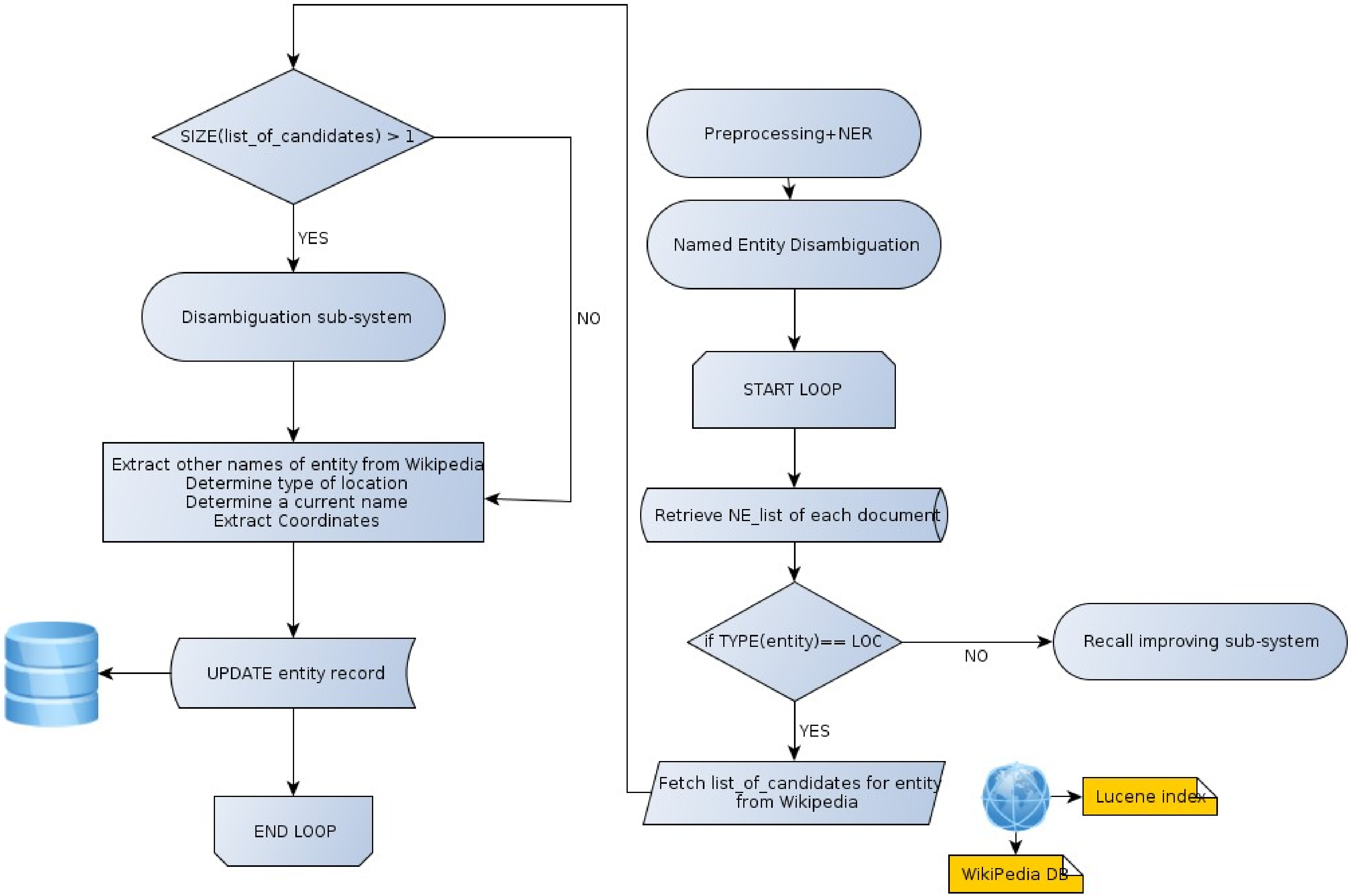
  3) Minimal distance set of name

Using Wikipedia as Knowledge Base

- Employing linking structure of entries to find the current name.
- Employing entries context for disambiguating.
- Extracting Coordinates from entries.
- We used dump of English Wikipedia database and JWPL to exploit Wikipedia information. (expert suggestion: do it on server!)

Location Disambiguation

Diagram



Preprocessing+NER
Named Entity Disambiguation
START LOOP
Retrieve NE_list of each document
if TYPE(entity)== LOC
NO
Recall improving sub-system
YES
Fetch list_of_candidates for entity from Wikipedia
Lucene index
WikiPedia DB
SIZE(list_of_candidates) > 1
YES
Disambiguation sub-system
NO
Extract other names of entity from Wikipedia
Determine type of location
Determine a current name
Extract Coordinates
UPDATE entity record
END LOOP

Visualization

- Dynamic Google Maps API
- Web Framework Django
  - Access the database
  - Create dynamic HTML pages to fill with the data

Evaluation

Next Episode

Future Work

- In order to improve the accuracy of our method:
  - OCR Correction
  - Train our own language model for the NER
  - Apply string correction and spell checking to the list of locations to avoid spelling variation (*Marseilles* → *Marseille*)
- Other improvements:
  - Extract types of locations (using another NER system: **SuperSense Tagger**)
  - Look for other different strategies to find candidates and disambiguating them (not considering locations as independent entities)
  - Use other search engines such as Google or Yahoo to help Wikipedia finding and extracting the disambiguated coordinates
  - Set a hierarchy of locations so that the user can see all the locations inside a country or a continent

References

- **Image Sources:**
  - Bristol Blenheim, RAF Museum Hendon
  - Churchill Picture from http://charlespaolino.files.wordpress.com/2011/12/war-churchill.jpg
  - St. Paul's Cathedral during the Blitz, Daily Mail 31 December 1940
  - The iconic photo taken on V-J Day in 1945, Alfred Eisenstaedt/Time & Life Pictures, via Getty Images
- **Other:**
  - Cabinet War Room Museum: http://www.iwm.org.uk/exhibitions/the-cabinet-war-rooms
  - British Cabinet Papers from the National Archives: http://www.nationalarchives.gov.uk/cabinetpapers/
  - Stanford NER: http://nlp.stanford.edu/software/CRF-NER.shtml
  - SuperSense Tagger NER: http://medialab.di.unipi.it/wiki/SuperSense_Tagger
  - English Wikipedia: http://en.wikipedia.org/wiki/Main_Page
  - Dynamic Google Maps API: http://googlemapsapi.blogspot.com/2007/03/creating-dynamic-client-side-maps.html
  - Web Framework Django: https://www.djangoproject.com/
  - MySQL: http://www.mysql.com/

THE END

Thank you for your attention!