

# **SaariStory: A framework to represent the medieval history of Saarland**

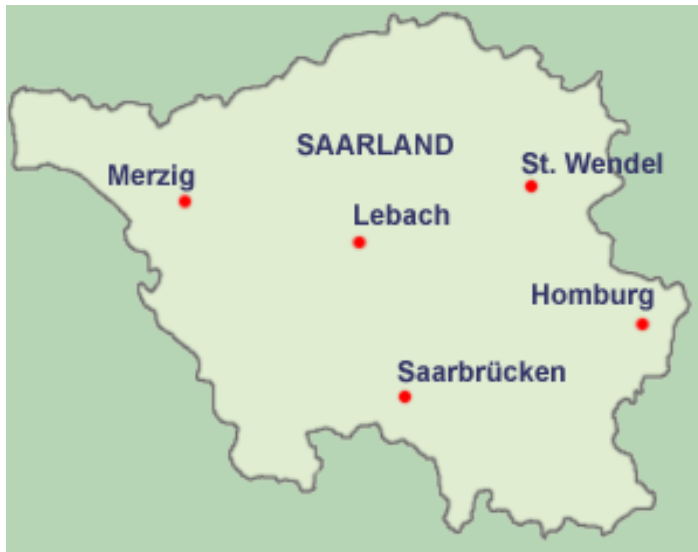
Group: SaariStorians

Michael Barz, Jonas Hempel, Cornelius Leidinger, Mainack Mondal  
Course supervisor: Dr. Caroline Sporleder

Text Mining for Historical Documents  
WS 2011/12

# Motivation

- Medieval History of Saarland



## Regesten zur Geschichte der Stadt Saarbrücken (bis 1545)

Bearbeitet unter Verwendung  
von Vorarbeiten von Hanns Klein †

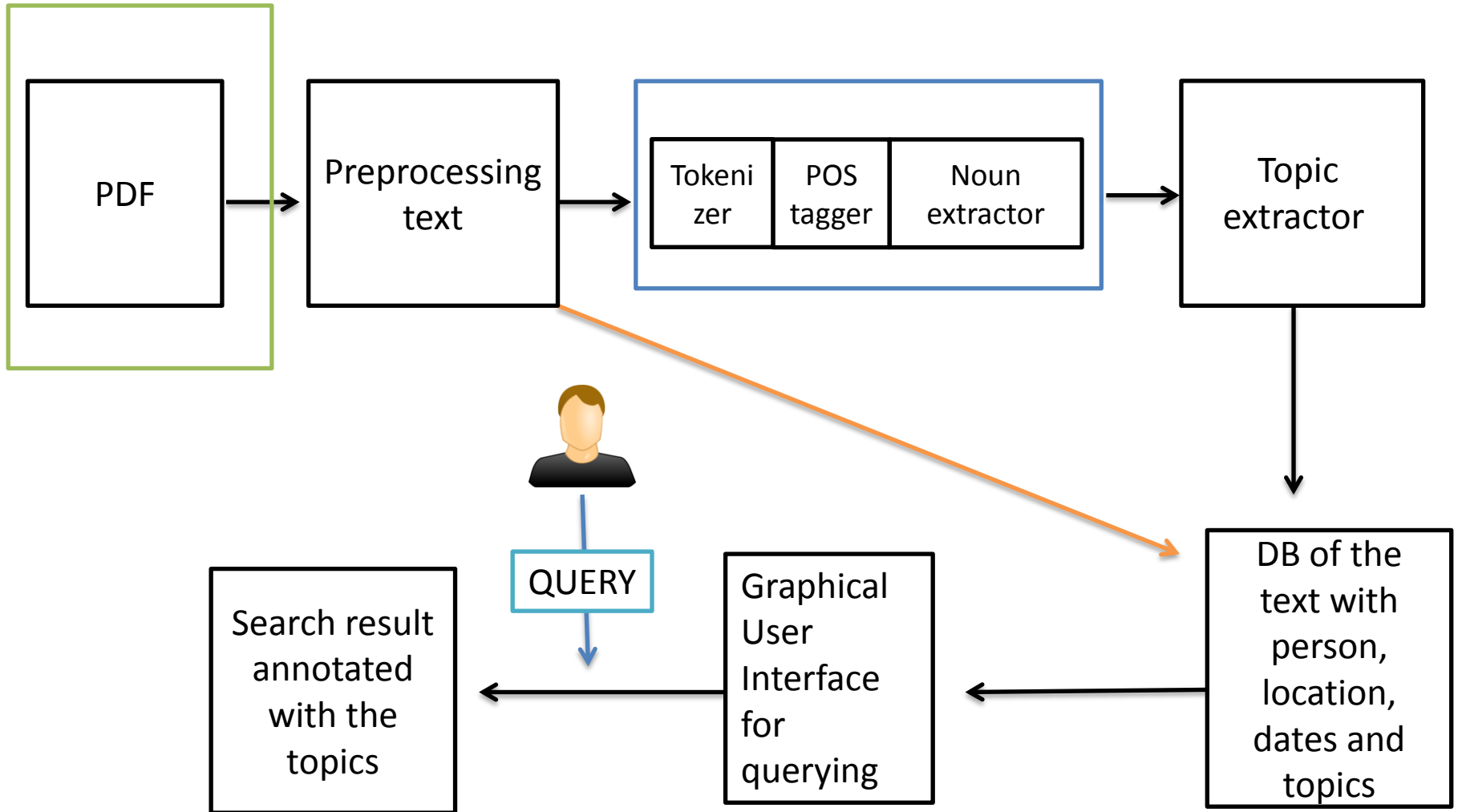
von  
Irmtraut Eder-Stein

- **Query:** Which records talked about financial matters involving St. Annual in the year 600 to 1300

# SaariStory

- Enabling keyword based search
- Answering complex queries
- Providing topic based search
- Showing temporal changes in the number of results for a query

# Workflow of SaariStory



# Description of the data

- Two parts of the data
  - **Data block:** Chronologically sorted records of events
  - **Index block:** index for all these data blocks with alphabetically sorted keywords

# Components of a data block

1275-03-04

timestamp

Simon [IV.], Graf zu Saarbrücken-Commercy, verbessert das Lehen seines Vasallen, des Ritters Friedrich von Kofel (*Ferrico dicto de Cofle*) von Saarbrücken, um jährlich 100 Metzger Schillinge, nämlich mit der Mühle zu Neumünster und Abgaben aus dem Forst zu Völklingen und dem Dekanat von *Harreswilre*, wofür dieser auf der Burg zu Saarbrücken wohnen und Burghut halten soll.

*MCCLXX quarto feria secunda post buras mense martii*<sup>22</sup>

S.: A.

1. HessHStAWi - Original (Jungk), hier nicht mehr ermittelt
2. LASb Best. Nassau-Saarbrücken II Nr. 1238 (ehem. HessHStAWi Abt. 121 Urkunden Cofle, um 1900/1910 nach Koblenz extradiert) - Abschrift (14. oder Anfang 15. Jh.), Papier

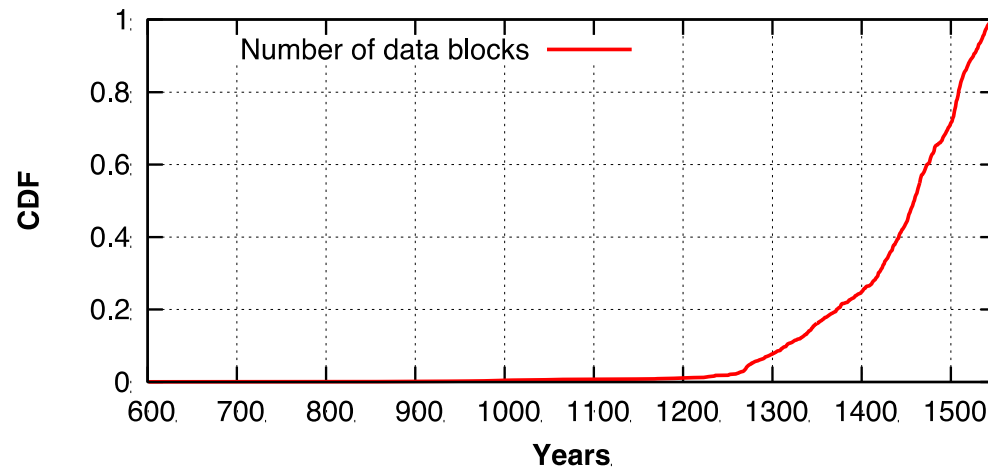
Druck: Kremer (1785) II S. 346 (*MCCLXX feria 2 post marci*); Schönemann I (1800) S. 208 (*MCCLXX feria 2 post marci*); Hardt, MRUB IV (2007), S. 845, Nr. 1246 (*MCCLXX, feria secunda post Marci*); Hardt MRUB V (2007), S. 260 Nr. 391 (mit richtiger Datierung nach Abschrift in Saarbrücken)

Regest: MRR III (1881) S. 565 Nr. 2504; Jungk (1914/19) S. 146 Nr. 501 (alles nach der fehlerhaften Transkription der Datierung bei Kremer und Schönemann, die von Hardt zunächst wiederholt wird); Jungk (1914/19) S. 160 Nr. 540 (richtig nach Vorlage ehem. in Wiesbaden) - (Ed)

data

# Characteristics of the data blocks

<b>Number of words</b>	<b>200,646</b>
Number of lines	15,021
Unique data blocks	1,490
Pages	612



# Components of a index block

*Arnheim, Jacob* von 1435-03-06

**Arnual**, Bf. von Metz (599-615) 0857

**Aschaffenburg** (*Oschoffenburg*), Stadt (By)

Einwohner

- *Peter Vox von Oschoffenburg*, Bürger zu Saarbrücken, oo *Madlen*

1529-05-31

keywords

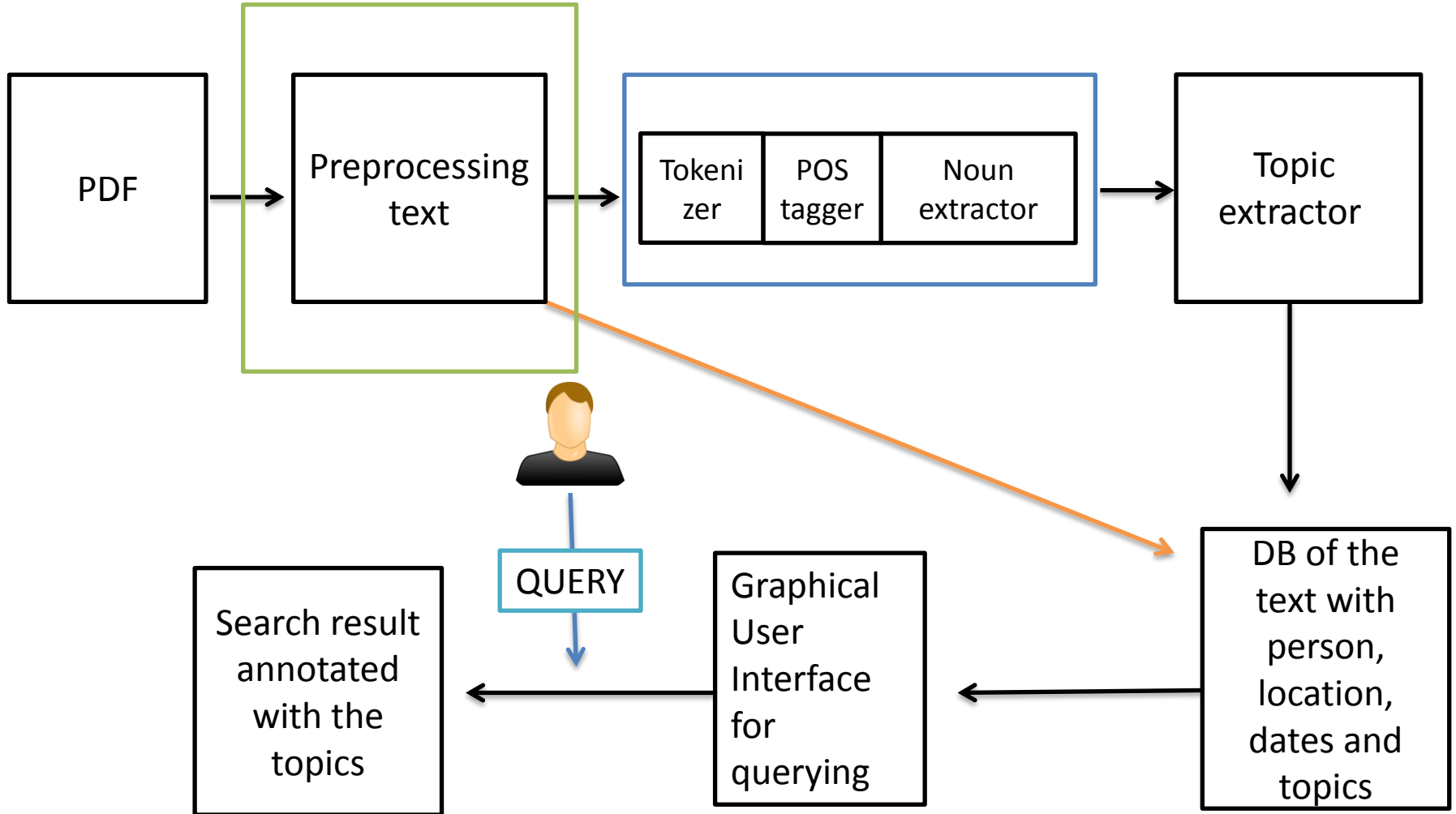
Dates connecting to index block



# Characteristics of the index blocks

<b>Number of words</b>	<b>86,485</b>
Number of lines	10,803
Unique index blocks	934
Pages	277

# SaariStory



# Preprocessing of the data block

- Basic strategy:
  - Convert pdf to text using nitro pdf
  - Parse text to separate the data blocks
- Problem: How to separate data blocks?
  - New lines do not indicate starting of data blocks
  - Distinguish between start of a page and start of a data block

# Preprocessing of the data block

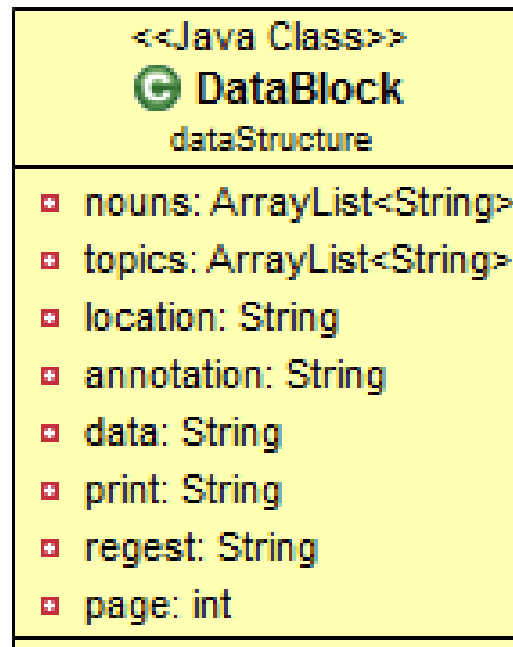
- Solution: Regular expression
  - Each data block starts with a date:
  - yyyy-mm-dd, yyyy-mm, yyyy/mm, yyyy

```
// is current line a title?  
else if (lastLineEmpty &&  
        (line.matches("\\d{4}-\\d{2}-\\d{2}.*" ) ||  
         line.matches("\\d{4}-\\d{2}.*" ) ||  
         line.matches("\\d{4}[-/]\\d{4}.*" ) ||  
         line.matches("\\d{4}.*" ))) {  
  
    blockCount++;  
  
    if (currentBlock != null) {  
        // extract regest and print  
        currentBlock.extractSourceInfo();  
    }  
}
```

- Use them to search “Regest” and “Druck” too

# Preprocessing of the data block

- Data structure to present the processed data



# Preprocessing of the index block

- Problem: How to separate index blocks?
  - The only way to separate them is to use the fact: titles are in bold text

*Arnheim, Jacob* von 1435-03-06

**Arnual**, Bf. von Metz (599-615) 0857

**Aschaffenburg** (*Oschoffenburg*), Stadt (By)




Einwohner

- *Peter Vox von Oschoffenburg*, Bürger zu Saarbrücken, oo *Madlen*

1529-05-31

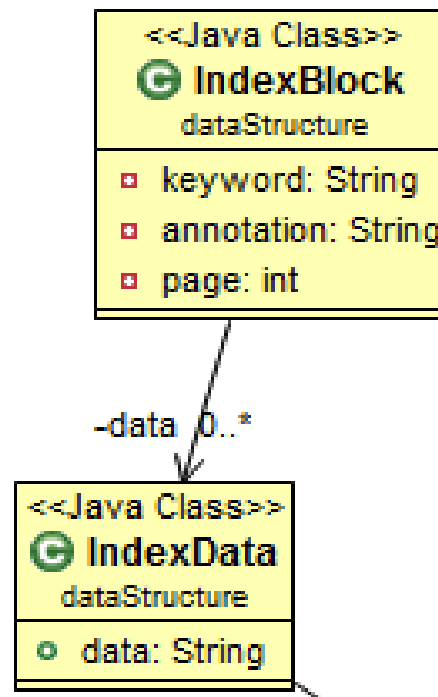
- The bold annotation is lost in pdf to text conversion

# Preprocessing of the index block

- Solution: pdf  doc  html
  - The bold annotations are preserved
  - Use regular expression to search for the `<b></b>` tag
  - Took care of broken lines for line breaks
    - 0830-05- <next line> 10  -0830-05-10

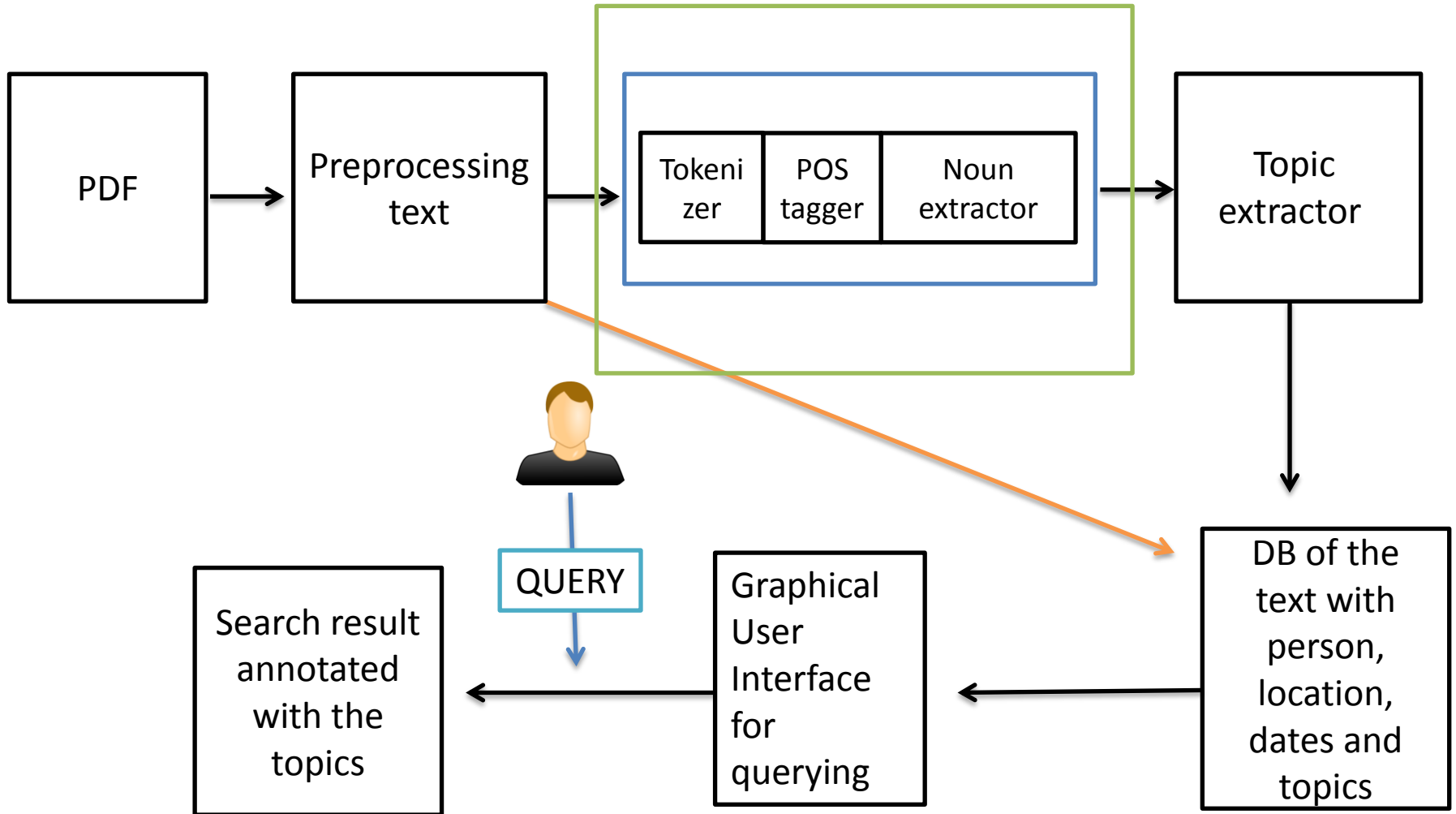
# Preprocessing of the index block

- Data structure to present the processed index data





# SaariStory



# Tokenizer and POS tagger

- Need to process data from each data block
- We use openNLP for this purpose
  - Open source
  - Easy to use
  - Pre-trained for German

## Noun extraction

- Straightforward to get from POS tagged text
- ~6000 nouns from 1490 data blocks
- Problem: 22 minutes for only 6000 nouns

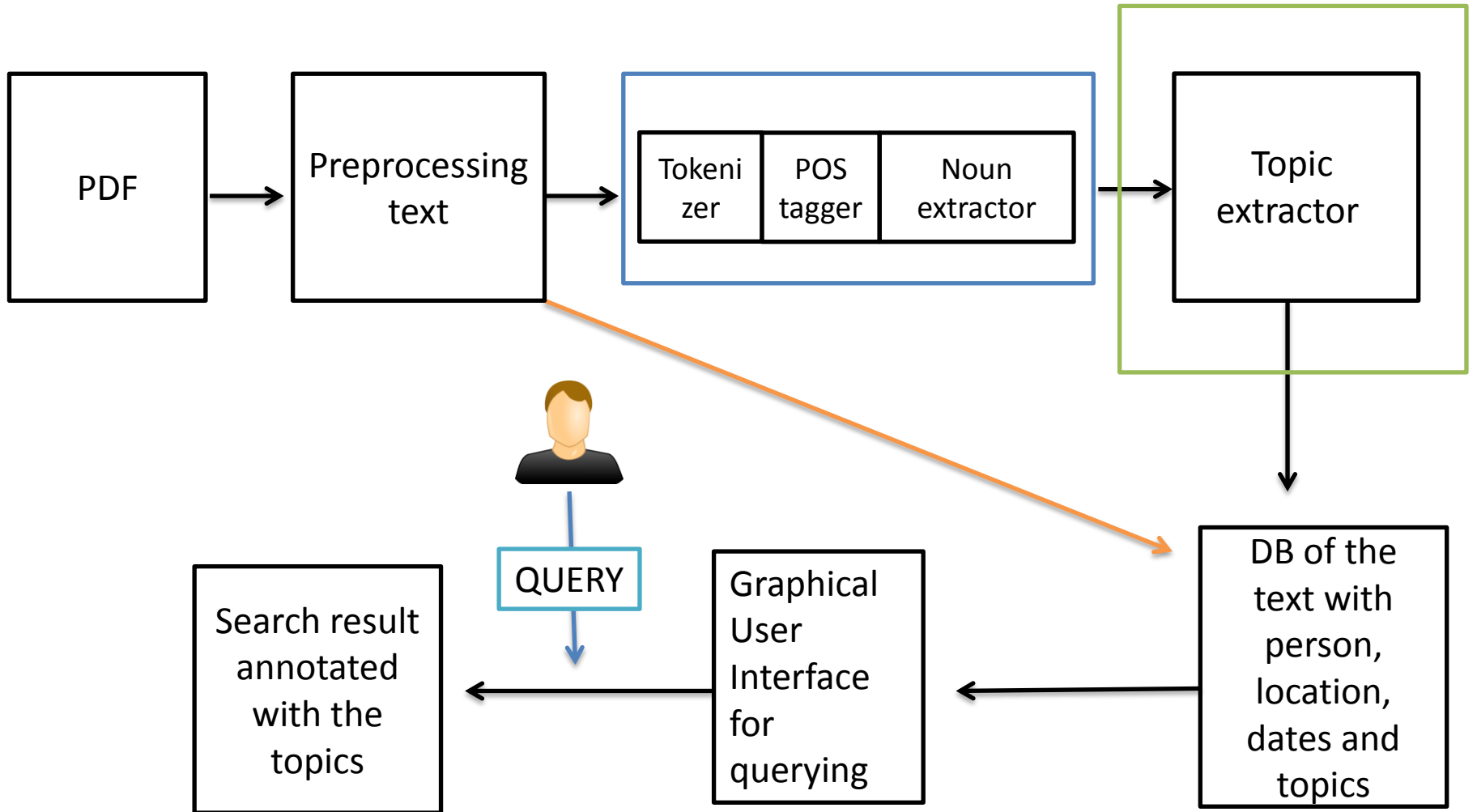
# Noun extraction

- Solution:
  - Run the POS tagger concurrently over data blocks
  - Feed only the tokens which *could be* nouns

Method	Time (seconds)	Speed up
Sequential	1322.6	-
Concurrent	12.4	100x
Concurrent with optimization	12.4	100x

we use the optimized version to have better quality of nouns

# SaariStory



# Topic extraction

- We use Latent Dirichlet allocation (LDA)
  - Proposed by Blei et al. in 2002
  - Assumption: Each document is a mixture of small number of topics
  - Needed to set two parameters: used trial and error to get meaningful topics

# Topic extraction

- Using LDA, try 1: Use blindly
  - Some very frequent not so meaningful words
    - E.g. Saarbrücken
  - They are in every topic and every data block
  - Every topic is assigned to every data block

# Topic extraction

- Using LDA, try 2: Remove some nouns
  - We remove the 15 most frequent nouns
    - Covered 24% of all nouns
  - Problem: some of these nouns *may be* relevant to some documents



# Topic extraction

- Using LDA, try 3: Only keep *essential* nouns
  - Calculate tf-idf score for word  $w$  in a data block  $d$

$tf_{w,d} = \text{\#of occurrences of } w \text{ in } d$

$$idf_w = \log_c \frac{|Data\_blocks|}{1 + |Data\_blocks\_having\_word\_w|}$$

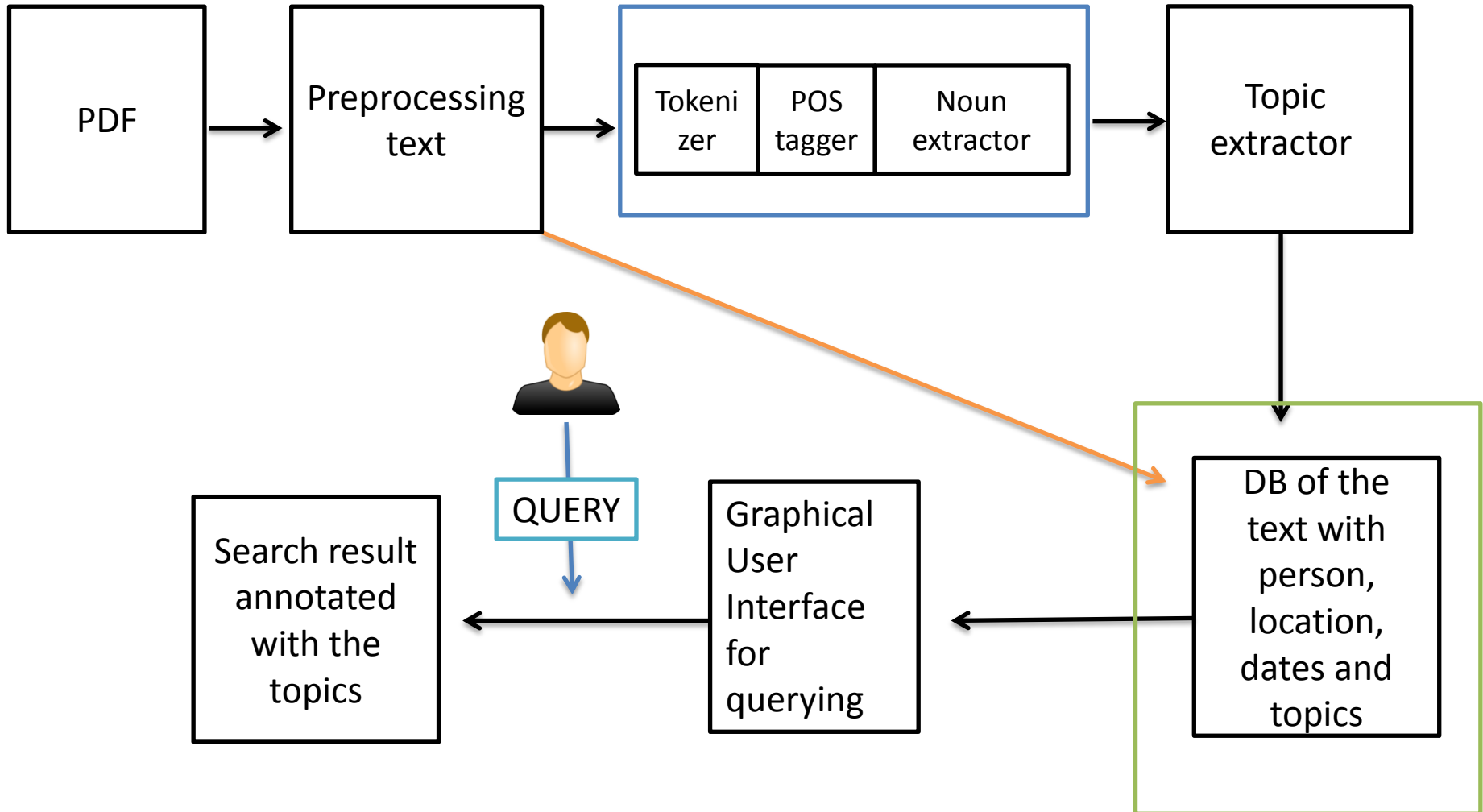
$$tf-idf_{w,d} = tf_{w,d} \times idf_w$$

- If  $tf-idf_{w,d} < 3.0$  we remove  $w$  from  $d$
- Use LDA on the modified data blocks

# Topic extraction

- We use LDA on the modified data blocks after using tf-idf score
- LDA identifies 7 meaningful topics with 10 words each
  - Bekanntmachung, Besitz, Finanzen, Vereinbarungen, Familie, Schulden, Recht

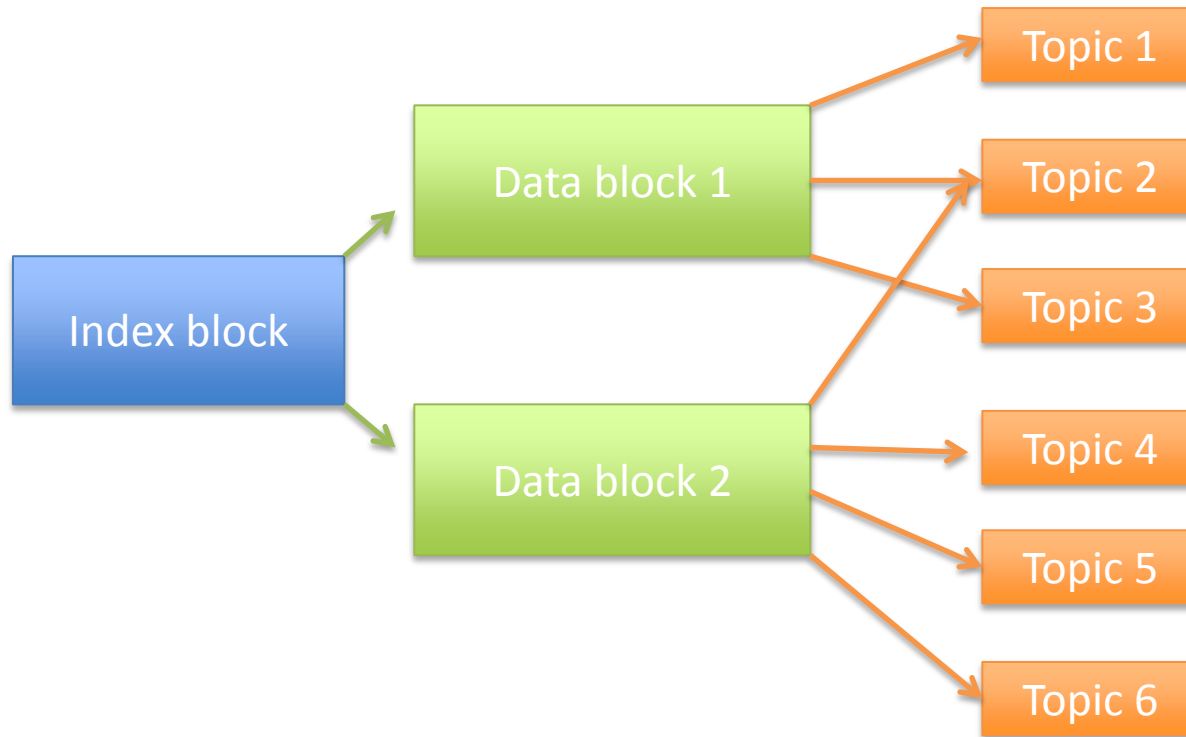
# SaariStory



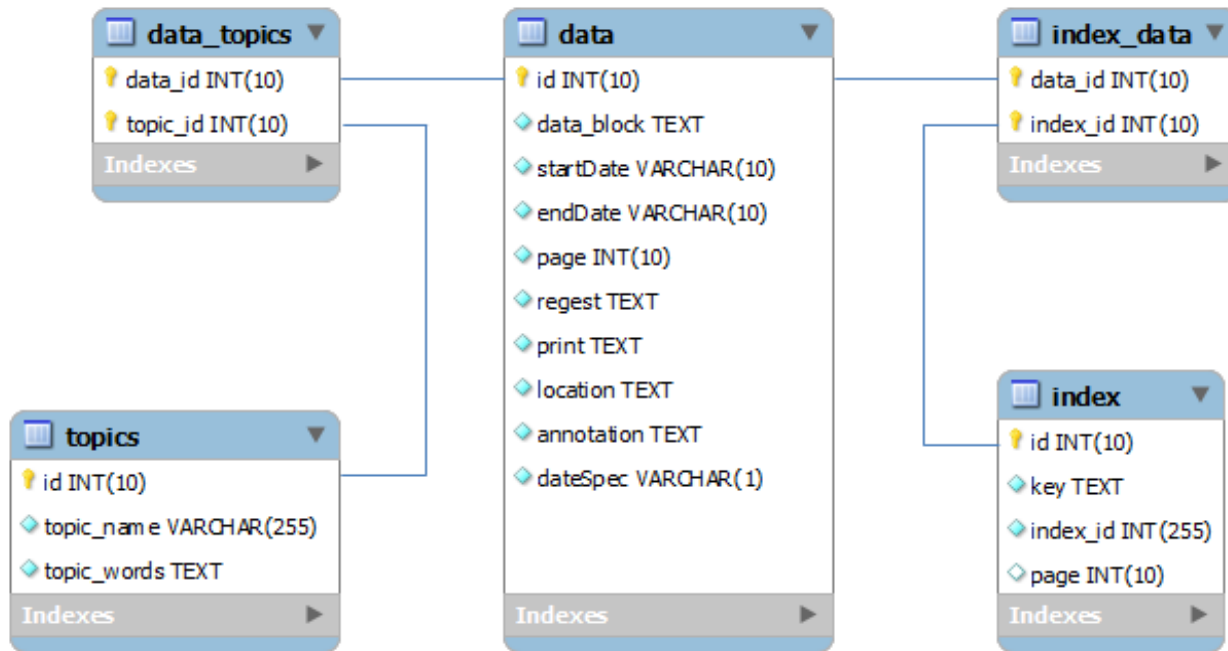
# Putting the data into database

- Type of features we want to have
  - Enabling keyword based search
  - Answering complex queries
  - Providing topic based search
  - Showing temporal changes in the number of results for a query

# Dependency within index and data block



# Database schema



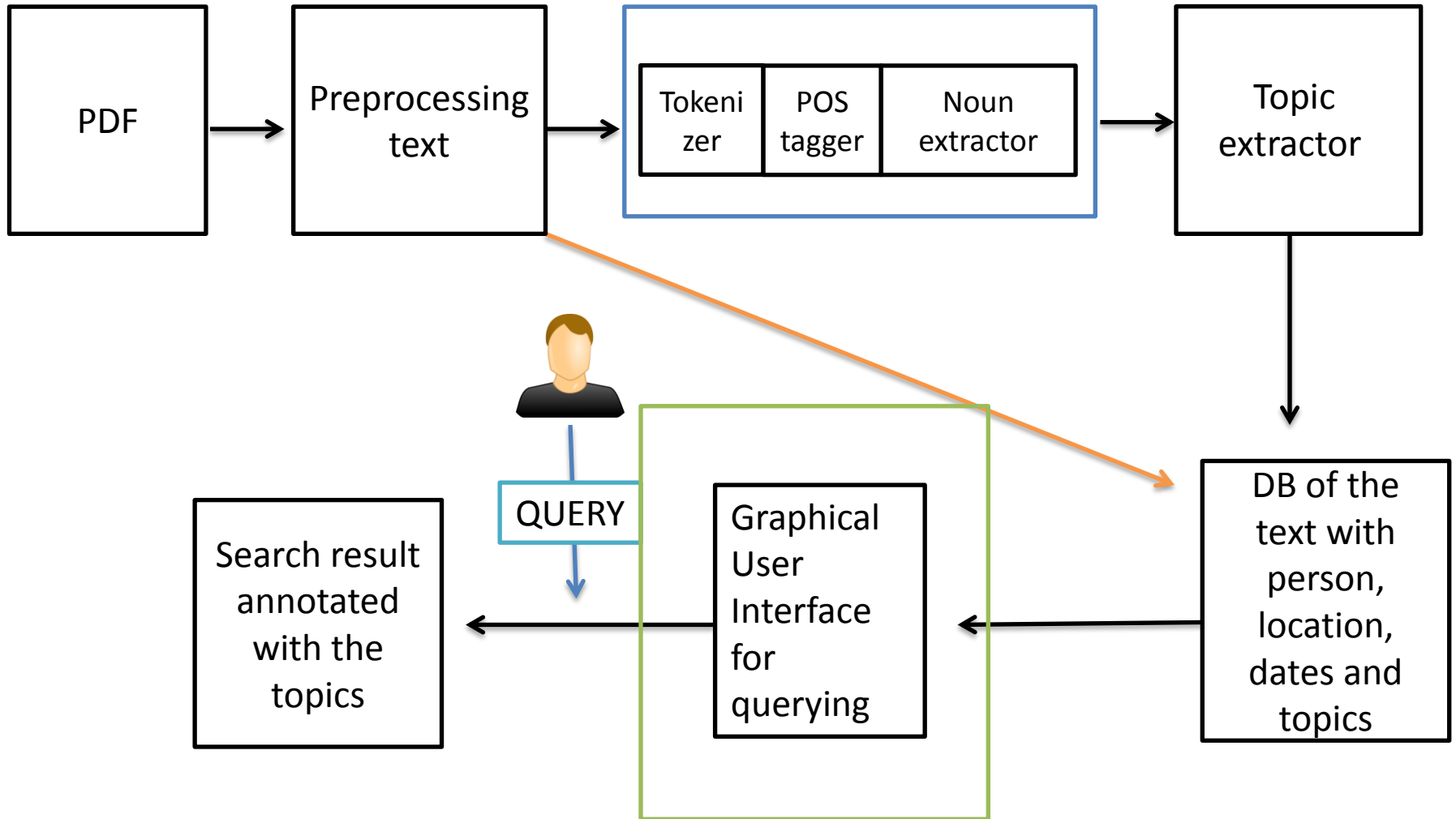
```
SELECT d.*, GROUP_CONCAT(DISTINCT t.topic_name) AS topic_names FROM (`data` AS d LEFT OUTER JOIN `data_topics` AS dt ON d.id = dt.data_id) LEFT OUTER JOIN `topics` AS t ON dt.topic_id = t.id WHERE d.startDate >= '0600-00-00' AND d.endDate <= '1600-00-00' AND dt.topic_id = 1 OR dt.topic_id = 6 AND d.data_block LIKE '%Keyword%' GROUP BY d.id ORDER BY d.startDate
```

# Database settings

- We use MySQL database
- First we used a web based sql provider
  - Not always live
  - Too slow in times of high load
- We set up a local database

Database	# Rows filled	Time to fill the DB (seconds)	Average time for query (seconds)
Remote	25,274	4,800	10
Local	25,527	300	3

# SaariStory





# Graphical user interface

Suche ✖

Suchbegriff

Suchort

Datenblöcke  Indizes

Filter

Ort der Veröffentlichung

Datum der Veröffentlichung

von

bis

Regesten

Druck

Themen

Bekanntmachung  Familie

Besitz  Schulden

Finanzen  Recht

Vereinbarungen

Suche starten

Information ✖

Optionen

Datenblock in pdf anzeigen

von 0960-00-03 bis 0960-00-03  
in Köln

König Otto I. bestätigt den Nonnen von St. Peter zu Metz ihre Besitzungen, darunter die Kirchen (aecclesias) von Malstatt (de Mathalstat) und anderen Orten, und verleiht ihnen das Recht der freien Wahl der Äbtissin und des Vogtes. 3 non. iun. S. : A. 1. AD Metz H 3904 - Ausf., Perg., S. ab, lat.; teilweise Abb. in: SaariStory Nr. 14 (1960) S. 299. BN-Bischof, lat. 4

SaariStory - □ ✖

Datei Datenbank Statistiken

ID	Seite	Von	Bis	Spez	Anmerkung	Veröffentlicht in	
13	19	0601-00-00	0609-00-00		(?)		Kg. ▲
2	19	0857-00-00	0857-00-00		(?)		Ave ☰
23	19	0935-00-00	1000-00-00		(ca.)		Auf
22	20	0960-06-03	0960-06-03			Köln	Kör
122	20	0977-05-11	0977-05-11			Diedenhofen	Kai
14	20	0993-03-26	0993-03-26			Lüttich	Kai
3	21	0999-04-14	0999-04-14			Rom	Kai
1	21	1009-00-00	1009-00-00				Kör
7	22	1046-05-25	1046-05-25			Aachen	Kör
9	22	1065-04-03	1065-04-03			Mainz	Kör
4	23	1065-08-28	1065-08-28		(nach)	Saarbrücken	Ada
5	23	1147-06-22	1147-06-22		(um)		Als
6	23	1168-00-00	1168-00-00				Die
10	24	1171-09-04	1171-09-04			Aachen	Kai
8	24	1190-12-20	1190-12-20				Ebe
11	24	1200-00-00	1200-00-00	a	(um) (a)		Hin
12	24	1200-00-00	1200-00-00	b	(um) (b)		JoH
15	25	1212-00-00	1212-00-00				Hei
16	25	1223-09-18	1223-09-18			Saarbrücken	Ma
51	25	1227-00-00	1227-00-00				Sin
18	27	1227-05-31	1227-05-31				JoH
17	27	1228-00-00	1228-00-00				Hei
19	27	1230-01-25	1230-01-25			Perugia	Pa
20	28	1234-03-16	1234-03-16			Kaiserslautern	Sin
21	28	1235-04-00	1235-04-00				JoH
27	28	1236-03-03	1236-03-03				JoH
24	29	1236-06-05	1236-06-05		(vor)		Ge
25	29	1250-12-07	1250-12-07			Saarbrücken	Grä
30	30	1251-00-00	1251-00-00				Grä
28	30	1252-00-00	1252-00-00				JoH
32	30	1252-12-13	1252-12-13			Saarbrücken	Phi ▼

1490 (100,00 %) Ergebnisse aus insgesamt 1490 Datenblöcken

# Graphical user interface

**Suche** ✕

Suchbegriff

---

Suchort

Datenblöcke     Indizes

Filter

Ort der Veröffentlichung

Datum der Veröffentlichung

von: 1400

bis: 1600

Regesten: alle

Druck

Themen

Bekanntmachung     Familie

Besitz     Schulden

Finanzen     Recht

Vereinbarungen

Suche starten

---

**Information** ✕

Optionen

zabenernach Gedanken bawen. Die Schuldner müssen die Baukosten bei der Einlösung zusätzlich zahlen. Die Auslösung der Pfandschaft ist ein Viertel Jahr zuvor anzukündigen. Können die Schuldner dann die 100 Gulden nicht zurückzahlen, so kann das Kloster das Haus weiter verpfänden und versetzen nach Stadtrecht. Uf Sameſtag nach Unser Frowent ag Assumpcionis 1400 S.: A. und Johann, Dechant von St. Arnal AD Metz H 3897 - Ausf., Perg., alle S. ab; RV: Phandbriefhaus zu Sand Johan. N. (Anfang 16. Jh.); Unterschiedliche Brief von Sarbrucken und St. Johan (17. Jh.)

SaariStory

Datei	Datenbank	Statistiken				
ID	Seite	Von	Bis	Spez	Anmerkung	Veröffentlicht in
376	160	1400-06-29	1400-06-29			Saarbrücken
369	160	1400-08-21	1400-08-21			Saarbrücken
366	162	1401-05-24	1401-05-24			Saarbrücken

**Themen pro Zeit**

424	184	1416-12-24	1416-12-24			Saarbrücken
446	187	1419-02-09	1419-02-09			Saarbrücken
473	189	1419-04-01	1419-04-01			Saarbrücken

422 (28,32 %) Ergebnisse aus insgesamt 1490 Datenblöcken

# Evaluation

- Precision and recall is 100% for keyword based random queries
  - We compared manual results with results from our system
- The topic detection works! The following text is labeled as “Familie” and “Finanzen”

Gerbod verzichtet mit seiner Frau und seinen Kindern nach erhaltenen 40 Metzer Schillinge auf seinen Anteil an der Wetzelsmühle zu Gunsten des Klosters Fraulautern. Zeugenliste. Zur besseren Bekräftigung hat Gerbod noch in Saarbrücken (*Sareponti*) vor den Kirchtüren<sup>15</sup> unter Berührung der Reliquien einen Eid geschworen, wobei der Propst von Fraulautern, Rorich d.J., Ludwig vom Berg, Petrus von Bebelsheim und Liebwin von Malstatt zugegen waren.

## Future work

- Improving our pre-processing step to include more corner cases
- Extract and save footnotes from the text
- Possibility to add more data blocks to the database
- Taking care of line breaks in data blocks
- Keyword based search in complete index block

# SaariStory

