

# Combining Concept Based and Text Based Indexes for CLIR

**CLEF09: Ad-hoc (TEL) Session, Corfu, Greece**

Institute AIFB – University of Karlsruhe

**Philipp Sorg**

Institute AIFB, Universität Karlsruhe  
sorg@kit.edu

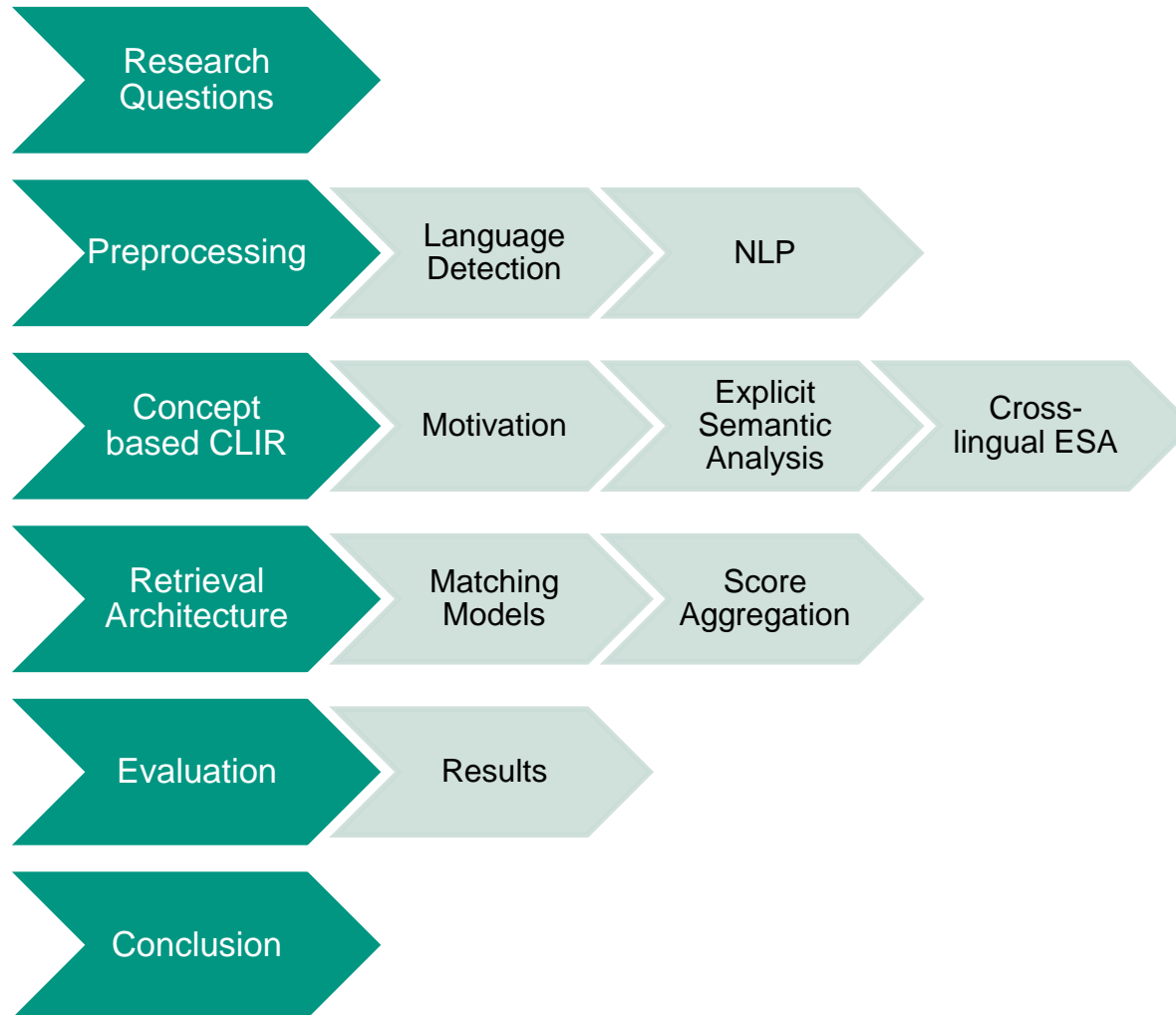
**Philipp Cimiano**

Web Information Systems Group, Delft University of Technology  
p.cimiano@tudelft.nl

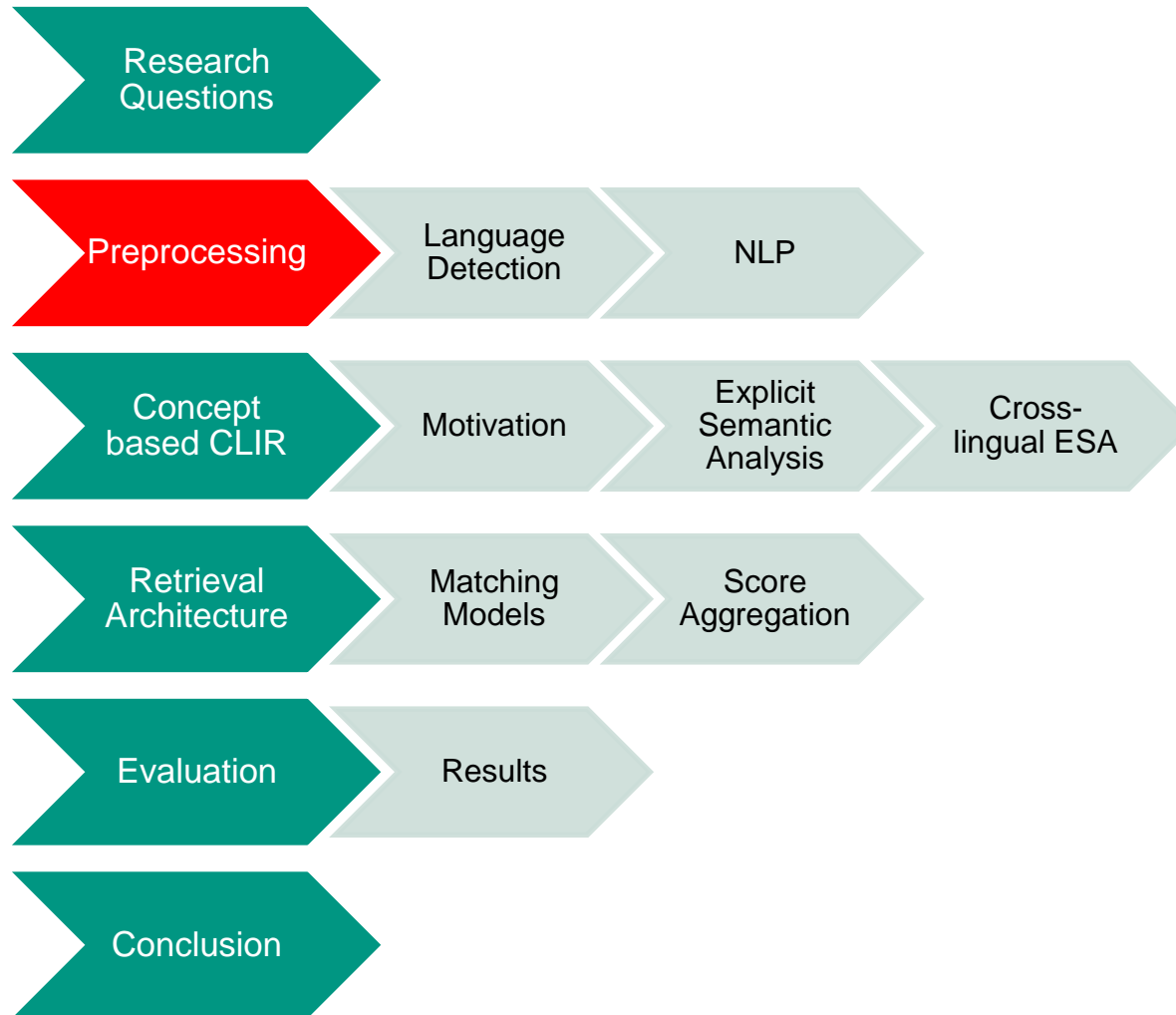
# Research Questions

- Can multi-lingual information be used to improve retrieval on the TEL dataset?
  - Queries in different languages
  - Documents in different languages
  - Fields of documents in different languages
  
- Can text based (= Machine Translation based) retrieval be combined with concept based retrieval?
  - Representation of documents in concept space
    - Explicit Semantic Analysis (ESA)
  - Score aggregation problem

# Agenda



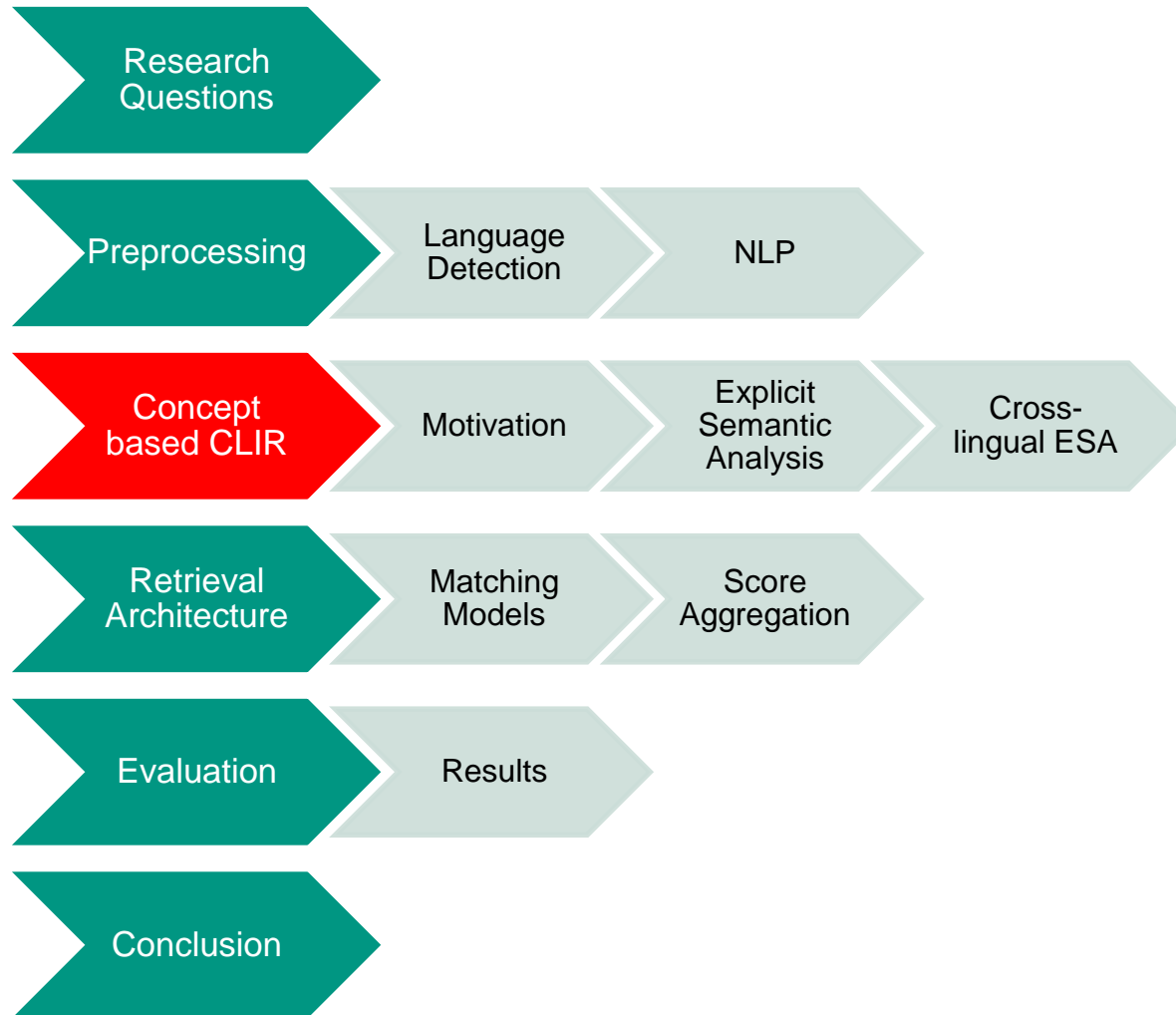
# Agenda



# Preprocessing of Dataset

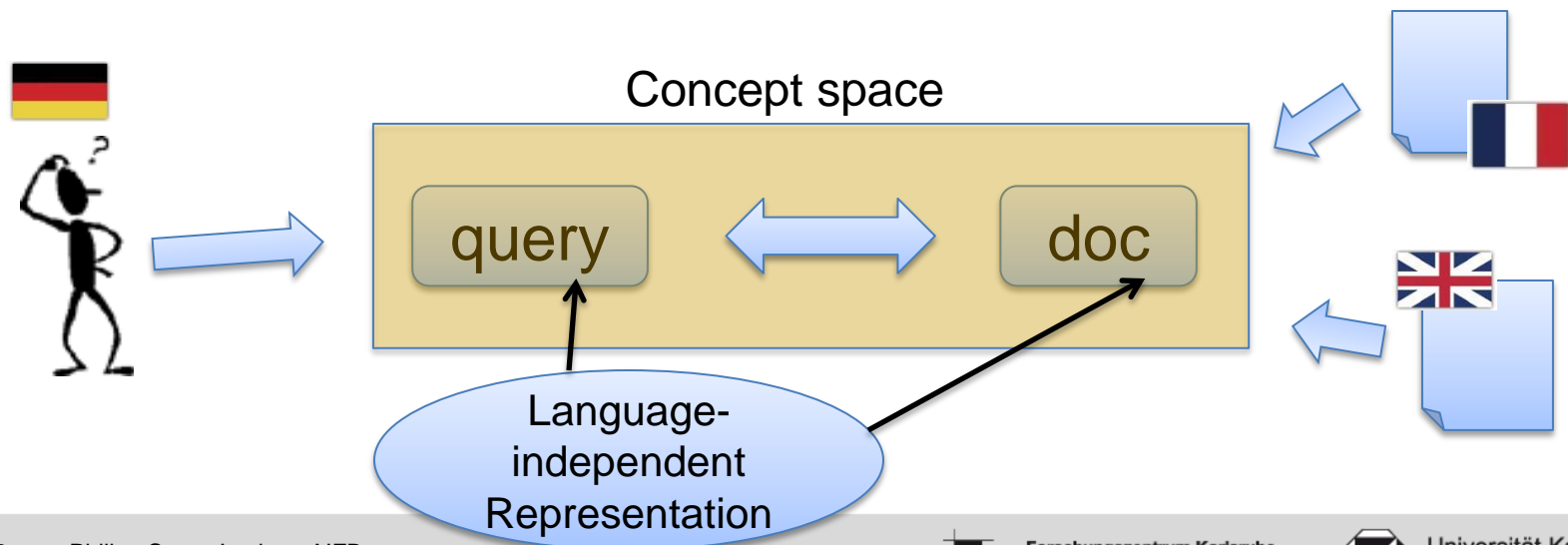
- Selection of content fields
  - Title, subject, alternative, abstract
- Language Detection
  - Character n-gram model for language detection
    - Ling Pipe Identification Tool
  - Each field is classified
    - Based on language tag and language detection
    - Results in documents with multi-lingual fields
- NLP
  - Stemming in all languages supported by Snowball stemmer
  - Language specific stopword removal

# Agenda



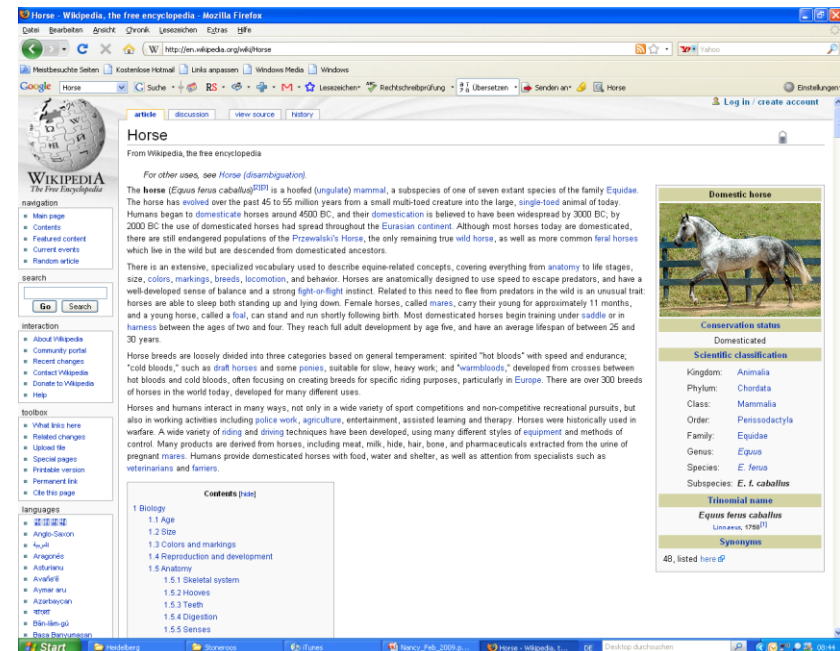
# Motivation of Concept Based CLIR

- Traditional approach to Multi-lingual IA
  - Translation of queries or documents
  - Problems
    - MT is not available for many language pairs
    - Propagation of error, inherits all problems of mono-lingual retrieval
  
- Alternative approach:



# Explicit Concept Model

- Idea: Use Web 2.0 resources to define concepts
  - Pragmatic definition of concepts
    - Wikipedia articles, tagged web sites, products, ...
  - Cover a broad range of topics and languages
  - Freely available
- Example
  - Wikipedia articles as concepts
- We use Explicit Semantic Analysis (Cross-lingual ESA)
  - Gabrilovich and Markovitch IJCAI 2007
  - Potthast et al. ECIR 2008, Sorg and Cimiano CLEF 2008

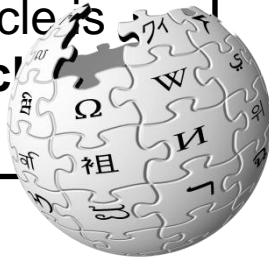




# Idea of ESA

## Bicycle

A **bicycle**, **bike**, or **cycle** is a pedal-driven, human-powered vehicle with two wheels attached to a frame, one behind the other. A person who rides a bicycle is called a **cyclist** or **bicyclist**.



WIKIPEDIA  
Die freie Enzyklopädie

TF.IDF Function

“The transport of bicycles on trains”

1.52	<Road_bicycle>
1.18	<Bicycle>
1.12	<Velorama>
0.92	<Cycling>
0.92	<Biker>
0.92	<Bianchi_(bicycle_manufacturer)>
0.79	<Train_(disambiguation)>
0.77	<Transport>
...	...

# Example Cross-lingual ESA Concept Vector

“The transport of bicycles on trains”



- 1.52
- 1.18
- 1.12
- 0.92
- 0.92
- 0.92
- 0.79
- 0.77
- ...

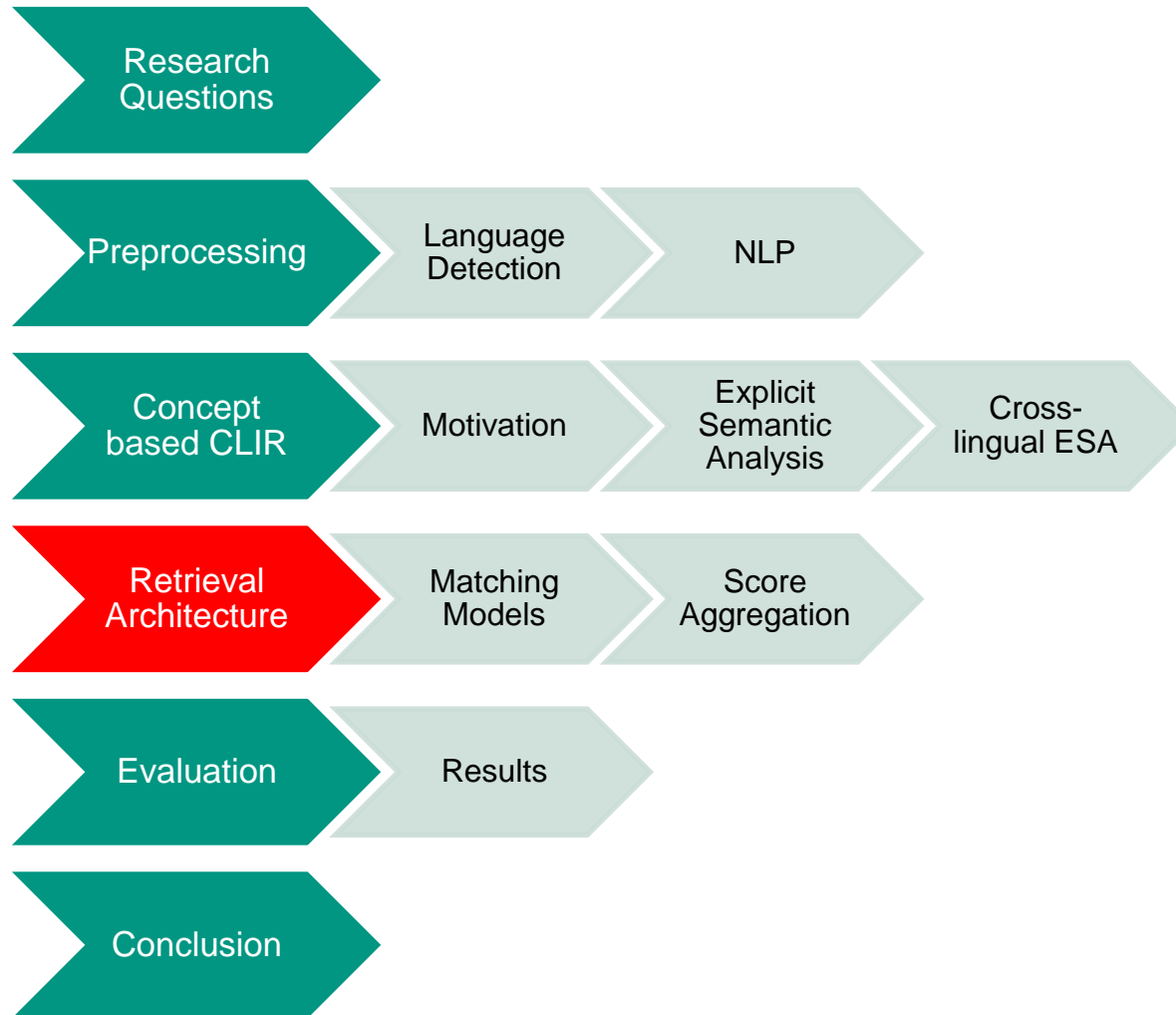
A1	<Road_bicycle>
A2	<Bicycle>
A3	<Velorama>
A4	<Cycling>
A5	<Biker>
A6	<Bianchi_(bicycle_manufacturer)>
A7	<Train_(disambiguation)>
A8	<Transport>
...	...

English interpretation

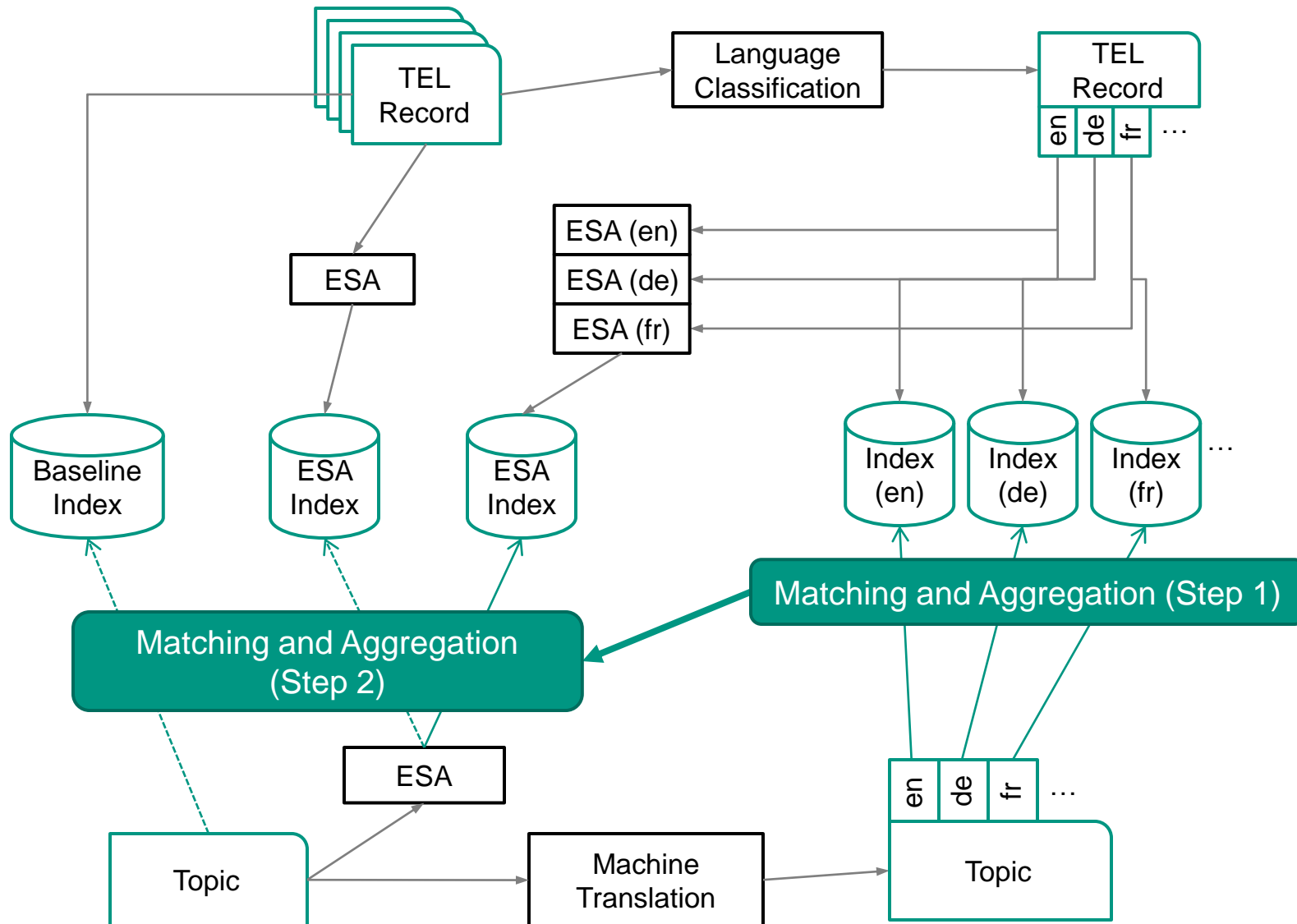
<Radrennen>
<Fahrrad>
<Velorama>
<Fahrradfahren>
<Biker>
<Bianchi_(Unternehmen)>
<Train>
<Verkehr>
...

German interpretation

# Agenda



# Retrieval Architecture



Indexing

Search

# Matching and Aggregation (Step 1)

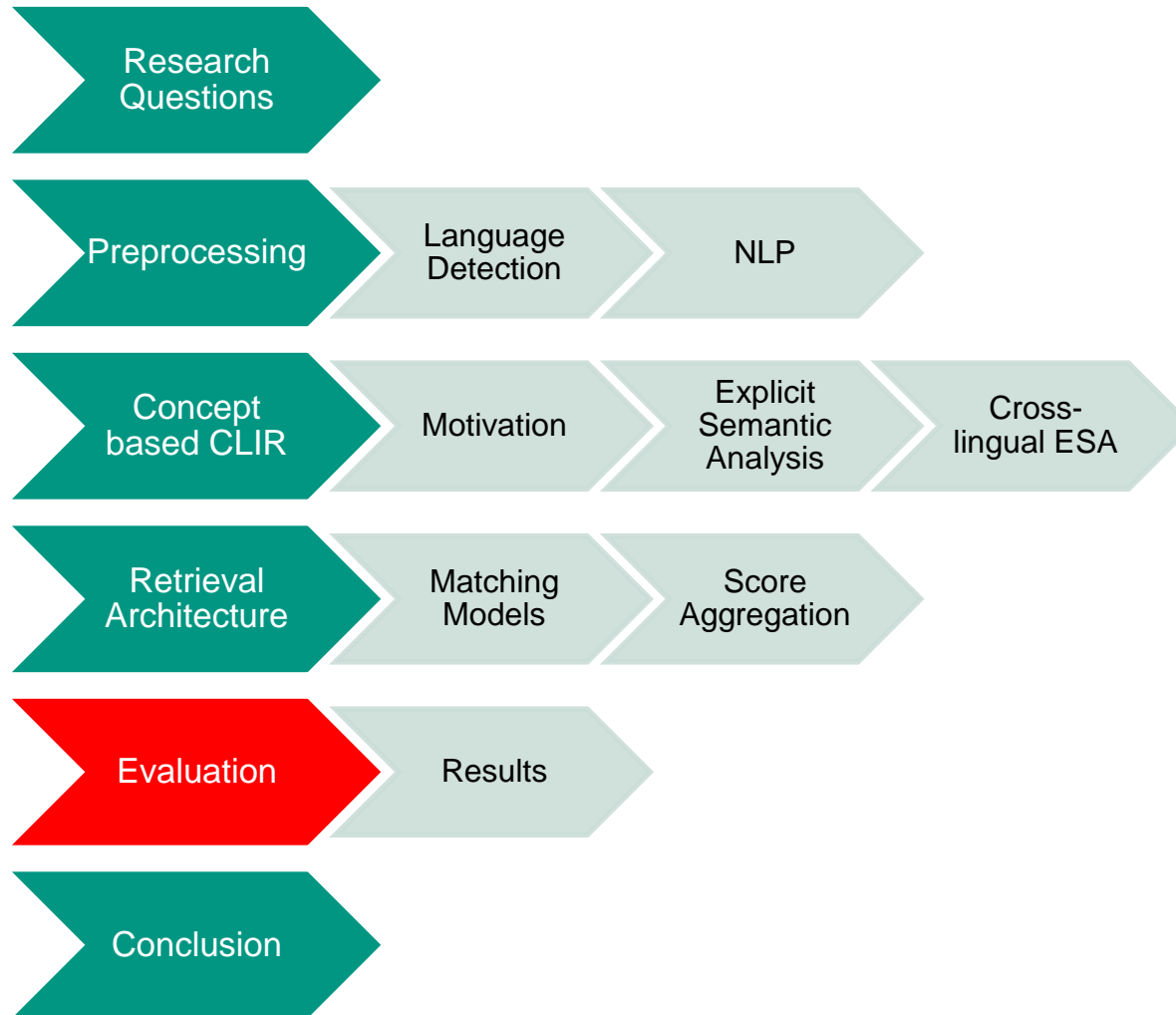
- Optimization of matching model
  - Using CLEF2008 topics and relevance assessments
  - Models provided by the Terrier framework
    - BL: DLH13, ONB: LemurTF\_IDF, BNF: BB2
- Linear aggregation of scores
  - Each document has a score for each index (=language)
  - Different normalization functions
    - Based on maximal score in each ranking
    - Based on the number of retrieved documents of each ranking
    - Based on a priori weights
      - Language distribution of text in corpus

$$score(t, d) := \sum_{r \in R} \delta(r) score_r(t, d)$$

## Matching and Aggregation (Step 2)

- ESA retrieval using cosine similarity
  - Implementation based on inverted concept index
- Linear aggregation of concept based scores and text based scores
  - Using the aggregated score from text based retrieval (Step 1)
  - Weight factor to modify influence of concept based retrieval
    - Optimized on CLEF2008 topics
- Evaluation measures
  - MAP: Mean Average Precision
  - P@10: Precision at cutoff level of 10 documents

# Agenda



# Evaluation

Topic lang.	Retrieval Method	BL		ONB		BNF	
		MAP	P@10	MAP	P@10	MAP	P@10
En	Baseline (single index)	35	51	16	26	25	39
	Multiple Indexes	33	50	15	24	22	34
	Concept + Baseline	35	52	<b>17</b>	27	25	39
De	Baseline (single index)	33	49	23	35	24	35
	Multiple Indexes	31	48	23	34	22	32
	Concept + Baseline	33	49	<b>24</b>	35	24	36
Fr	Baseline (single index)	31	48	15	22	27	38
	Multiple Indexes	29	45	14	20	25	35
	Concept + Baseline	32	<b>51</b>	15	22	27	37



# Evaluation

Topic lang.	Retrieval Method	BL		ONB		BNF	
		MAP	P@10	MAP	P@10	MAP	P@10
En	Baseline (single index)	<b>35</b>	51	16	26	25	39
	Multiple Indexes	33	50	15	24	22	34
	Concept + Baseline	35	52	<b>17</b>	27	25	39
De	Baseline (single index)	33	49	<b>23</b>	35	24	35
	Multiple Indexes	31	48	23	34	22	32
	Concept + Baseline	33	49	<b>24</b>	35	24	36
Fr	Baseline (single index)	31	48	15	22	<b>27</b>	38
	Multiple Indexes	29	45	14	20	25	35
	Concept + Baseline	32	<b>51</b>	15	22	27	37

# Evaluation

Topic lang.	Retrieval Method	BL		ONB		BNF	
		MAP	P@10	MAP	P@10	MAP	P@10
En	Baseline (single index)	35	51	16	26	25	39
	Multiple Indexes	<b>33</b>	50	<b>15</b>	24	<b>22</b>	34
	Concept + Baseline	35	52	<b>17</b>	27	25	39
De	Baseline (single index)	33	49	23	35	24	35
	Multiple Indexes	<b>31</b>	48	<b>23</b>	34	<b>22</b>	32
	Concept + Baseline	33	49	<b>24</b>	35	24	36
Fr	Baseline (single index)	31	48	15	22	27	38
	Multiple Indexes	<b>29</b>	45	<b>14</b>	20	<b>25</b>	35
	Concept + Baseline	32	<b>51</b>	15	22	27	37

# Evaluation

Topic lang.	Retrieval Method	BL		ONB		BNF	
		MAP	P@10	MAP	P@10	MAP	P@10
En	Baseline (single index)	35	51	16	26	25	39
	Multiple Indexes	33	50	15	24	22	34
	Concept + Baseline	<b>35</b>	52	<b>17</b>	27	<b>25</b>	39
De	Baseline (single index)	33	49	23	35	24	35
	Multiple Indexes	31	48	23	34	22	32
	Concept + Baseline	<b>33</b>	49	<b>24</b>	35	<b>24</b>	36
Fr	Baseline (single index)	31	48	15	22	27	38
	Multiple Indexes	29	45	14	20	25	35
	Concept + Baseline	<b>32</b>	<b>51</b>	<b>15</b>	22	<b>27</b>	37

# Conclusion

- Baseline is very strong
- Can multi-lingual information be used to improve retrieval on the TEL dataset?
  - Use of multi-lingual indexes based on language detection did not improve retrieval
    - Problem of score aggregation
    - Linear aggregation model with (simple) normalization is not working
- Can text based (= Machine Translation based) retrieval be combined with concept based retrieval?
  - Combination of concept and text based indexes yields only small improvements
    - We could not reconstruct the large improvements reported on mono-lingual collections
    - Not enough context in short TEL records for concept mapping?

# Thank you!

## ■ Questions?

### ■ Joint work with

- Philipp Cimiano (Universität Bielefeld)
- Marlon Braun, David Nicolay (Universität Karlsruhe)

### ■ Acknowledgments

- Multipla Project  
DFG grant 38457858

