Semantic URL Analytics to Support Efficient Annotation of Large Scale Web Archives

Tarcísio Souza[1], Elena Demidova[1], Thomas Risse[1], Helge Holzmann[1], Gerhard Gossen[1] and Julian Szymanski[2]

L3S Research Center, Hannover, Germany[1]

Gdansk University of Technology, Poland[2]

1st International **KEYSTONE** Conference

8-9 September 2015

Coimbra-Portugal

Introduction and motivation

Web Archives

- Large data
- Important source for communication and media history and within historiography in general
- Existing web archives are very difficult to use

URL level analysis

URL	Entities
http://www.wg-gesucht.de:80/wohnungen-in-**Berlin**-**Prenzlauer**-**Berg**.1529789.html	Berlin, Prenzlauer Berg

Related Work

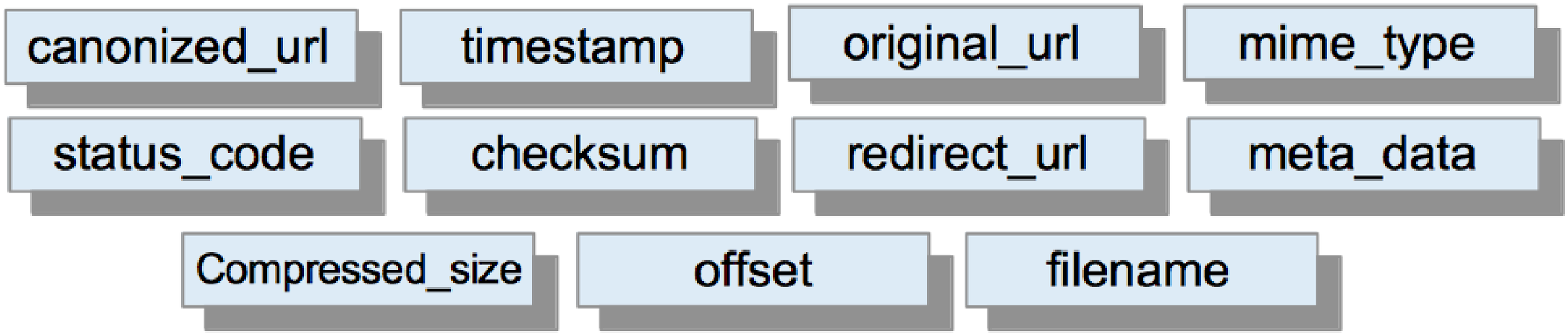
- Classification of a web document
  - Baykan et al. detect the topic of a Web document.
    - Precision around 0.86 and a recall between 0.36 and 0.4
- Special applications of URL classification
  - Detection of the document language (Baykan et al., 2013)
  - Genre classification (Myriam Abramson et al., 2012)
  - Locational relevance (Anastacio et al., 2009)
  - Detect malicious content (Peilin Zhao and Steven C.H. Hoi, 2013)
  - Online advertising (Santosh Raju and Raghavendra Udupa, 2012)

The Popular German Web: a dataset description

Dataset description

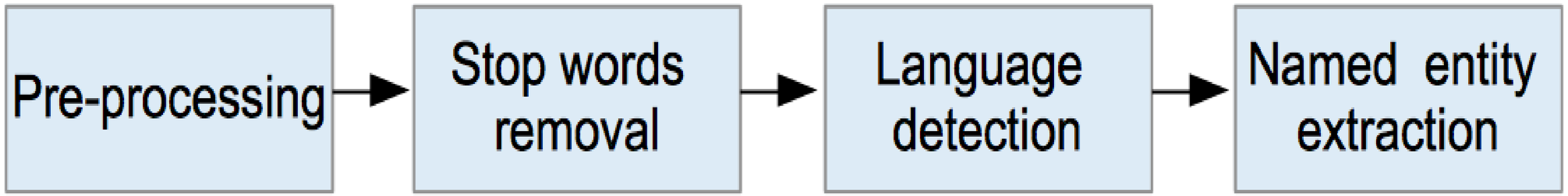
Provided in the context of ALEXANDRIA project

- We generated a subset named Popular German Web
- The subset contains 17 categories from 2000 to 2012 according to Alexa ranking
- URL (uniform resource locator) and captures stored as CDX files.



Dataset cleaning and pre-processing

- Focus on the captures of URLs with .htm and .html extensions
- Discard all captures of the URLs that never returned a successful status code (starting with ``2'').
- URL Tokenization



Dataset statistics

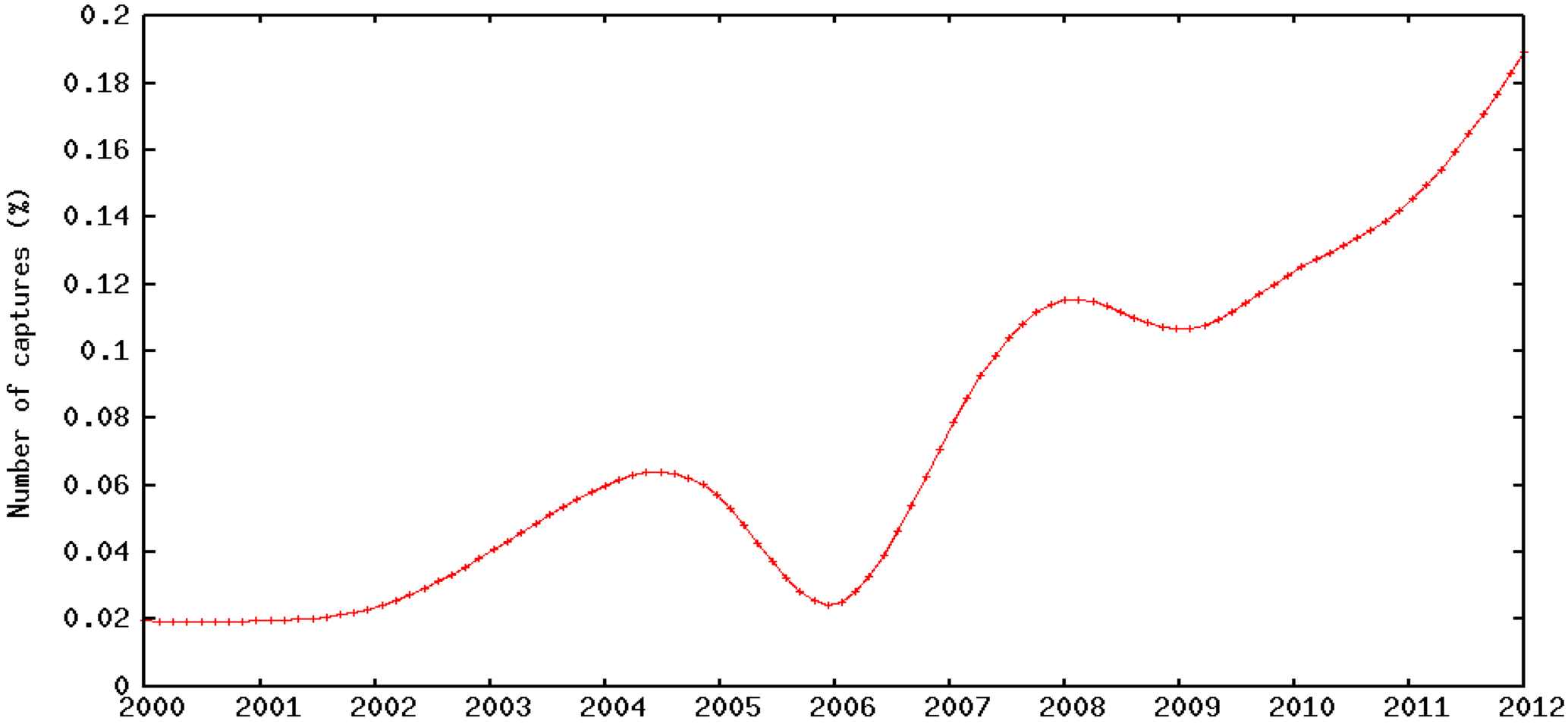
Category	# Domains	# Captures	Entities(%)
Education	100	12,406,130	73.36
Regional	100	34,204,862	44.79
Sports	100	17,358,130	39.33
Business	100	25,457,639	36.39
Recreation	100	8,260,029	30.95
Media	100	11,277,003	28.20
Universities	100	14,299,856	25.09
News	40	41,710,500	23.13
Shopping	100	33,045,310	20.14
Culture	100	6,822,986	19.69
Society	100	9,968,534	18.37
Games	99	13,518,500	16.40
Computer	100	26,298,534	15.90
Home	100	45,488,255	14.07
Kids & Teens	10	1,682,848	10.45
Health	100	6,260,340	9.31
Science	100	13,651,913	7.86
TOTAL	***1444***	***321,711,369***	

Education dominant domains:
wer-weiss-was.de
stayfriends.de

Temporal dimension

Most frequent domains

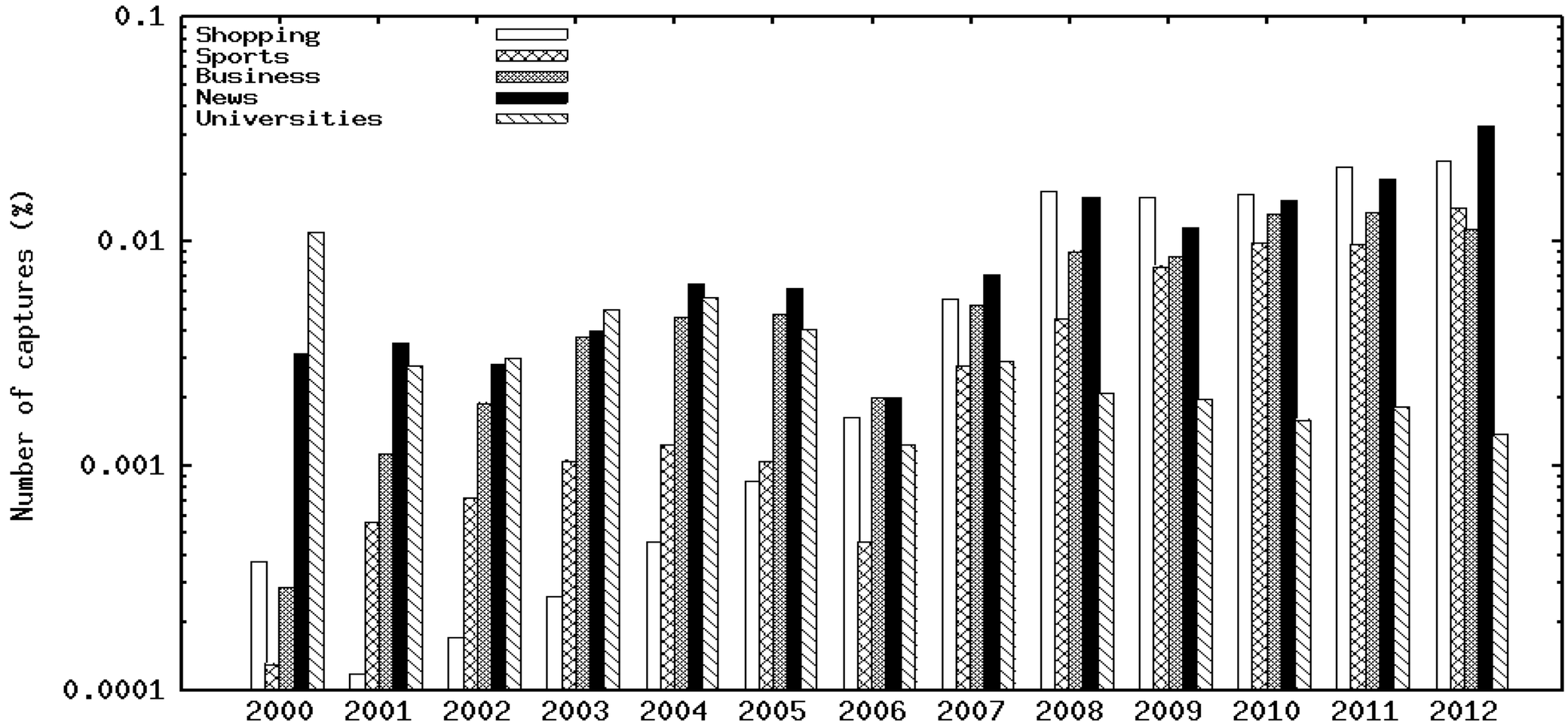
- spiegel.de (2001-2012): 7.72%
- tu-berlin (2000): 42%



Captures within selected domain categories

Majority of captures

- 2002-2003: universities domains (140) and news (40)
- 2008-2011: shopping (532) and news (136)



URL analytics

Language detection statistics

- State-of-the-art techniques to language detection using n-grams
- URL Splitting and removal of URL-specific stop words to increase precision
- 52.89% are in German 27.96% in English and 19.14% in other languages.
- 89% of precision for language detection after filtering steps

Precision of NER for URLs

- Named entity recognition
  - State-of-the-art named entity recognition are language dependent
  - Restriction to German and English (cover more than 80% of URLs in our subset)
  - Manually evaluation of a random sample of 100 URLs
- Initially: 60% for German; 56% for English
- Post-filtering steps
  - Removal of the entities with long labels (more than 2 terms)
  - Removal of entities that rarely occur in the archive (less than 3)
- Increased to 85% for German; 82% for English

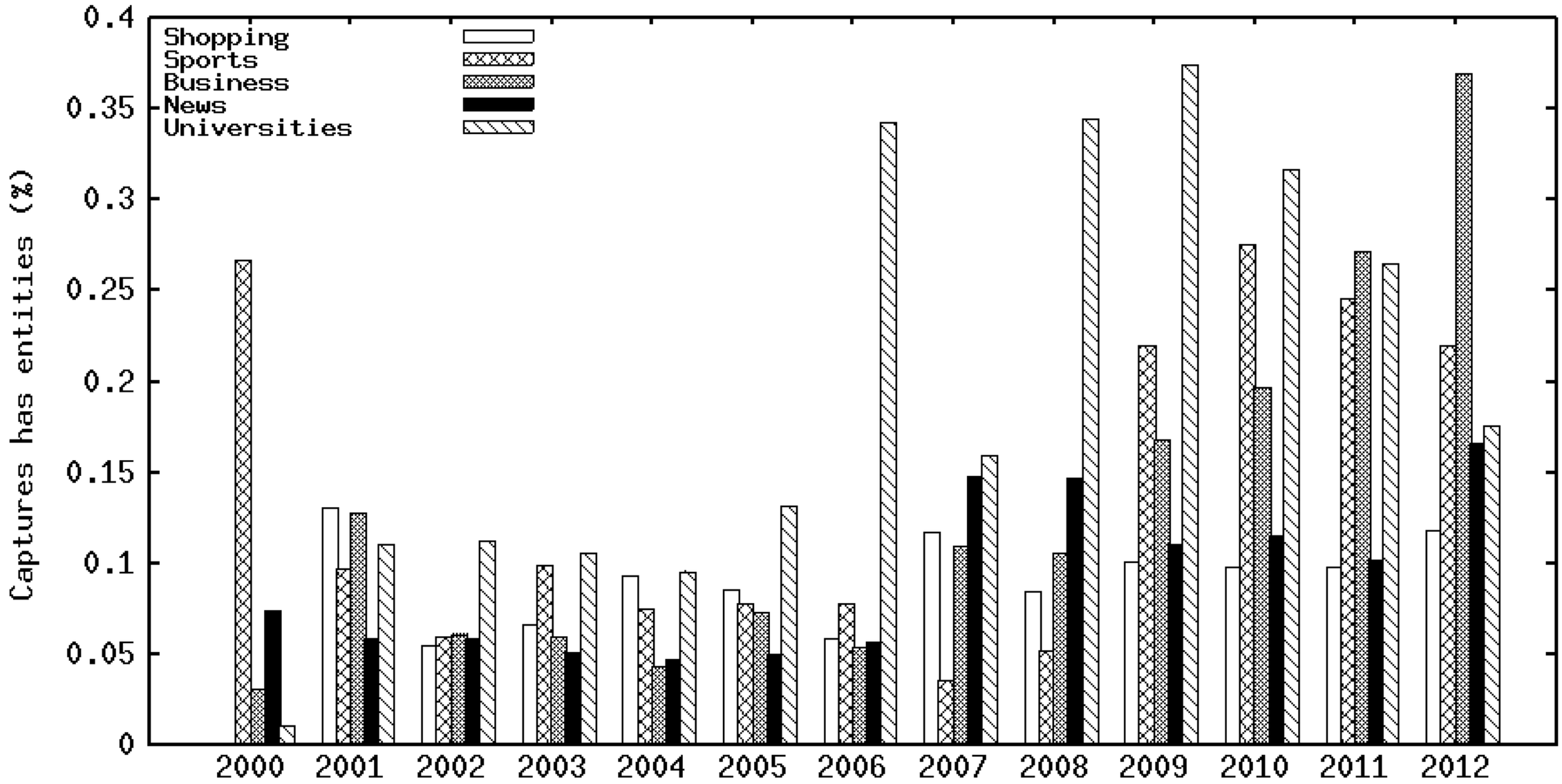
Domain and temporal coverage of NER

- Overall 42,547,734 captures containing named entities have been identified by the extractor
- Frequency range: from 2,301,917 to 3

Label	Type	Frequency
deutschland	location	2,301,917
berlin	location	628,300
hamburg	location	557,000
nordrhein	location	430,939
muenchen	location	405,845

Label	Type	Frequency
michael jackson	person	30,210
tommy hilfiger	person	25,943
harald schmidt	person	25,176
heidi klum	person	21,291
merkel	person	17,835

Distribution of entities by domain category

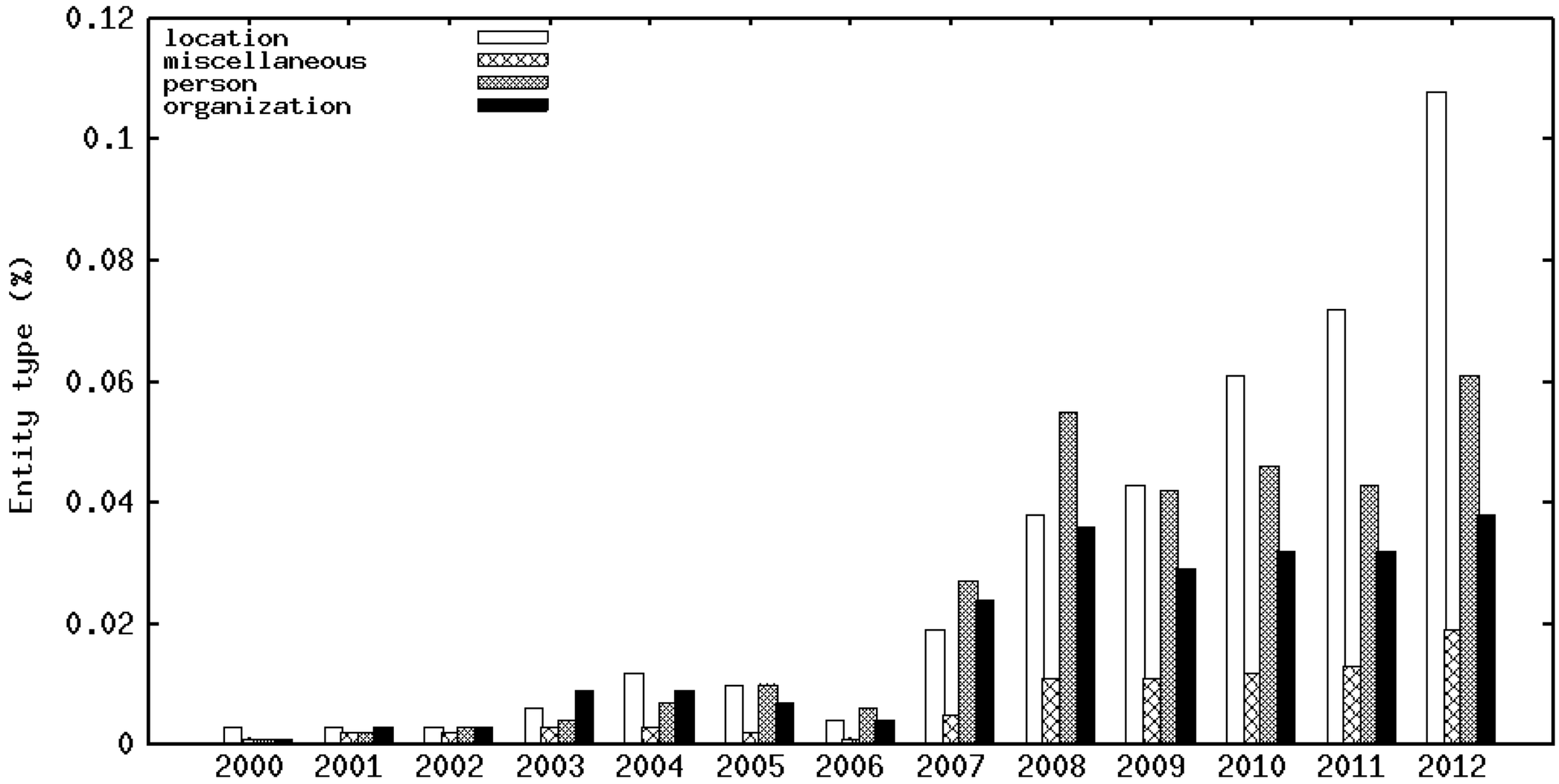


Dominant Domains

- Universities
  - uni-leipzig.de (19.81% in 2005)
  - dblp.uni-trier.de (42.73% in 2006 and 6.48% in 2007)
  - dict.tu-chemnitz.de (decreases from 2008 to 2011)
- News
  - openpr.de (from 200k pages in 2006 to 700k in 2007)
- Sports
  - transfermarkt.de (from 500k in 2007 to 1.5 million in 2010)
- Business
  - postbank.de (680k in 2008 to 1.1 million in 2011)

Distribution of entities by type

- Entity-rich sites increased from 2006 onwards (postbank.de, openpr.de, transfermarkt.de)



Conclusion

- URL analytics towards providing efficient semantic annotations to large-scale Web archives
- named entity recognition techniques can be effectively applied to URLs of the Web documents in order to provide an efficient way of initial document annotation
- Future Work
  - Analyze the correlation between the URLs and document content
  - Temporal expressions in URLs
  - Seed URL selection for focused sub-collection

Thank You!

Tarcísio Souza
Forschungszentrum L3S
Appelstraße 9a
30167 Hannover

E-Mail: souza@L3S.de