Semantic Analysis and Structuring of German Legal Documents using Named Entity Recognition and Disambiguation

Ingo Glaser, 25.09.2017, Final Presentation Master's Thesis

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
wwwmatthes.in.tum.de

Outline

Motivation

Legal Technology is rising [4]

- Digitalisation of legal documents [13]
- Increasing number of startups
- New and changing business models [7]

Unstructured and semi-structured data [16]

- Modelling and structuring of legal documents
- Understanding the content of documents
- Creating added value

Capability of systems and algorithms [18]

- Computational power increases continuously
- Technologies such as Apache Spark or Hadoop allowing even more powerful clusters
- Natural Language Processing
- Machine Learning and Data Mining

Outline

Motivation

Research Questions

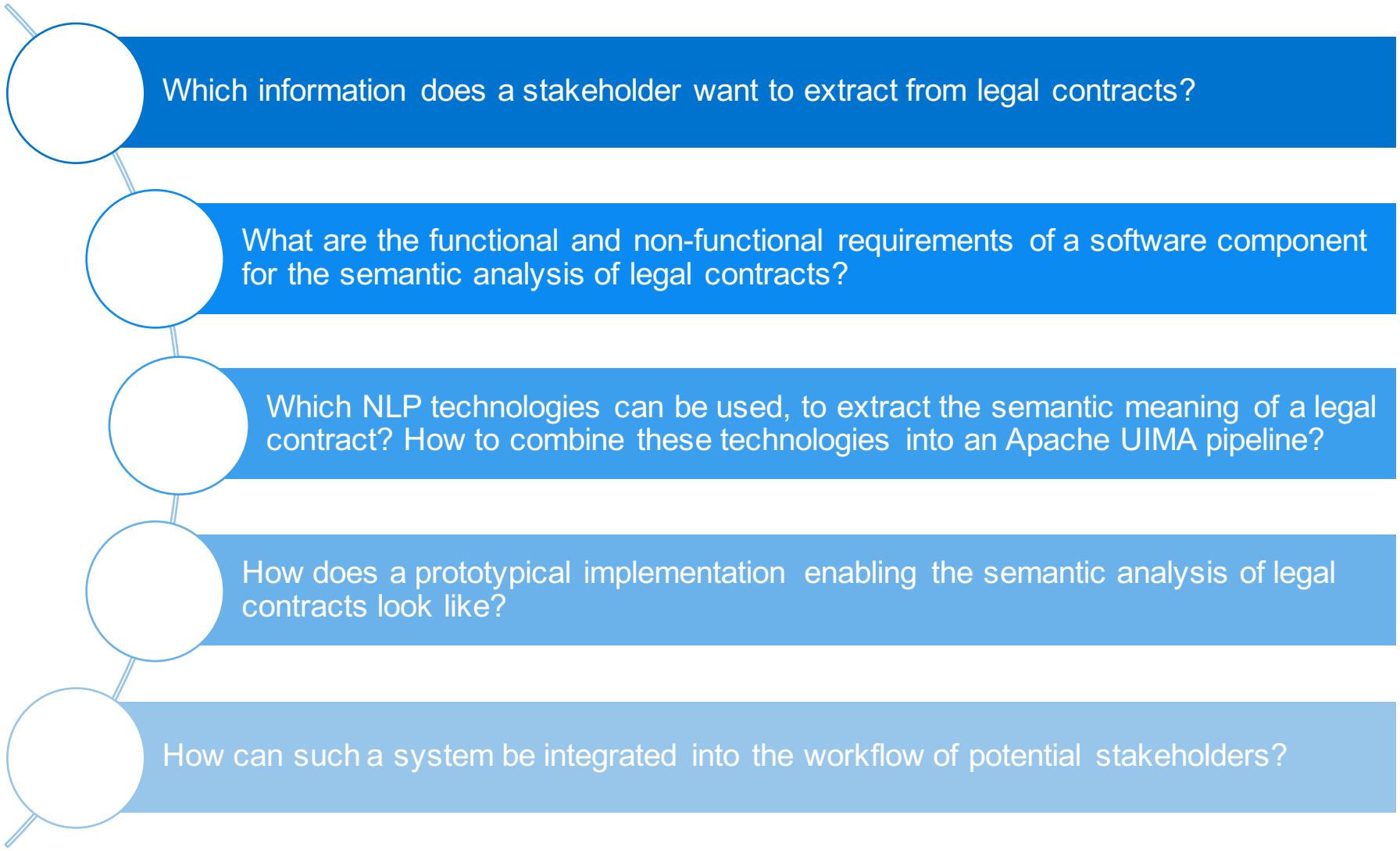
Research Approach & Objectives

Implementation

Evaluation

Conclusion

Research Questions



Outline

Motivation

Research Questions

Research Approach & Objectives

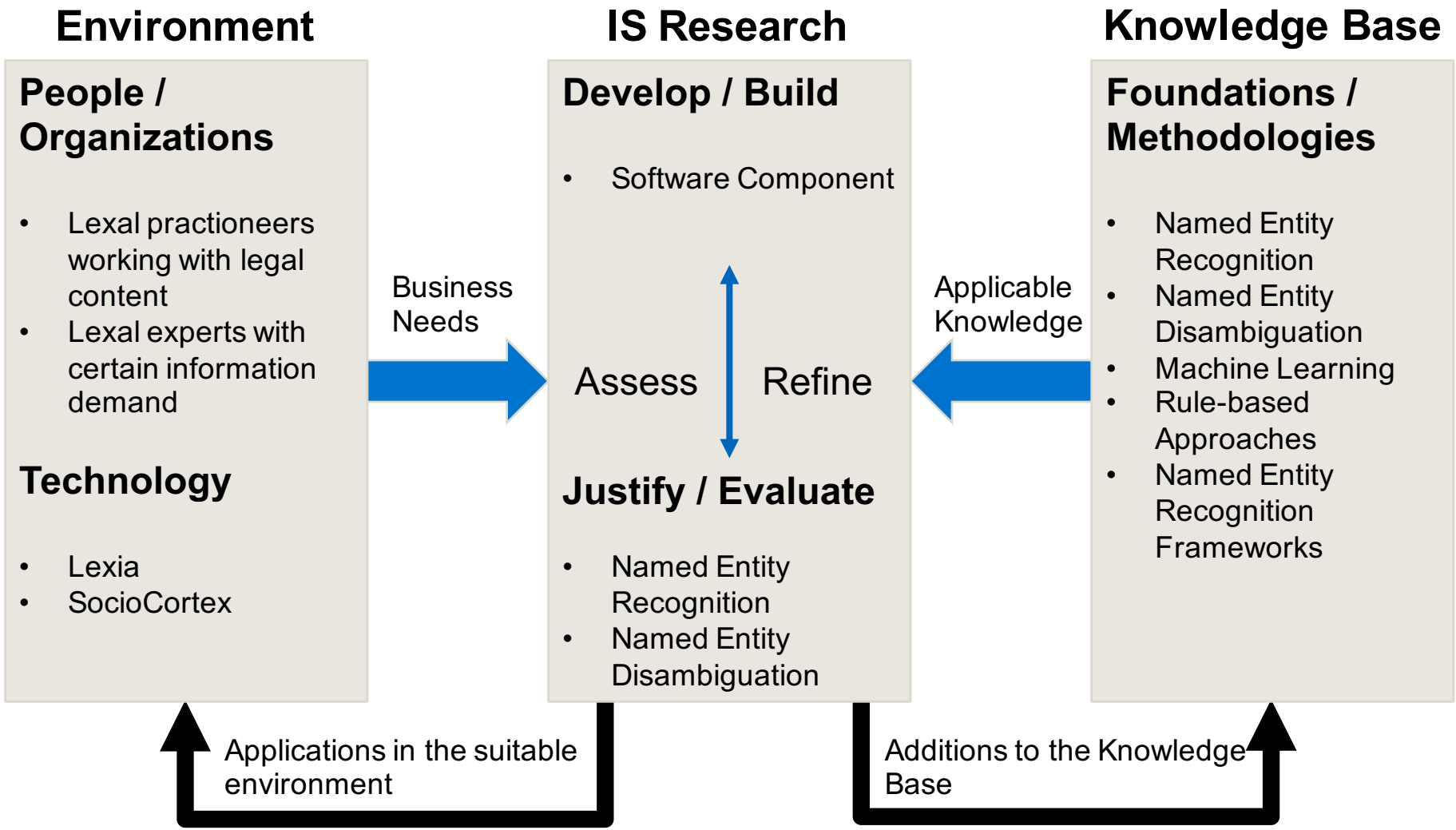
Implementation

Evaluation

Conclusion

Research Approach & Objectives

Design Science [17]



Research Approach & Objectives

Knowledge Base

Money values Dates	Rule-based Approaches (e.g. Apache Ruta, RegEx)
References	Rule-based Approaches (e.g. Apache Ruta, RegEx)
Named Entities: • Persons • Organizations • Locations	Statistical (Machine Learning) Approaches (e.g. GermaNer, Stanford NER) Knowledge Bases (e.g. DBPedia, OpenCalais) Template-based NER

Research Approach & Objectives

Knowledge Base (II)

Monetary Values

- Absolute: 1.234 Euro;
- Relative: „50 % der Miete“;

Dates

- Absolute:
  - „15. September bis 15. Mai“

- Relativ:
  - „12 Monate nach Ende des Abrechnungszeitraums“
  - „4 Wochen nach XX“
  - „3 Monate vor Beginn der Bauarbeiten“

References

- „Teilkündigung und Verwertungskündigung §§ 573, 573a, 573b BGB"

§ 5 Versorgung mit Heizung und Warmwasser

1. Der Vermieter muss die Sammelheizung, soweit es die Witterung erfordert, mindestens aber in der Zeit vom 15. September bis 15. Mai in Betrieb halten. Eine Temperatur von mindestens 20°C bis 22°C zwischen 6.00 und 24.00 Uhr in den beheizbaren Räumen ist zu erreichen. In der übrigen Nachtzeit sind 18°C ausreichend.

6. Der Vermieter kann eine Nachzahlung auf die Heiz- und Betriebskosten nur verlangen, sofern er spätestens 12 Monate nach Ende des Abrechnungszeitraumes dem Mieter durch schriftliche Abrechnung nachweist, dass die Vorauszahlungen auf die Betriebskosten nicht ausgereicht haben. Ergibt sich ein Guthaben aus der Abrechnung für den Mieter, wird dies unverzüglich ausgezahlt. Eine Aufrechnung mit bestrittenen oder nicht rechtskräftig festgestellten Forderungen darf der Vermieter nicht vornehmen. Einwendungen des Mieters gegen die Abrechnung müssen dem Vermieter spätestens 12 Monate nach Zugang der Abrechnung mitgeteilt werden.

7. Nachforderungen des Vermieters werden 4 Wochen nach Zugang der ordnungsgemäßen Abrechnung fällig. Der Vermieter gewährt dem Mieter Einsicht in die Berechnungsunterlagen. Gegen Erstattung angemessener Kopier- und Portokosten kann der Mieter verlangen, dass ihm Kopien der Berechnungsunterlagen zugesandt werden.

§ 2 Mietzeit

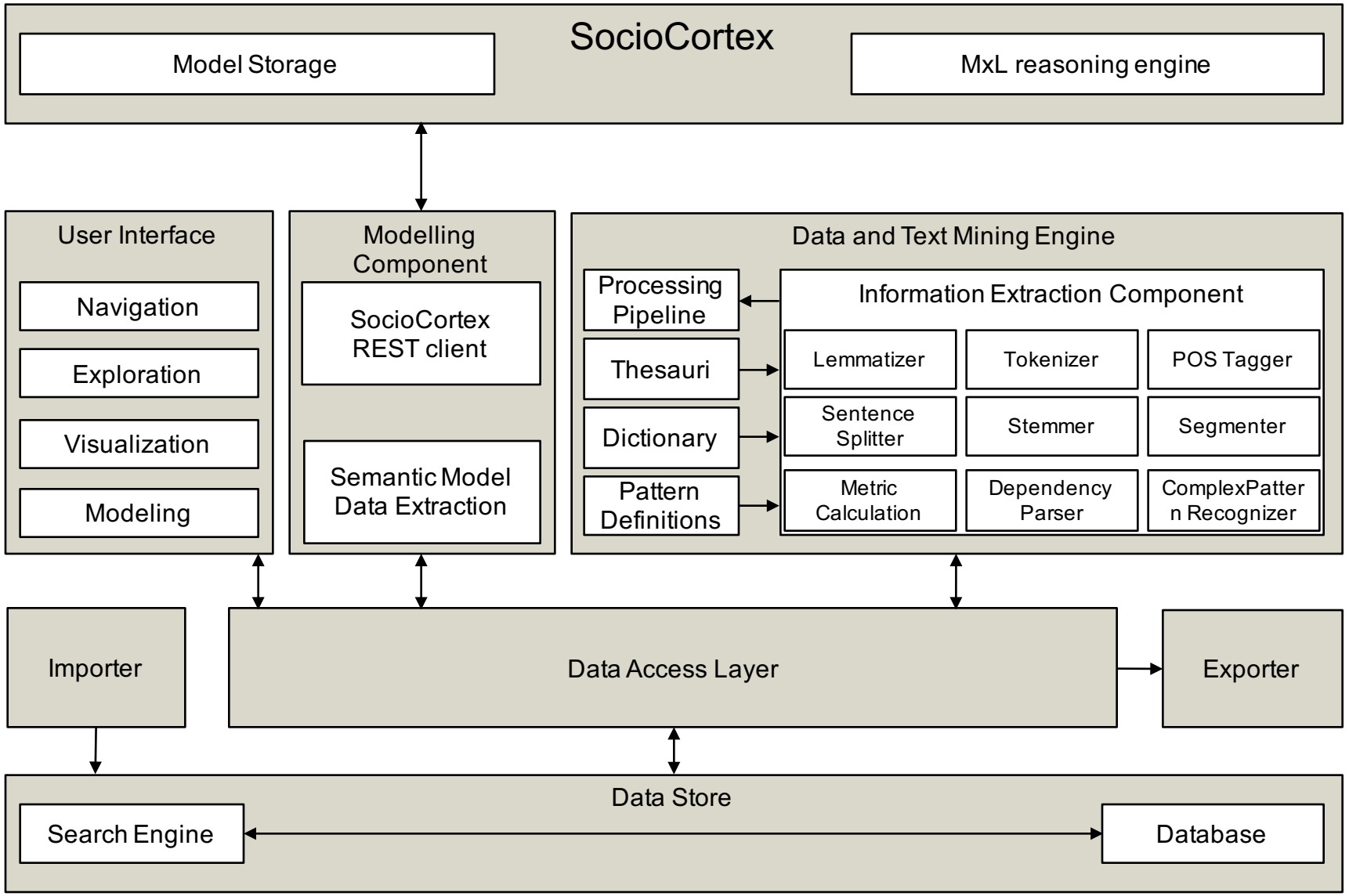
Das Mietverhältnis beginnt am: _____, es läuft auf unbestimmte Zeit.

Die Vertragspartner streben ein längerfristiges Mietverhältnis an. **Der Vermieter verzichtet für einen Zeitraum von 3 Jahren und 9 Monaten ab Vertragsabschluss auf das Recht zur ordentlichen Kündigung (Kündigung wegen Eigenbedarf, als Einliegerwohnung, Teilkündigung und Verwertungskündigung §§ 573, 573a, 573b BGB). Die Kündigung kann somit frühestens zum Ablauf dieses Zeitraums ausgesprochen werden.** Die Kündigungsvoraussetzungen richten sich im Übrigen nach den gesetzlichen Vorschriften und den vertraglichen Absprachen (siehe §§ 8, 17 – 22 dieses Vertrages).

Hinweis: Die Mietvertragsparteien können unter § 22 dieses Mietvertrages auch einen dauerhaften oder längerfristigeren Kündigungsverzicht des Vermieters vereinbaren.

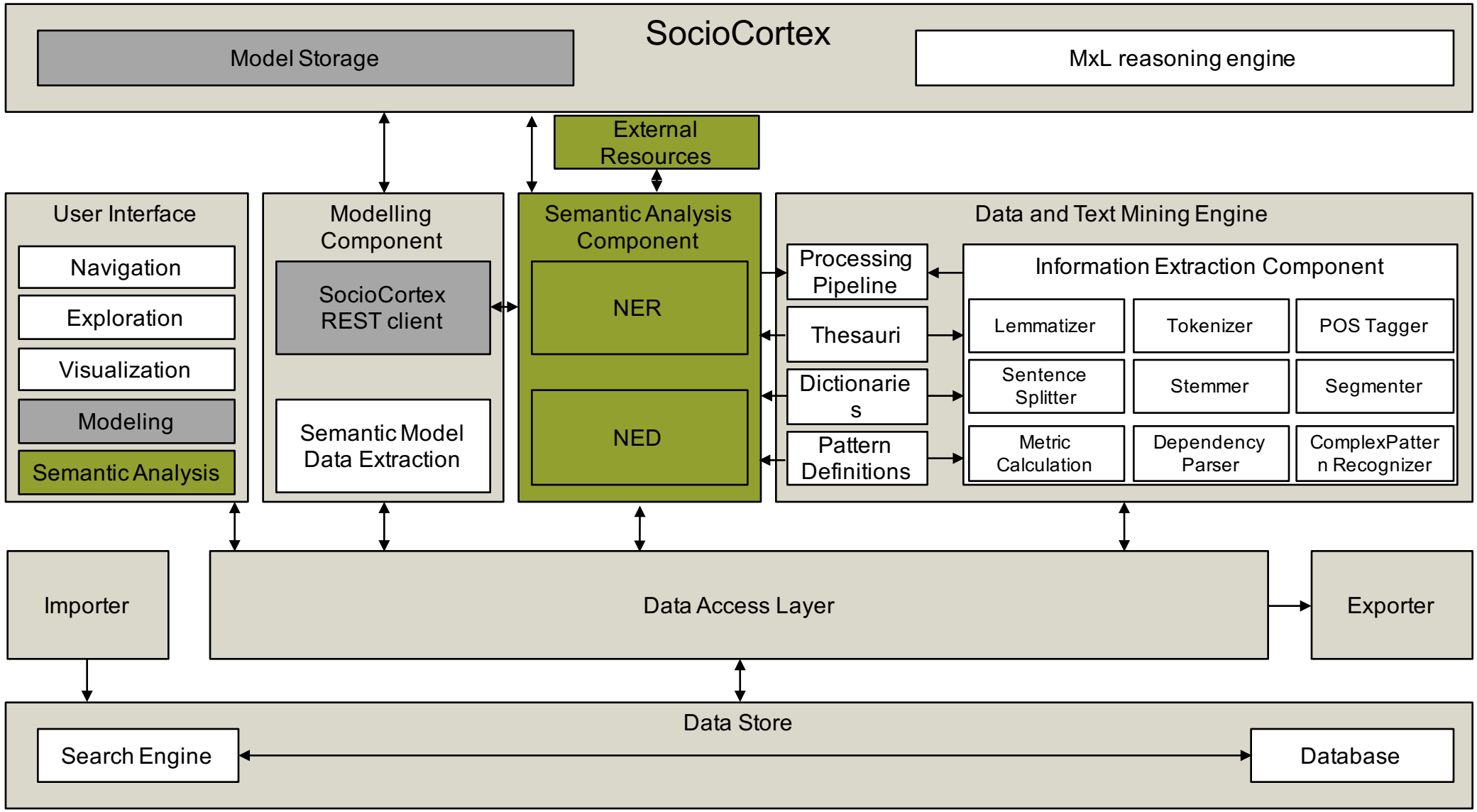
Research Approach & Objectives

Environment & IS Research



Research Approach & Objectives

Environment & IS Research (II)



Outline

Motivation

Research Questions

Research Approach & Objectives

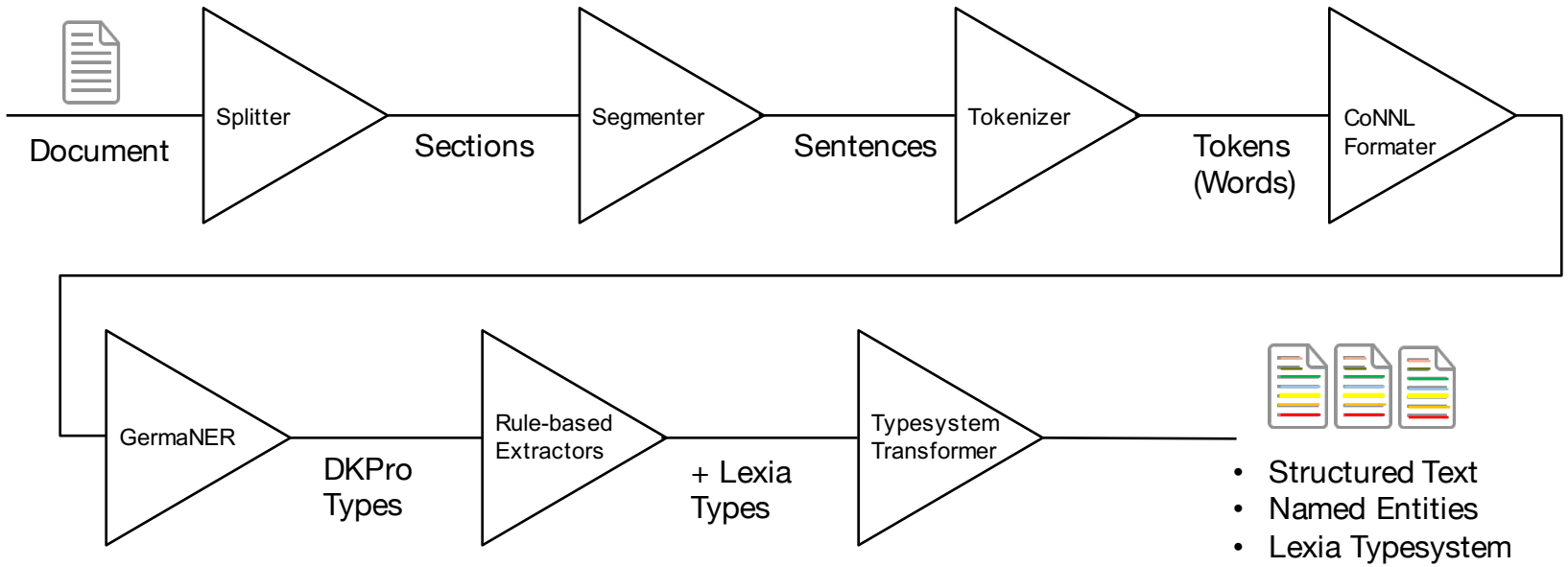
Implementation

Evaluation

Conclusion

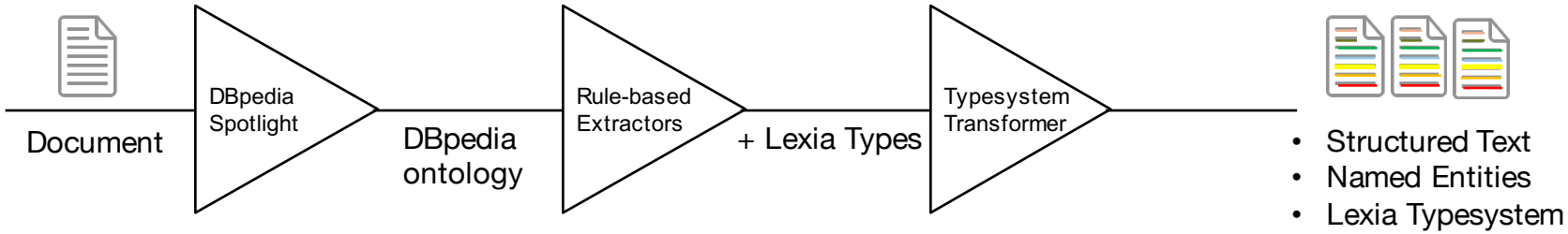
Implementation

GermaNER



Implementation

DBpedia



Implementation

Template-based NER

Mietvertrag für Wohnräume

zwischen xxxxx xxxxxxx Vermieter

xxxxxxxxxx xx, xxxxx xxxxxxxxx Telefon xxx xxxxxxxx
(Anschrift)

xxxxx@xxxxx.xx Mobil xxxx xxxxxxxxxx
(E-Mail-Adresse)

xxxxxx xxxxxxxx
(Bevollmächtigte/r)

xxxxxxx xx, xxxxx xxxxxxxxx Telefon xxx xxxxxxxxx
(Anschrift)

Mobil
(E-Mail-Adresse)

und xxxxxxxxxx xxxxxxxxxxx Mieter
(Vor- und Zuname, Geburtsdatum, Nummer des Personalausweises/Reisepasses)

xxxxxx xx, xxxxx xxxxxxxxxx
(Anschrift)

Telefon xxx xxxxxxx

max@mieter.de Mobil
(E-Mail-Adresse)

xxxxxxxxxxxxxx
(Bankverbindung des Mieters, Name und Ort der Bank, Name des Kontoinhabers)

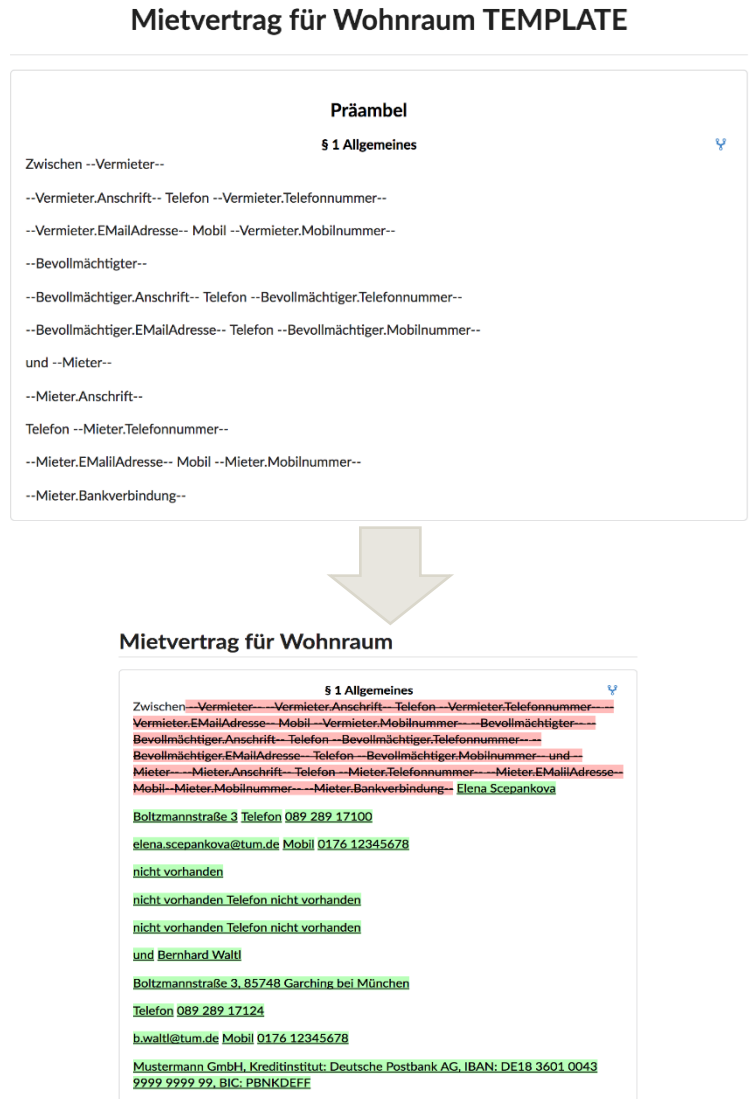
BLZ Kto.-Nr.

IBAN DExxxxxxxxxxxxxxxxxx BIC xxxxxxxxxxx

1. Transformation of contract model into template

2. Analysis via placeholder
   - --Vermieter--
   - --Vermieter.Anschrift--
   - --Mieter-
   - --Mieter.Mobilnummer--

3. Resolution of the placeholder



Mietvertrag für Wohnraum TEMPLATE

Präambel

§ 1 Allgemeines

Zwischen --Vermieter--

--Vermieter.Anschrift-- Telefon --Vermieter.Telefonnummer--

--Vermieter.EMailAdresse-- Mobil --Vermieter.Mobilnummer--

--Bevollmächtigter--

--Bevollmächtiger.Anschrift-- Telefon --Bevollmächtiger.Telefonnummer--

--Bevollmächtiger.EMailAdresse-- Telefon --Bevollmächtiger.Mobilnummer--

und --Mieter--

--Mieter.Anschrift--

Telefon --Mieter.Telefonnummer--

--Mieter.EMalilAdresse-- Mobil --Mieter.Mobilnummer--

--Mieter.Bankverbindung--

Mietvertrag für Wohnraum

§ 1 Allgemeines

Zwischen ~~--Vermieter-- --Vermieter.Anschrift-- Telefon --Vermieter.Telefonnummer-- --Vermieter.EMailAdresse-- Mobil --Vermieter.Mobilnummer-- --Bevollmächtigter-- --Bevollmächtiger.Anschrift-- Telefon --Bevollmächtiger.Telefonnummer-- --Bevollmächtiger.EMailAdresse-- Telefon --Bevollmächtiger.Mobilnummer-- und --Mieter-- --Mieter.Anschrift-- Telefon --Mieter.Telefonnummer-- --Mieter.EMalilAdresse-- Mobil --Mieter.Mobilnummer-- --Mieter.Bankverbindung--~~ Elena Scepankova

Boltzmannstraße 3 Telefon 089 289 17100

elena.scepankova@tum.de Mobil 0176 12345678

nicht vorhanden

nicht vorhanden Telefon nicht vorhanden

nicht vorhanden Telefon nicht vorhanden

und Bernhard Waltl

Boltzmannstraße 3, 85748 Garching bei München

Telefon 089 289 17124

b.waltl@tum.de Mobil 0176 12345678

Mustermann GmbH, Kreditinstitut: Deutsche Postbank AG, IBAN: DE18 3601 0043 9999 9999 99, BIC: PBNKDEFF

Implementation

Template-based NER: Example

Geschäftsraum-Mietvertrages

§ 1 Parteien

Zwischen

BERNHARD WALTL

und

SINC GmbH

wird folgender Geschäftsraummietvertrag geschlossen:

§ 2 Mieträume

1. Vermietet werden im Haus Rheingaustraße 182, 65203 Wiesbaden folgende Räume:

   a. Erdgeschoss: 1
   b. 1. Etage: 3
   c. Keller: 1
   d. Dachboden: 1

Die Mietfläche beträgt 75 qm.

2. Für die oben genannten Räume erhält der Mieter folgende Schlüssel:

2 x Haustüre, 1 x Keller, 1 x Dachboden

3. Schäden an diesen Räumen sind dem Vermieter unverzüglich anzuzeigen.

4. Der Mieter ist verpflichtet, eine Glasversicherung für sämtliche Fenster-, Schaufenster- und Türscheiben der Mieträume (oder sonstige Versicherungen nach Vereinbarung, wobei eine Doppelversicherung durch Mieter und Vermieter vermieden

Geschäftsraum-Mietvertrages TEMPLATE

§ 1 Parteien

Zwischen

~~BERNHARD WALTL und SINC GmbH~~--Vermieter--

und

--Mieter--

wird folgender Geschäftsraummietvertrag geschlossen:

§ 2 Mieträume

1. Vermietet werden im Haus ~~Rheingaustraße 182, 65203 Wiesbadenfolgende Räume: Erdgeschoss:1 1. Etage:3 Keller:1 Dachboden:1~~--Adresse-- folgende Räume:

   a. Erdgeschoss: --Mietobjekt.Erdgeschoss--
   b. 1. Etage: --Mietobjekt.1Etage--
   c. Keller: --Mietobjekt.Keller--
   d. Dachboden:--Mietobjekt.Dachboden--

Die Mietfläche beträgt~~75~~ --Mietobjekt.Größe-- qm.

2. Für die oben genannten Räume erhält der Mieter folgende Schlüssel:

~~2 x Haustüre, 1 x Keller, 1 x Dachboden~~--Mietobjekt.Schlüssel--

3. Schäden an diesen Räumen sind dem Vermieter unverzüglich anzuzeigen.

4. Der Mieter ist verpflichtet, eine Glasversicherung für sämtliche Fenster-, Schaufenster- und Türscheiben der Mieträume (oder sonstige Versicherungen nach

Implementation

Template-based NER: Negative Example

Mietvertrag für Wohnraum TEMPLATE

§ 1 Allgemeines

Zwischen --Vermieter--

--Vermieter.Anschrift-- Telefon --Vermieter.Telefonnummer--

--Vermieter.EMailAdresse-- Mobil --Vermieter.Mobilnummer--

--Bevollmächtigter--

--Bevollmächtiger.Anschrift-- Telefon --Bevollmächtiger.Telefonnummer--

--Bevollmächtiger.EMailAdresse-- Telefon --Bevollmächtiger.Mobilnummer--

und --Mieter--

--Mieter.Anschrift--

Telefon --Mieter.Telefonnummer--

--Mieter.EMalilAdresse-- Mobil --Mieter.Mobilnummer--

--Mieter.Bankverbindung--

Mietvertrag für Wohnraum

§ 1 Allgemeines

Zwischen ~~--Vermieter-- --Vermieter.Anschrift-- Telefon --Vermieter.Telefonnummer-- --Vermieter.EMailAdresse-- Mobil --Vermieter.Mobilnummer-- --Bevollmächtigter-- --Bevollmächtiger.Anschrift-- Telefon --Bevollmächtiger.Telefonnummer-- --Bevollmächtiger.EMailAdresse-- Telefon --Bevollmächtiger.Mobilnummer-- und --Mieter-- --Mieter.Anschrift-- Telefon --Mieter.Telefonnummer-- --Mieter.EMalilAdresse-- Mobil --Mieter.Mobilnummer-- --Mieter.Bankverbindung--~~ Elena Scepankova

Boltzmannstraße 3 Telefon 089 289 17100

elena.scepankova@tum.de Mobil 0176 12345678

nicht vorhanden

nicht vorhanden Telefon nicht vorhanden

nicht vorhanden Telefon nicht vorhanden

und Bernhard Waltl

Boltzmannstraße 3, 85748 Garching bei München

Telefon 089 289 17124

b.waltl@tum.de Mobil 0176 12345678

Mustermann GmbH, Kreditinstitut: Deutsche Postbank AG, IBAN: DE18 3601 0043 9999 9999 99, BIC: PBNKDEFF

Problem

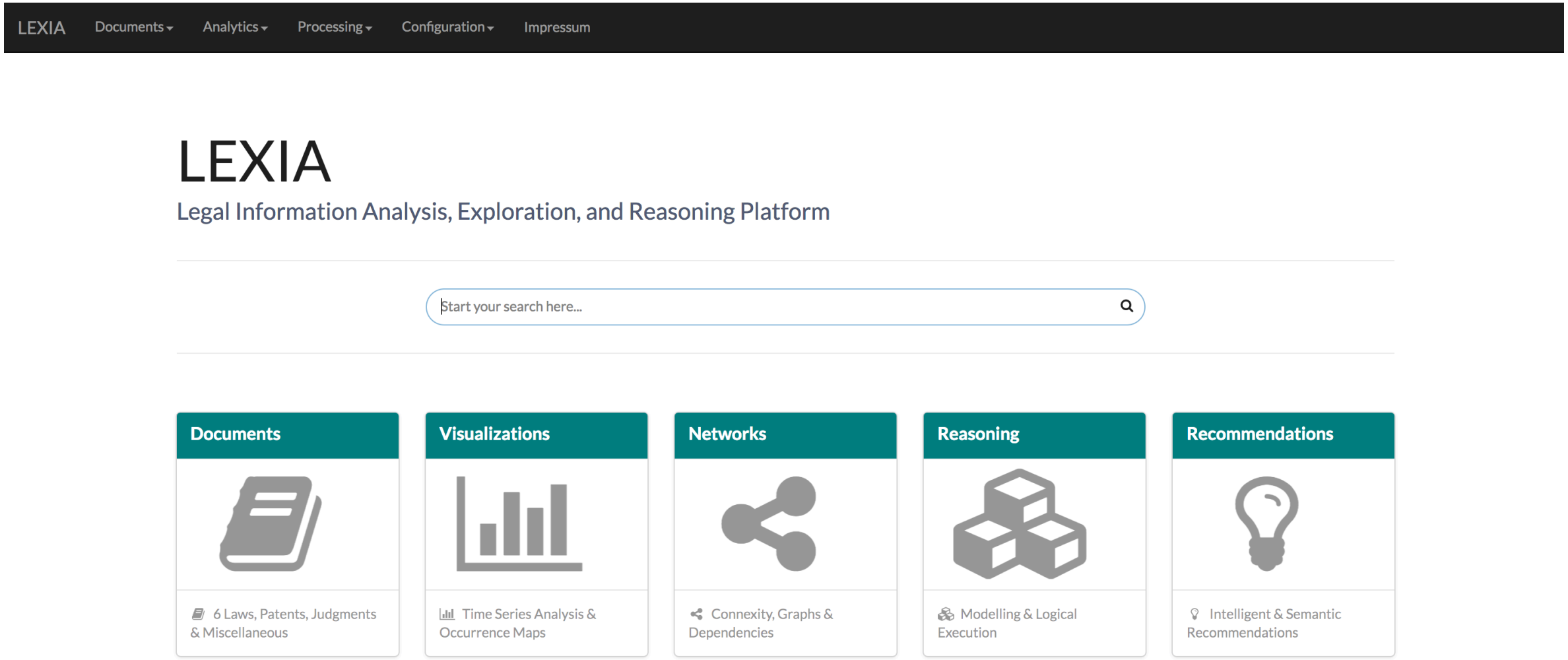
- Resolution of the placeholders, respectively greedyness of the comperative operators
- e.g. two consecutive placeholders

Next Steps

- Current implementation based on Myers‘ diff algorithm
- Integration into a pipeline

Implementation

Live-Demo



Outline

Motivation

Research Questions

Research Approach & Objectives

Implementation

Evaluation

Conclusion

Evaluation

Qualitative

Semi-structured interviews

- Three areas: personal, data, technological
- Optionally additional demo

Conducted via

- Face-to-face
- Phone

Outcome

- Recognizing and highlighting Named Entities can be helpful
- Detecting more complex structures
- First implementation looks pretty promising
- Structured view of a contract would be very helpful
- A lot of lawyers still prefer printed versions

Evaluation

Quantitative

Evaluation on judgments from the *8th Zivilsenat* of the *German BGH*

- 500 judgments from the law of tenancy
- 20 randomly selected judgments
- 25.423 tokens

Template-based NER on contracts

- 5 different contracts
- 7.790 tokens

	Per-entity F1							Overall		
System	PER	ORG	LOC	DAT	MV	REF	OTH	Precision	Recall	F1
Templated	0.88	0.77	0.82	0.86	0.88	0.93	0.71	0.94	0.91	0.92
GermaNER	0.35	0.71	0.45	0.91	0.89	0.91	0.33	0.98	0.68	0.80
DBpedia	0.51	0.76	0.52	0.91	0.86	0.91	0.59	0.87	0.87	0.87

Outline

Motivation

Research Questions

Research Approach & Objectives

Implementation

Evaluation

Conclusion

Conclusion

Named **E**ntity **R**ecognition was successfully **applied to the legal domain**

Incorporation of three different approaches

- Statistical machine learning based on CRF
- Knowledge-based approach with DBpedia
- Template-based approach

Lack of legal data

- Gold standards
- Training sets
- Ontologies

Named **E**ntity **D**isambiguation needs **further research**

B.Sc.

Ingo Glaser

Technische Universität München
Faculty of Informatics
Chair of Software Engineering for Business Information Systems

Boltzmannstraße 3
85748 Garching bei München

Tel +49.89.289.17138
Fax +49.89.289.17136

ingo.glaser@tum.de
wwwmatthes.in.tum.de

Bibliography

[1]
Aizawa, A., "An information-theoretic perspective of tf–idf measures", Information Processing & Management, vol. 39, no. 1, pp. 45-65, 2003.

[2]
Bauer, L., "Introducing linguistic morphology", 2nd ed. Washington, D.C.: Georgetown University Press, 2003, pp. x, 366 p.

[3]
Benikova, D., Muhie, S., Prabhakaran, Y., and Biemann, S.C., "C.: GermaNER: Free Open German Named Entity Recognition Tool", in In: Proc. GSCL-2015, 2015: Citeseer.

[4]
Boston Consulting Group, "How Legal Technology Will Change The Business of Law", 2016, Bucerius Law School

[5]
Beeferman, D., Berger, A., and Lafferty, J., "Statistical Models for Text Segmentation", Machine Learning, journal article vol. 34, no. 1, pp. 177-210, 1999.

[6]
Choi, F. Y. Y., "Advances in domain independent linear text segmentation", presented at the Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, Seattle, Washington, 2000.

[7]
Deloitte, "Digitisation of Documents and Legal Archiving", 2014

[8]
Habash, N., Rambow, O., and Roth, R., "MADA+ TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization", in Proceedings of the 2nd international conference on Arabic language resources and tools (MEDAR), Cairo, Egypt, 2009, vol. 41, p. 62.

[9]
Hashmi, M., "A Methodology for Extracting Legal Norms from Regulatory Documents", 2015, IEEE 19th International Enterprise Distributed Object Computing Workshop

Bibliography (II)

[10]
Hakimov, S., Oto, S. A. & Dogdu, E., “Named Entity Recognition and Disambiguation using Linked Data and Graph-based Centrality Scoring”, 2012, SIGMOD

[11]
Lavrac, N., Mladenic, D. & Erjavec, T., "Ripple Down Rule learning for utomated word lemmatisation”, AI Communications, vol. 21, no. 1, pp. 15-26, 2008.

[12]
Rajaraman, A.U., "JD (2011)." Data Mining”, Mining of Massive Datasets, pp. 1-17

[13]
Saravanan, M., Ravindran, B. & Raman, S., “Improving Legal Information Retrival Using an Ontological Framework”, 2009, Artificial Intelligence and Law

[14]
Singh, J. & Gupta, V., "Text Stemming: Approaches, Applications, and Challenges”, ACM Comput. Surv., vol. 49, no. 3, pp.

[15]
Stevenson, M., & Wilks, Y., "Word sense disambiguation”, The Oxford Handbook of Comp. Linguistics, pp. 249-265, 2003.

[16]
Svyatkovskiy, A., Imai, K., Kroeger, M., & Shiraito, Y., “Large-scale Text Processing Pipeline with Apache Spark”, 2016, IEEE International Conference on Big Data

[17]
Von Alan, H. R., March, S. T., Park, J., Ram, S., “Design science in information systems research”, 2004, MIS quarterly, vol. 28, pp. 75-105.

[18]
Waltl, B., Landthaler, J., Scepankova, E., Matthes, F., Geiger, F., Stocker, C., Schneider, C., “Automated Extraction of Semantic Information from German Legal Documents”, 2017, Internationales Rechtsinformatik Symposium, Salzburg

Backup Slides

Evaluation

Quantitative: Evaluation Data Set Judgments

NE Types	PER	ORG	LOC	DAT	MV	REF	OTH	O
Count	114	106	45	267	78	310	182	24314

Evaluation

Quantitative: GermaNER Confusion Matrix

	Gold							
Tokens	PER	ORG	LOC	DAT	MV	REF	OTH	O
PER	24	0	0	0	0	0	0	0
ORG	0	59	2	0	0	0	0	0
LOC	0	0	14	0	0	2	0	1
DAT	0	0	9	229	0	0	0	0
MV	0	0	0	0	62	0	0	0
REF	0	0	2	0	2	261	0	0
OTH	6	9	9	2	7	1	44	7
O	84	38	9	36	7	46	138	24306

Evaluation

Quantitative: DBpedia Confusion Matrix

	Gold							
Tokens	PER	ORG	LOC	DAT	MV	REF	OTH	O
PER	40	0	0	0	0	0	0	0
ORG	0	67	2	0	0	0	0	0
LOC	10	12	28	0	0	0	1	11
DAT	1	0	9	229	0	0	0	0
MV	0	0	0	0	62	0	0	0
REF	0	0	1	0	0	261	0	0
OTH	32	25	3	2	8	3	165	137
O	34	4	2	31	8	46	16	24193

Evaluation

Quantitative: GermaNER and DBpedia Evaluation Results

	GermaNER			DBpedia		
	Precision	Recall	F1	Precision	Recall	F1
PER	1.00	0.21	0.35	1	0.34	0.51
ORG	0.97	0.56	0.71	0.97	0.62	0.76
LOC	0.82	0.31	0.45	0.45	0.62	0.52
DAT	0.96	0.86	0.91	0.96	0.86	0.91
MV	1	0.79	0.89	1	0.79	0.86
REF	0.98	0.84	0.91	0.99	0.84	0.91
OTH	0.52	0.24	0.33	0.44	0.91	0.59
Overall	0.98	0.68	0.8	0.87	0.87	0.87

Evaluation

Quantitative: Evaluation Data Set Contracts

NE Types	PER	ORG	LOC	DAT	MV	REF	OTH	O
Count	14	8	23	38	23	25	46	7614

Evaluation

Quantitative: Template-based NER Confusion Matrix

	Gold							
Tokens	PER	ORG	LOC	DAT	MV	REF	OTH	O
PER	11	0	0	0	0	0	0	0
ORG	0	5	0	0	0	0	0	0
LOC	0	3	16	0	0	0	1	0
DAT	0	0	0	31	0	0	5	2
MV	0	0	0	0	18	0	1	1
REF	0	0	0	0	0	19	0	0
OTH	0	0	4	6	4	3	34	8
O	3	0	3	1	1	3	5	7603