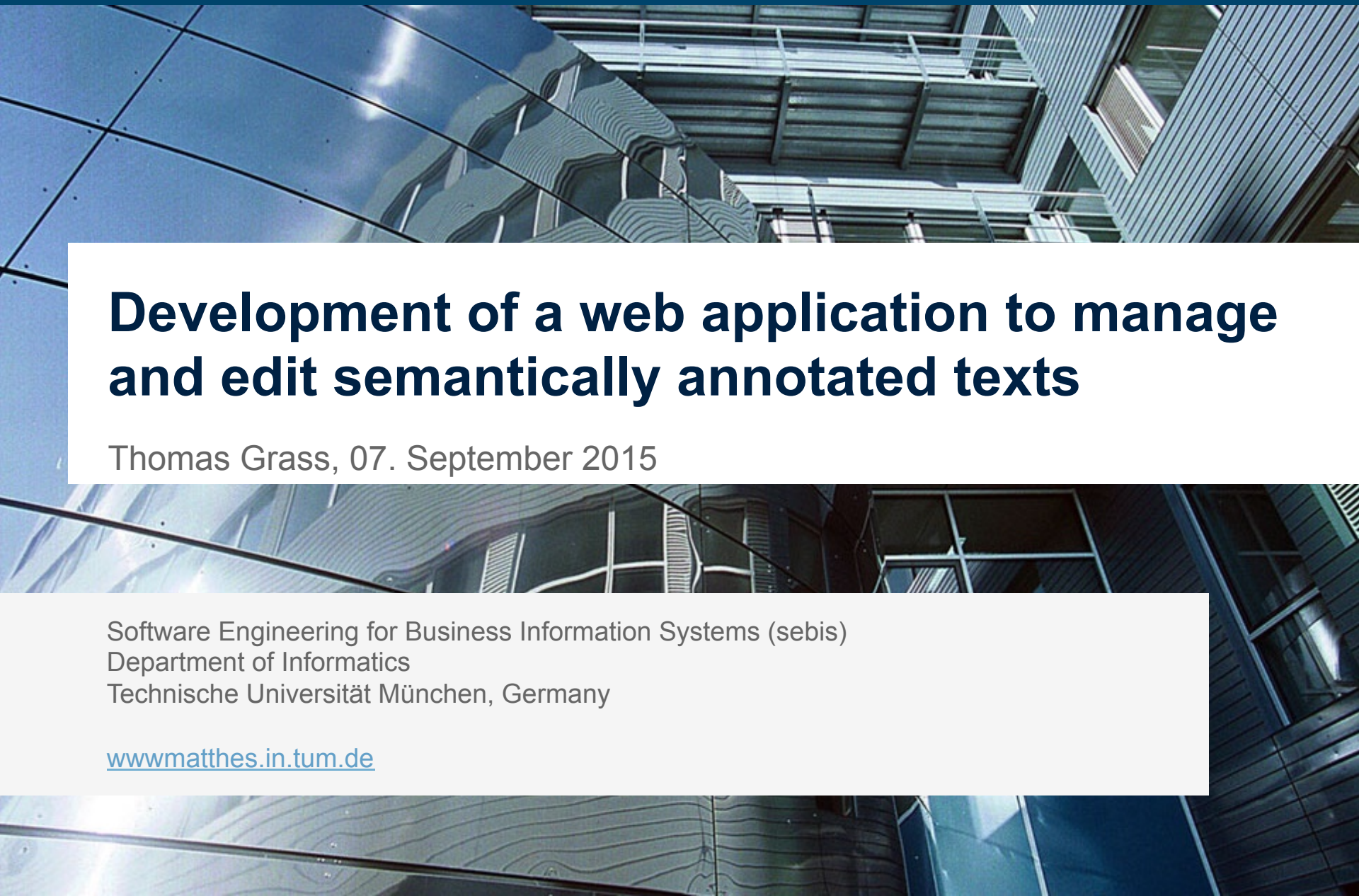


Development of a web application to manage and edit semantically annotated texts

Thomas Grass, 07. September 2015

Software Engineering for Business Information Systems (sebis)
Department of Informatics
Technische Universität München, Germany

wwwmatthes.in.tum.de

Key Facts

Master‘s Thesis – Information Systems

Development of a web application to manage and edit semantically annotated texts

Masterarbeit - Wirtschaftsinformatik

Entwicklung einer Web-Anwendung zur Verwaltung und Bearbeitung von semantisch annotierten Textsammlungen

Project

LexAlyze - Analysis of Legal Texts

Student
Thomas Grass

Advisor
Bernhard Waltl

Supervisor
Prof. Dr. Florian Matthes

Date

15.03.2015 – 15.09.2015

Agenda – An Overview

sebis

1. Introduction, Problem & Basic Theory
2. Scope & Research Questions
3. Semantic Text Annotations
4. Architecture & Implementation
    1. Overview
    2. Legal Documents
    3. Generic Importer
5. Demonstration
6. Conclusion

Problem / Basic Theory

sebis

Legal domain

Nowadays, legal texts are hard to read and understand

+

Advanced text analysis

Methods for performing advanced text analysis rapidly evolve

=

Usage of quantitative methods of structural network analysis and linguistics provide the possibility to do high class text analysis and comparison of different legal texts

Research Questions

What kind of legal texts exist in the German legislation?

What is a way to implement a generic importer for legal texts that can easily be adapted?

What kind of semantic text annotations exist and what are benefits and drawbacks of those?

How to persist semantic text annotations in order to access them for further semantic processing?

Semantic text annotations (1/3)

sebis

Semantic text annotations

- Add markups to raw text
- Fit raw text with additional information
- Can be done in-line or stand-off

In-line annotation

- Single file
- Add semantic text annotations in raw text file

Stand-off annotation

- Two files
  - Raw text file
  - Annotation file
- Annotation file points at locations in raw text file

Semantic text annotations (2/3)

sebis

In-line semantic text annotation

- Add semantic text annotations in raw text file

```
<sentence>
  <subject syllables="2">Homer</subject>
  <verb>likes</verb>
  <adjective type="color">blue</adjective>
  <object size="XXL">jeans</object>.
</sentence>
```

Benefits

- Single file usage
- Interpretable by human
- Easy to implement

Drawbacks

- Manipulation in raw text file
- Overlapping annotations not possible
- Analysis needs a bit of work

Semantic text annotations (3/3)

sebis

Stand-off semantic text annotation

- Add second file that contains the annotations

```
Homer likes blue jeans.
```

```
<sentence>
  <subject start="0" end="5" syllables="2" />
  <verb start="6" end="11" />
  <adjective start="12" end="16" type="color" />
  <object start="17" end="22" size="XXL" />
</sentence>
```

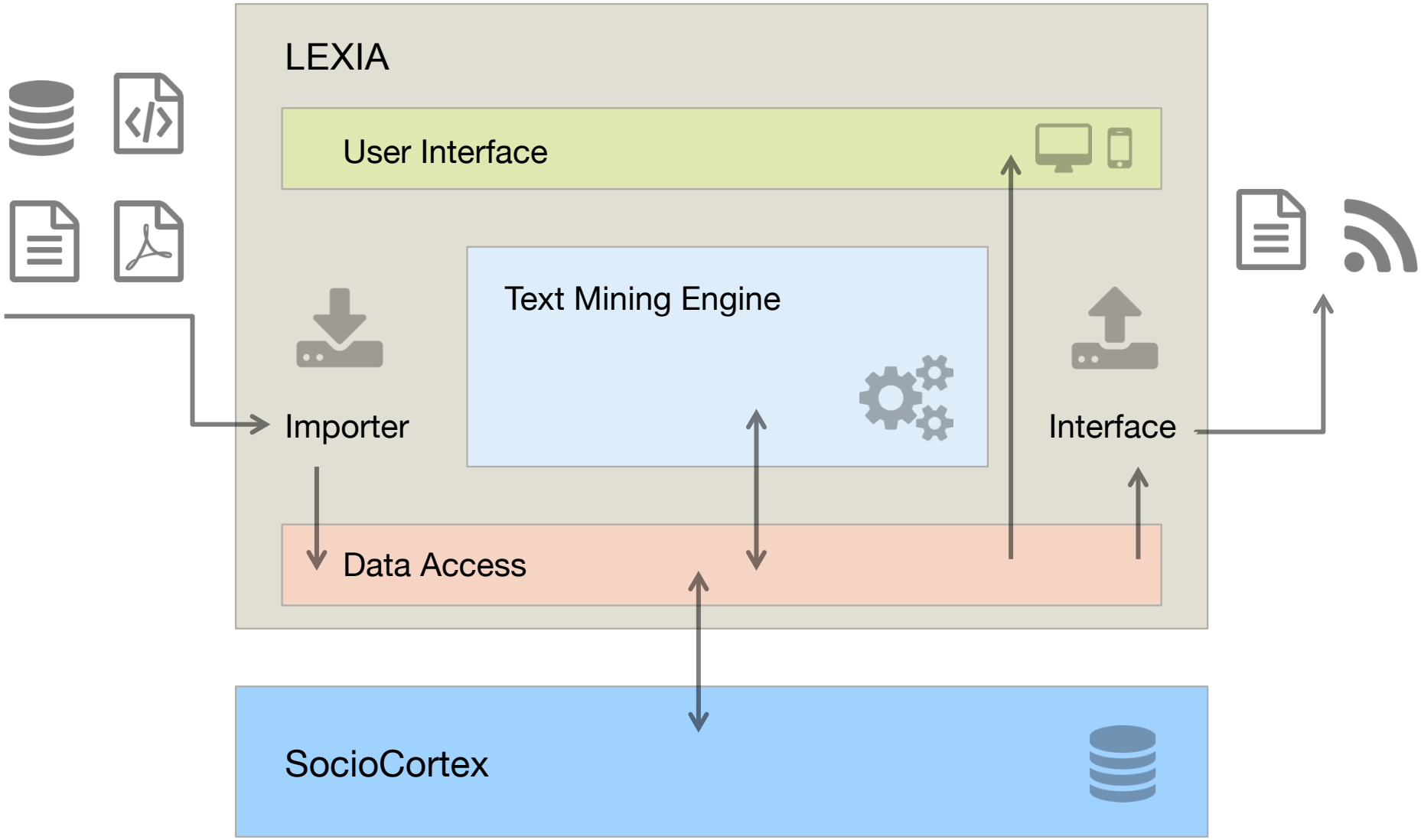
Benefits

- Raw text will not manipulated
- Easy to perform analysis
- Overlappings are possible

Drawbacks

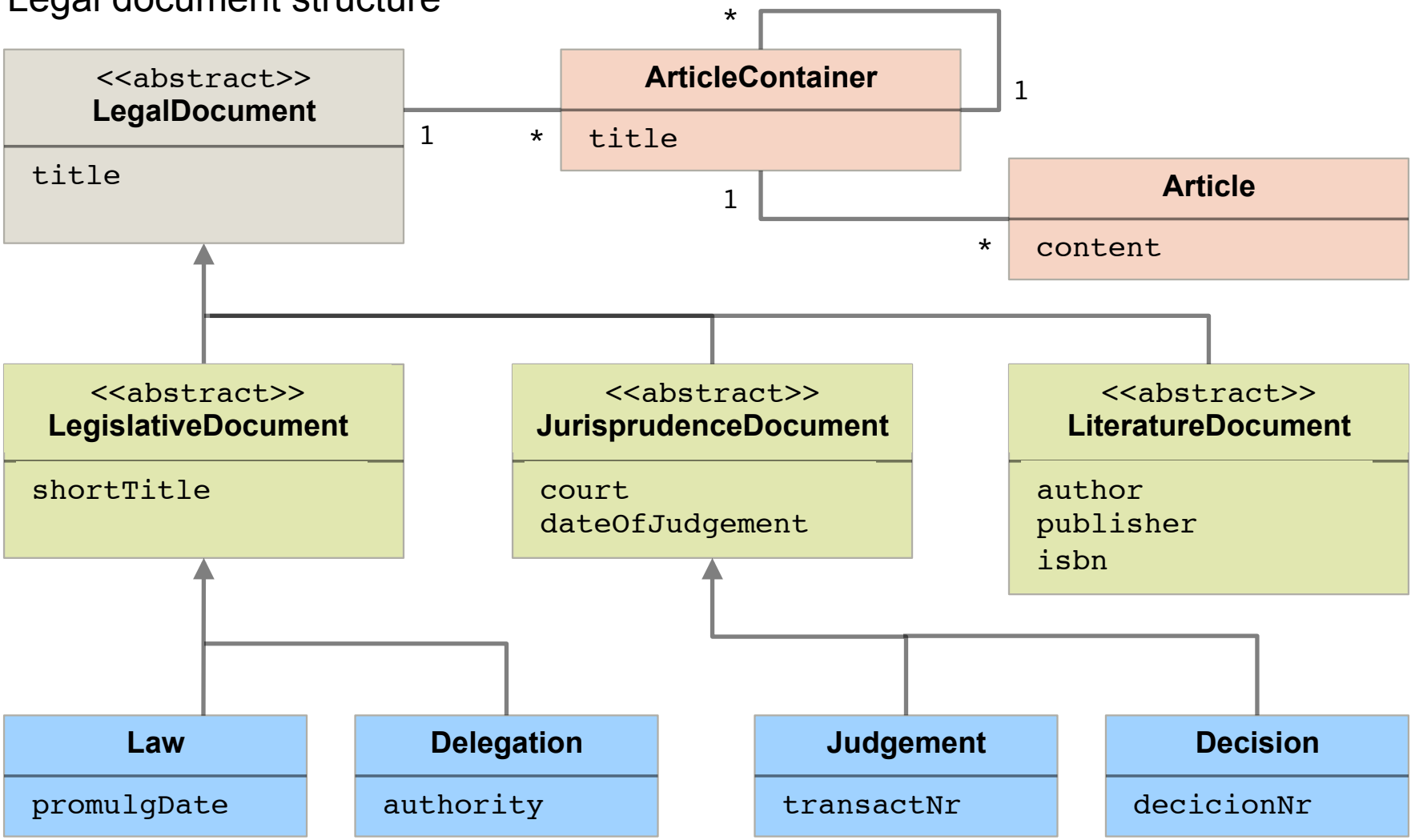
- Two files needed
- Update of raw text needs update of annotation file
- Not interpretable by human

Architecture & Implementation (1/3)



Architecture & Implementation (2/3)

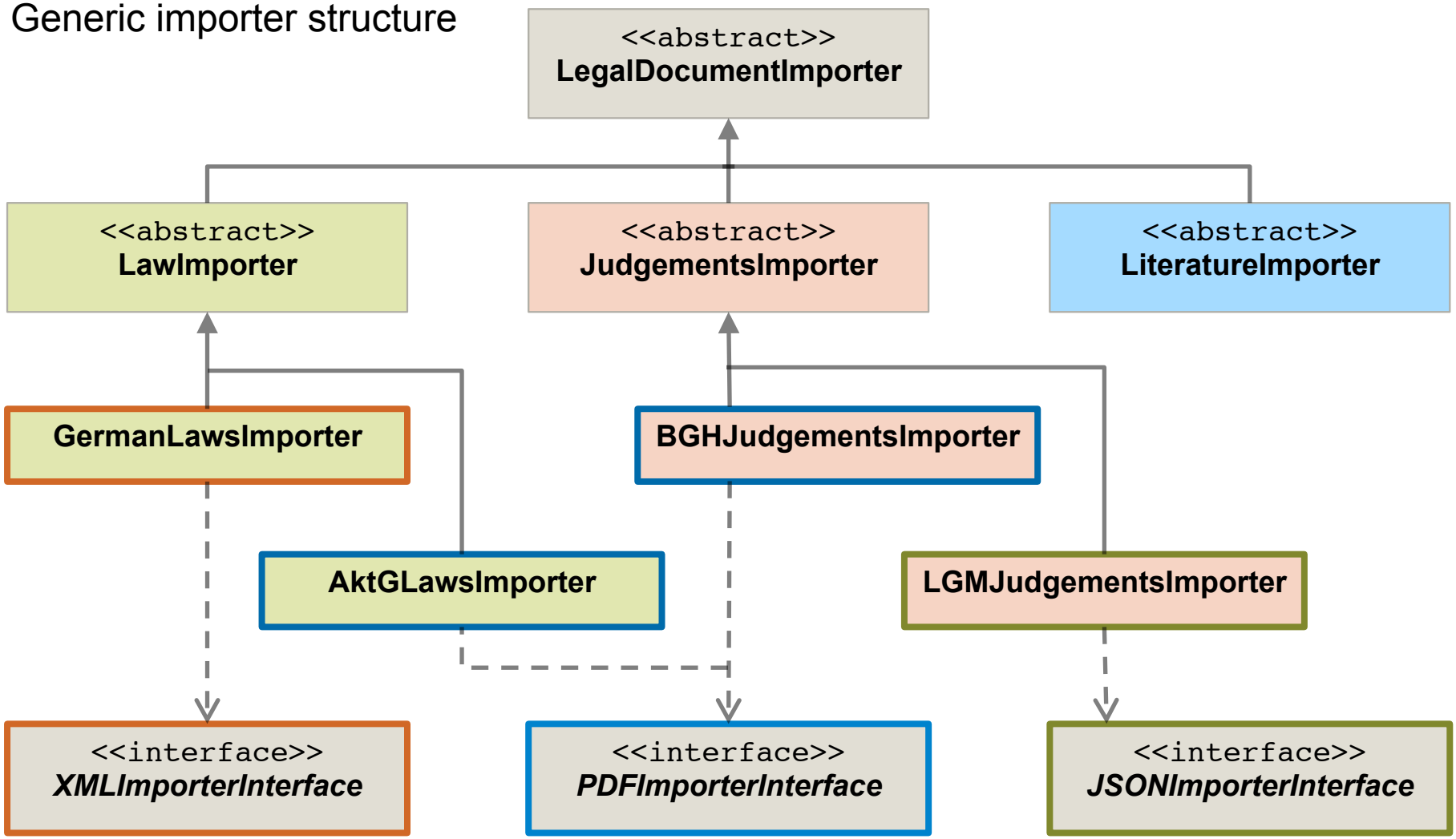
Legal document structure



Architecture & Implementation (3/3)

sebis

Generic importer structure



Demonstration

Conclusion and Outlook (1/2)

sebis

Summary

- German Legislation exists of various legal document types
  - e.g. Laws, Decisions, Judgements, …
  - existing model can easily be extended & adapted
- Generic importer for different document- and file-types
- Two types of semantic text annotations
  - In-line & Stand-off semantic text annotation
  - Prototypical implementation of Stand-off
- SocioCortex for persisting raw texts and annotations
  - MXL for selecting and quering legal documents

Open for further work

- Adding new sources
- Adding any annotation
- Extendable

Open issues & restrictions

- Slow interaction with SocioCortex (Bulk-Load)
- Success depends on sources

Conclusion and Outlook (2/2)

sebis

Upcoming

- Adding of advanced text analysis functionality
  - (October 2015, Tobias Waltl)
- Integration of new data sources, (e.g. contracts)
- Foundation for further advanced text analysis functionality
  - Determination of use cases

Thank you for your attention!

Thomas Grass
B.Sc.

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

Boltzmannstraße 3
85748 Garching bei München

Tel +49.89.289.17124
Fax +49.89.289.17136

thomas.grass@in.tum.de
wwwmatthes.in.tum.de