Vincent Kríž

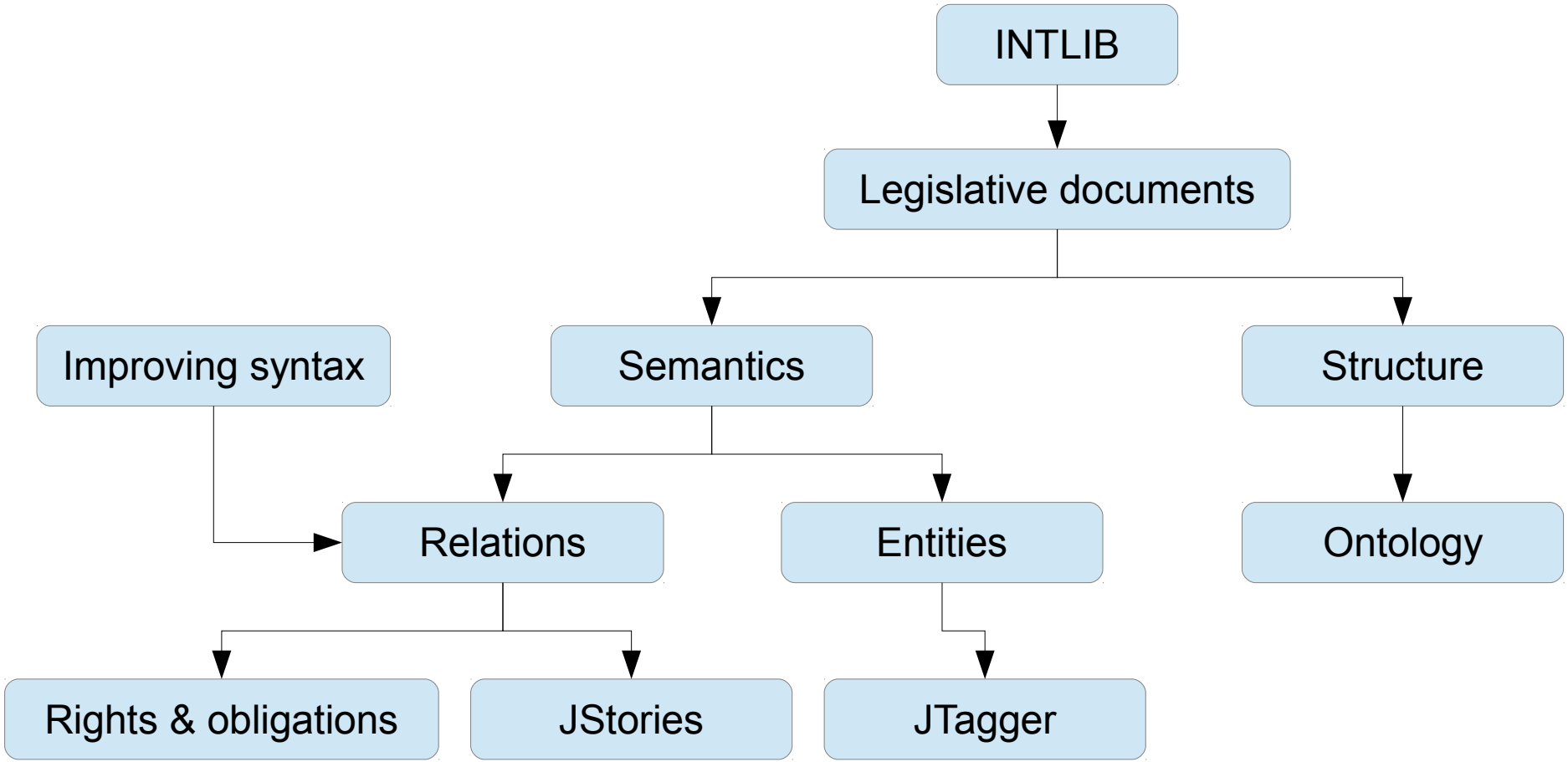
Detecting Semantic Relations in Texts and Their Integration with External Data Resources

Week of Doctoral Students 2013

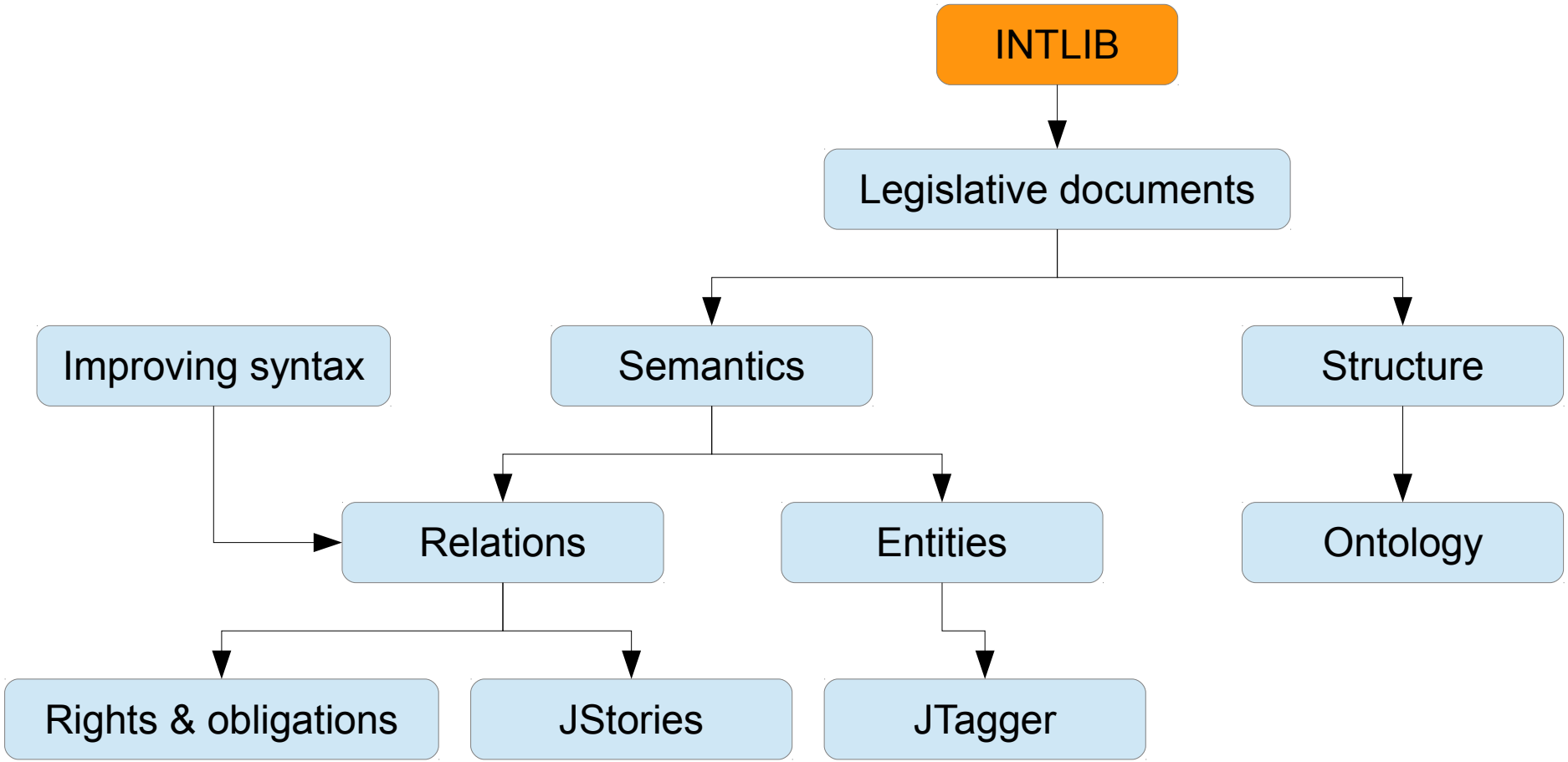
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University in Prague
Czech Republic

kriz@ufal.mff.cuni.cz
http://ufal.mff.cuni.cz/~kriz

Outline


INTLIB
Legislative documents
Improving syntax
Semantics
Structure
Relations
Entities
Ontology
Rights & obligations
JStories
JTagger

Outline

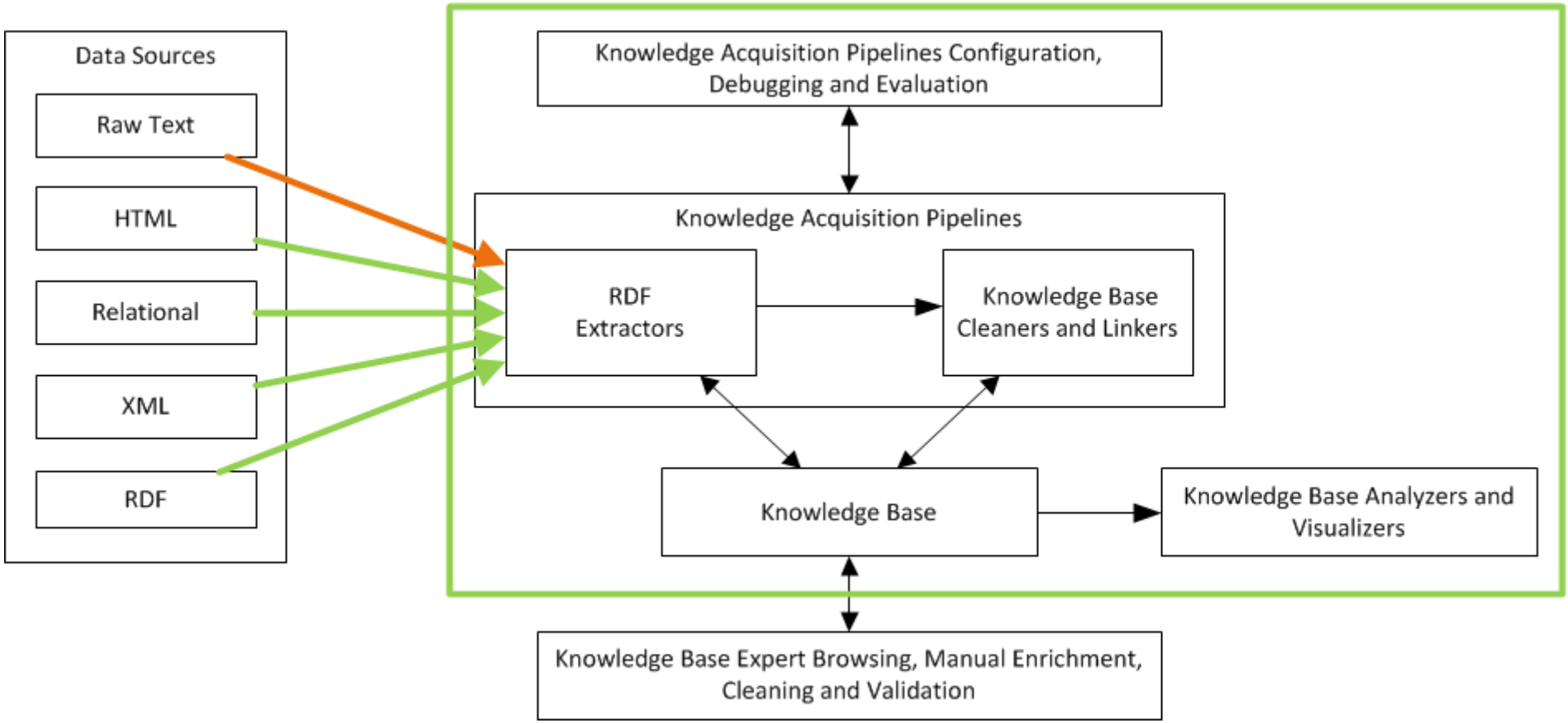

INTLIB
Legislative documents
Improving syntax
Semantics
Structure
Relations
Entities
Ontology
Rights & obligations
JStories
JTagger

Motivation

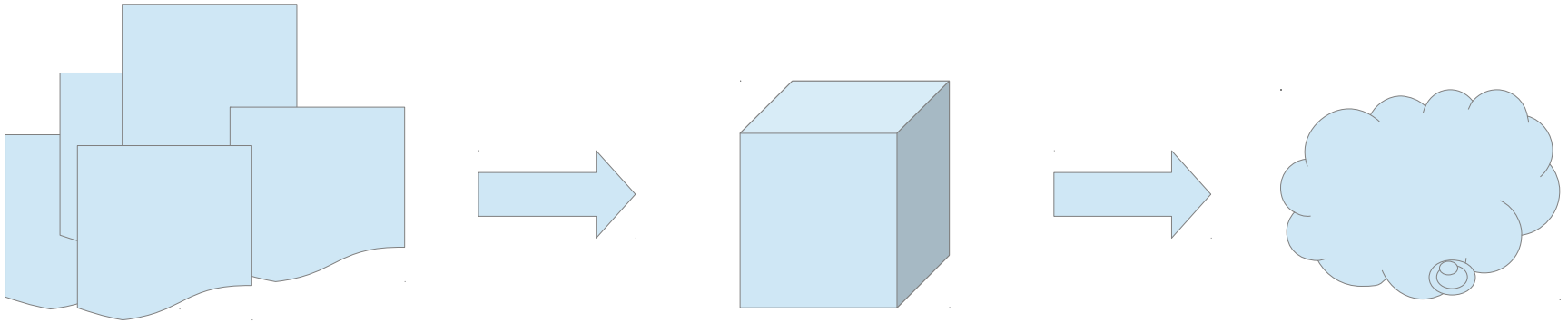
- large collections of documents
- efficient browsing & querying
- typical approaches
  - full-text search
  - metadata search

- interpret the semantics of the documents → suitable DB & query language → user-friendly browsing & querying

INTLIB


Data Sources
Raw Text
HTML
Relational
XML
RDF
Knowledge Acquisition Pipelines Configuration, Debugging and Evaluation
Knowledge Acquisition Pipelines
RDF Extractors
Knowledge Base Cleaners and Linkers
Knowledge Base
Knowledge Base Analyzers and Visualizers
Knowledge Base Expert Browsing, Manual Enrichment, Cleaning and Validation

NLP Group



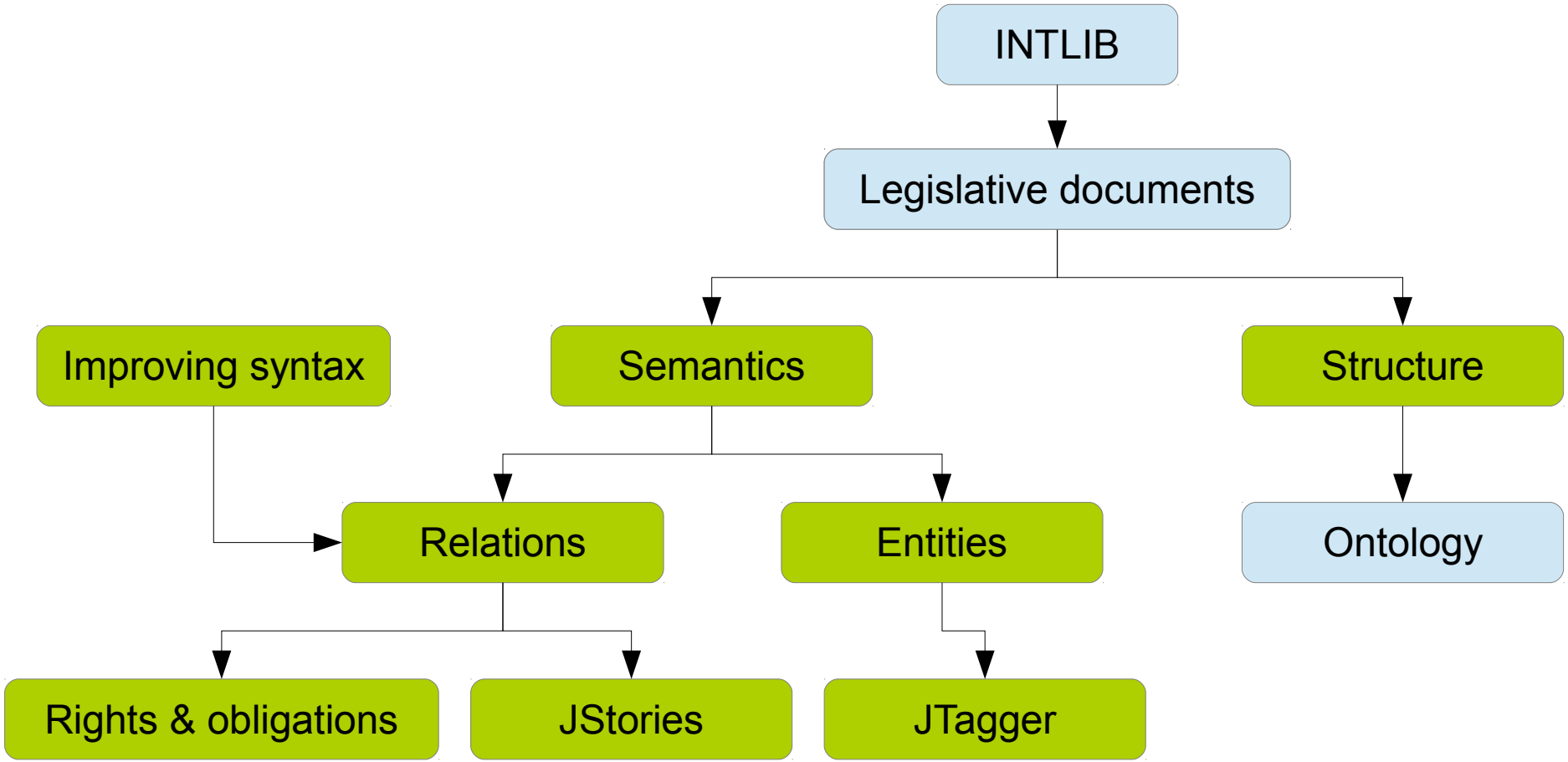
Documents Extractor Linked Data

Documents
- semi-structured documents from some domain
- legislative documents, project/medical documentation

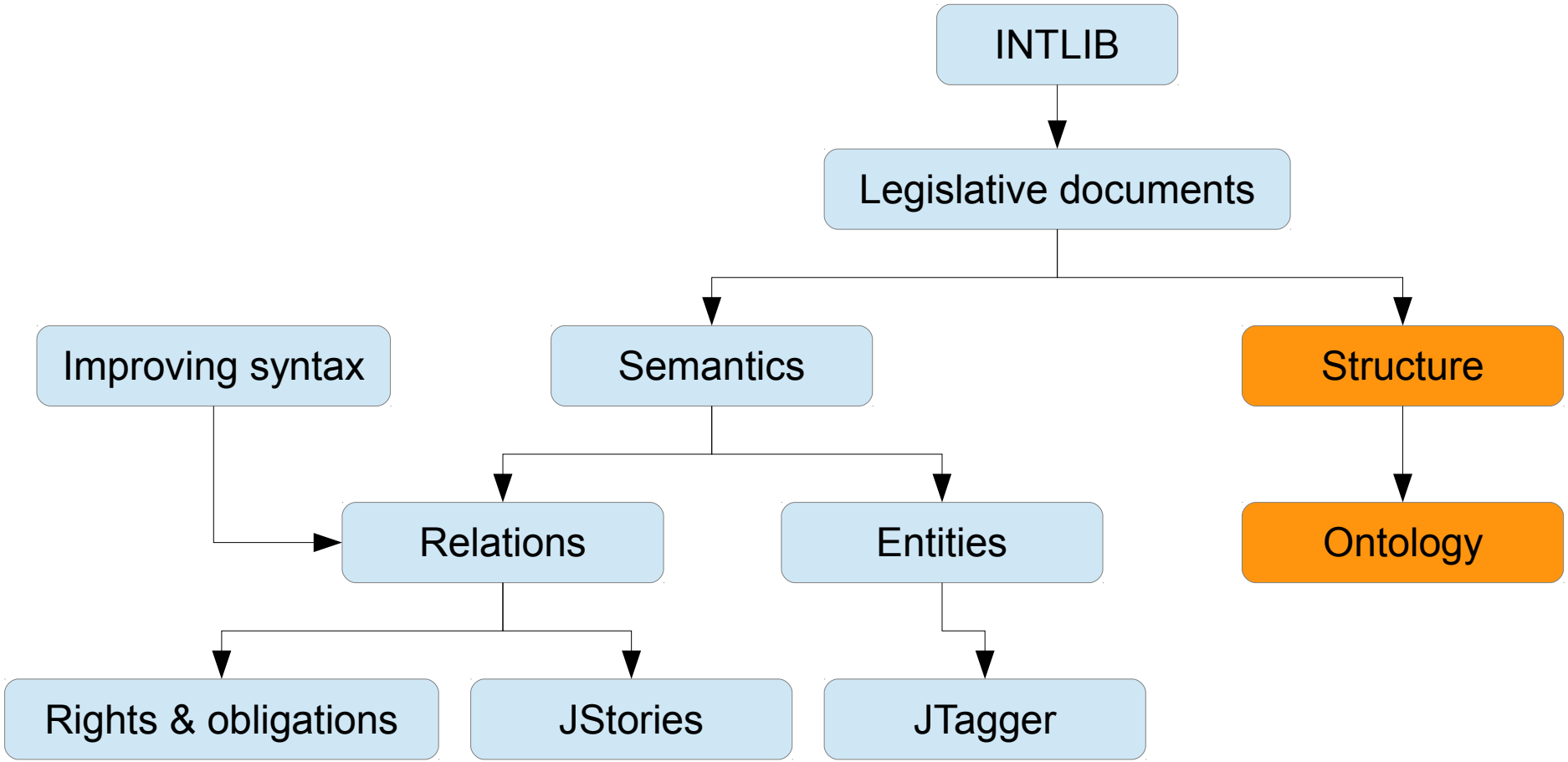
Extractor
- NLP techniques

Linked data
- automatically interconnected with other related data and with the original documents

Contribution



Outline


INTLIB
Legislative documents
Improving syntax
Semantics
Structure
Relations
Entities
Ontology
Rights & obligations
JStories
JTagger

Legislation domain - structure

What we have done

- ontology of legislative documents
- metadata and structure of acts, regulations and decrees represented as Linked Data
  - metadata about each version of each act, regulation and decree since 1945
  - structured content of versions of all acts, regulations and decrees valid in 2011, 2012

Legislation domain - structure

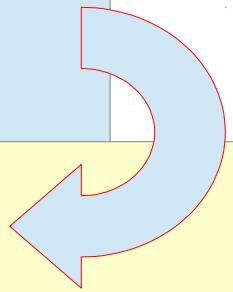
HLAVA I
ÚVODNÍ USTANOVENÍ

§ 1
Předmět úpravy
Tato vyhláška zapracovává příslušné předpisy Evropské unie a upravuje:
a) způsob vymezení hydrogeologických rajonů, vymezení útvarů podzemních vod,
b) způsob hodnocení stavu podzemních vod a
c) náležitosti programů zjišťování a hodnocení stavu podzemních vod.

Legislation domain - structure

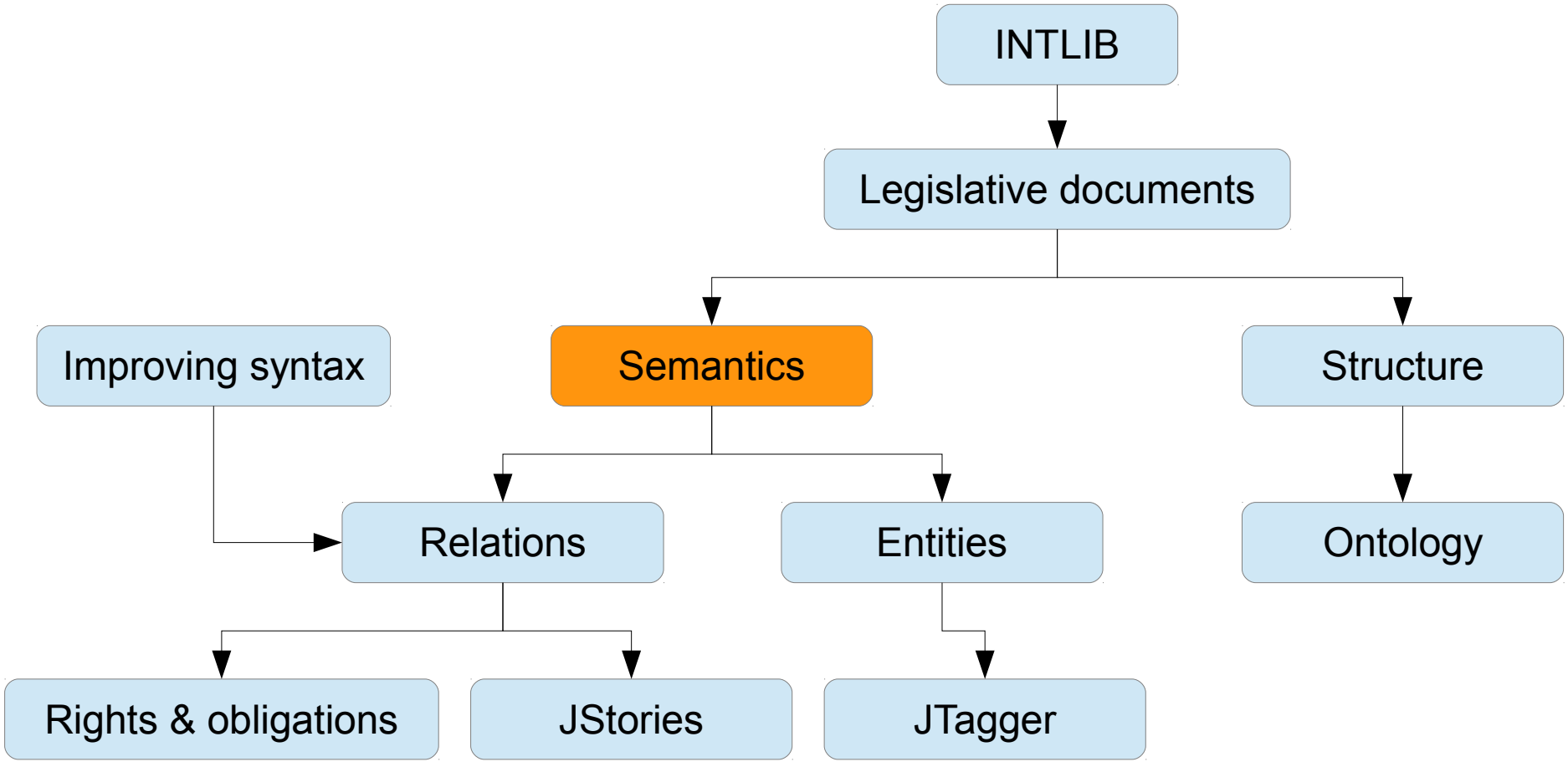
HLAVA I
ÚVODNÍ USTANOVENÍ

§ 1
Předmět úpravy
Ta
a)
b)
c)



```
<head id="11" label="HLAVA I">
        <title>ÚVODNÍ USTANOVENÍ</title>
        <section id="12" label="§ 1">
                <title>Předmět úpravy</title>
                <text>Tato vyhláška zapracovává příslušné předpisy Evropské unie a upravuje:</text>
                <section id="13" label="a)">
                        <text>způsob vymezení hydrogeologických rajonů, vymezení útvarů podzemních vod,</text>
                </section>
                <section id="14" label="b)">
                        <text>způsob hodnocení stavu podzemních vod a</text>
                </section>
                <section id="15" label="c)">
                        <text>náležitosti programů zjišťování a hodnocení stavu podzemních vod.</text>
                </section>
        </section>
</head>
```

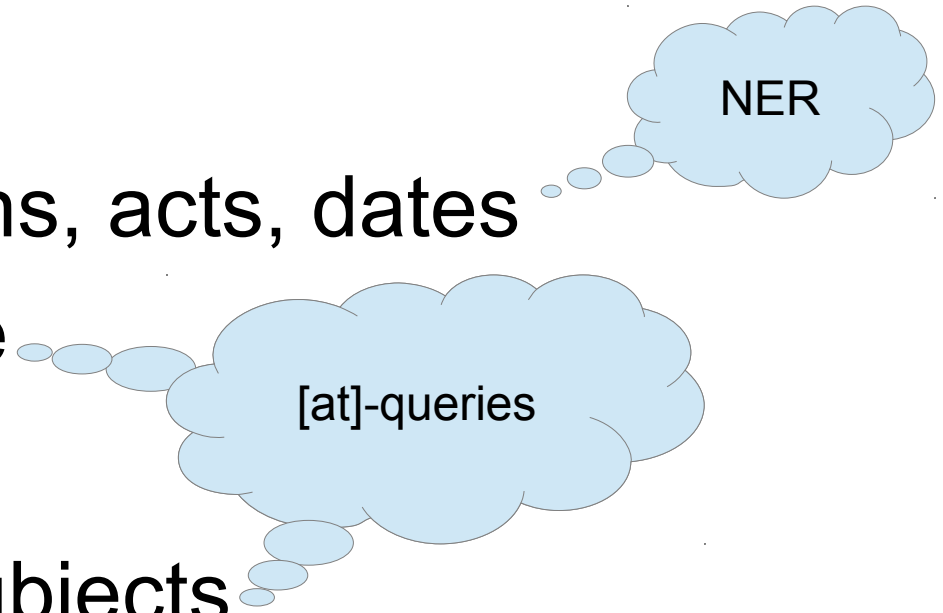
Outline


INTLIB
Legislative documents
Improving syntax
Semantics
Structure
Relations
Entities
Ontology
Rights & obligations
JStories
JTagger

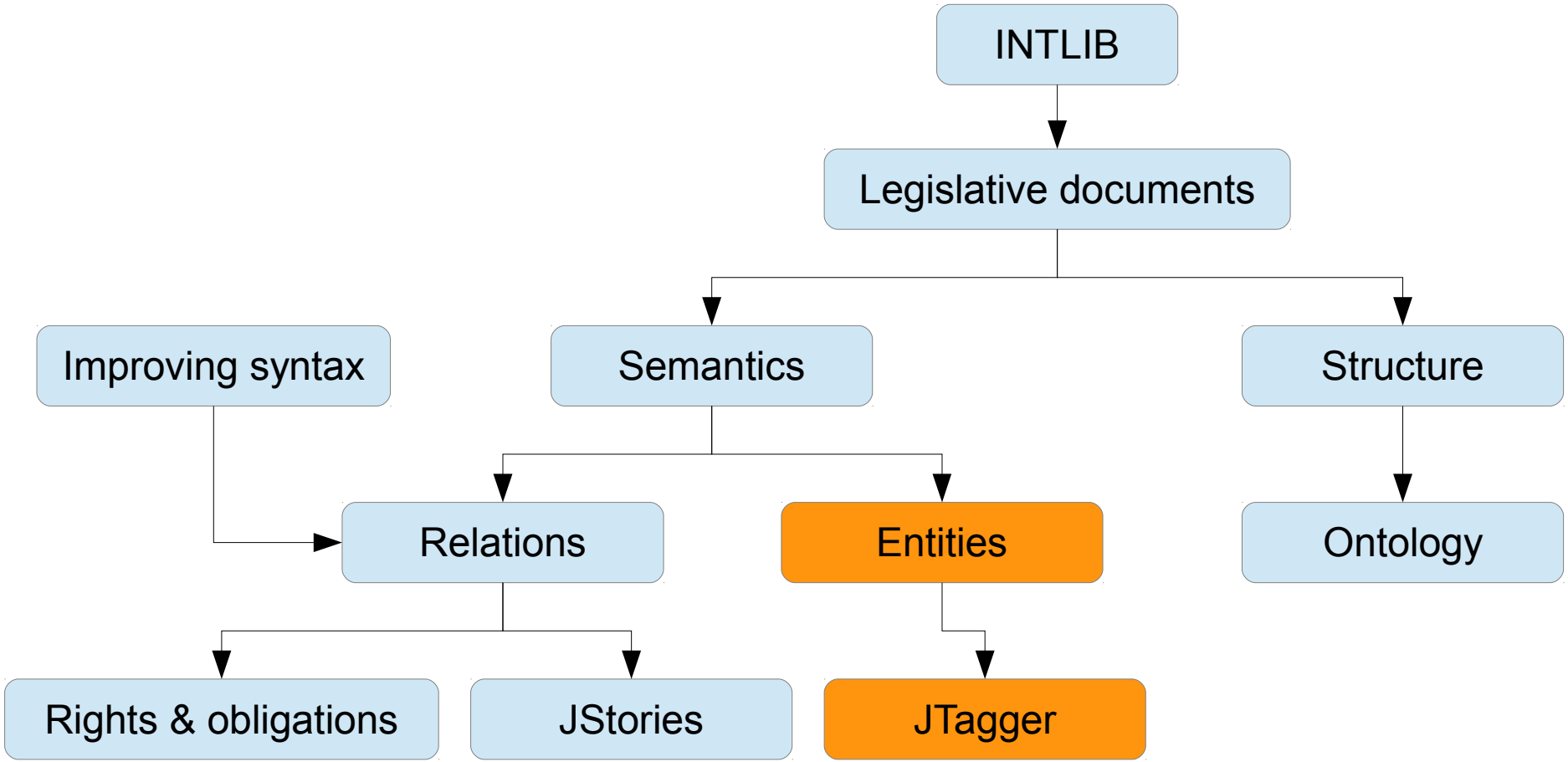
Legislation domain - semantics

Extracting concepts and relationships between them from documents

- court decisions
  - references, institutions, acts, dates
  - whole *story* of a case
- acts, regulations, ...
  - rights, obligations, subjects



Outline

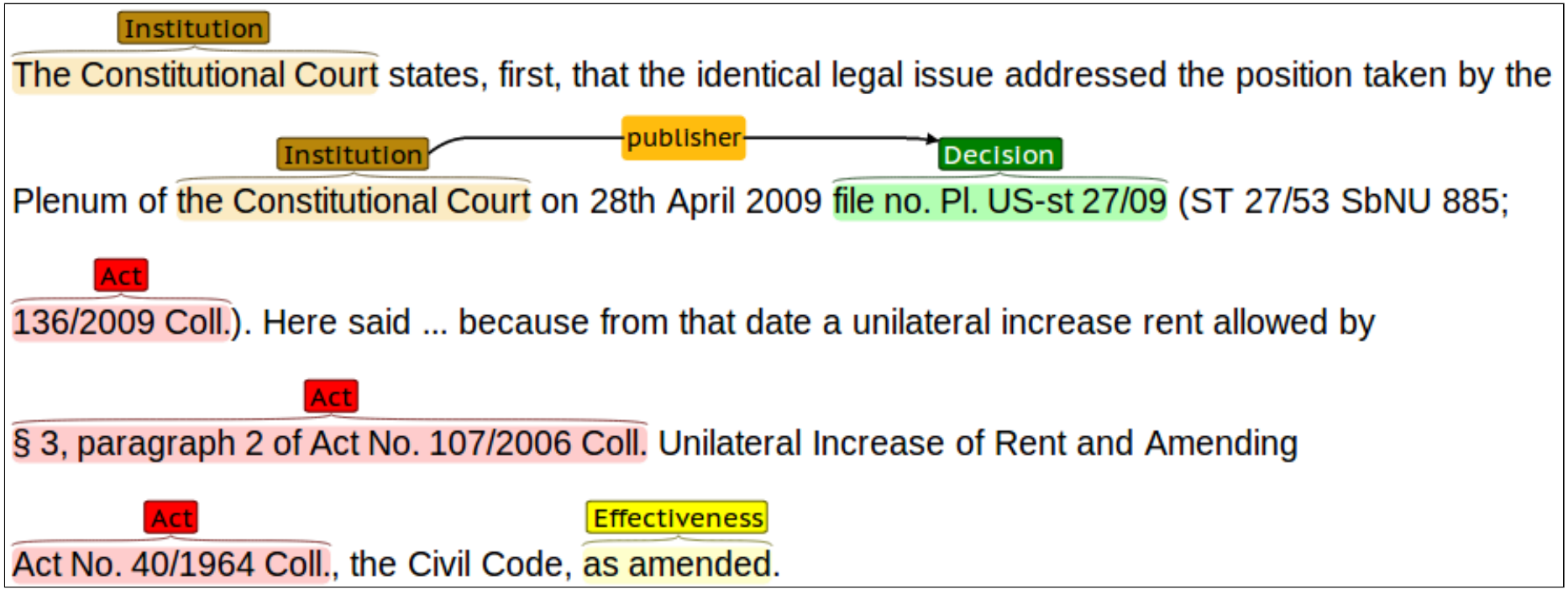

INTLIB
Legislative documents
Improving syntax
Semantics
Structure
Relations
Entities
Ontology
Rights & obligations
JStories
JTagger

JTagger

- Reference
  - Court decision (D)
  - Act (A)
- Effectiveness of Act (E)
- Institution (I)
- Relations
  - Publisher (Institution ~> Court Decision)
  - Abbreviation

JTagger

- Annotation in Brat (http://brat.nlplab.org)



JTagger

Corpus of manually annotated court decisions

- The Supreme Court (150)
- The Constitutional Court (150)

	SC			CC		
	# of docs	# of tokens	# of entities	# of docs	# of tokens	# of entities
Training set	135	332,535	8,487	135	312,191	7,910
Test set	15	36,999	943	15	34,701	879
Total	150	369,534	9,430	150	346,892	8,789

ufal.mff.cuni.cz/jtagger

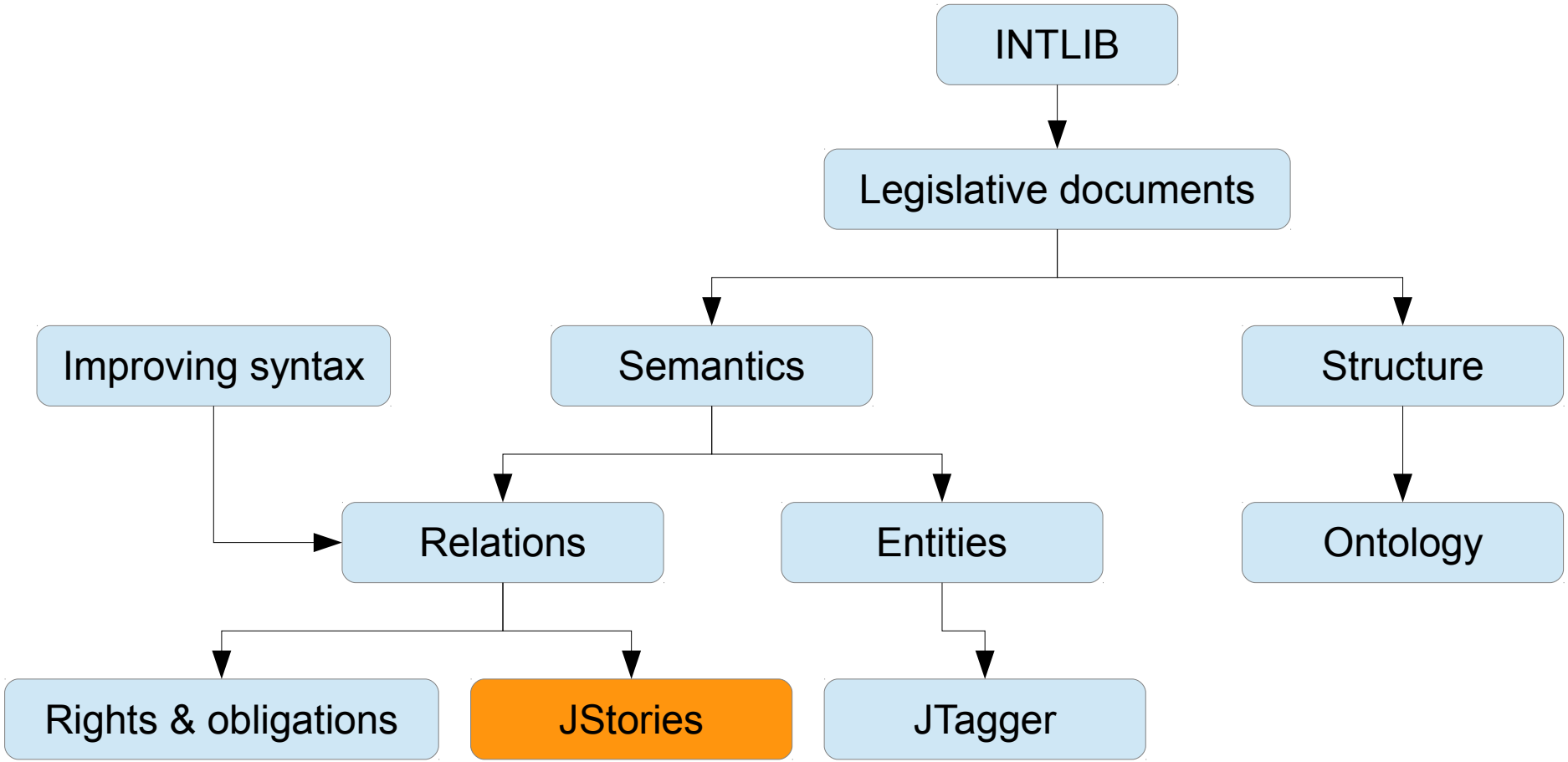
JTagger

Strict F1 on entities

	Entity	HMM	PM pos ext	PM pos	PM	PM small
SC	A	0,75±0.02	0,91±0,02	0,91±0,03	0,89±0,03	0,88±0,03
	D	0,82±0,08	0,97±0,02	0,96±0,02	0,95±0,03	0,94±0,02
	E	0,89±0,04	0,90±0,05	0,89±0,05	0,88±0,08	0,82±0,1
	I	0,92±0,03	0,96±0,02	0,96±0,02	0,95±0,02	0,96±0,02
CC	A	0,63±0,05	0,87±0,02	0,86±0,02	0,84±0,03	0,78±0,03
	D	0,83±0,05	0,95±0,03	0,95±0,03	0,93±0,03	0,92±0,03
	E	0,96±0,03	0,96±0,03	0,96±0,03	0,96±0,03	0,96±0,03
	I	0,91±0,02	0,93±0,02	0,93±0,02	0,92±0,01	0,92±0,01

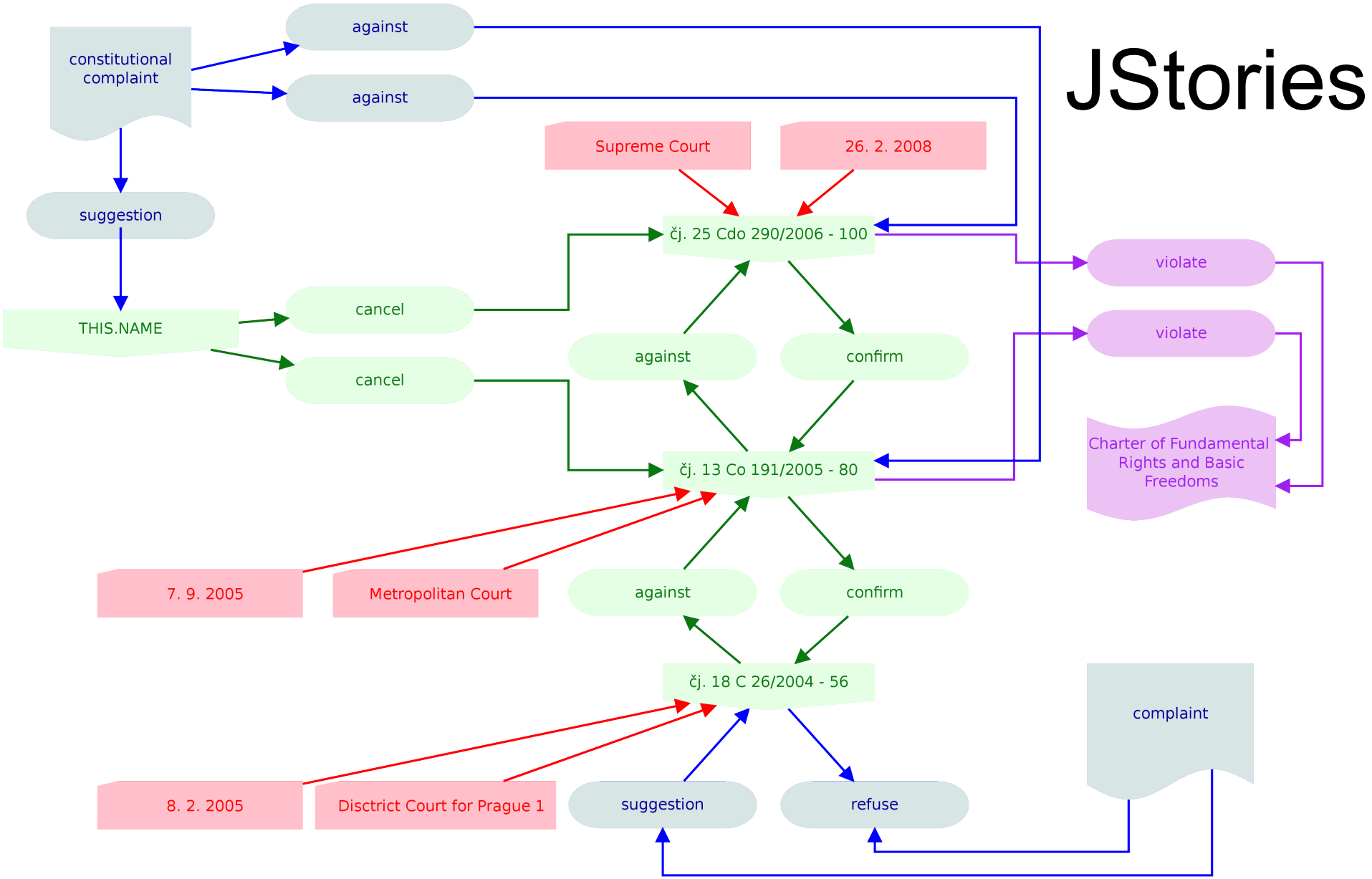
ufal.mff.cuni.cz/jtagger

Outline

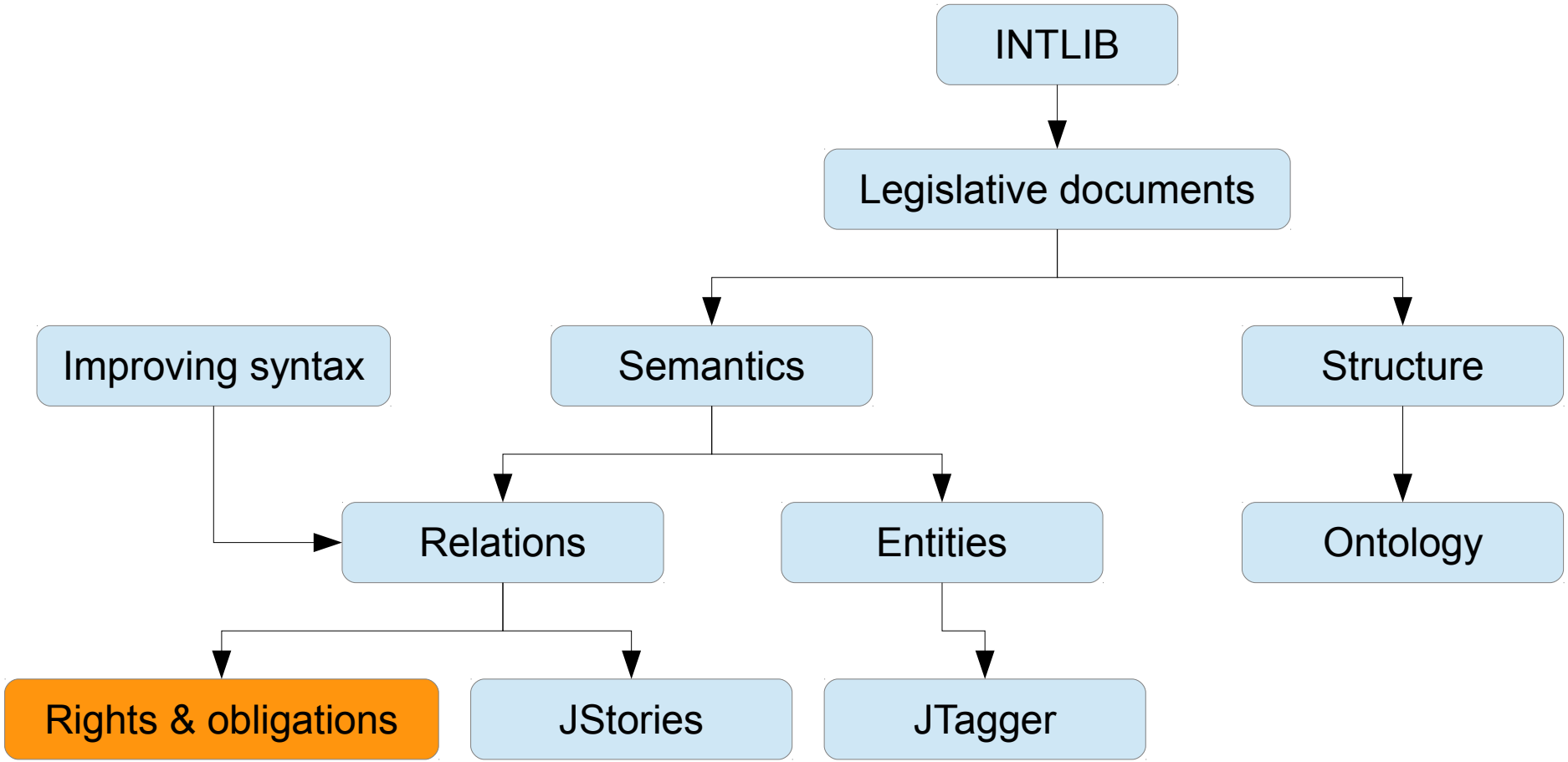

INTLIB
Legislative documents
Improving syntax
Semantics
Structure
Relations
Entities
Ontology
Rights & obligations
JStories
JTagger

JStories

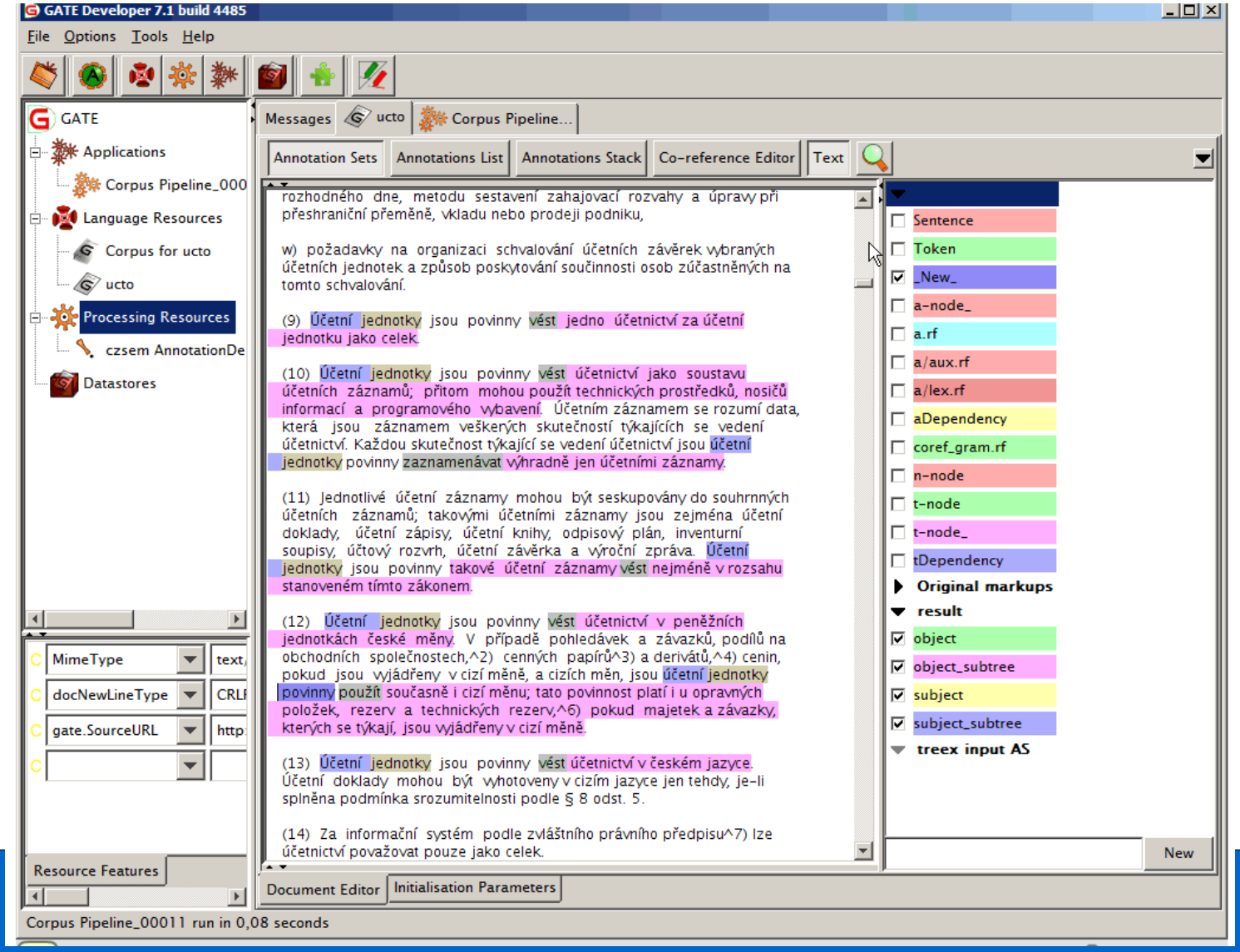
- Legal case retrospection
  - a story
  - begins when an accuser files a complaint to a court
  - ends when a final decision is rendered


JStories
constitutional complaint
against
against
suggestion
THIS.NAME
cancel
cancel
Supreme Court
26. 2. 2008
čj. 25 Cdo 290/2006 - 100
against
confirm
čj. 13 Co 191/2005 - 80
against
confirm
čj. 18 C 26/2004 - 56
7. 9. 2005
Metropolitan Court
8. 2. 2005
Disctrict Court for Prague 1
suggestion
refuse
violate
violate
Charter of Fundamental Rights and Basic Freedoms
complaint

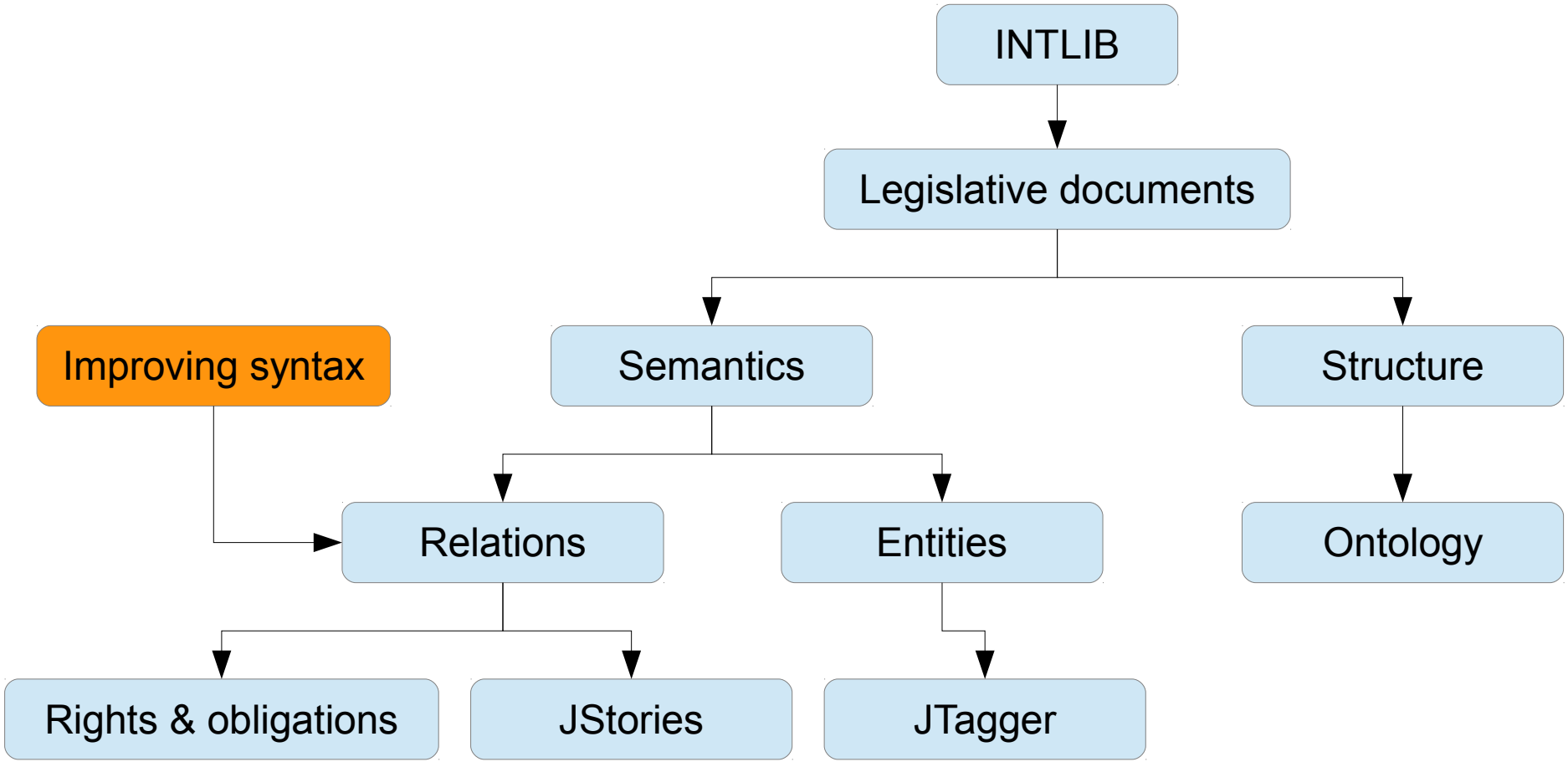
Outline


INTLIB
Legislative documents
Improving syntax
Semantics
Structure
Relations
Entities
Ontology
Rights & obligations
JStories
JTagger

Tree queries in GATE



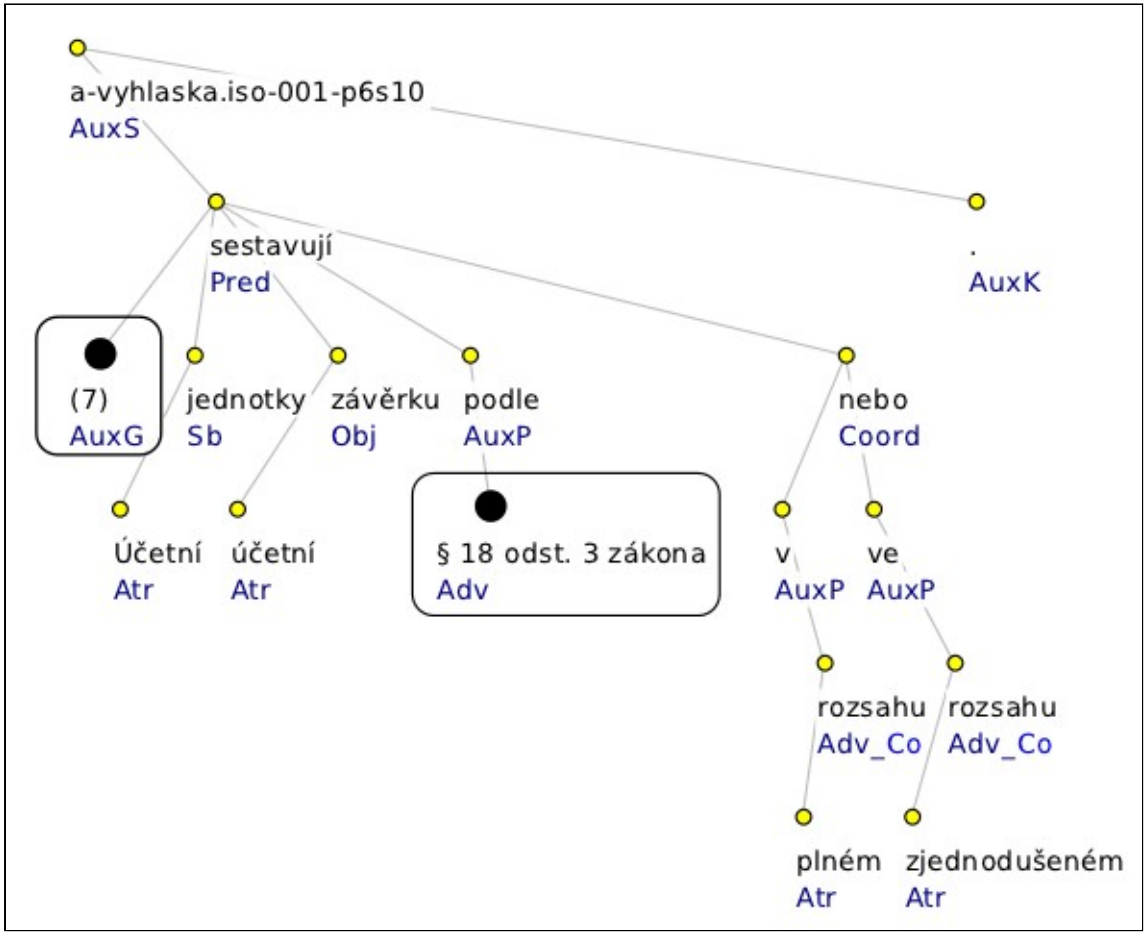
Outline


INTLIB
Legislative documents
Improving syntax
Semantics
Structure
Relations
Entities
Ontology
Rights & obligations
JStories
JTagger

Improving syntactic parsing

- Parsers trained on PDT data
  - How they work on legal texts?
  - Too many nodes make automatic parsing almost impossible
- Improving results
  - retokenization
  - new segmentation

Improving syntactic parsing



Improving syntactic parsing

New segmentation

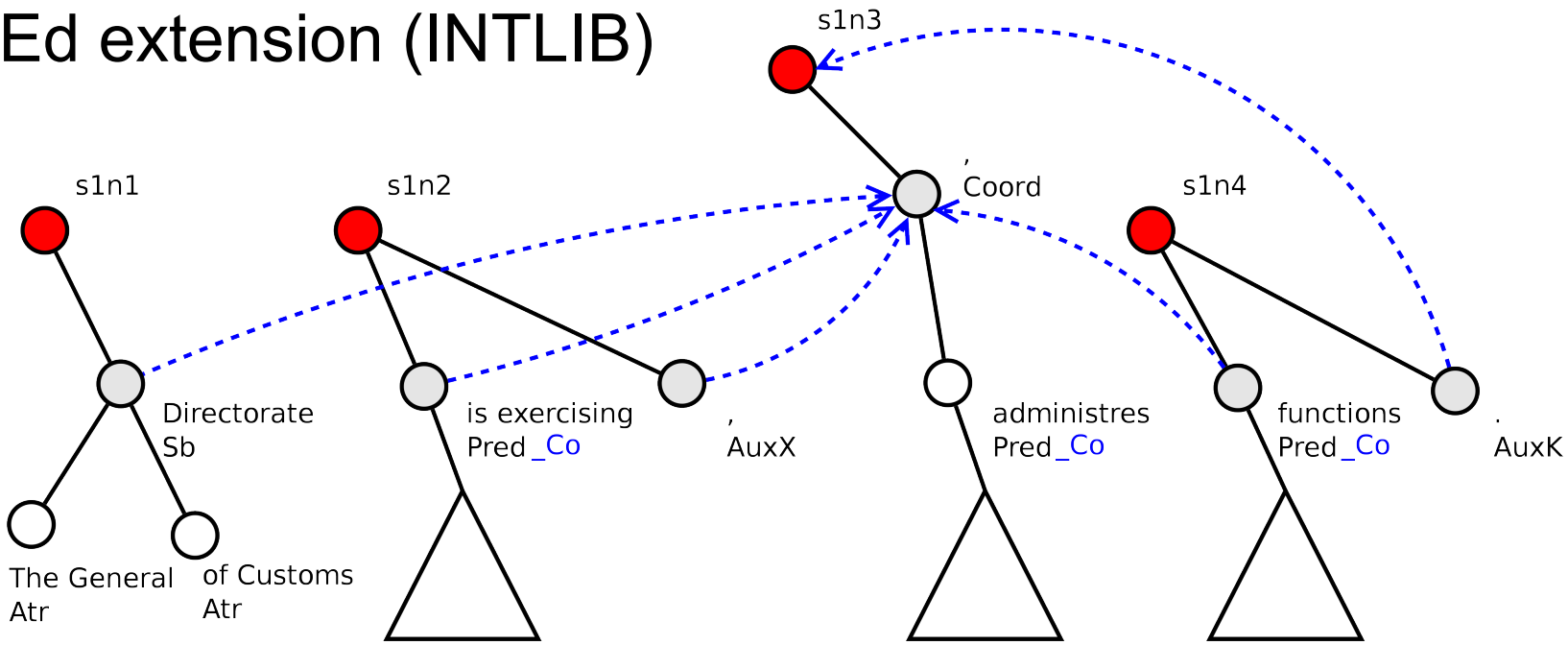
- try to split long coordination (complex sentences)
  - formal document layout
  - building simple sententences

(1) The General Directorate of Customs
a) is an administrative body exercising superior authority to customs offices,
b) administers the customs duty in compliance with the relevant EU regulation,
...
e) functions as a central analytical body analyzing risks.

Improving syntactic parsing

Formal document layout

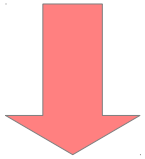
- new PML Schema (PML INTLIB)
- new TrEd extension (INTLIB)



Improving syntactic parsing

Building simple sentences

(1) The General Directorate of Customs
- a) is an administrative body exercising superior authority to customs offices,
- b) administers the customs duty in compliance with the relevant EU regulation,
- ...
- e) functions as a central analytical body analyzing risks.



The General Directorate of Customs is an administrative body exercising superior authority to customs offices.

The General Directorate of Customs administers the customs duty in compliance with the relevant EU regulation.

The General Directorate of Customs functions as a central analytical body analyzing risks.

Conclusion

JTagger

- manuall annotation
- references, entities

JStories

- specification

Legislative Ontology

- XML Schemas
- parsers TXT → XML

Syntactic parsing

- experiments with split sentences