



Vincent Kríž

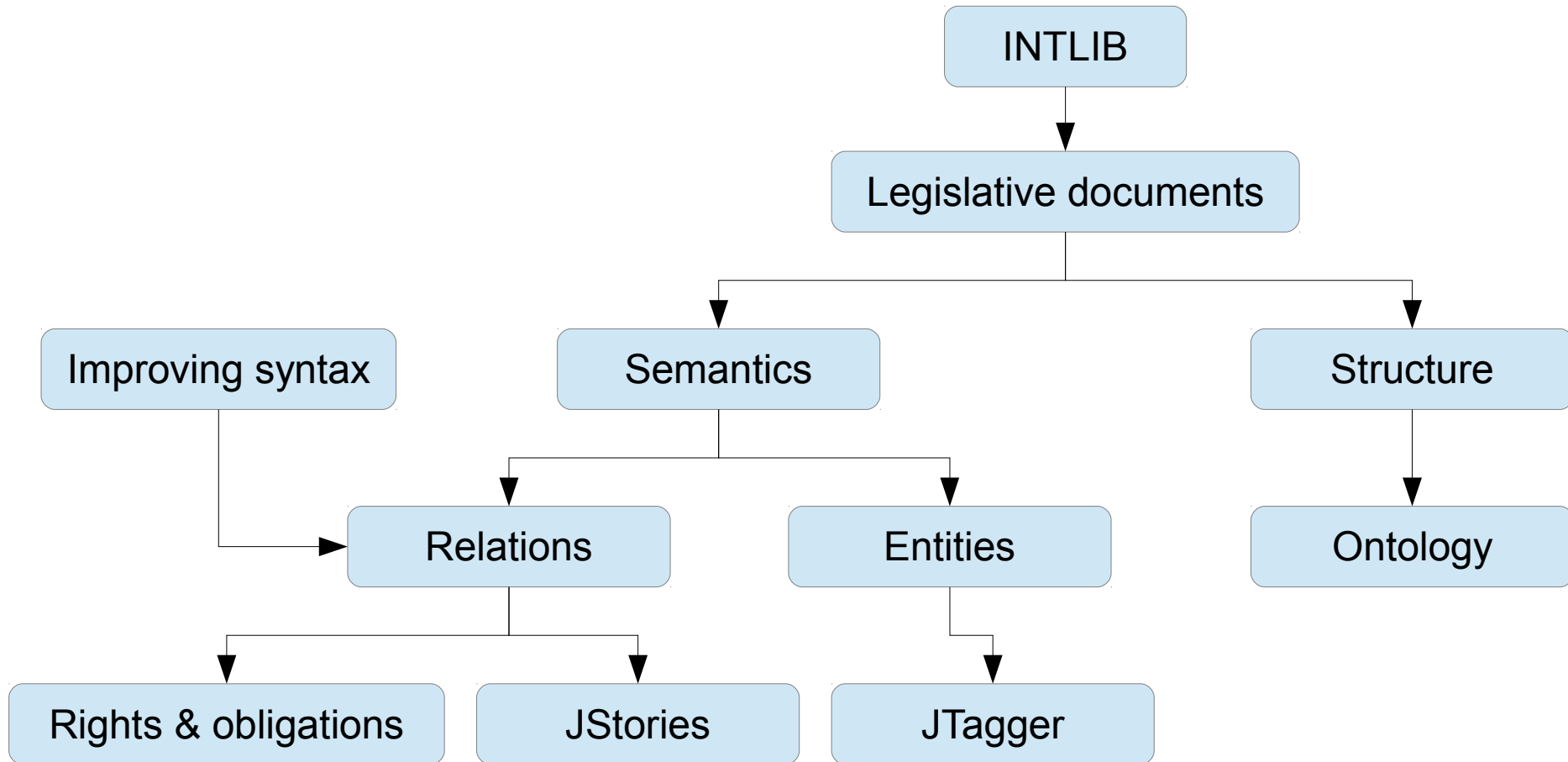
Detecting Semantic Relations in Texts and Their Integration with External Data Resources

Week of Doctoral Students 2013

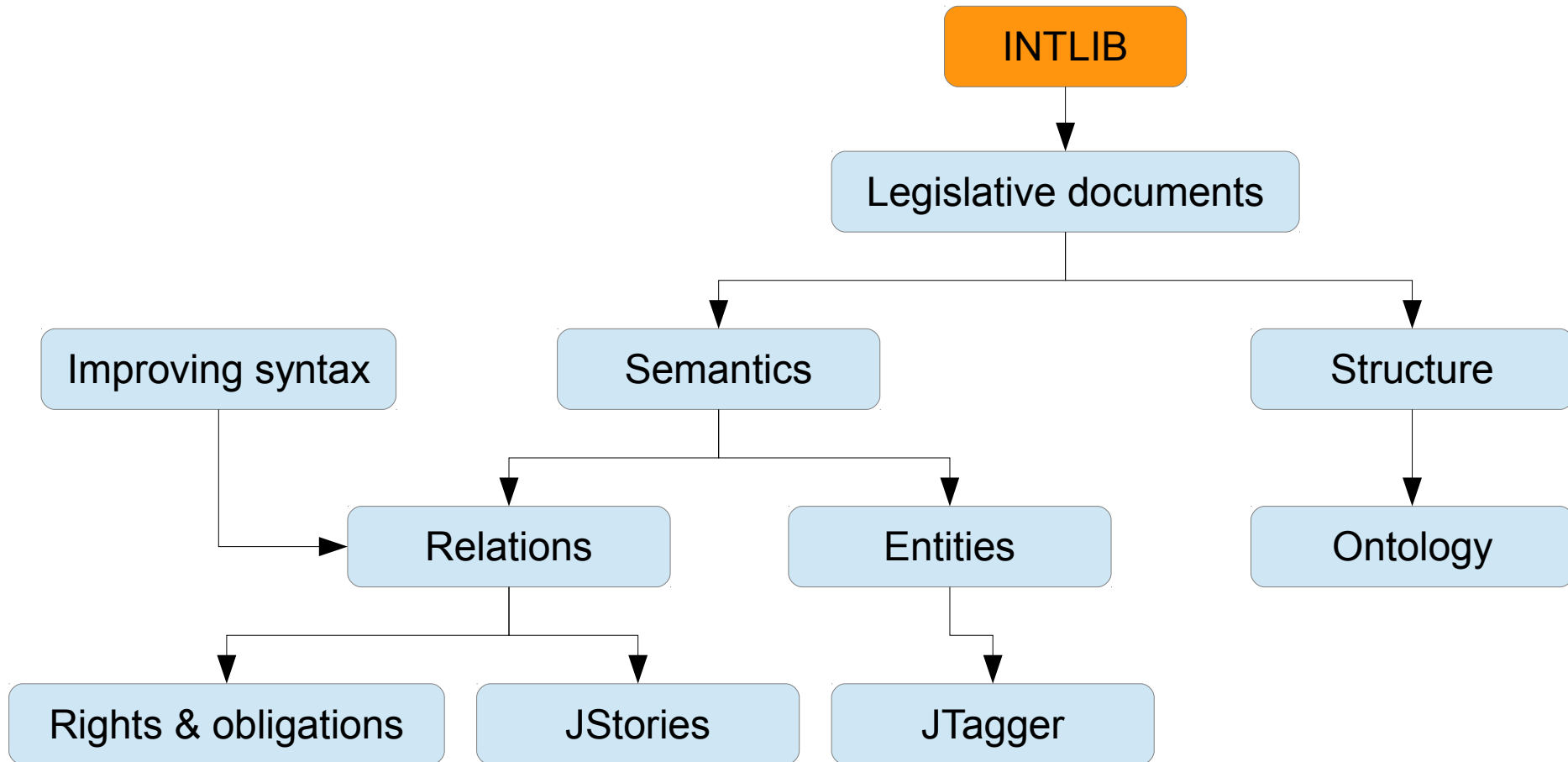
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University in Prague
Czech Republic

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

Outline



Outline

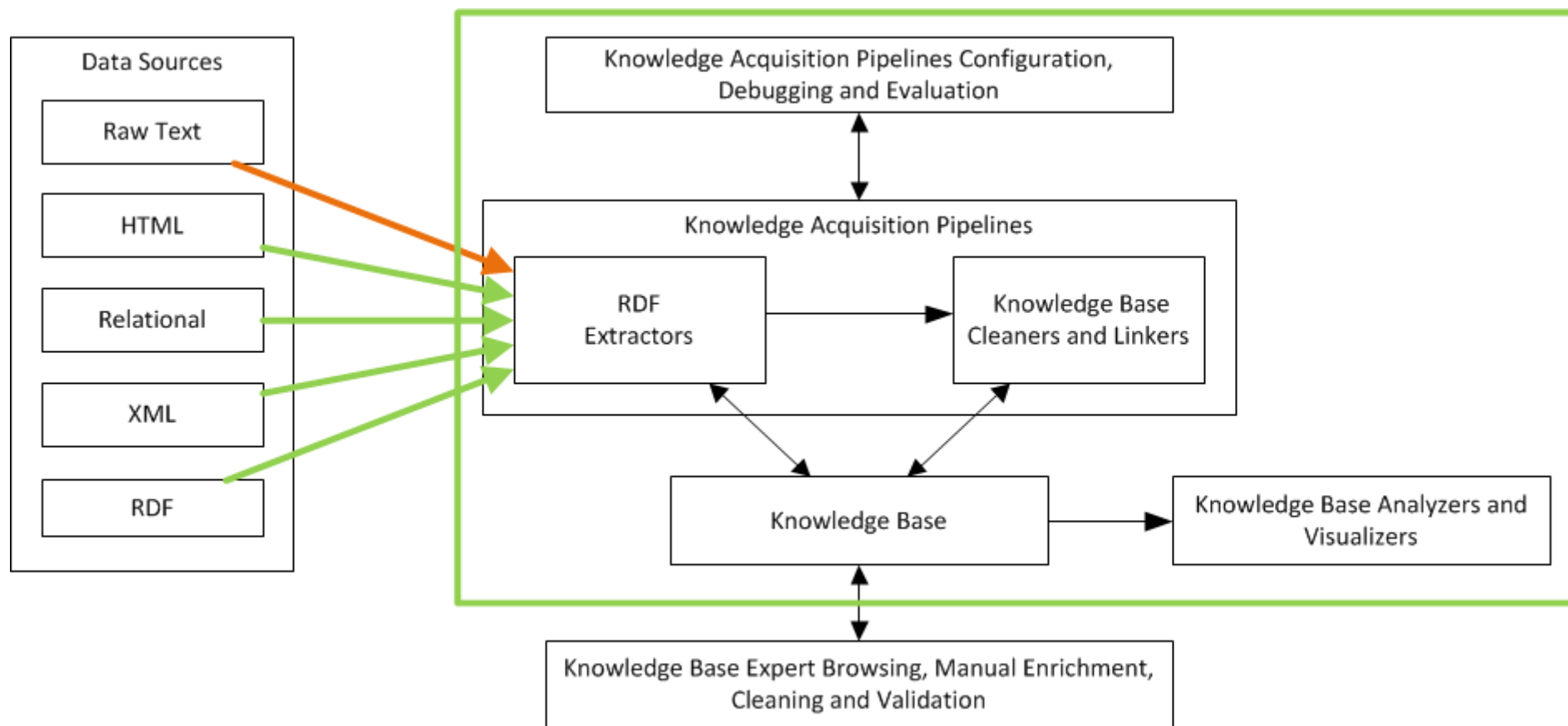


Motivation

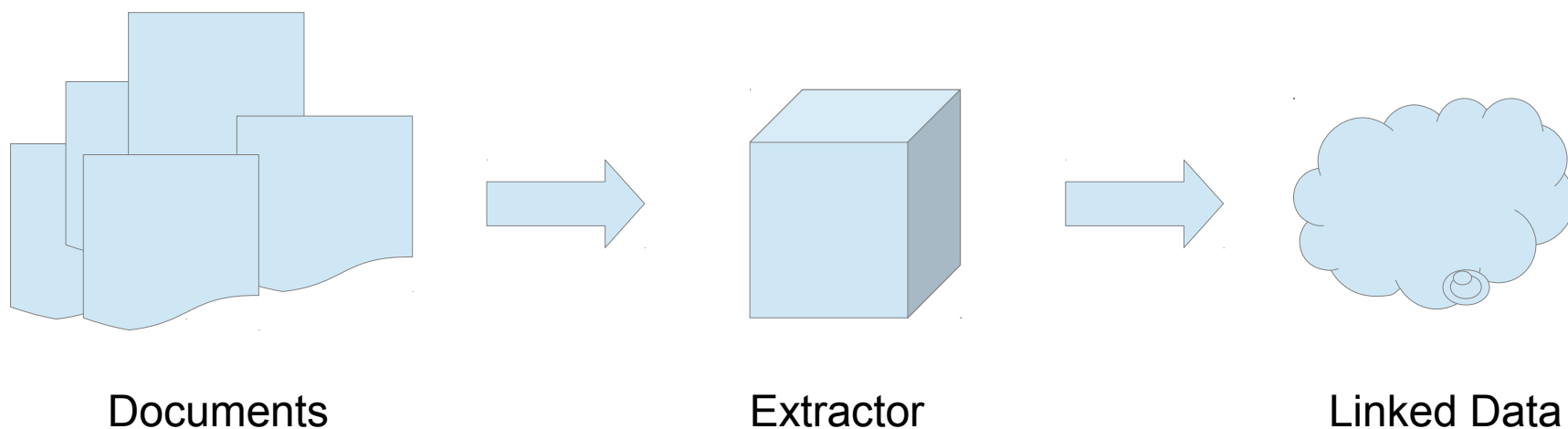
- large collections of documents
- efficient browsing & querying
- typical approaches
 - full-text search
 - metadata search
- interpret the semantics of the documents →
suitable DB & query language →
user-friendly browsing & querying



INTLIB



NLP Group



Documents

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

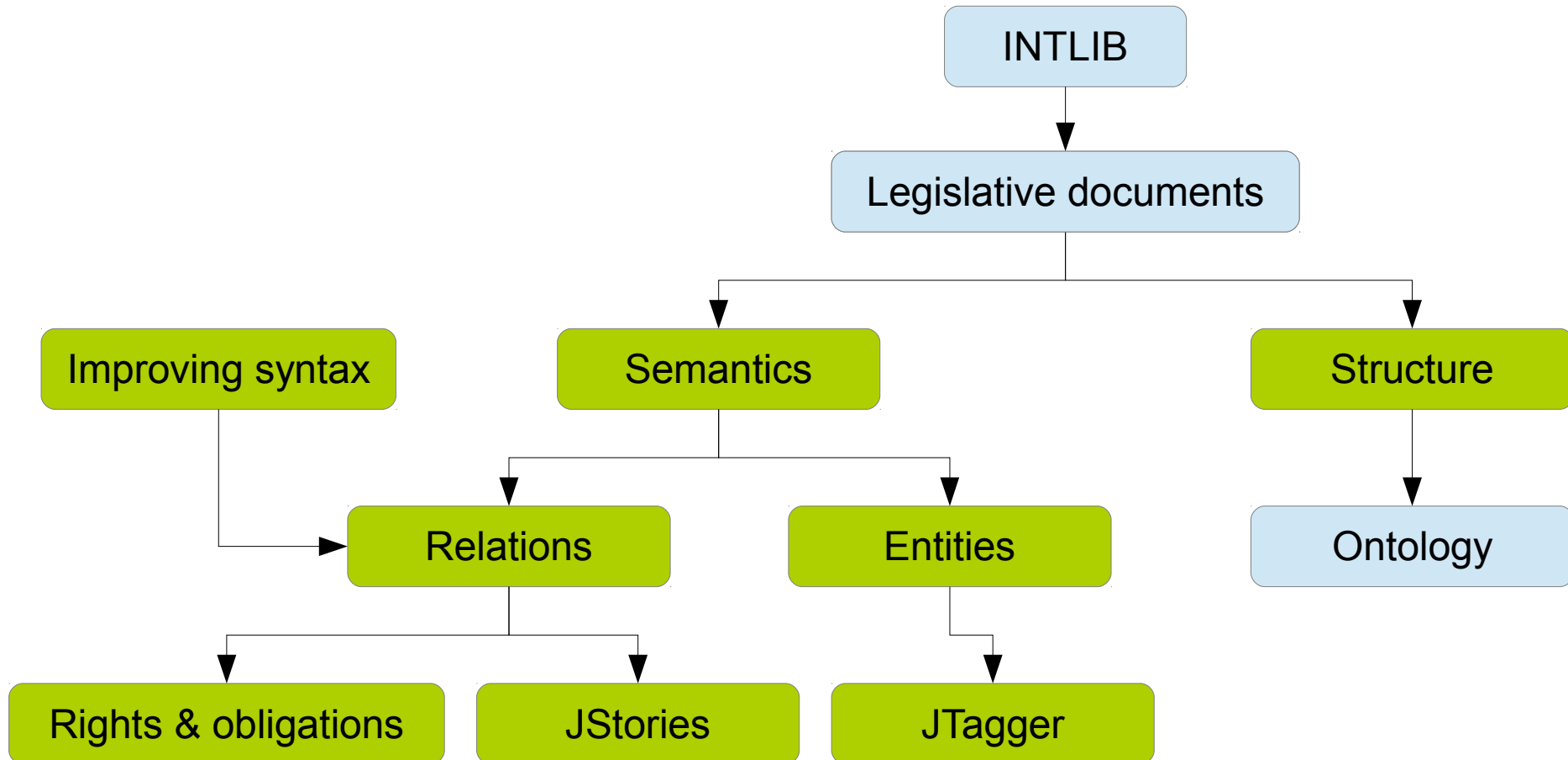
Extractor

- NLP techniques

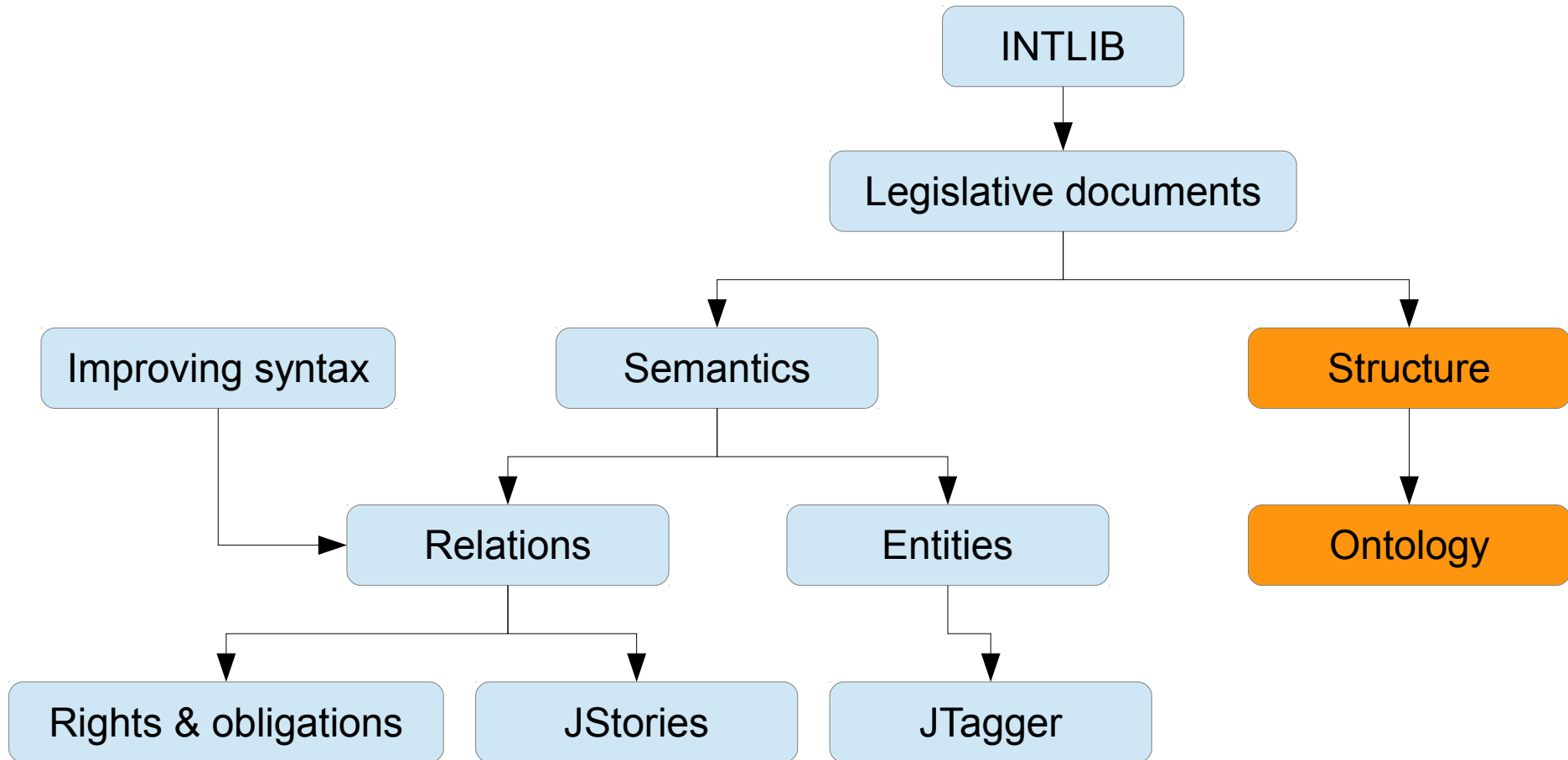
Linked data

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

Contribution



Outline



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

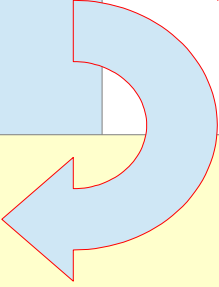
Tato vyhláška zpracovává příslušné předpisy Evropské unie a upravuje:

- 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

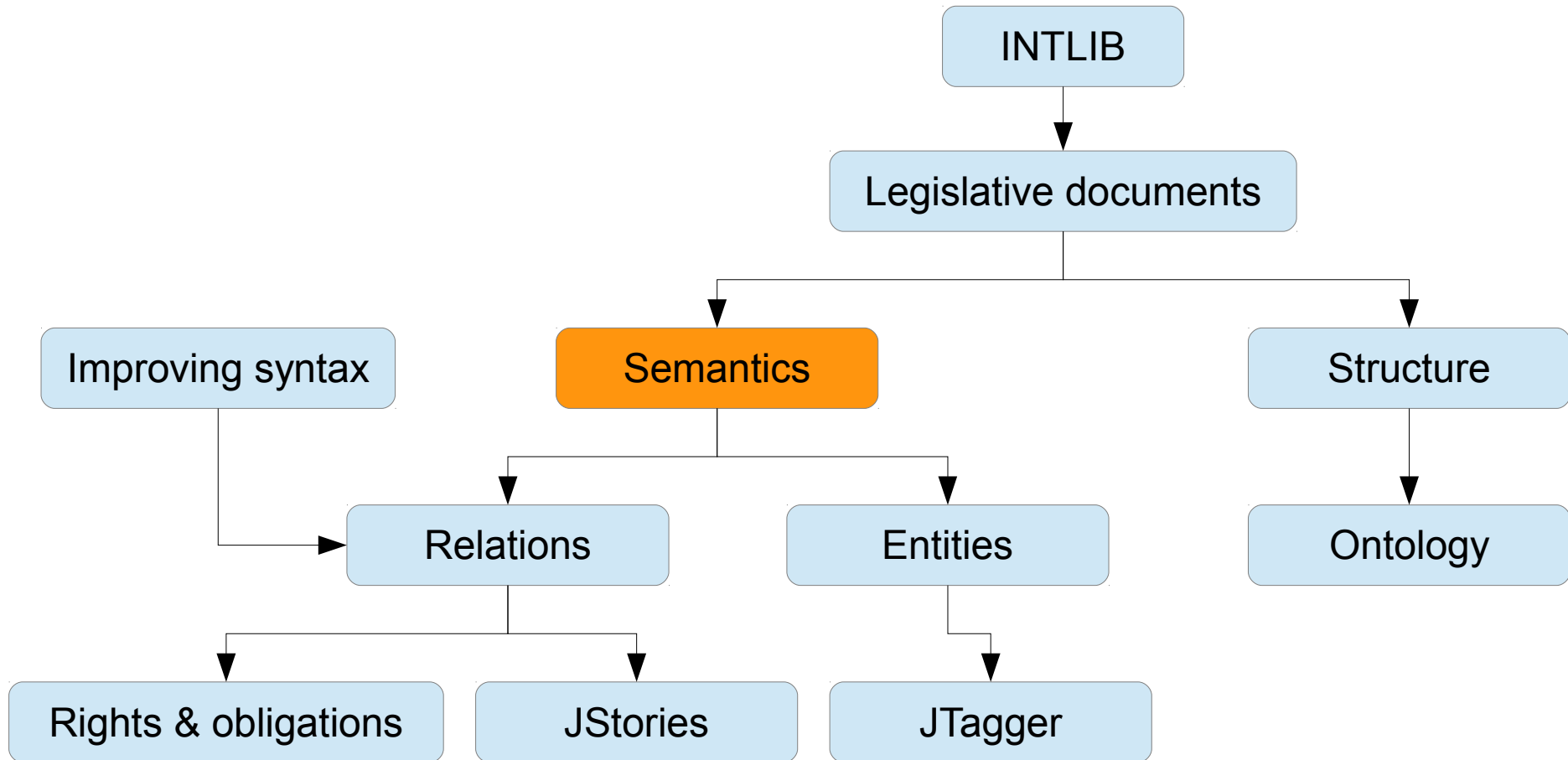
HLAVA I
ÚVODNÍ USTANOVENÍ

§ 1
Předmět úprav



```
<head id="11" label="HLAVA I">
a) <title>ÚVODNÍ USTANOVENÍ</title>
b) <section id="12" label="§ 1">
c) <title>Předmět úpravy</title>
<text>Tato vyhláška zpracovává 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



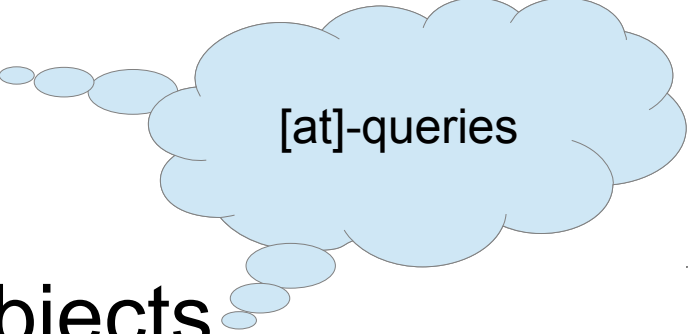
Legislation domain - semantics

Extracting concepts and relationships between them from documents

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

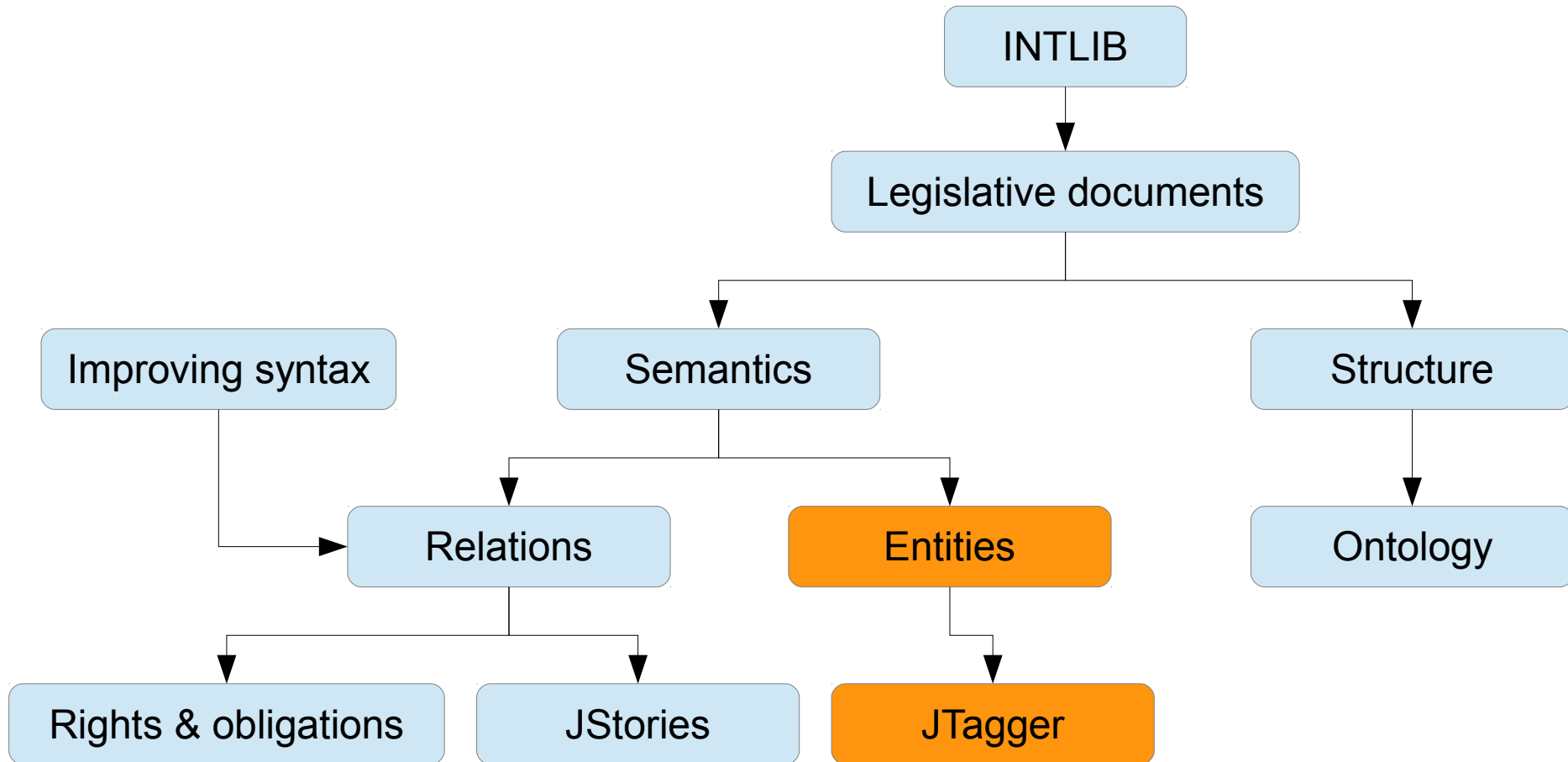


NER



[at]-queries

Outline



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>)

The Constitutional Court states, first, that the identical legal issue addressed the position taken by the Plenum of the Constitutional Court on 28th April 2009 file no. Pl. US-st 27/09 (ST 27/53 SbNU 885; 136/2009 Coll.). Here said ... because from that date a unilateral increase rent allowed by § 3, paragraph 2 of Act No. 107/2006 Coll. Unilateral Increase of Rent and Amending Act No. 40/1964 Coll., the Civil Code, as amended.

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

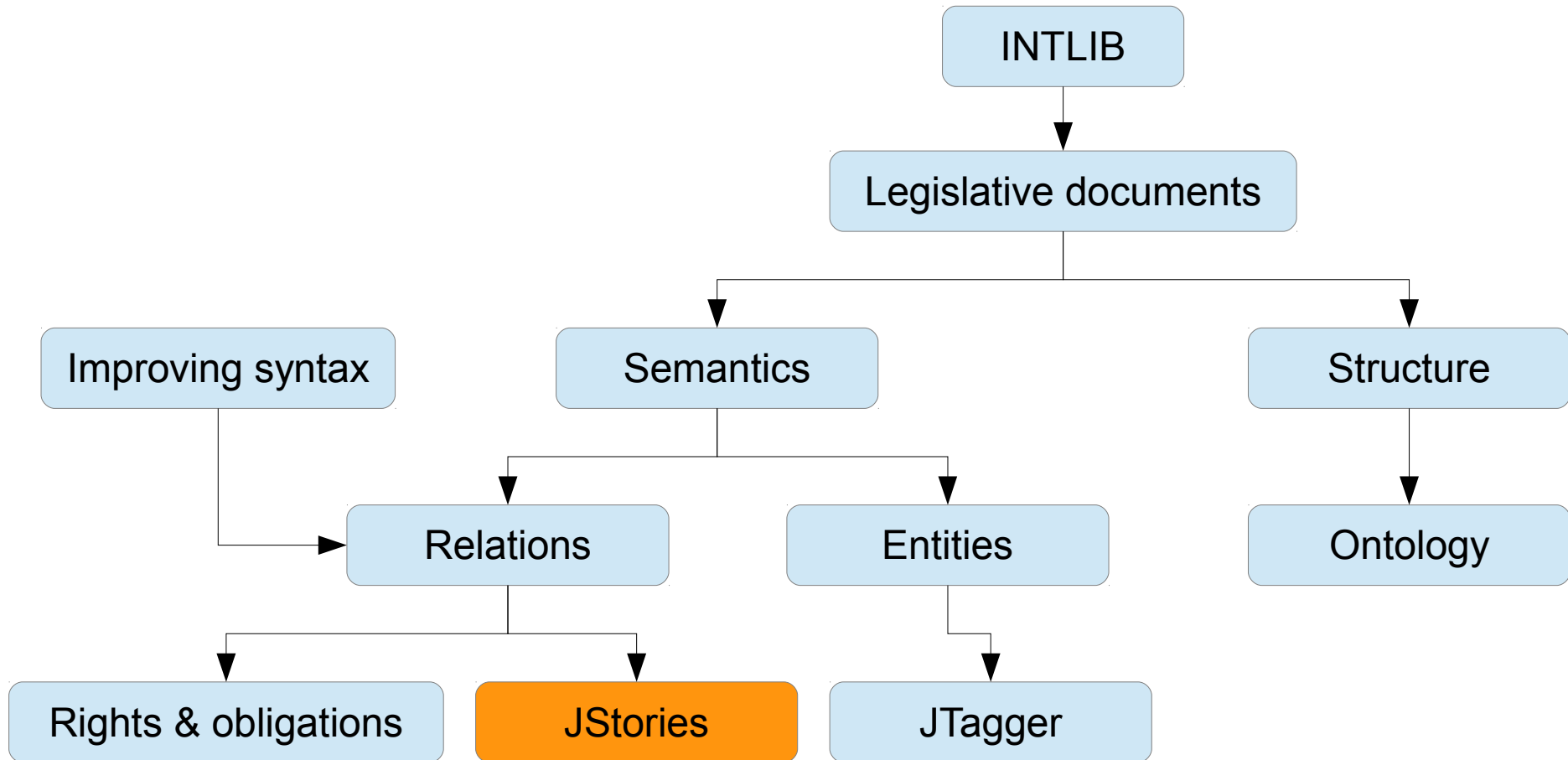
JTagger

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

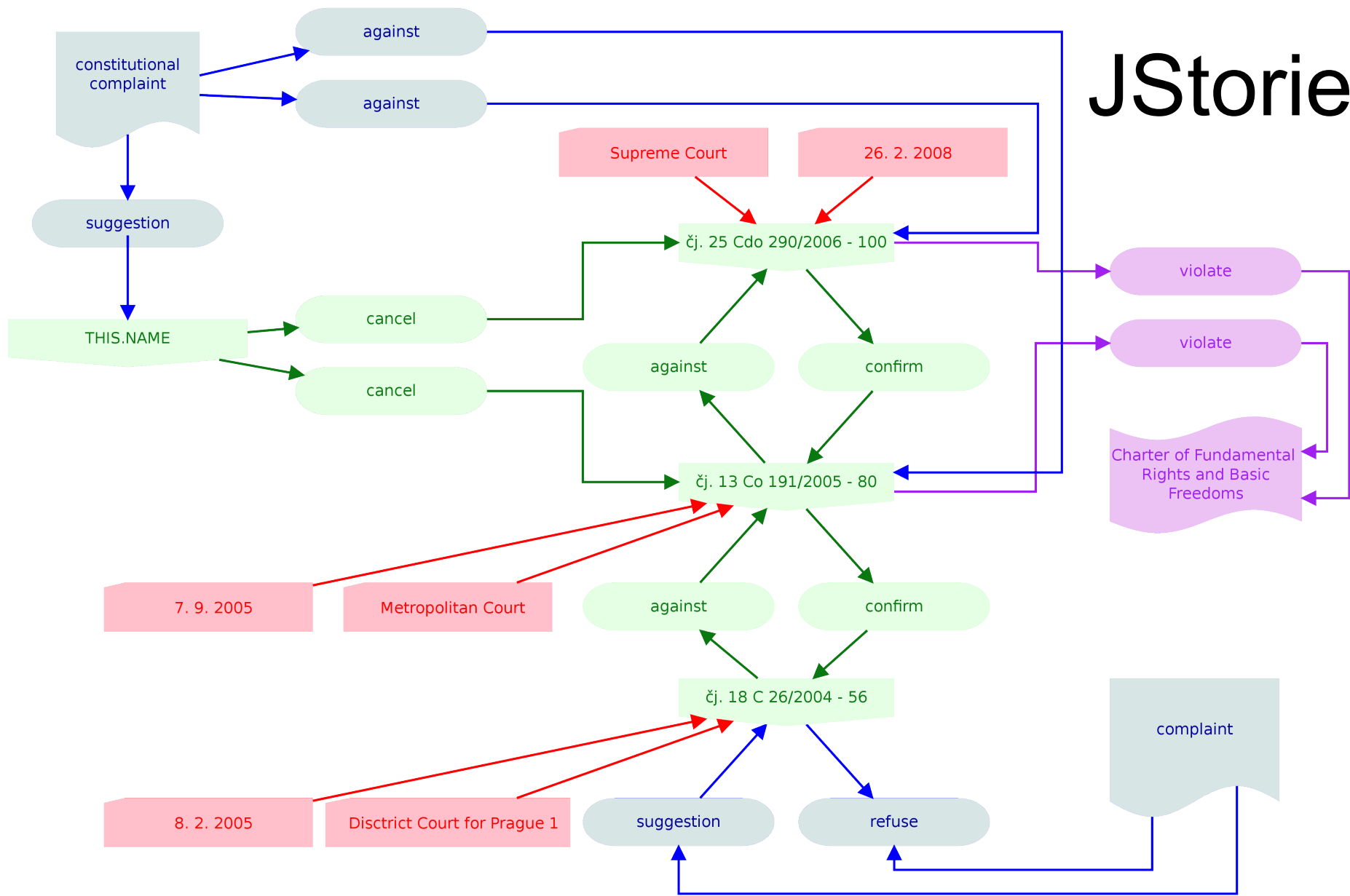
Outline



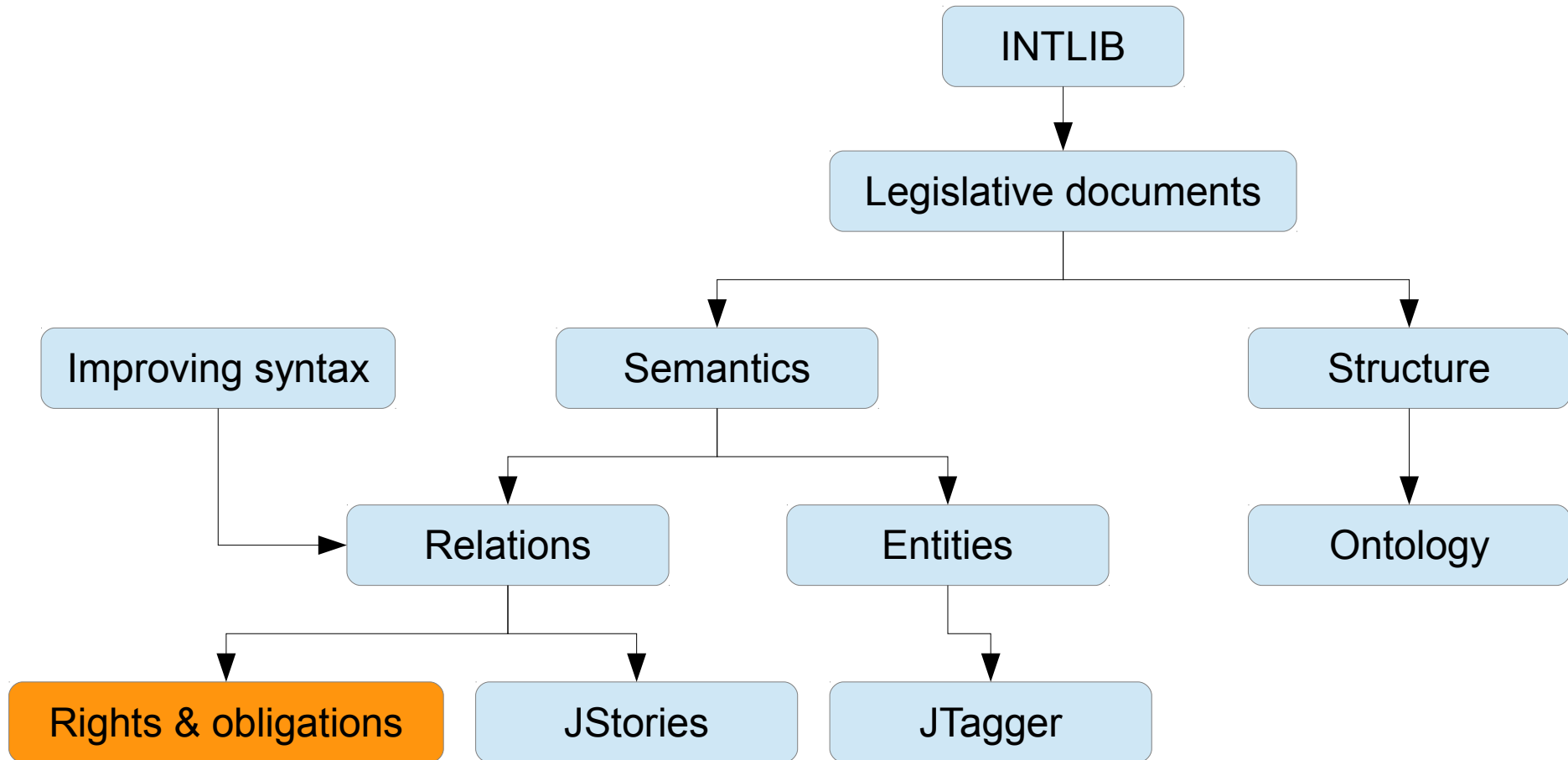
JStories

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

JStories



Outline



Tree queries in GATE

The screenshot displays the GATE Developer interface. The main window shows a text document with several paragraphs of text. The text is annotated with colored highlights, indicating the results of a tree query. The right-hand panel shows the tree query configuration, with a list of nodes and their corresponding colors. The 'tree input AS' is selected, and the 'result' section is expanded to show the 'object', 'object_subtree', and 'subject' nodes.

Text Document Content:

rozhodného dne, metodu sestavení zahajovací rozvahy a úpravy při přeshraniční přeměně, vkladu nebo prodeji podniku,

w) požadavky na organizaci schvalování účetních závěrek vybraných účetních jednotek a způsob poskytování součinnosti osob zúčastněných na tomto schvalování.

(9) Účetní jednotky jsou povinny vést jedno účetnictví za účetní jednotku jako celek.

(10) Účetní jednotky jsou povinny vést účetnictví jako soustavu účetních záznamů; přitom mohou použít technických prostředků, nosičů informací a programového vybavení. Účetním záznamem se rozumí data, která jsou záznamem veškerých skutečností týkajících se vedení účetnictví. Každou skutečnost týkající se vedení účetnictví jsou účetní jednotky povinny zaznamenávat výhradně jen účetními záznamy.

(11) Jednotlivé účetní záznamy mohou být seskupovány do souhrnných účetních záznamů; takovými účetními záznamy jsou zejména účetní doklady, účetní zápisy, účetní knihy, odpisový plán, inventurní soupisy, účtový rozvrh, účetní závěrka a výroční zpráva. Účetní jednotky jsou povinny takové účetní záznamy vést nejméně v rozsahu stanoveném tímto zákonem.

(12) Účetní jednotky jsou povinny vést účetnictví v peněžních jednotkách české měny. V případě pohledávek a závazků, podílů na obchodních společnostech, cenných papírů a derivátů cenin, pokud jsou vyjádřeny v cizí měně, a cizích měn, jsou účetní jednotky povinny použít současně i cizí měnu; tato povinnost platí i u opravných položek, rezerv a technických rezerv, pokud majetek a závazky, kterých se týkají, jsou vyjádřeny v cizí měně.

(13) Účetní jednotky jsou povinny vést účetnictví v českém jazyce. Účetní doklady mohou být vyhotoveny v cizím jazyce jen tehdy, je-li splněna podmínka srozumitelnosti podle § 8 odst. 5.

(14) Za informační systém podle zvláštního právního předpisu lze účetnictví považovat pouze jako celek.

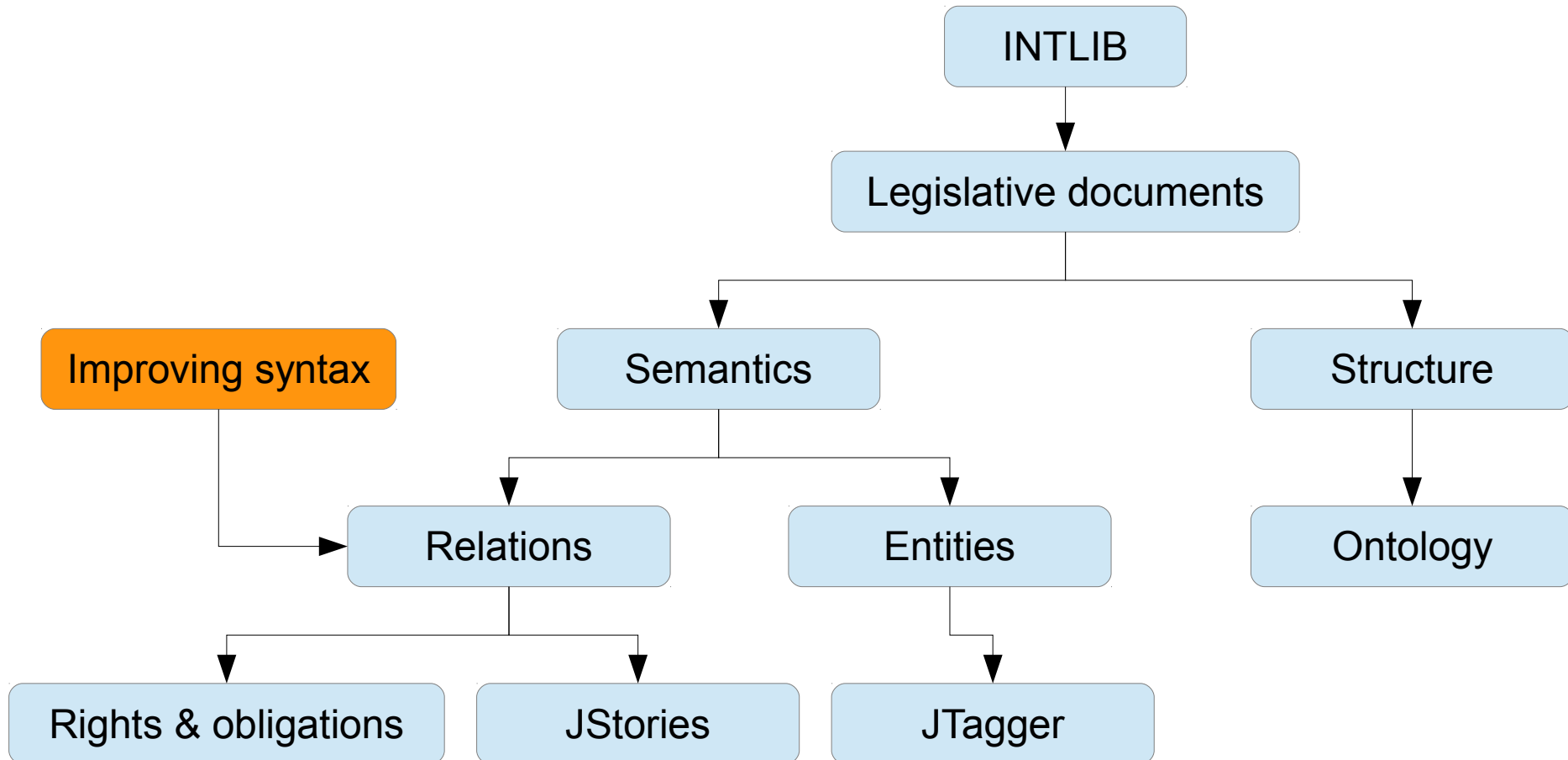
Tree Query Configuration:

- Sentence
- Token
- _New_
- a-node_
- a.rf
- a/aux.rf
- a/lex.rf
- aDependency
- coref_gram.rf
- n-node
- t-node
- t-node_
- tDependency
- Original markups
- result
 - object
 - object_subtree
 - subject
 - subject_subtree
- treex input AS

Resource Features: Mime type (text), docNewLineType (CRLF), gate.SourceURL (http).

Corpus Pipeline_00011 run in 0,08 seconds

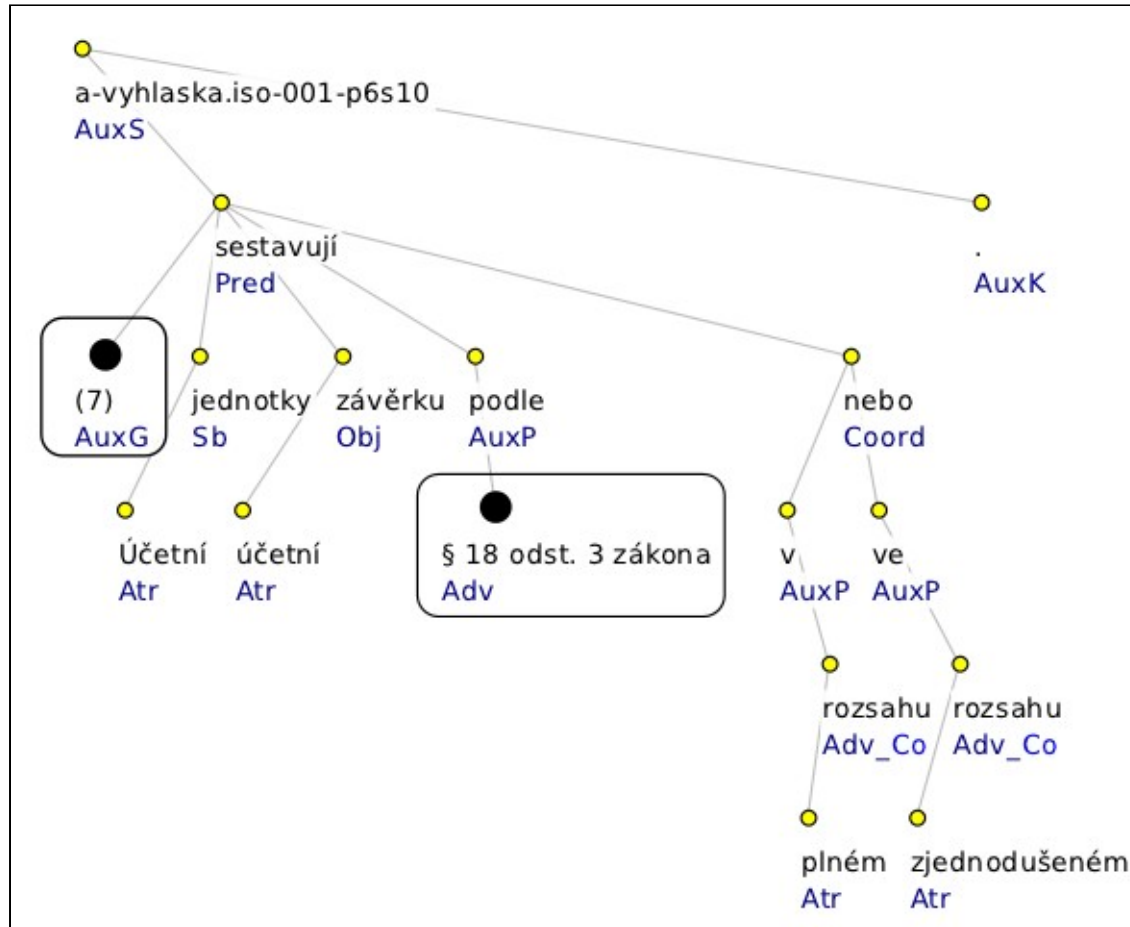
Outline



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 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,

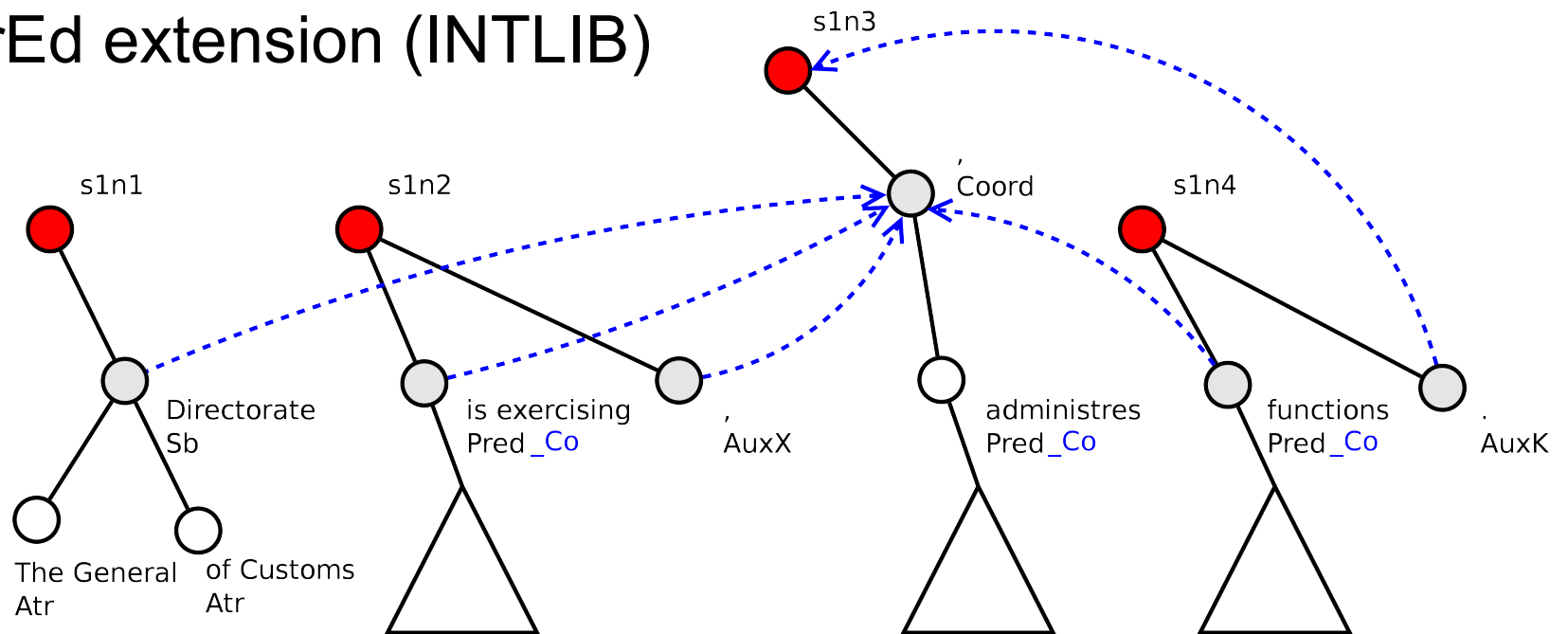
...

e) functions as a central analytical body analyzing risks.

Improving syntactic parsing

Formal document layout

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



Improving syntactic parsing

Building simple sentences

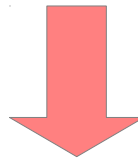
(1) The General Directorate of Customs

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

- manual annotation
- references, entities

JStories

- specification

Legislative Ontology

- XML Schemas
- parsers TXT → XML

Syntactic parsing

- experiments with split sentences