From Ontologies to Information Extraction and Back

LAMA SAEEDA

DEPARTMENT OF CYBERNETICS AND ARTIFICIAL INTELLIGENCE

FACULTY OF ELECTRICAL ENGINEERING

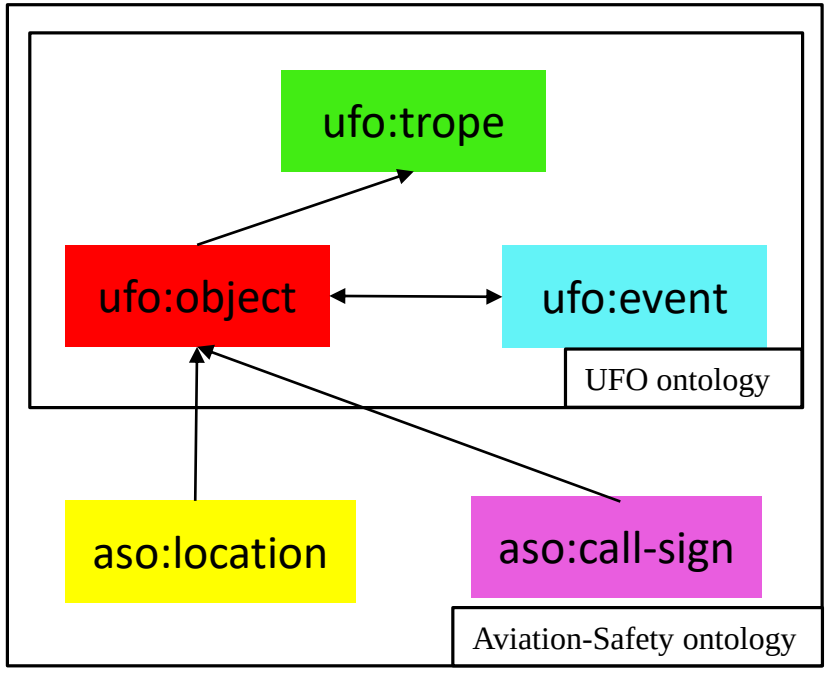
CZECH TECHNICAL UNIVERSITY IN PRAGUE

Content

- IE and OBIE
- Motivation
- Information extraction and Ontology learning
- Applications
  - Aviation safety
  - Dataset exploration
- Conclusion

Information extraction and ontology-based information extraction

Flight had a prolonged loss of communication over Swiss territory. Zurich radar informed at 09.28.39 about loss of contact and that also no contact on 121,5 MHz could be established. Geneva informed 09.46.35 that ABC1234 has contacted them. Length of loss of comm. is approx. 11 minutes.



“A little semantics goes a long way”

Ontology-based information extraction

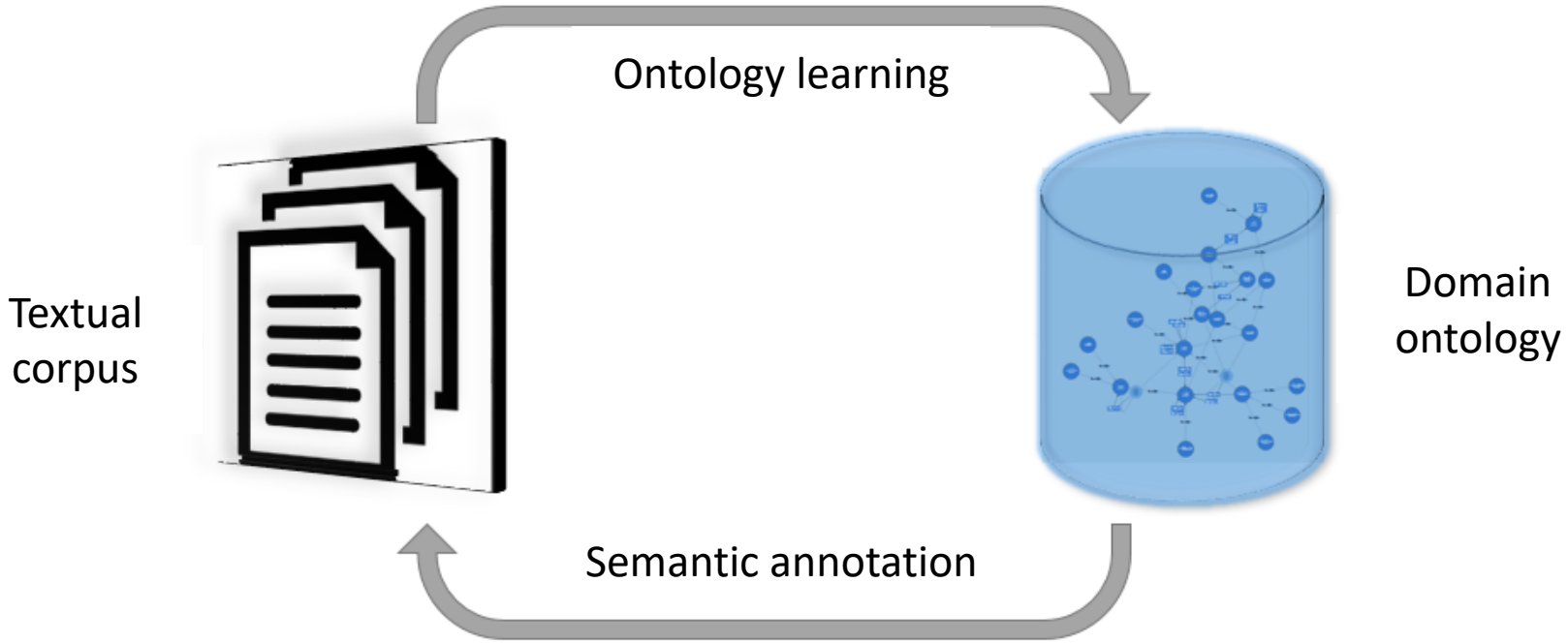
Challenges:

- Availability of domain ontology
- Lack of complete knowledge graph that would serve all purposes (DBpedia)

Ontology is essential in IE from textual corpus, and enriching the domain ontology will definitely enhance the process!

Information extraction and semi-structured ontology learning iterative methodology

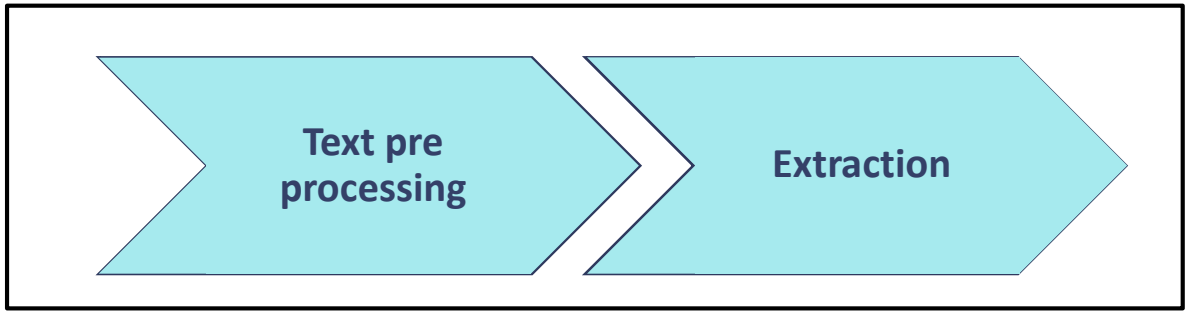
The iterative methodology



More extracted information leads to a richer ontology

Richer ontology leads to extracting more information

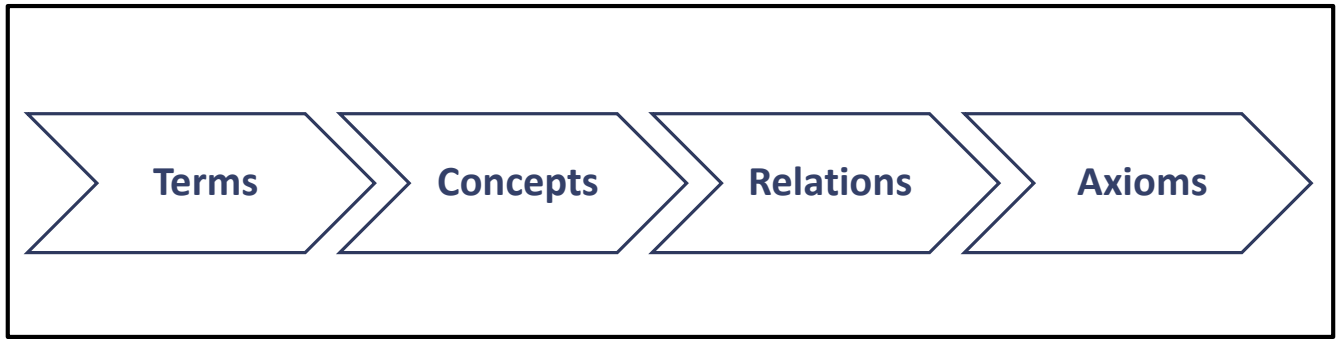
Information extraction



Pre-processing by removing stop words, POS tagging

Rule-based / Machine Learning-based extraction

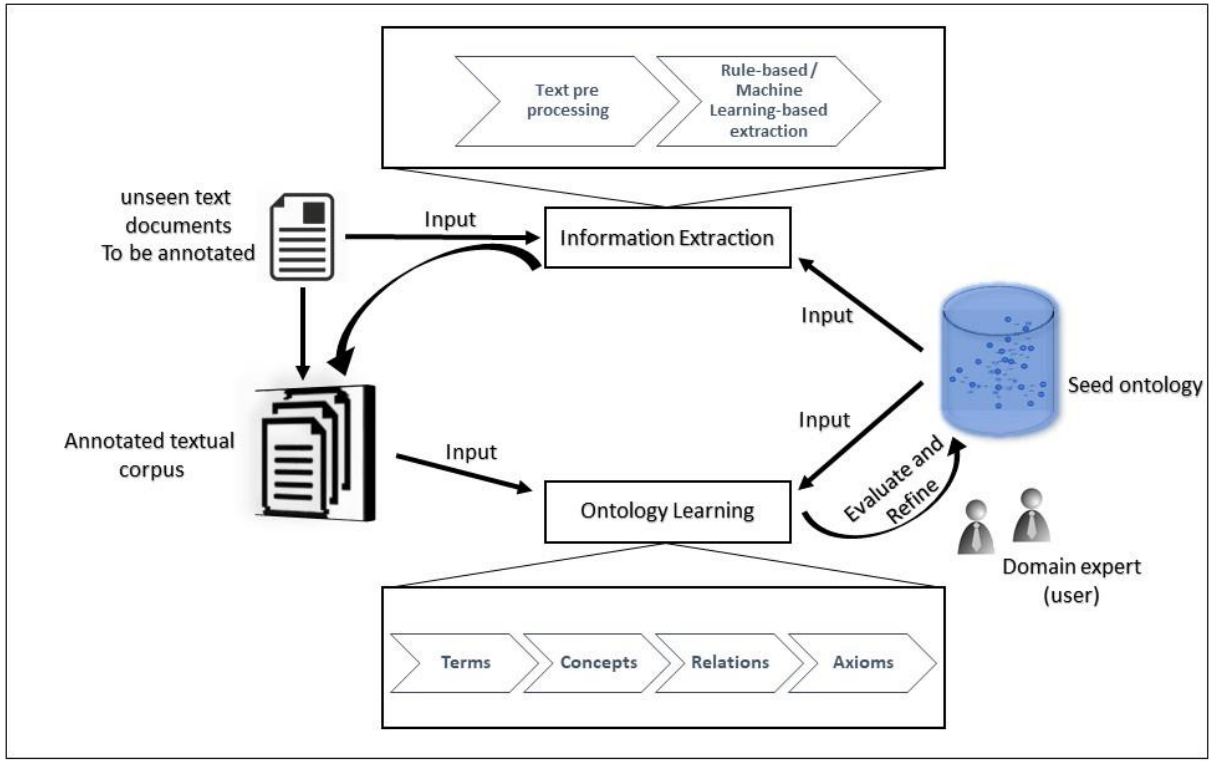
Ontology learning

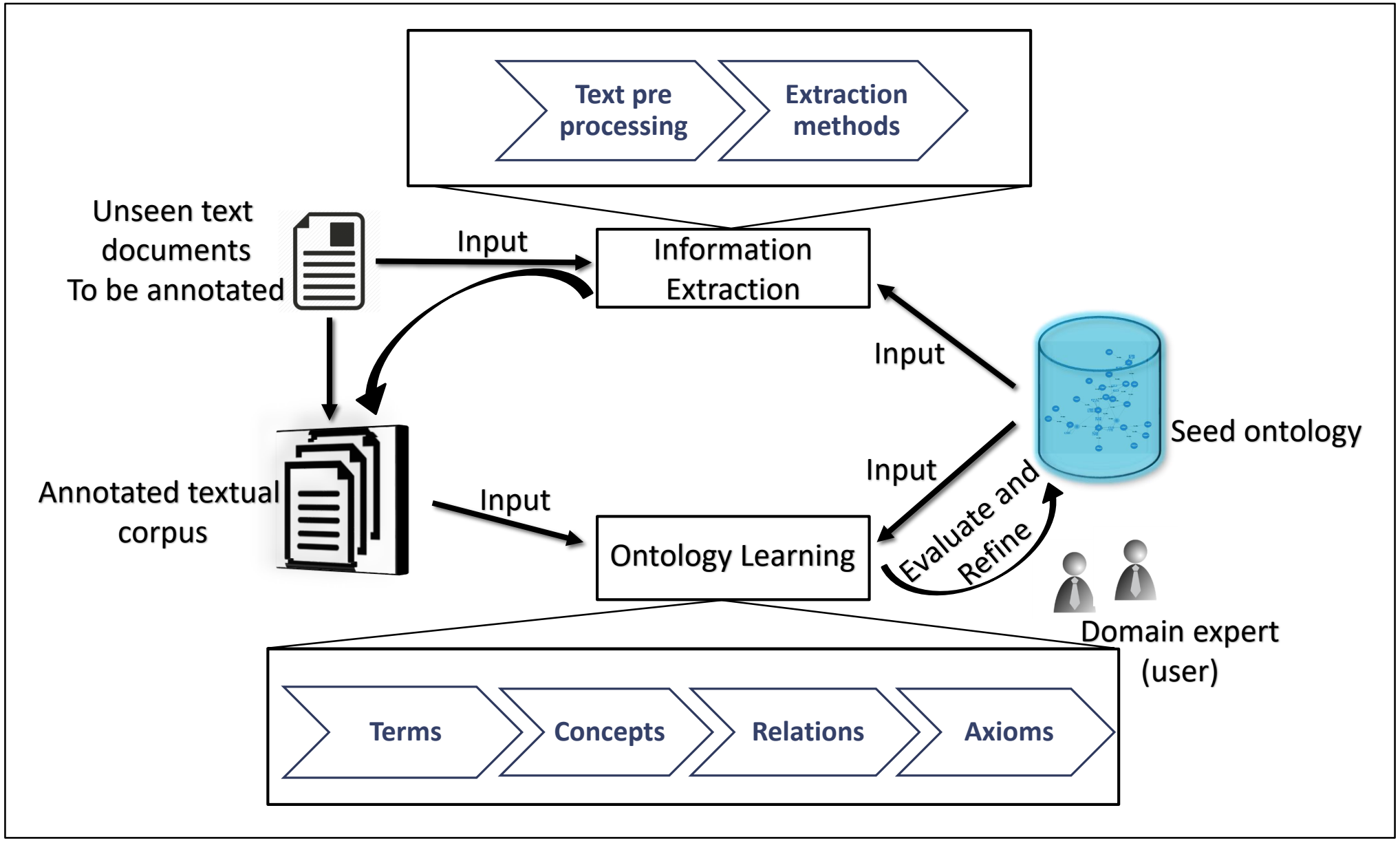


Ontology learning

		Terms	Concepts	Taxonomic relations	Non-taxonomic relations	Axioms
statistic methods	Text pre-processing	X				
	POS tagging	X				
	Sentence parsing	X				
	Latent semantic		X			
	Cooccurrence	X	X			
	Clustering		X	X		
	Term subsumption			X		
	Association rules					
Linguistic methods	Seed words	X				
	Semantic lexicon		X	X	X	
	Sub-categorization frames	X	X			
	Syntactic structure	X			X	
	Dependency analysis	X			X	
	Semantic templates			X	X	
	Lexico-syntactic paterns			X	X	
	Axiom templates					X
Logical methods	Logical inference			X	X	
	Inductive Logic					X

The iterative methodology




Text pre processing
Extraction methods
Unseen text documents To be annotated
Input
Information Extraction
Input
Seed ontology
Annotated textual corpus
Input
Input
Evaluate and Refine
Ontology Learning
Domain expert (user)
Terms
Concepts
Relations
Axioms

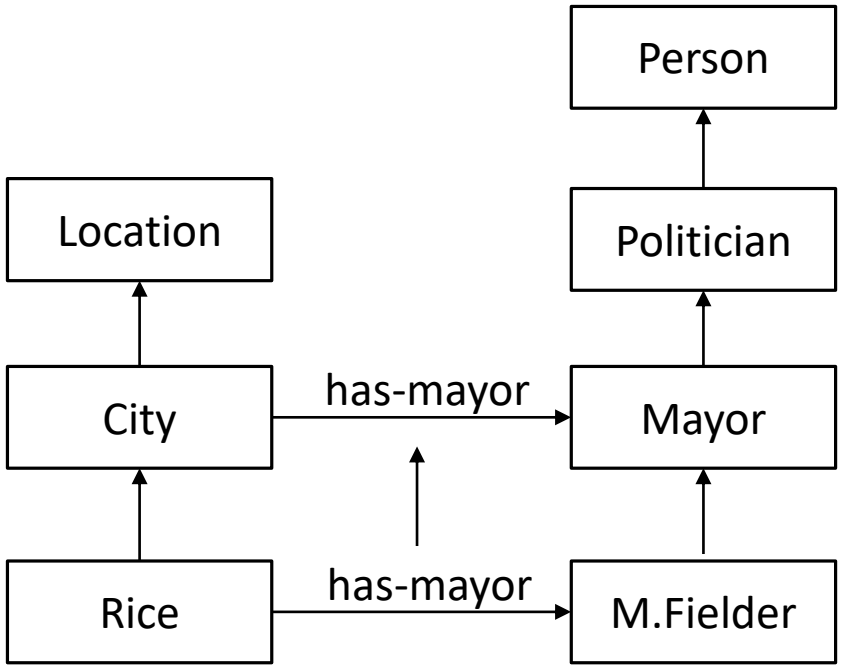
Information extraction and semi-structured ontology learning iterative methodology - Example

Doc1: Rice, a city in Minnesota

Doc2: Mitch Fielder is mayor of Rice

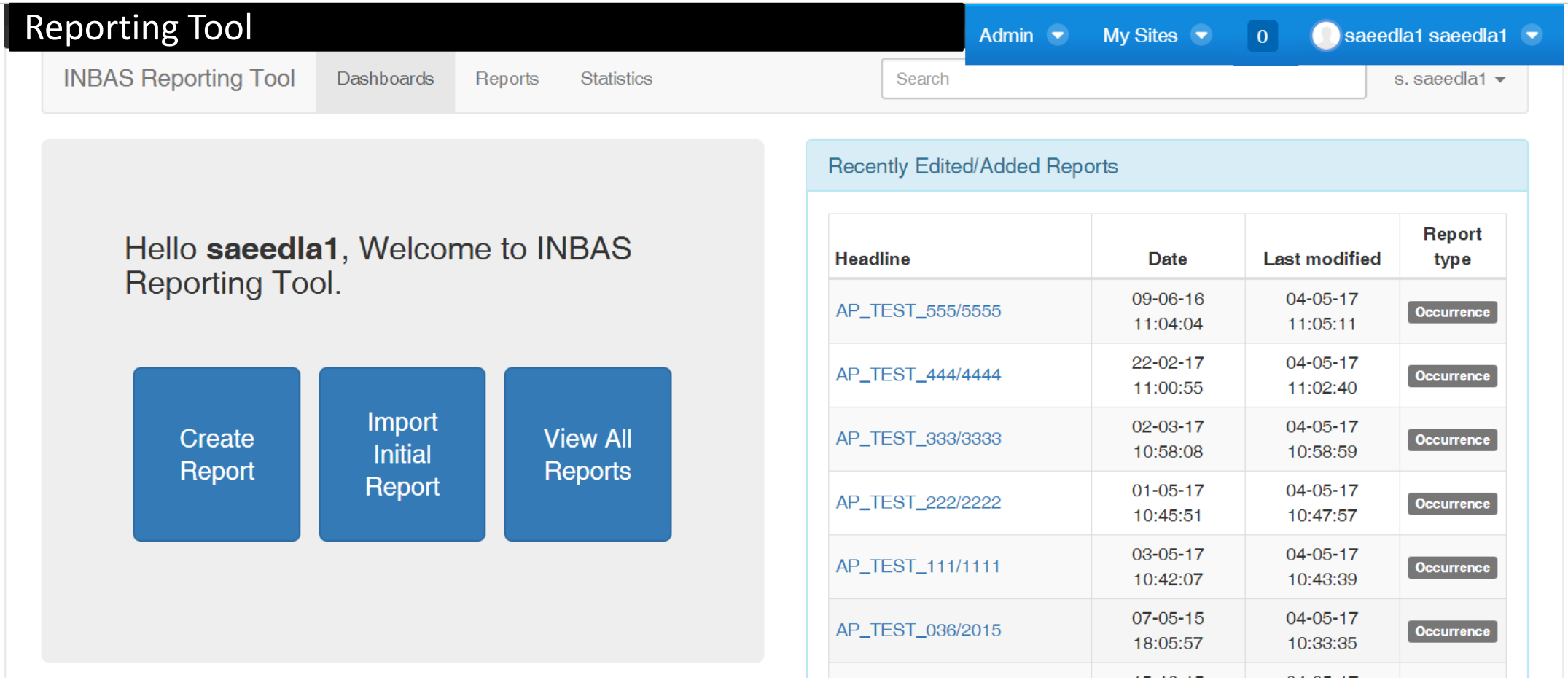
Doc3: Krnačova, mayor of Prague

Doc4: Tong wrote Rice

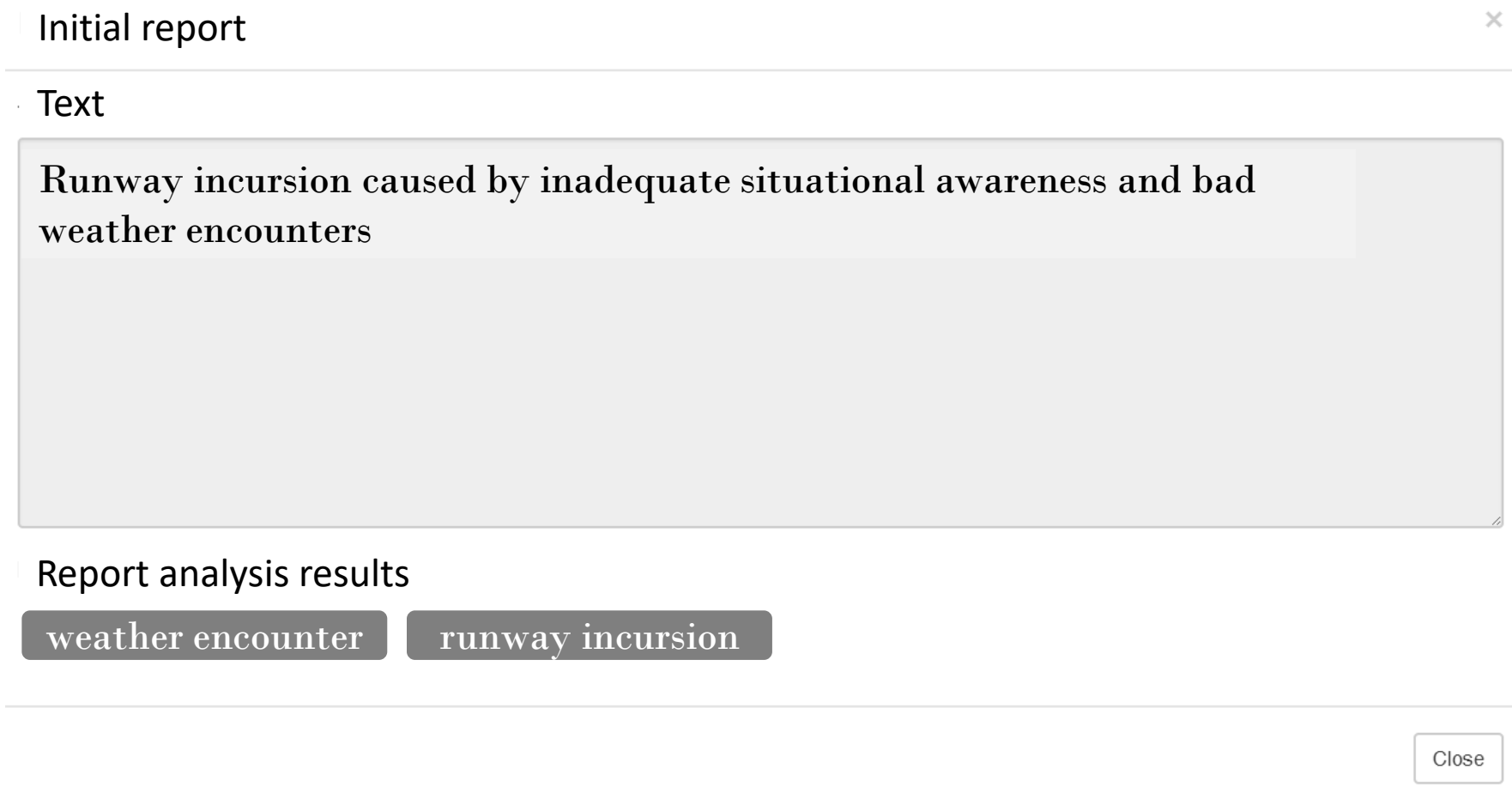


Ontology-based information extraction from aviation safety reports

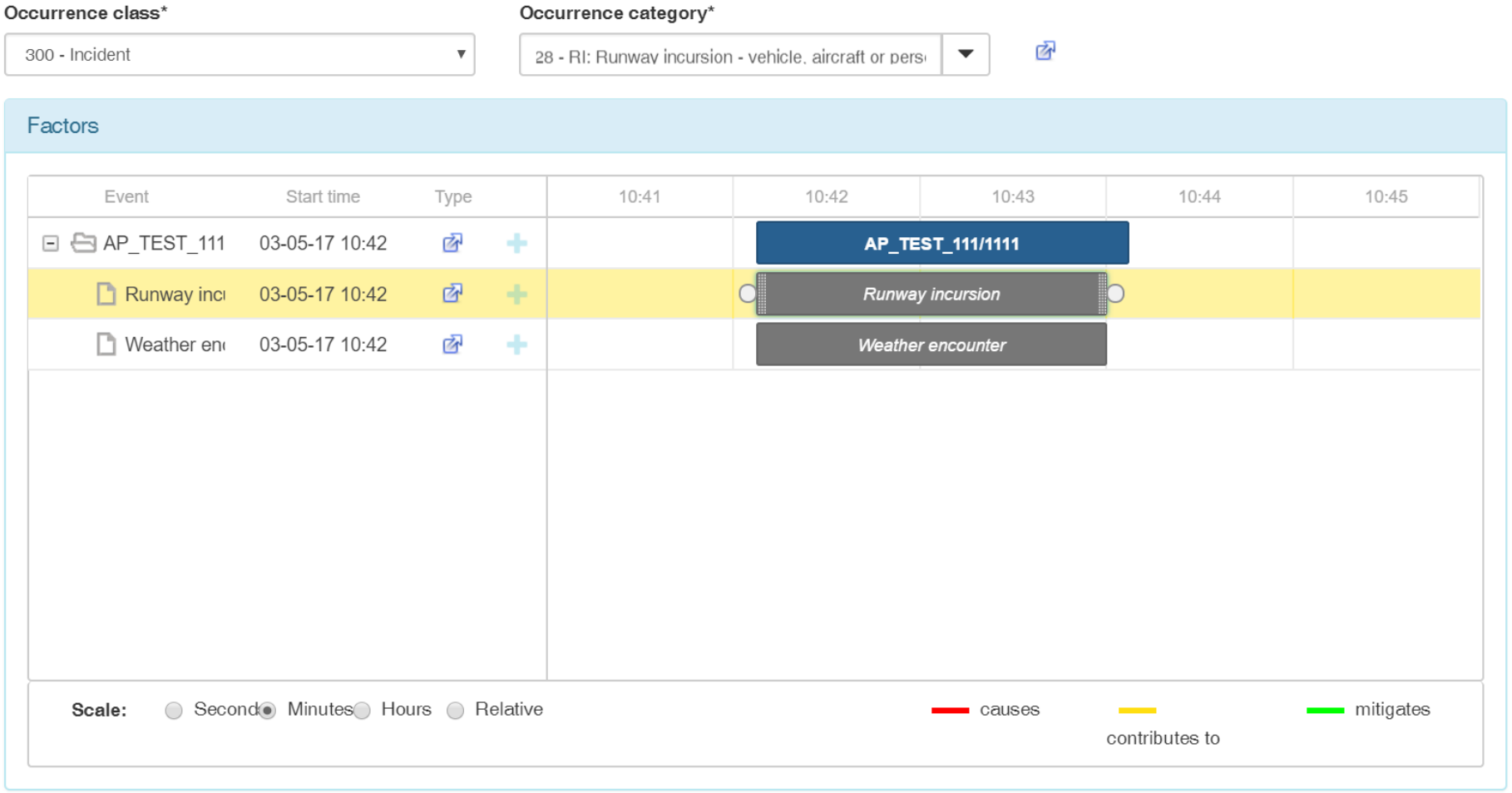
OBIE from aviation safety reports - Motivation

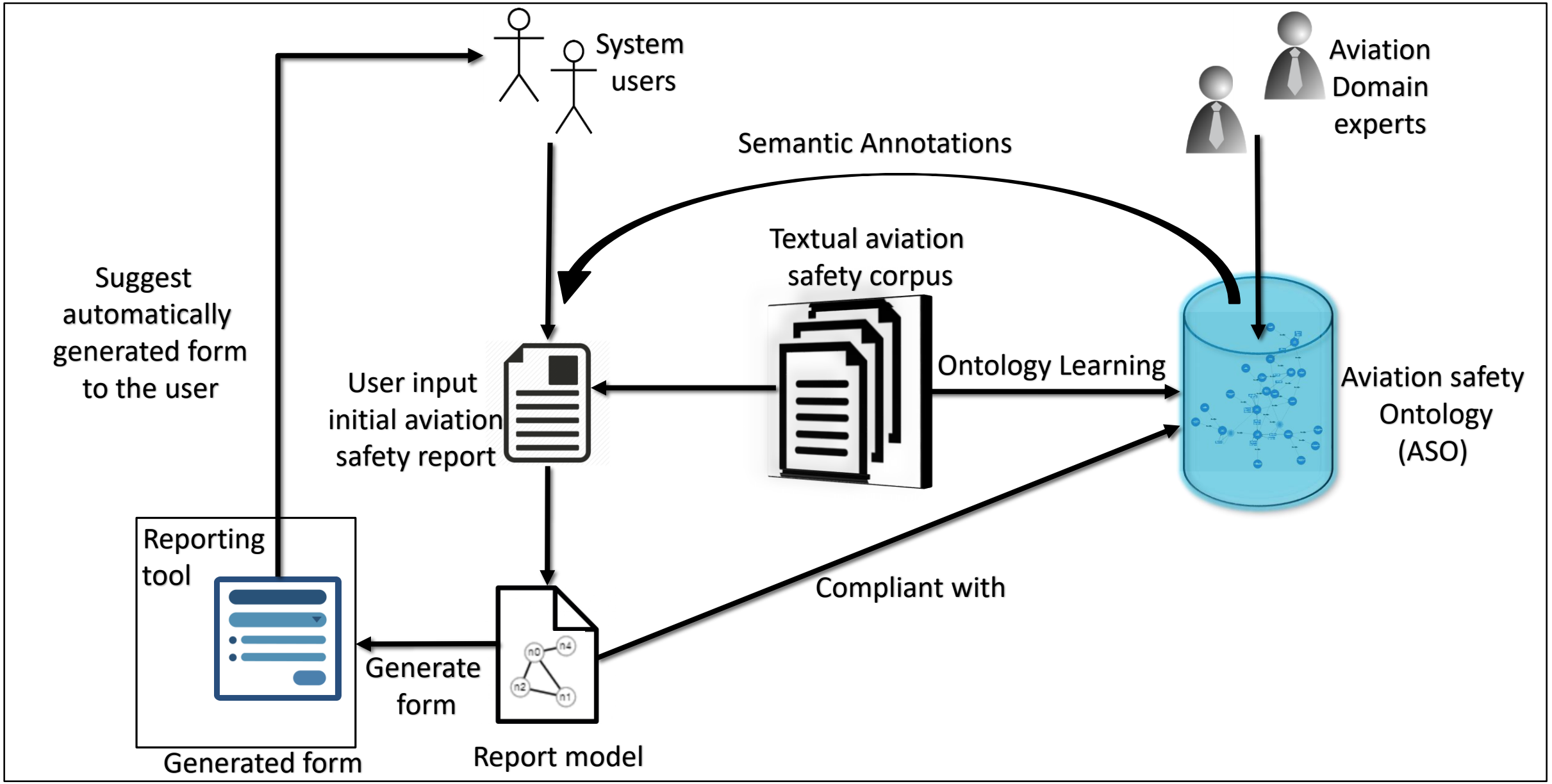


OBIE from aviation safety reports - Motivation



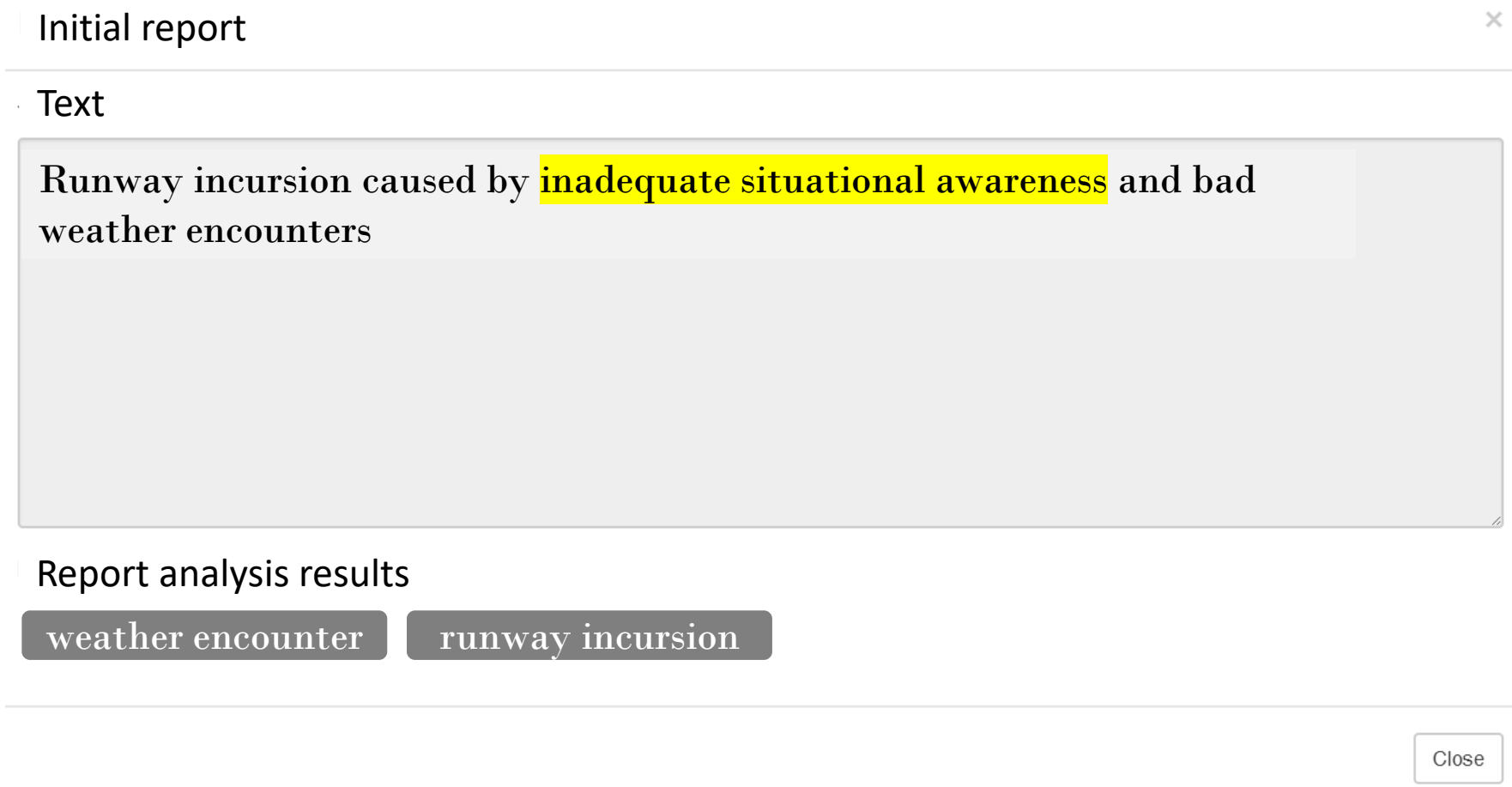
OBIE from aviation safety reports - Motivation

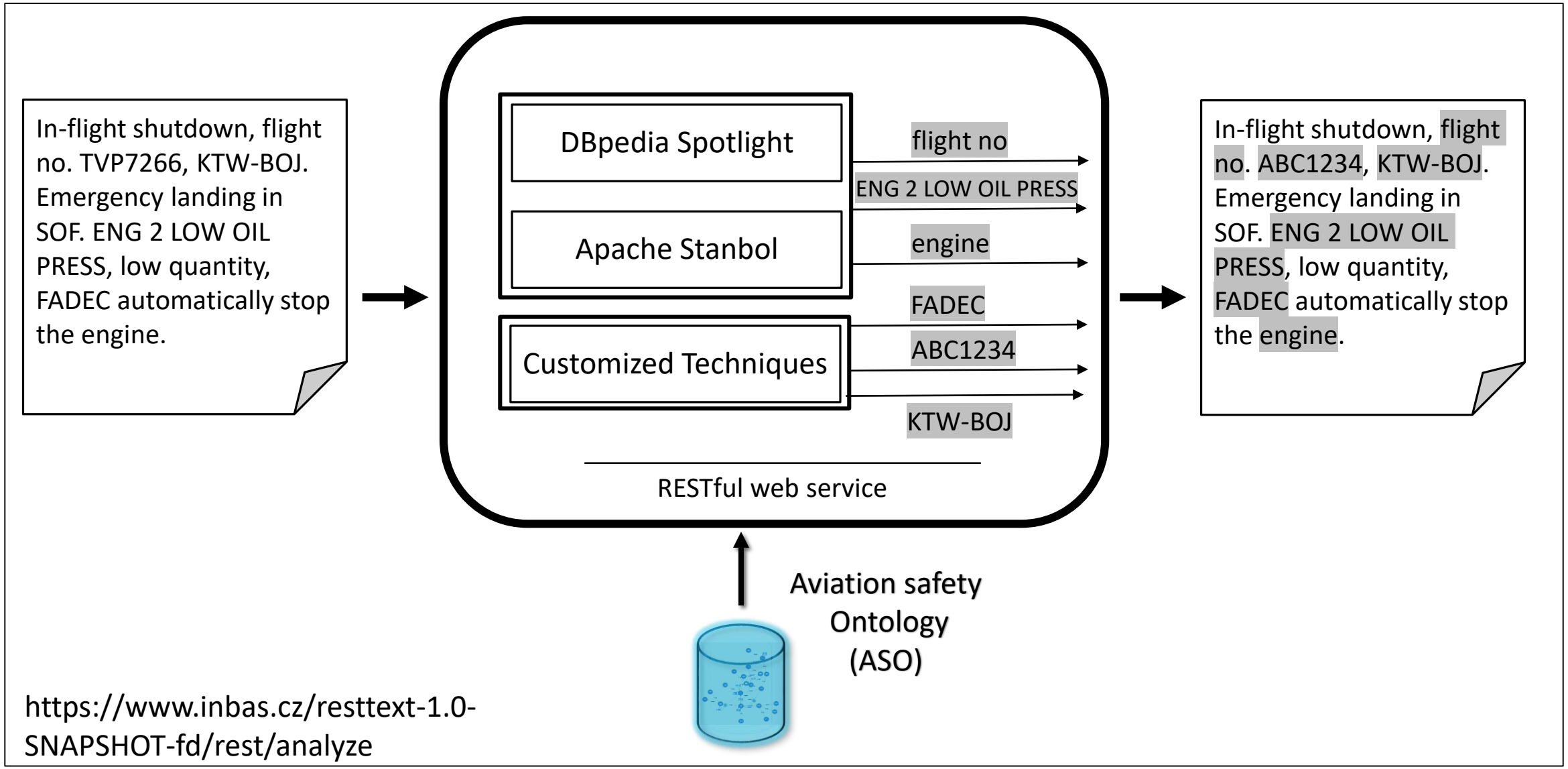




Information extraction and ontology learning from aviation safety reports

OBIE from aviation safety reports - Motivation

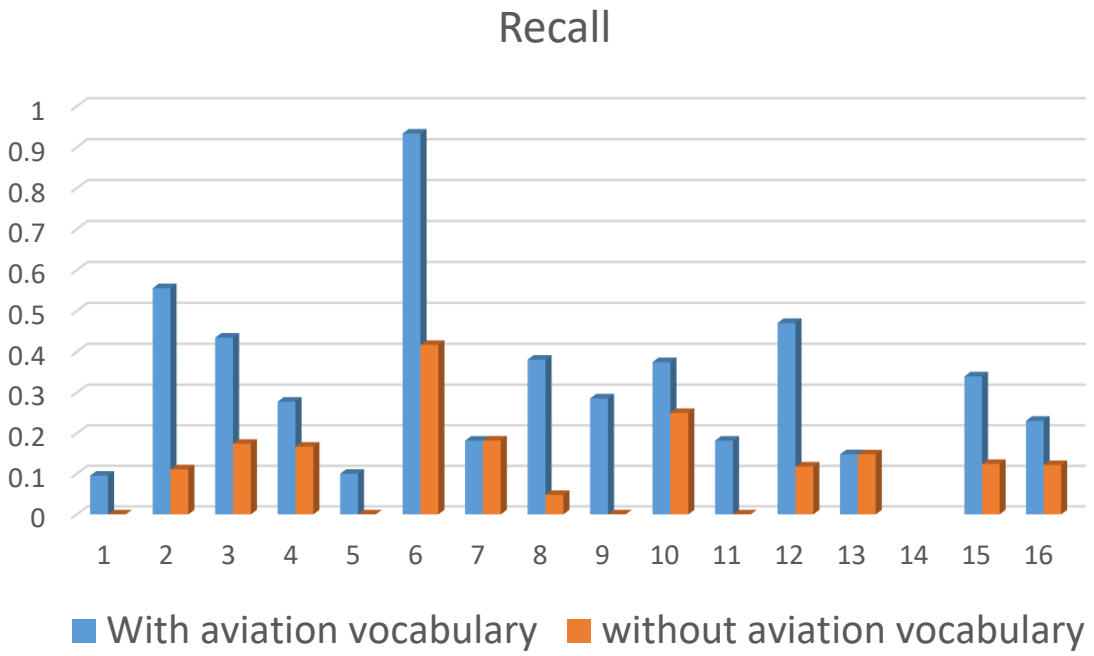


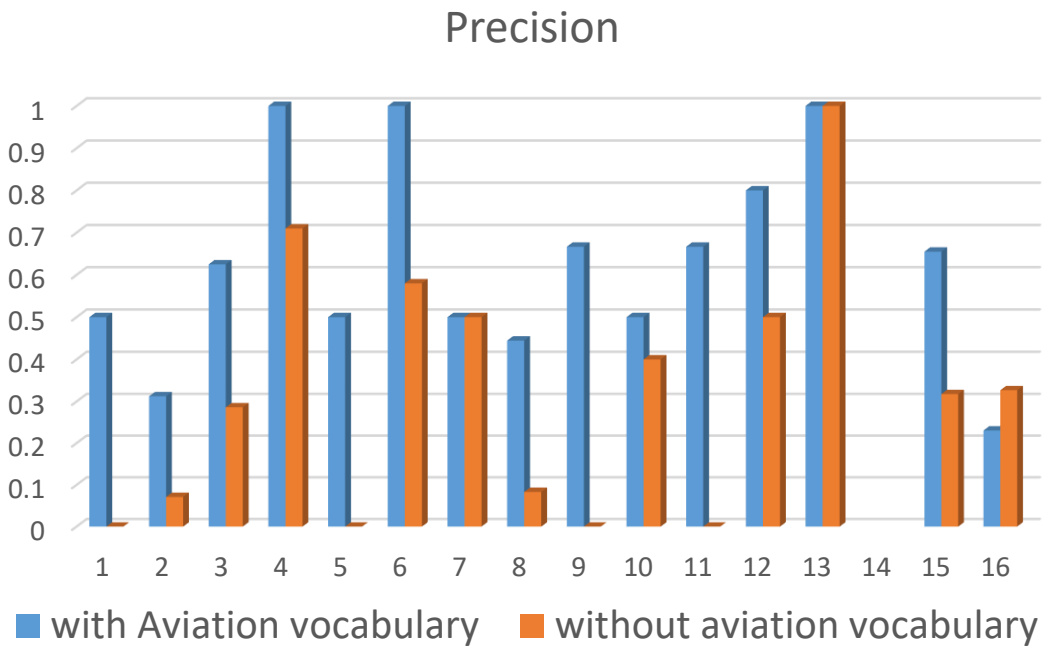


OBIE from aviation safety reports – Extraction pipeline

OBIE from aviation safety reports - Evaluation

- GATE (General Architecture for Text Engineering)
- 80 high quality annotated reports





OBIE from aviation safety reports – Next steps

- Ontology is essential for information extraction
- Incompleteness of the ontology
- Experimenting with various ontology learning techniques and tools
- Creating the ontology learning pipeline

Temporal descriptors for dataset exploration

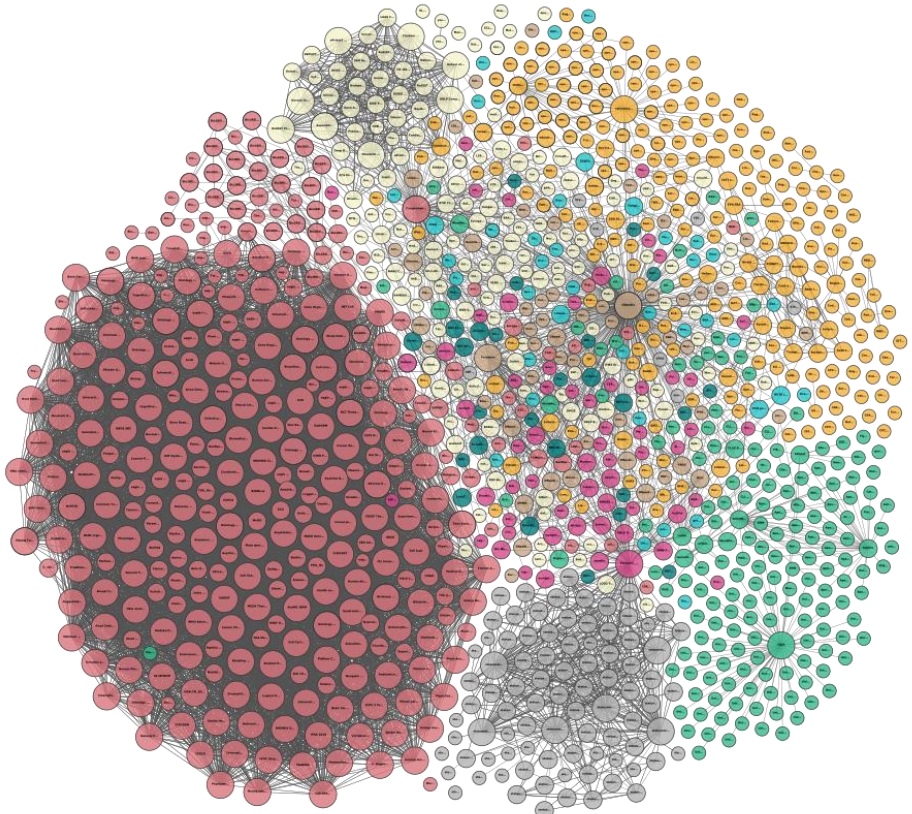
Temporal Knowledge extraction - Motivation

- Linked open data cloud
  - Data size vs. data quality
  - Dataset exploration

  Finding useful dataset without prior knowledge is difficult

- Temporal information
  - Dataset management
  - Temporal filters for search

  VoID descriptions could be enriched with a temporal dimension



Source: "Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/"

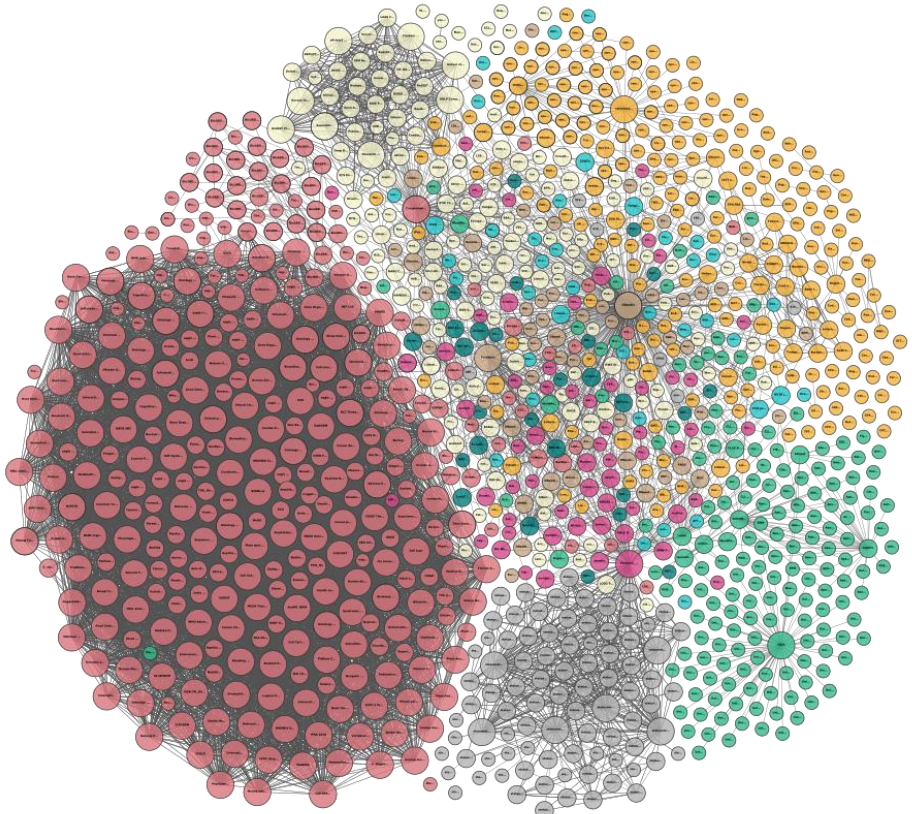
Temporal Knowledge extraction - Motivation

- Structured data

  no defined schema

- Dataset summaries:

Describe → Explore



Source: "Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/"

Temporal knowledge extraction - Temporal analysis

- Temporal data structure in datasets

  - Structured data:

    e.g. *xsd:date, xsd:time, xsd:gyear*

    is-temporal

  - Semi-structured: e.g. *dcterms: description, dcterms: title*

    has-temporal-potential

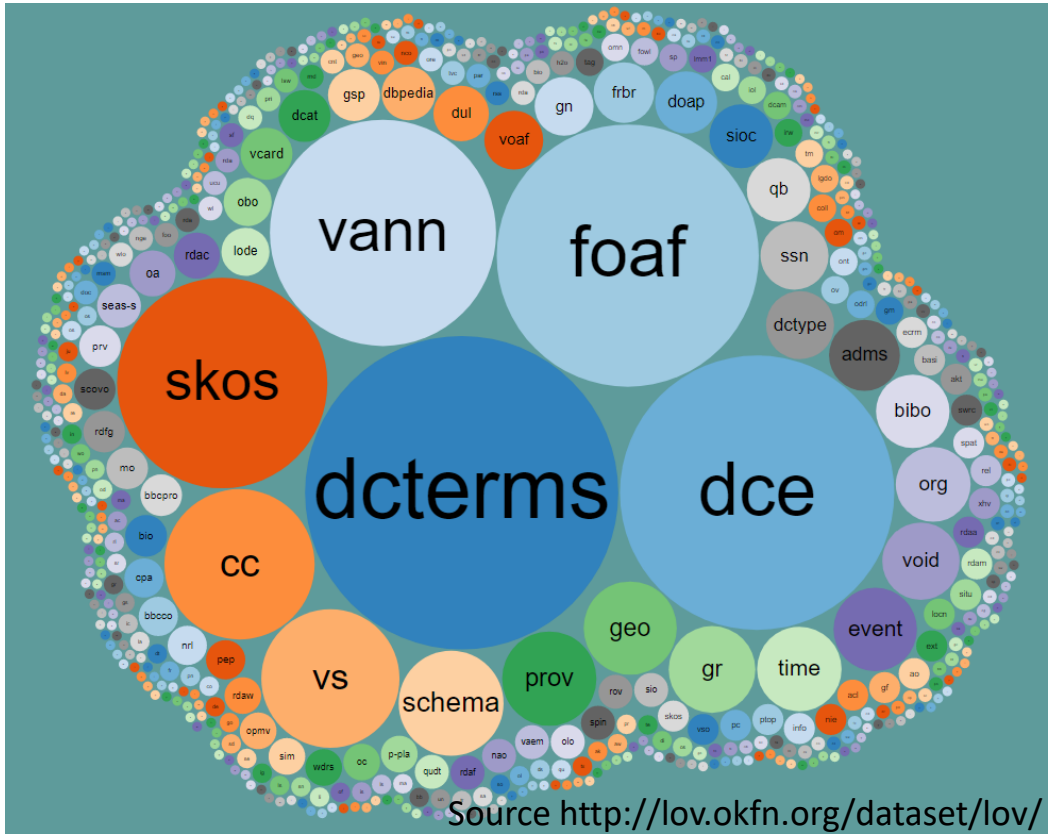
Temporal knowledge extraction - Temporal analysis

- Temporal data in Czech Linked Open Data Cloud (Czech LOD)

	is-temporal		has-temporal-potential	
	number of properties	number of triples	number of properties	number of triples
dcterms	12	7955431	2	7734043
schema.org	8	512781	4	149339
goodRelations	6	3110226	0	0
prov-o	2	146115	0	0
CzLOD	36	15268994	8	2597668

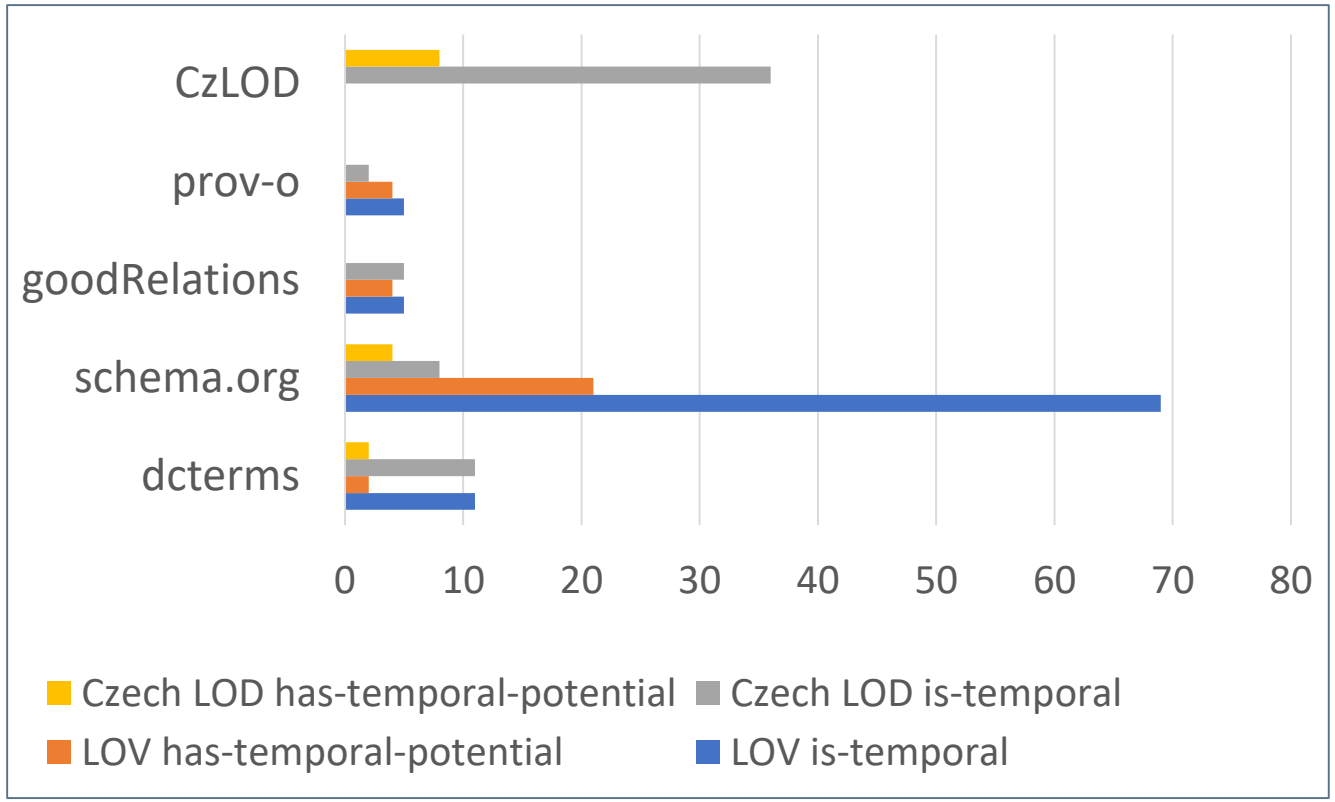
Temporal knowledge extraction - Temporal analysis

- Temporal data in Linked Open Data Vocabularies (LOV)



Temporal knowledge extraction - Temporal analysis

- Dataset-based vs vocabulary-based



is-temporal: Structured

has-temporal-potential: semi-Structured

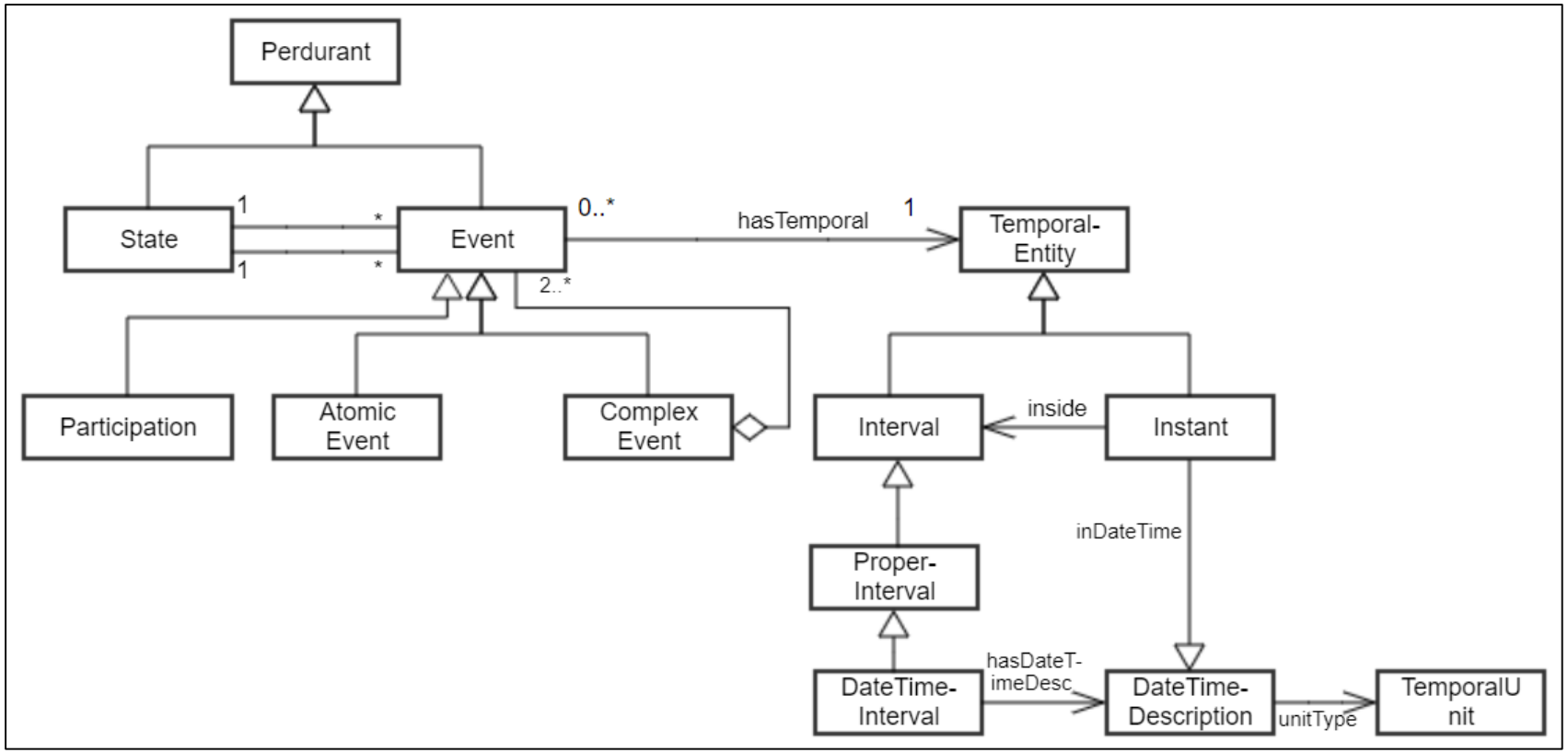
Temporal knowledge extraction - Temporal analysis

- *SUTime* (temporal taggers with 90.32 F-measure score in TempEval-3) *+*
- *extend* the default rule files for SUTime

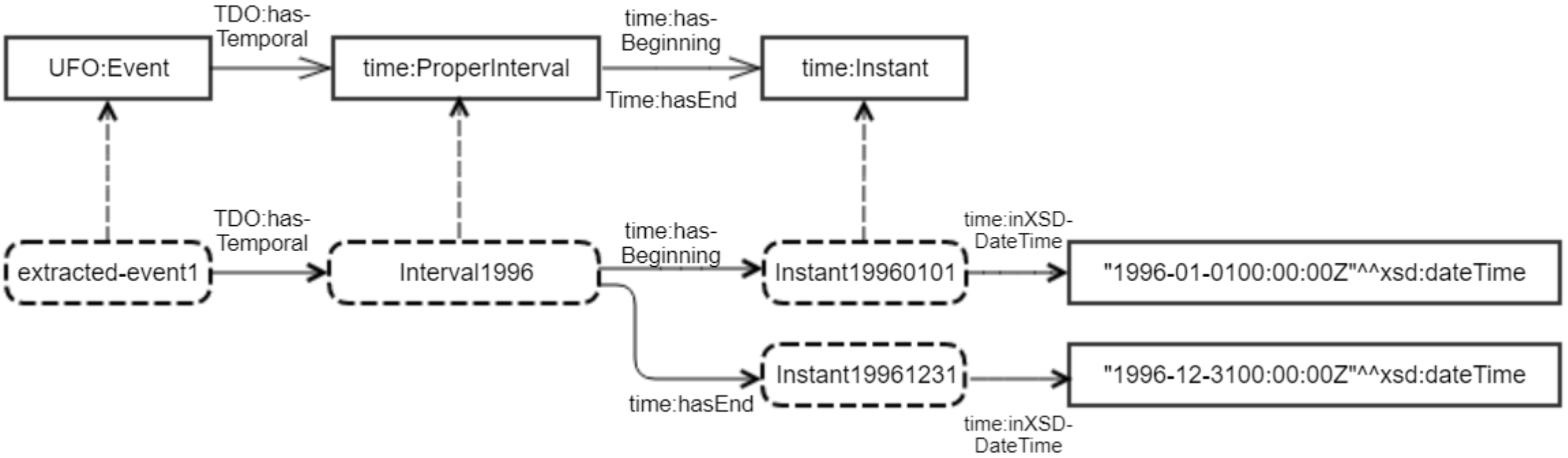
  - { ruleType : "time“, pattern : /dd/MM/yyyy/} → 22/10/2017
  - { ruleType : "time“, pattern : /dd-MM-yyyy/} → 22-10-2017
  - {ruleType : "time“, pattern : /[0-9]{1, 4}.?yyyy’Sb’/} → 253/1992 Sb

Temporal knowledge extraction - Temporal Descriptors Ontology (TDO)

- Temporal data in Czech Linked Open Data Cloud (Czech LOD)



Temporal knowledge extraction - Temporal Descriptors Ontology (TDO)



Temporal knowledge extraction – Evaluation & results

- Temporal coverage of dataset
  - DCAT voc → *dct:temporal* (startDate, endDate)
  - Temporal descriptor → considers only structured temporal data
  - Our approach → consider actual data (structured + unstructured)

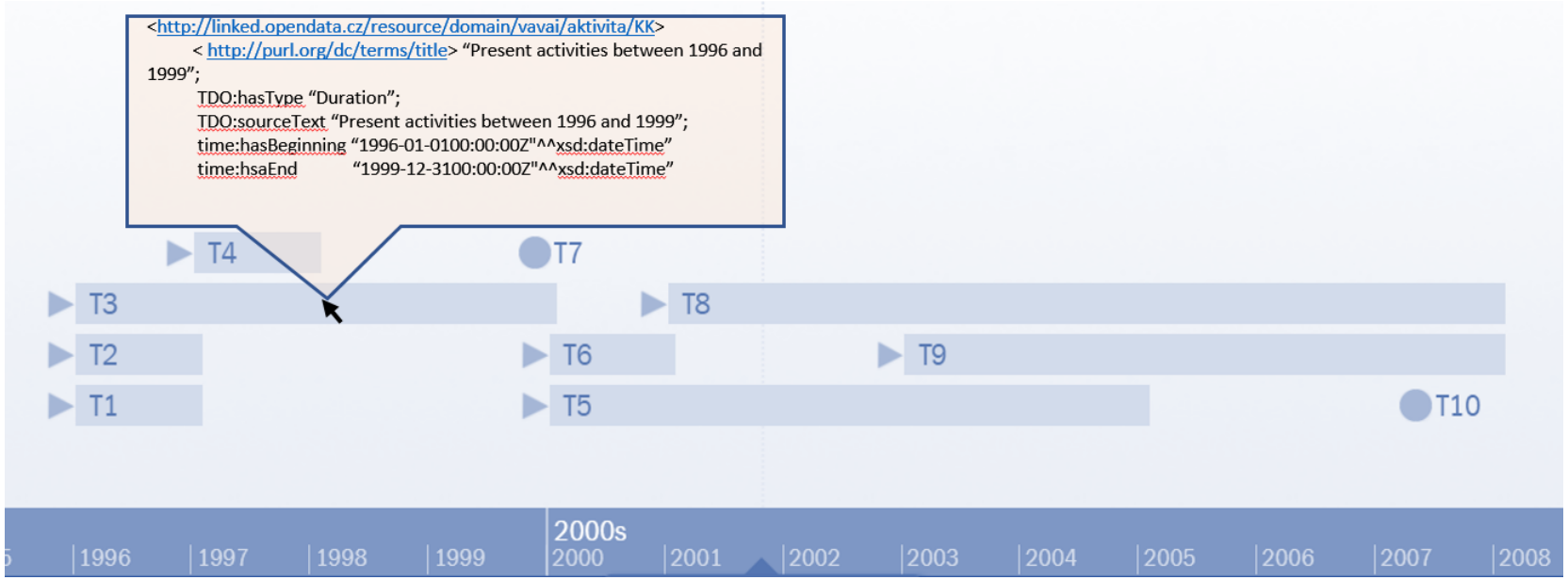
15 datasets: 46.6% of the cases matching (some cases there is no presence of TI description)

46% of the cases we were able to extract more accurate range

e.g. Dataset *ds:vavai/evaluation/2009* has no structured temporal info.

Temporal knowledge extraction – Conclusion

- Improve the temporal knowledge representation of the datasets
- Better exploring experience for users.
- Timeline visualization as a future work



Conclusion

- Support the information extraction process
- Incomplete ontology → Iterative approach
- Information extraction pipeline
- Ontology learning pipeline
- Temporal information extraction

- Evaluate the iterative approach to other approaches
- Evaluate involving experts (time, effort)
  - Reduced with advanced iterations?