Ontology Learning: Framework, Techniques and a Software Environment

MEANING WS Presentation, San Sebastian

Alexander Maedche

Forschungszentrum Informatik an der Universität Karlsruhe

Forschungsbereich Wissensmanagement (WIM)

http://www.fzi.de/wim

Agenda

- Introduction & Motivation
- Ontology Learning Framework & Techniques
- Text-To-Onto Tool-Environment
- Applications
- Conclusion

Introduction

- Semantics-driven processing of information has been recently become a hype (= Semantic Web).
- The global vision:
  - Allow machines to read and interpret information that is distributed and heterogeneous, stored in databases, semi-structured documents and free text documents.
  - Allow humans for „semantics-based“ access to information.
- This vision is not new, many communities have been working on it, e.g. the
  - Knowledge engineering & Representation Community
  - Natural Language Processing Community
  - Database Community (in the context of Information Integration)

Introduction

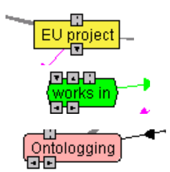
- Lexical and ontological resources are seen as the key for bringing this vision to reality.
- Extracting these resources from data (structured data, semi-structured and free text documents) on which they will be later applied on is promising.
- This presentation will present some work in the field of ontology learning, with specific focus on textual data as input for ontology learning.

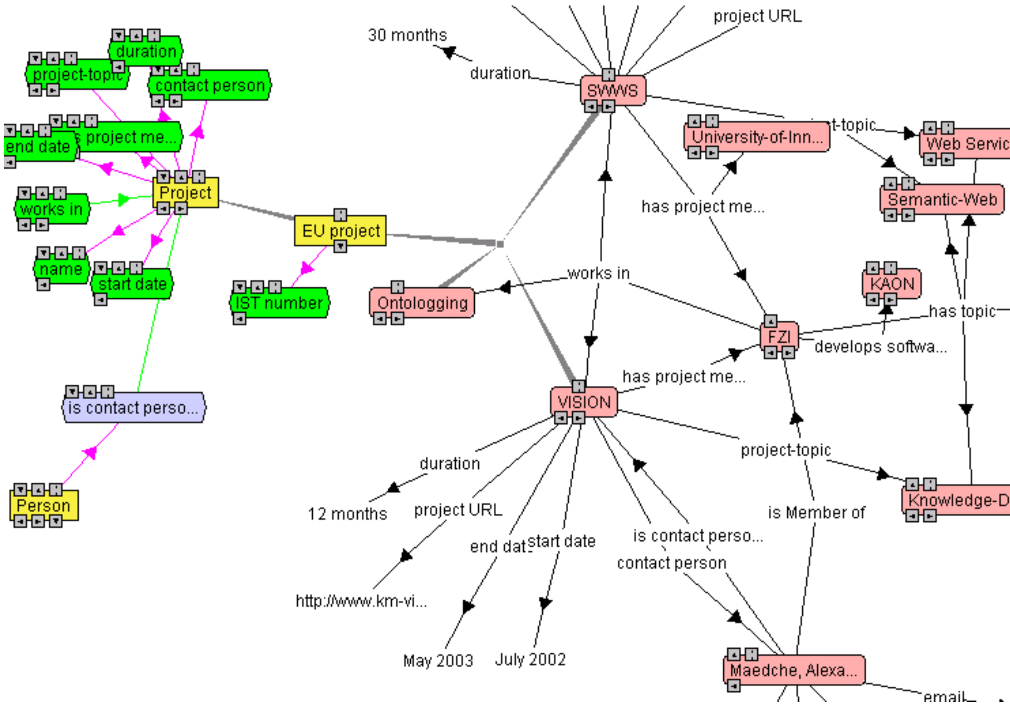
Agenda

- Introduction & Motivation
- Ontology Learning Framework & Techniques
- Text-To-Onto Tool-Environment
- Applications
- Conclusion

Ontologies

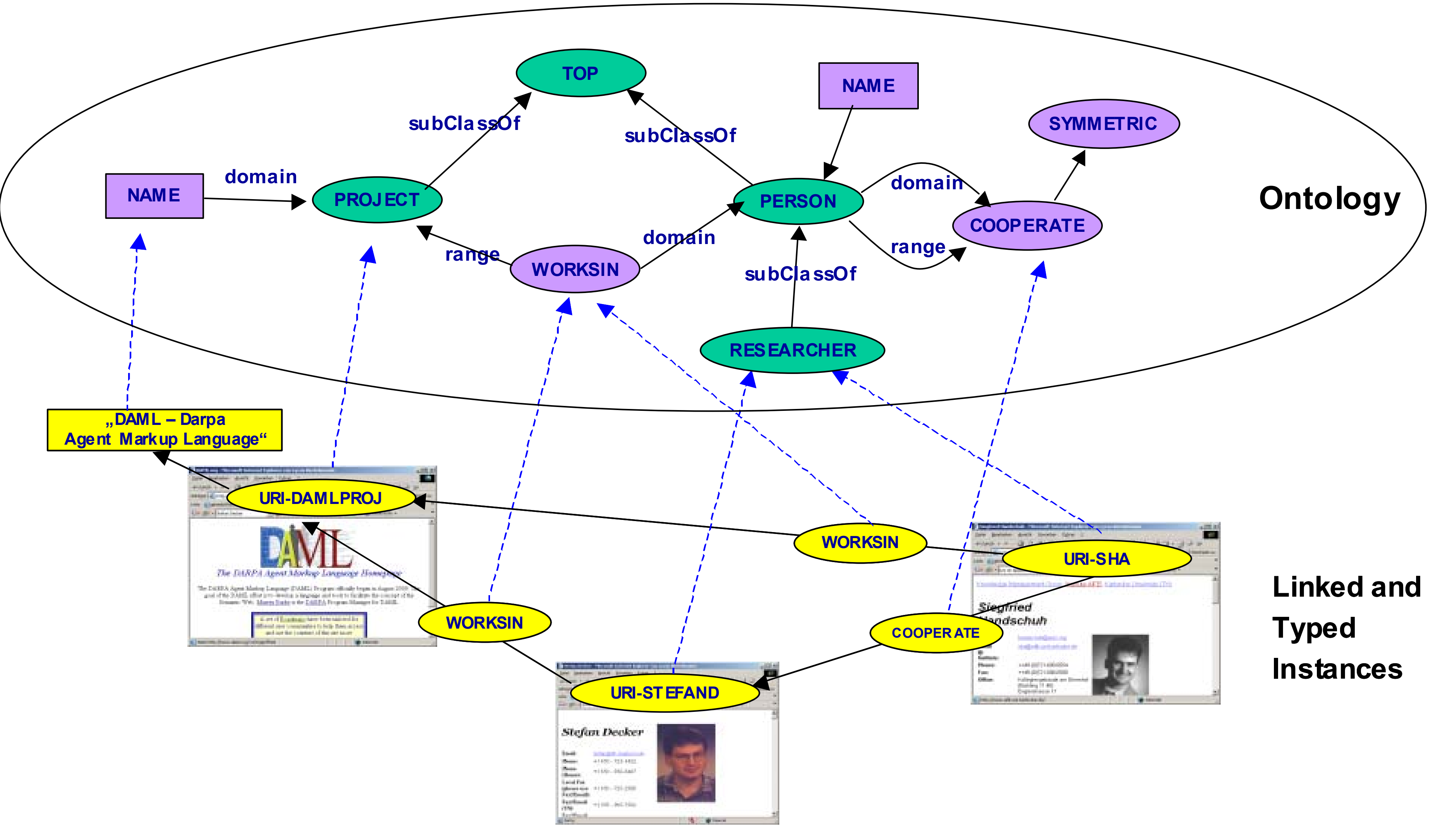
- Expressive conceptual models, no strict separation between schema and instance.
- OI-model (ontology-instance model) – elementary information container, contains ontology and instance data:
  - concepts
  - relations
  - instances
  - relation instances



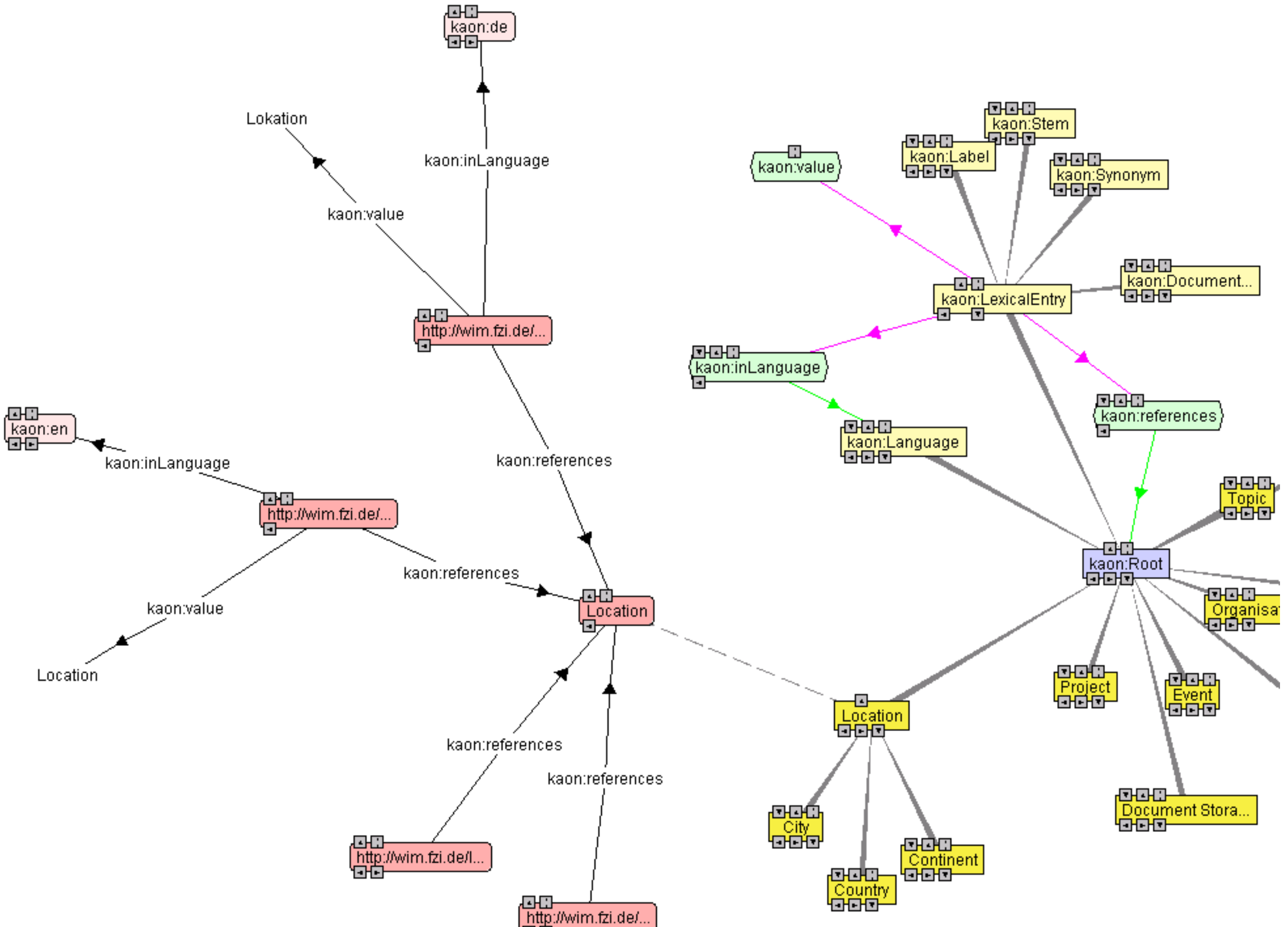


- Extensions of W3C's RDF-Schema, along the same lines of W3C OWL.
- Builds on an expressive hybrid knowledge representation mechanism, inspired by Description Logics paradigm, but executed using deductive database techniques.

Ontologies & Semantic Web

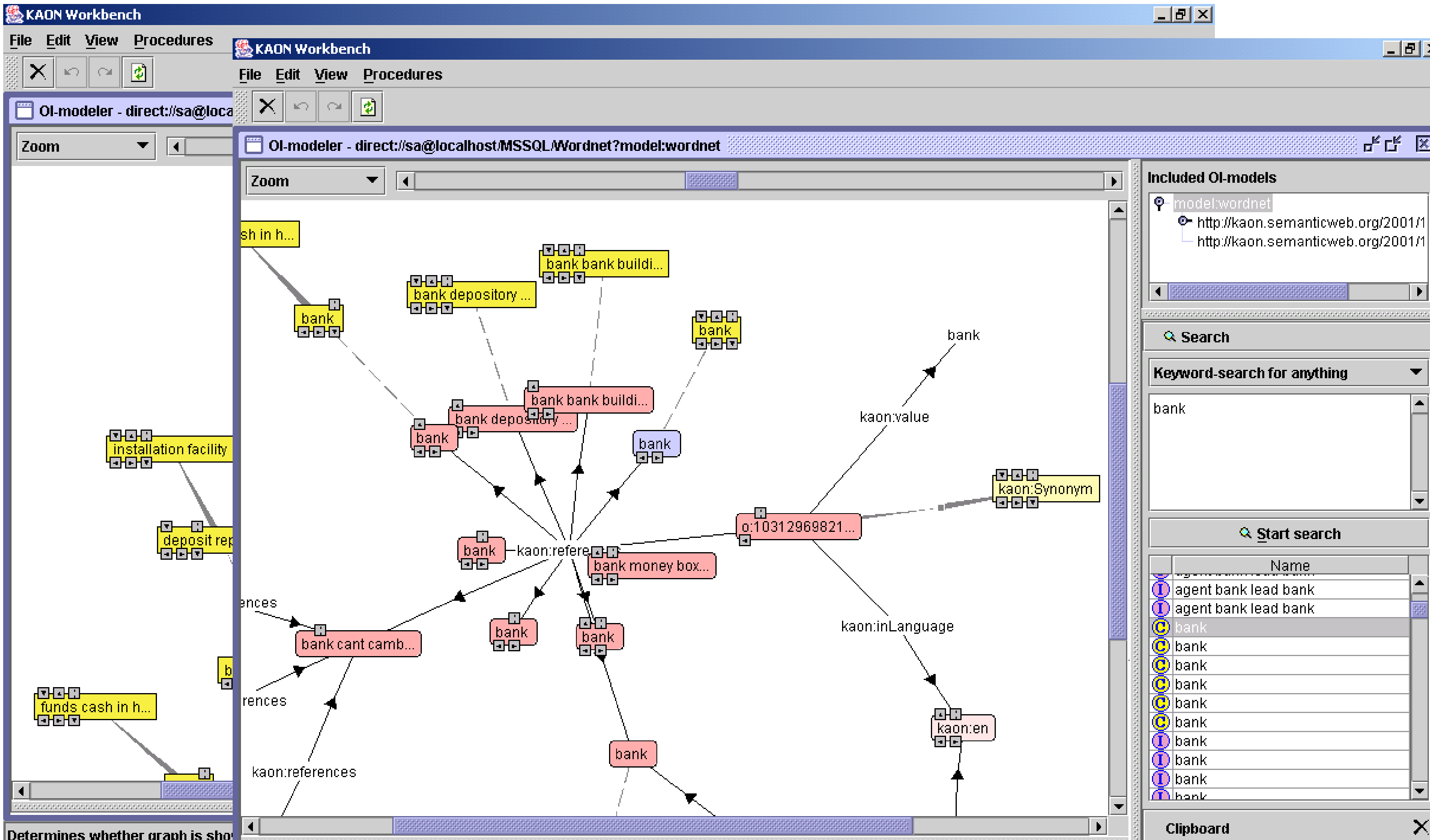


Ontology & Natural Language



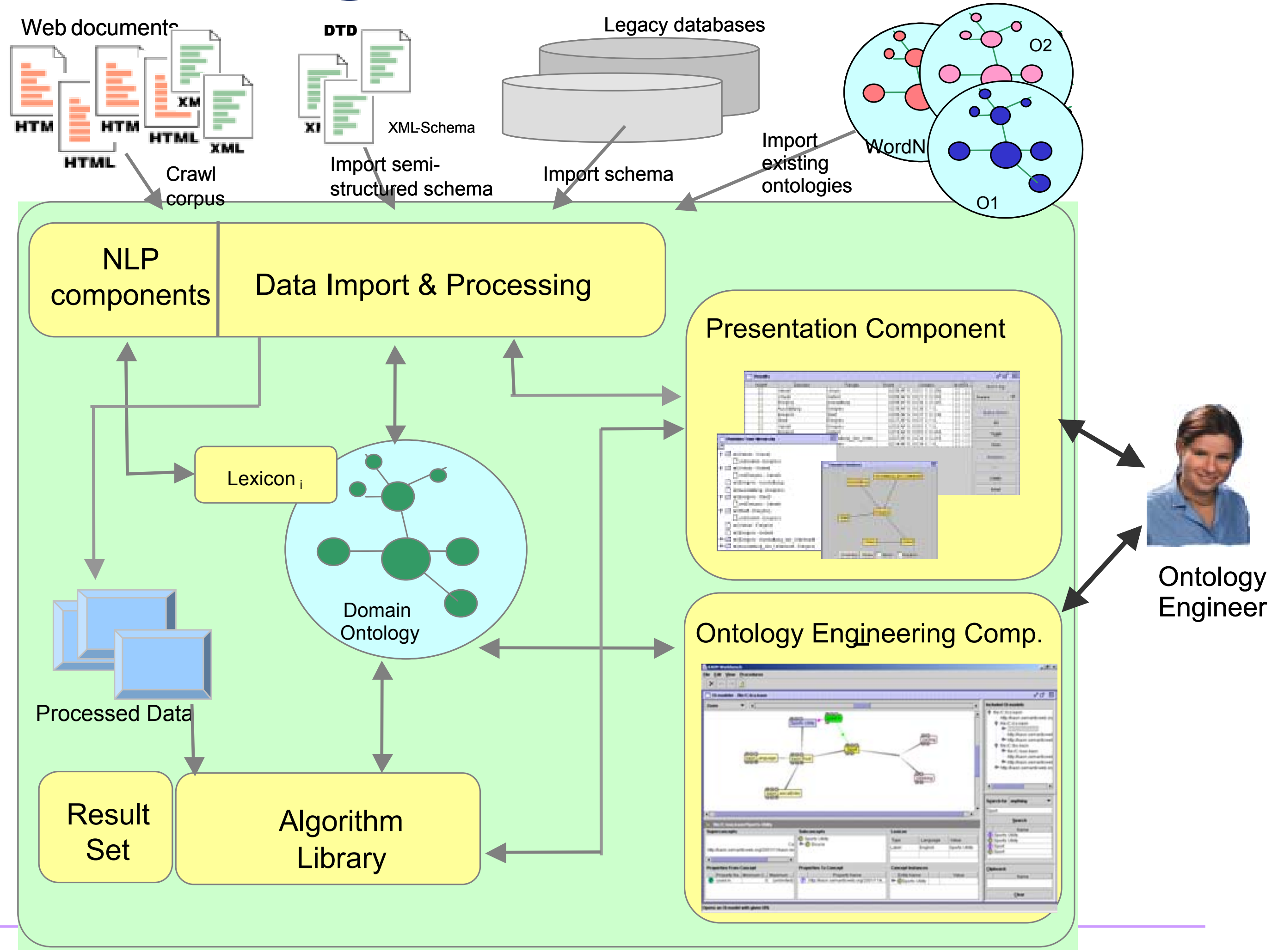
- The lexicon is part of the ontology.
- It is considered as a specific model within the ontology (lexical OI-Model) and is considered as meta-information.
- It allows to encode multilingual labels, synonyms, etc. etc.

WordNet seen as an OI-Model



Ontology Learning Framework

- Balanced cooperative modeling architecture
- Incremental and interactive
- Multiple resources
- Multiple algorithms



Ontology Learning Techniques

1. Concept Extraction

- Multi-Word-Term Extraction
- Multi-Word-Term Meaning Extraction

2. Concept Relation Extraction:

- Taxonomy Learning
- Non-taxonomic relation extraction

Beside these two core phases, ontology reuse via “ontology pruning“ is provided.

Concept Extraction

Extracting multi-word terms from a given corpus:

- Term extraction is a basic technology for ontology learning.
- Typically, relevancy measures like tf/idf are used to determine important terms of a corpus.
- Beside the relevancy measures, multi-words term recognition techniques are of importance.

Discovering the meaning of extracted terms:

- An extracted multi-word term has to be embedded into the ontology, where one typically has several possibilities, e.g. create a new concept, add it as a synonym to an existing concept, etc.
- Within our framework, we provide semi-automatic support for adding an extracted multi-word term to the ontology.
- The approach is based on measuring distributional similarity of the extracted term with existing entities in the ontology.

Multi-Word Term Extraction

- C-value method (*):
  - Domain-independent method for automatic extraction of multi-word terms, from machine-readable specific language corpora
  - Combines linguistic and statistical information

- Relevancy of terms is determined via the classical tf/idf technique.

(*) based on: Katerina Frantzi, Sophia Ananiadou, Hideki Mima: *Automatic recognition of multi-word terms: the C-value/NC-value method,* Int J Digit Libr (2000) 3: 115-130

Multi-Word Term Meaning Extraction

For each extracted term and also each concept in given ontology we create following vector:

$\{term(verb_1,freq),\ldots,(verb_n,freq),(noun_1,freq),\ldots,(noun_t,freq)\}$

Where verbs and nouns are considered if they are in the same sentence as the term and in the defined window size.

A distributional distance between each pair of vectors is computed. The smaller the distance is, the more similar terms or concepts (which are described by those vectors) should be.

Concept Relation Extraction

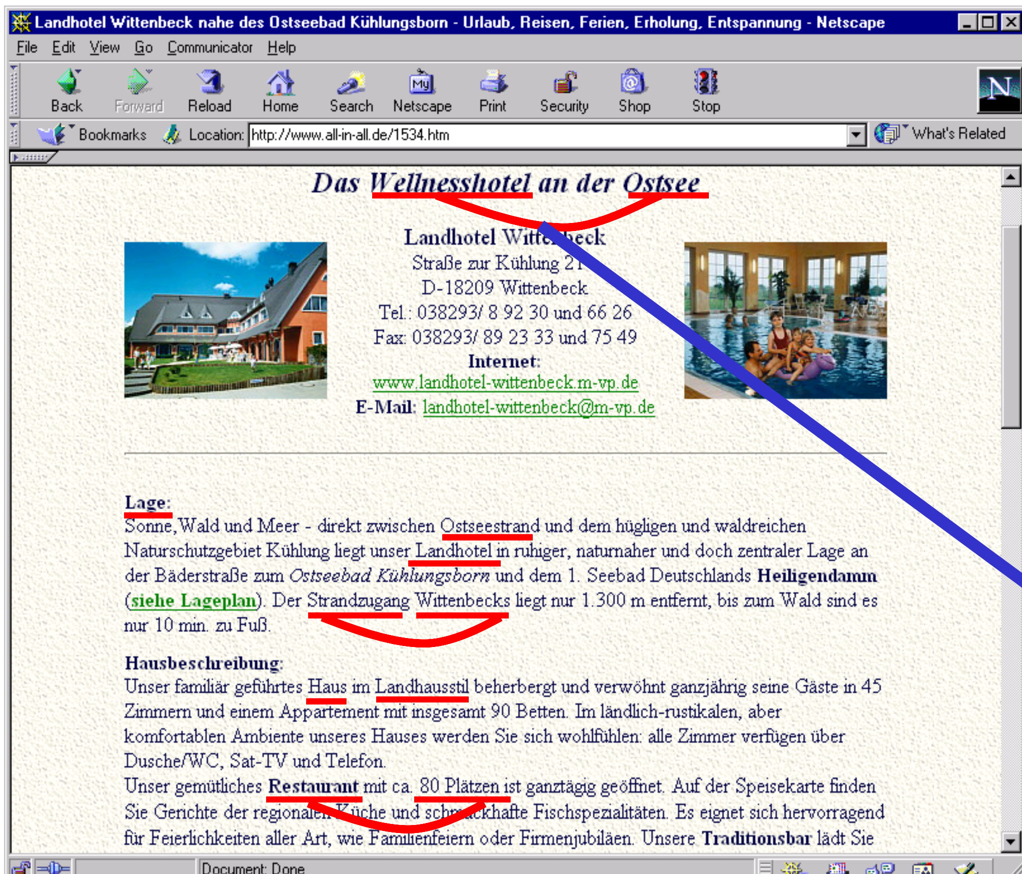
Concept Hierarchy Extraction

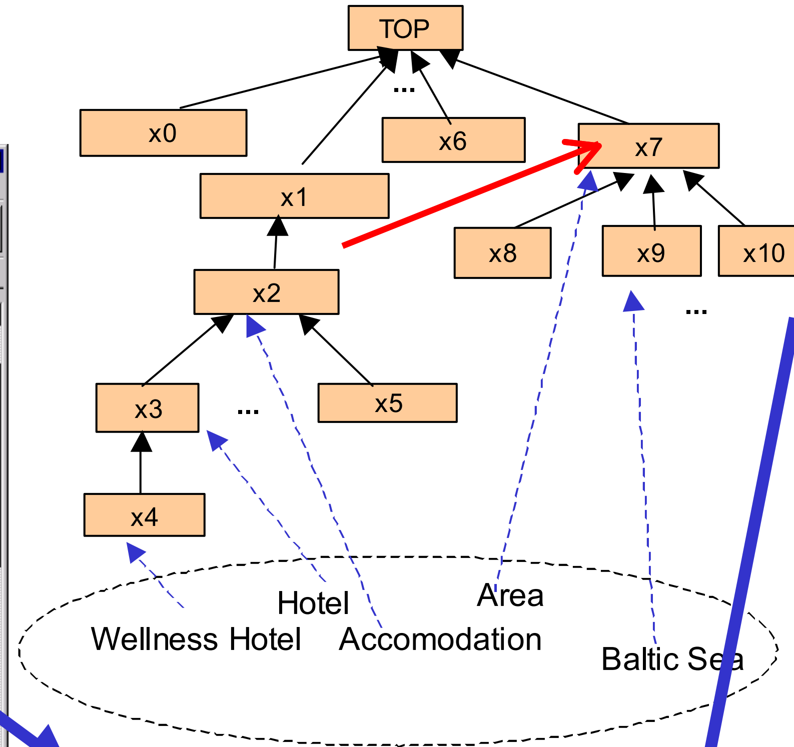
- Lexico-syntactic pattern-based extraction works fine for structured resources like dictionaries.
- Hierarchical clustering did not show a good performance in our experiments, labeling extracted super concepts is a problem.
- Verb-driven approaches seem to work well in some domains (e.g. cooking recipes).

Non-taxonomic Relation Extraction

- Linguistics and heuristic based association between concepts and the application of an association rule algorithm developed.
- Currently, this is extended with means for automatic relation labeling using a verb-driven approach.

Non-Taxonomic Relation Extraction



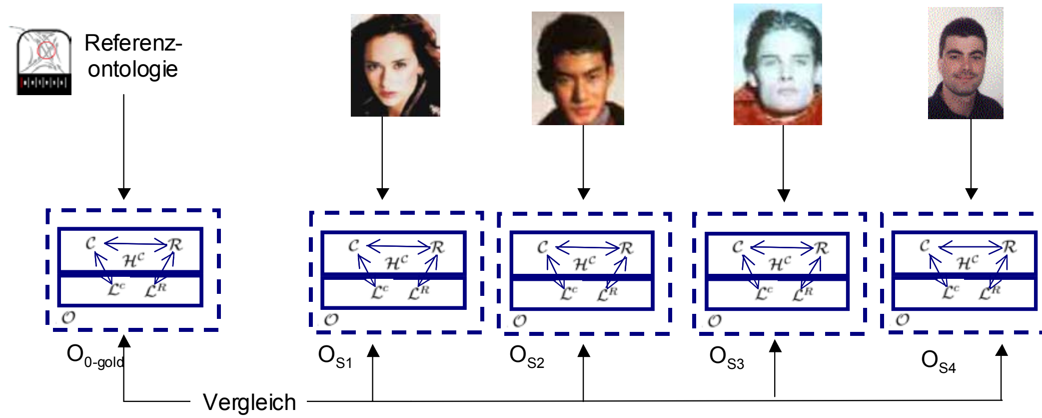


F(Wellness Hotel) = x4
F(Baltic Sea) = x9

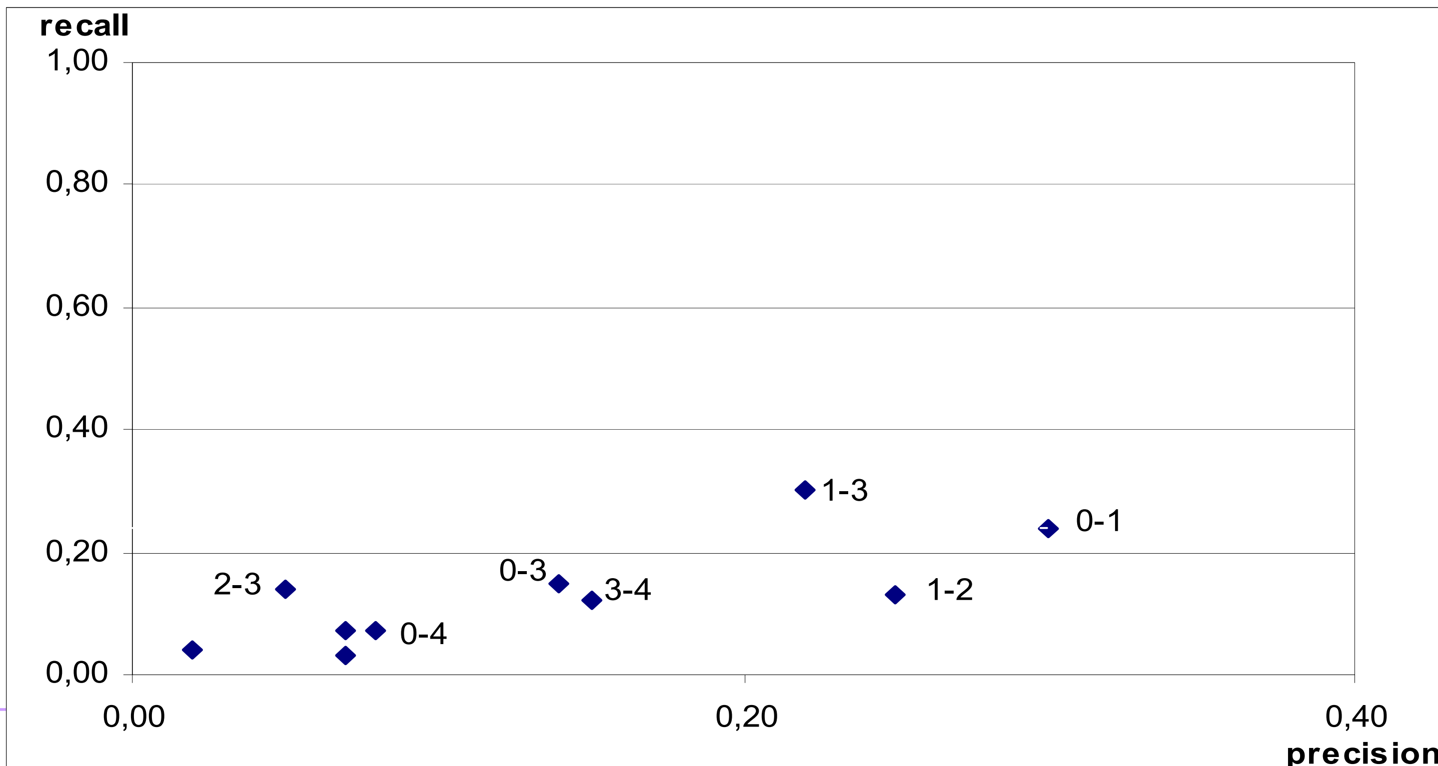
Concept pair (ling. transaction)
(x4,x9) bzw. (F(Wellness Hotel), F(Baltic See))

Generalized Association:
(F(Accomodation) -> F(Area)) (with label: G(locatedin))

Evaluation



Referenz-
ontologie
O0-gold
OS1
OS2
OS3
OS4
Vergleich



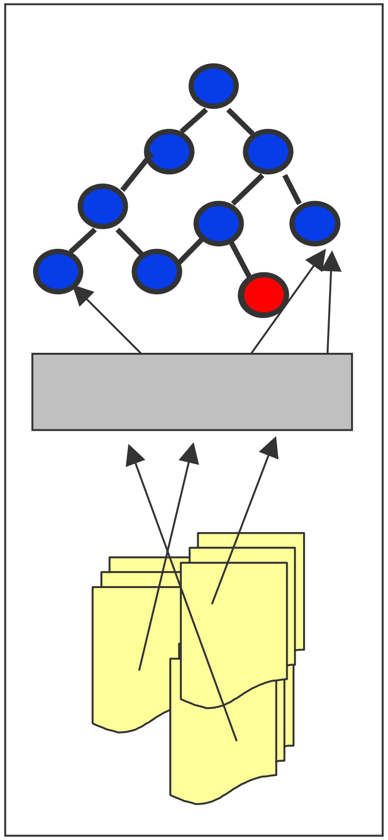
recall
1,00
0,80
0,60
0,40
0,20
0,00
1-3
0-1
2-3
0-3
3-4
1-2
0-4
0,00
0,20
0,40
precision

Non-Taxonomic Relation - Labeling

- **Problem: relations between concepts extracted via association rules are not labeled.**
- **Proposed extensions:**
  - Verbs are common representants of relations, based on information from POS-tagger
    1. Collect verb-concept pairs from corpus
    2. Score the verbs (use analogy of TFIDF measure for term-document occurences)
    3. Let the user select important verbs
  - Find and display verbs, which may be involved in relation between concepts, discovered by association rules, based on statistics of concept-verb occurrences of involved concepts

Pruning

- Given: An ontology (e.g. WordNet as OI-Model) and a set of domain-specific documents
- Approach: Delete all „unimportant“ concepts, means:
  - Based on the lexicon count weighted frequencies and propagate frequencies according to the taxonomy.
  - Define threshold and delete all concepts appearing less than the defined threshold
- A useful method to reuse existing resources (see UN application).



Agenda

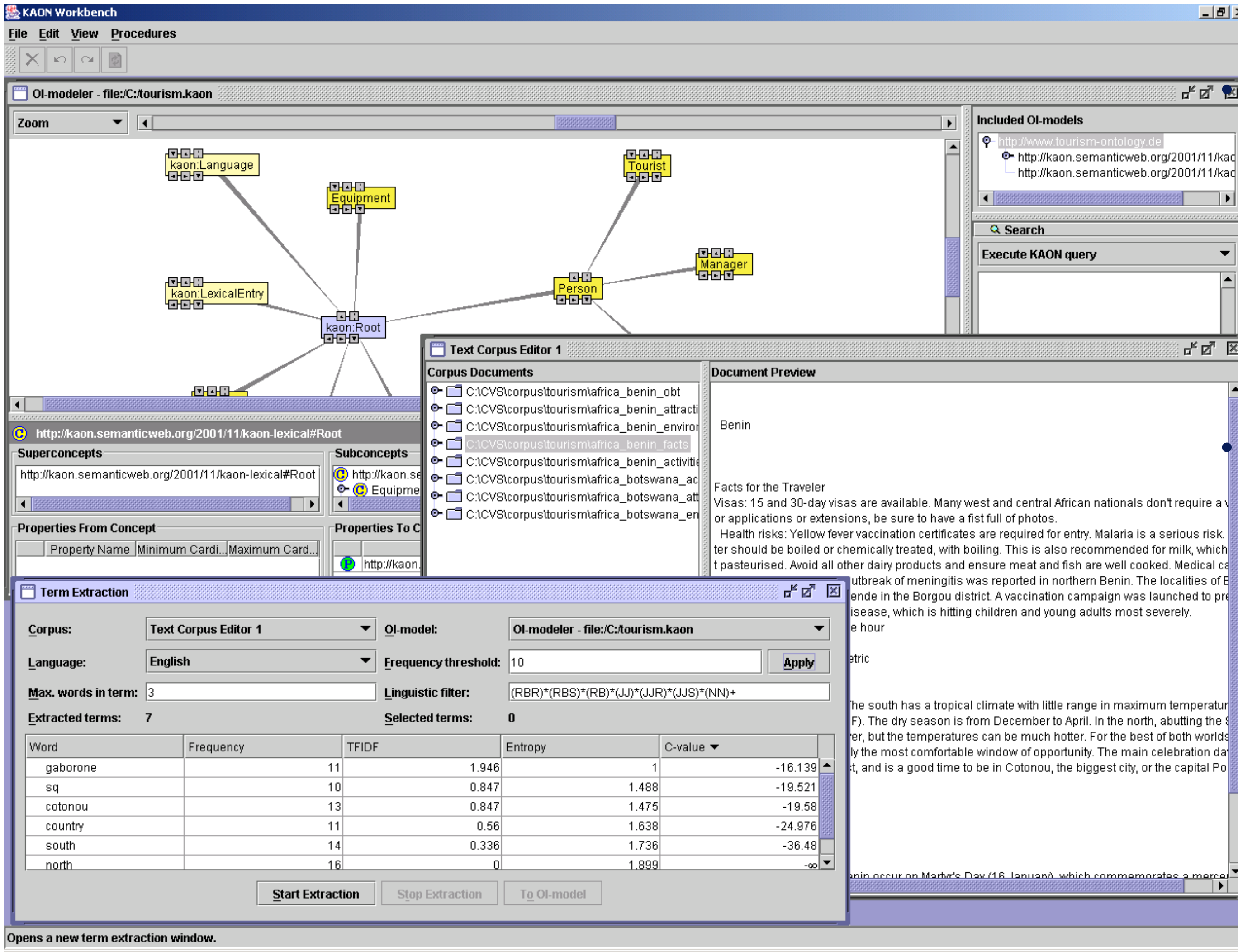
- Introduction & Motivation
- Ontology Learning Framework & Techniques
- Text-To-Onto Tool-Environment
- Applications
- Conclusion

KAON & Text-To-Onto

- KAON stands for Karlsruhe Ontology and Semantic Web Framework.
- Open Source platform for ontology-related tools, including
  - Ontology Modeling tools, including ontology learning
  - Scalable Ontology Server, including API, inference engine and query language.
- Open source under LGPL, available at:

  http://kaon.semanticweb.org

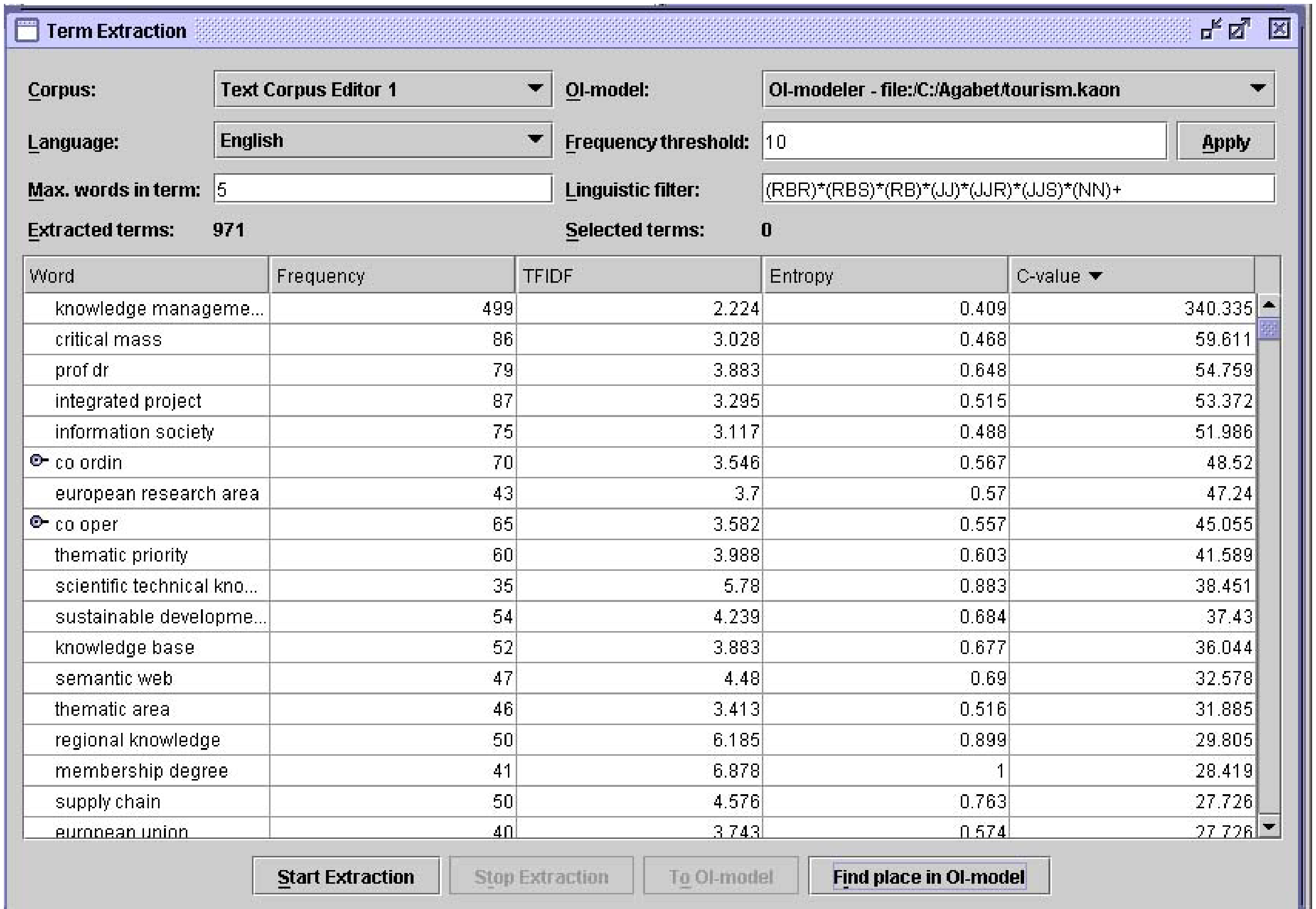
Text-To-Onto



Text-To-Onto is tightly integrated into the ontology management architecture KAON.

Balanced cooperative modeling approach, means that everything can be done manually, but automatic methods exist.

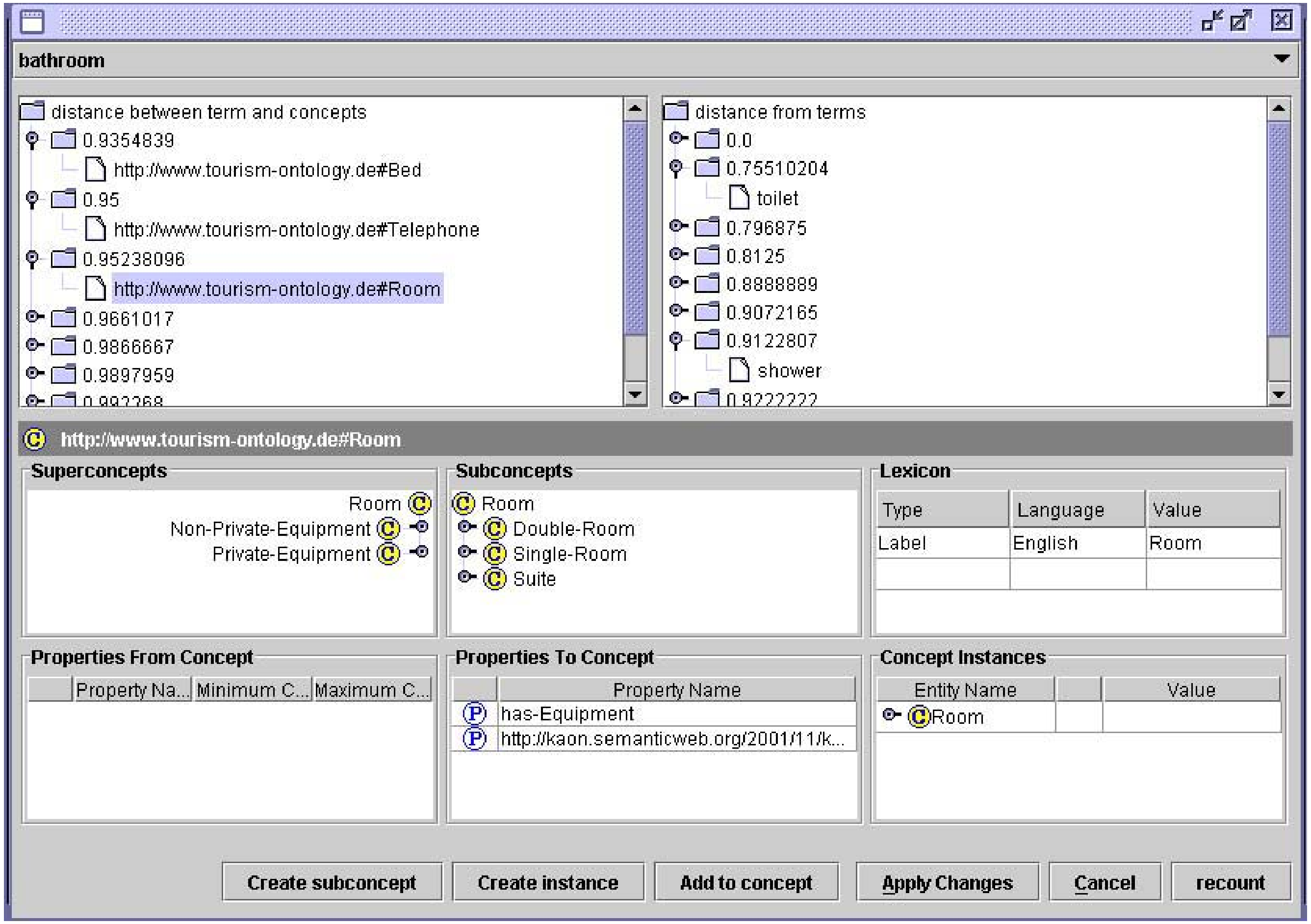
Multi-Word Term Extraction



Word	Frequency	TFIDF	Entropy	C-value
knowledge manageme...	499	2.224	0.409	340.335
critical mass	86	3.028	0.468	59.611
prof dr	79	3.883	0.648	54.759
integrated project	87	3.295	0.515	53.372
information society	75	3.117	0.488	51.986
co ordin	70	3.546	0.567	48.52
european research area	43	3.7	0.57	47.24
co oper	65	3.582	0.557	45.055
thematic priority	60	3.988	0.603	41.589
scientific technical kno...	35	5.78	0.883	38.451
sustainable developme...	54	4.239	0.684	37.43
knowledge base	52	3.883	0.677	36.044
semantic web	47	4.48	0.69	32.578
thematic area	46	3.413	0.516	31.885
regional knowledge	50	6.185	0.899	29.805
membership degree	41	6.878	1	28.419
supply chain	50	4.576	0.763	27.726
european union	40	3.743	0.574	27.726

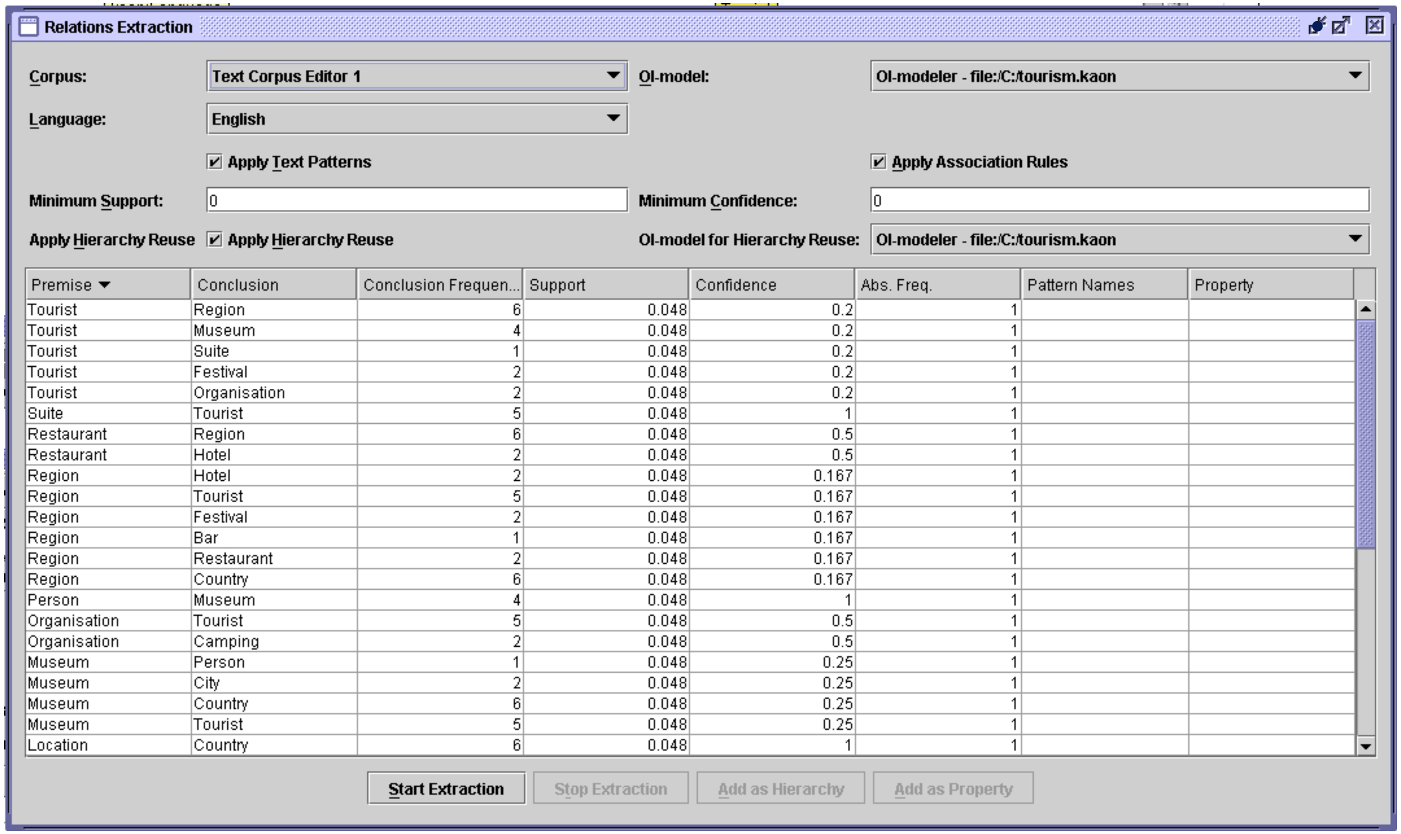
- **Baseline tool for multi-word term extraction.**

Multi-Word Term Meaning



- Supports the different process of classifying extracted terms into the ontology.

Concept Relation Extraction



Relations Extraction

Corpus: Text Corpus Editor 1
Ol-model: Ol-modeler - file:/C:/tourism.kaon
Language: English
☑ Apply Text Patterns
☑ Apply Association Rules
Minimum Support: 0
Minimum Confidence: 0
Apply Hierarchy Reuse ☑ Apply Hierarchy Reuse
Ol-model for Hierarchy Reuse: Ol-modeler - file:/C:/tourism.kaon

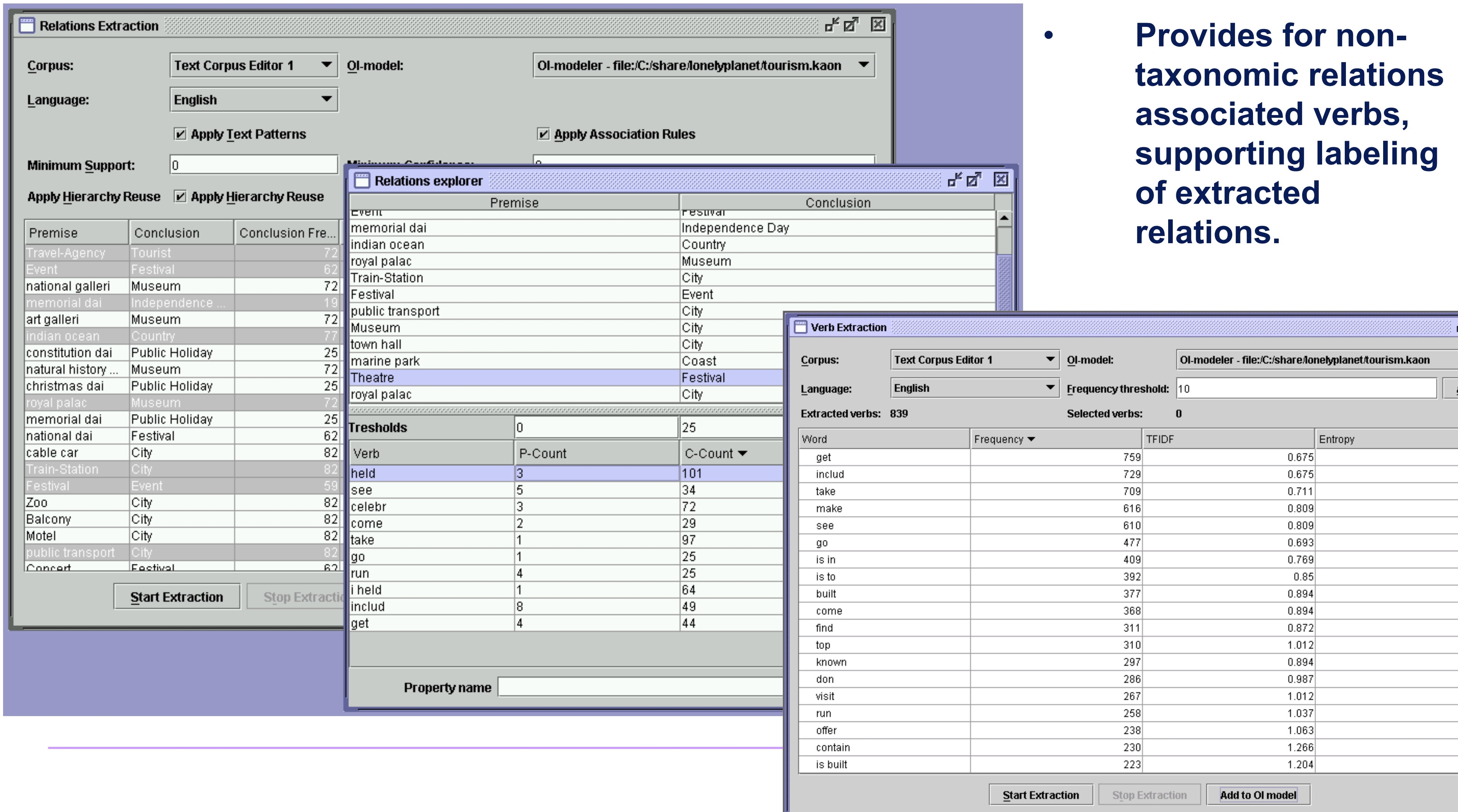
Premise	Conclusion	Conclusion Frequen...	Support	Confidence	Abs. Freq.	Pattern Names	Property
Tourist	Region	6	0.048	0.2	1		
Tourist	Museum	4	0.048	0.2	1		
Tourist	Suite	1	0.048	0.2	1		
Tourist	Festival	2	0.048	0.2	1		
Tourist	Organisation	2	0.048	0.2	1		
Suite	Tourist	5	0.048	1	1		
Restaurant	Region	6	0.048	0.5	1		
Restaurant	Hotel	2	0.048	0.5	1		
Region	Hotel	2	0.048	0.167	1		
Region	Tourist	5	0.048	0.167	1		
Region	Festival	2	0.048	0.167	1		
Region	Bar	1	0.048	0.167	1		
Region	Restaurant	2	0.048	0.167	1		
Region	Country	6	0.048	0.167	1		
Person	Museum	4	0.048	1	1		
Organisation	Tourist	5	0.048	0.5	1		
Organisation	Camping	2	0.048	0.5	1		
Museum	Person	1	0.048	0.25	1		
Museum	City	2	0.048	0.25	1		
Museum	Country	6	0.048	0.25	1		
Museum	Tourist	5	0.048	0.25	1		
Location	Country	6	0.048	1	1		

Start Extraction | Stop Extraction | Add as Hierarchy | Add as Property

Integrated view for extracting relations, including:

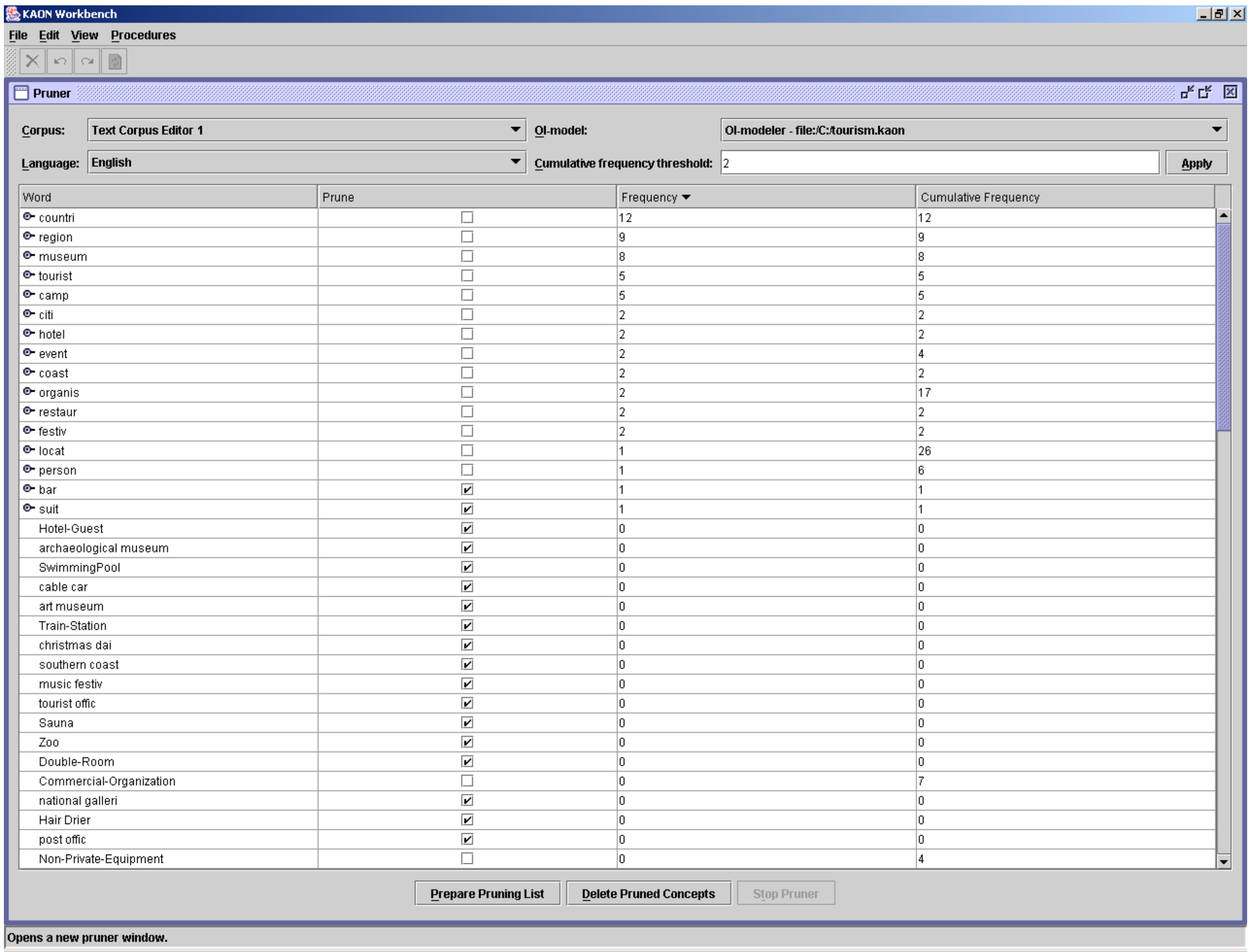
- **Association rules**
- **Pattern based extraction**
- **Taxonomy reuse**

Relation explorer



- Provides for non-taxonomic relations associated verbs, supporting labeling of extracted relations.

Ontology Pruning/Reuse



- Allows to user to prune existing ontologies according to a predefined corpus.

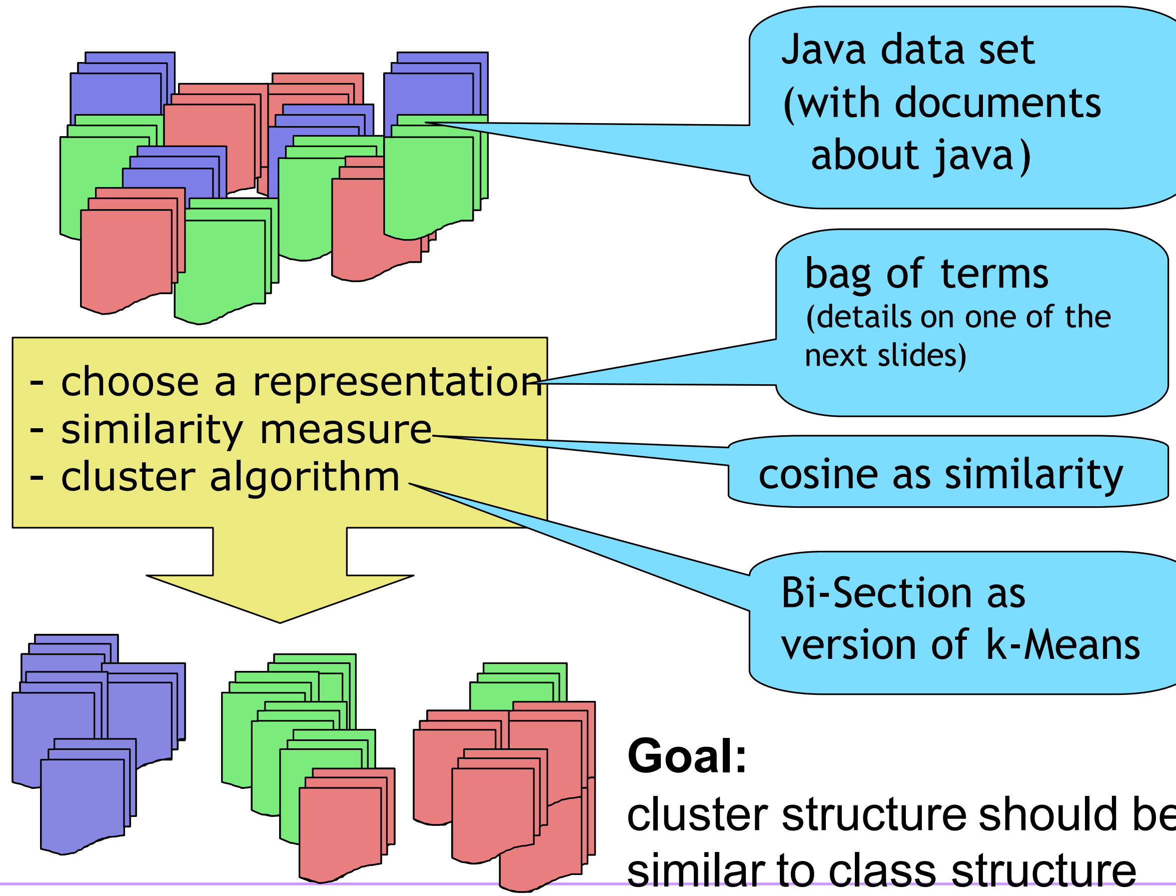
Agenda

- Introduction & Motivation
- Ontology Learning Framework & Techniques
- Text-To-Onto Tool-Environment
- Applications
- Conclusion

Applications of Ontology Learning

- Text Clustering
  - Exploit ontological background knowledge for document clustering
- Information Extraction
  - Use ontologies as templates for extracting information
- Document Search Application
  - Exploit ontologies for document search

Text Clustering with Background Knowledge(*)



Goal:
cluster structure should be similar to class structure

(*) work done by Andreas Hotho, University of Karlsruhe

Background Knowledge

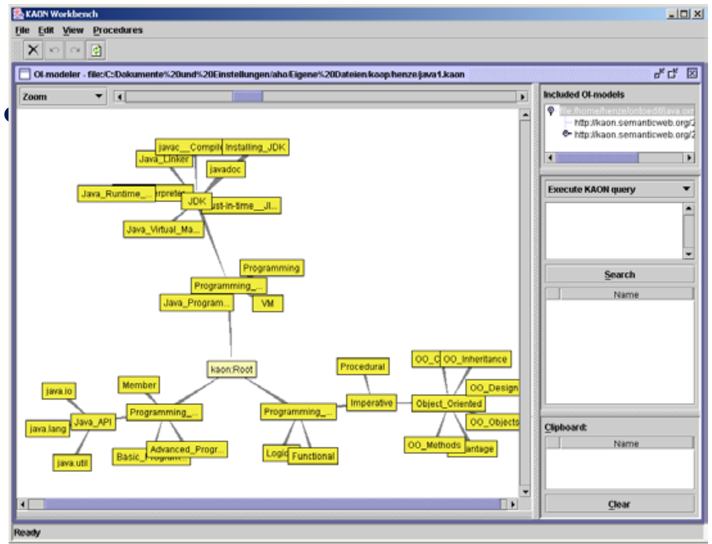
Bag of words model:

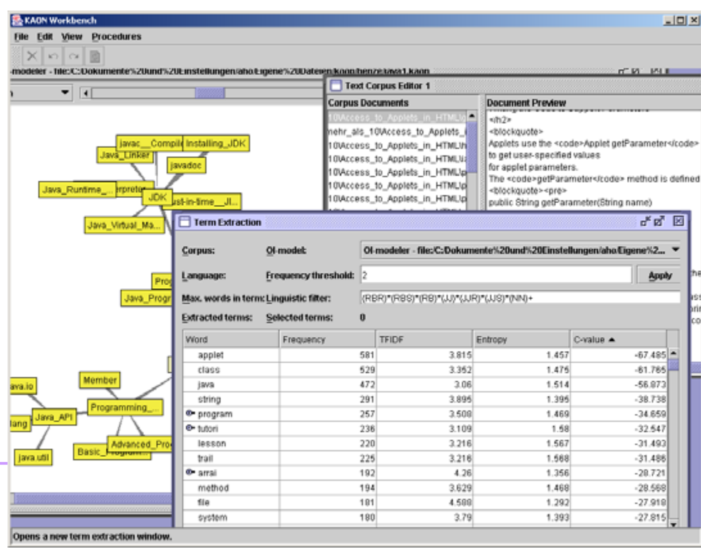
docid	term1	term2	term3	...
doc1	0	0	1	
doc2	2	3	1	
doc3	10	0	0	
doc4	2	23	0	
...				

Bag of concept model (term and concept vector):

docid	term1	term2	term3	...	concept1	concept2	concept3	..
doc1	0	0	1		0	1	1	
doc2	2	3	1		2	0	1	
doc3	10	0	0		10	0	0	
doc4	2	23	0		2	23	0	
...								

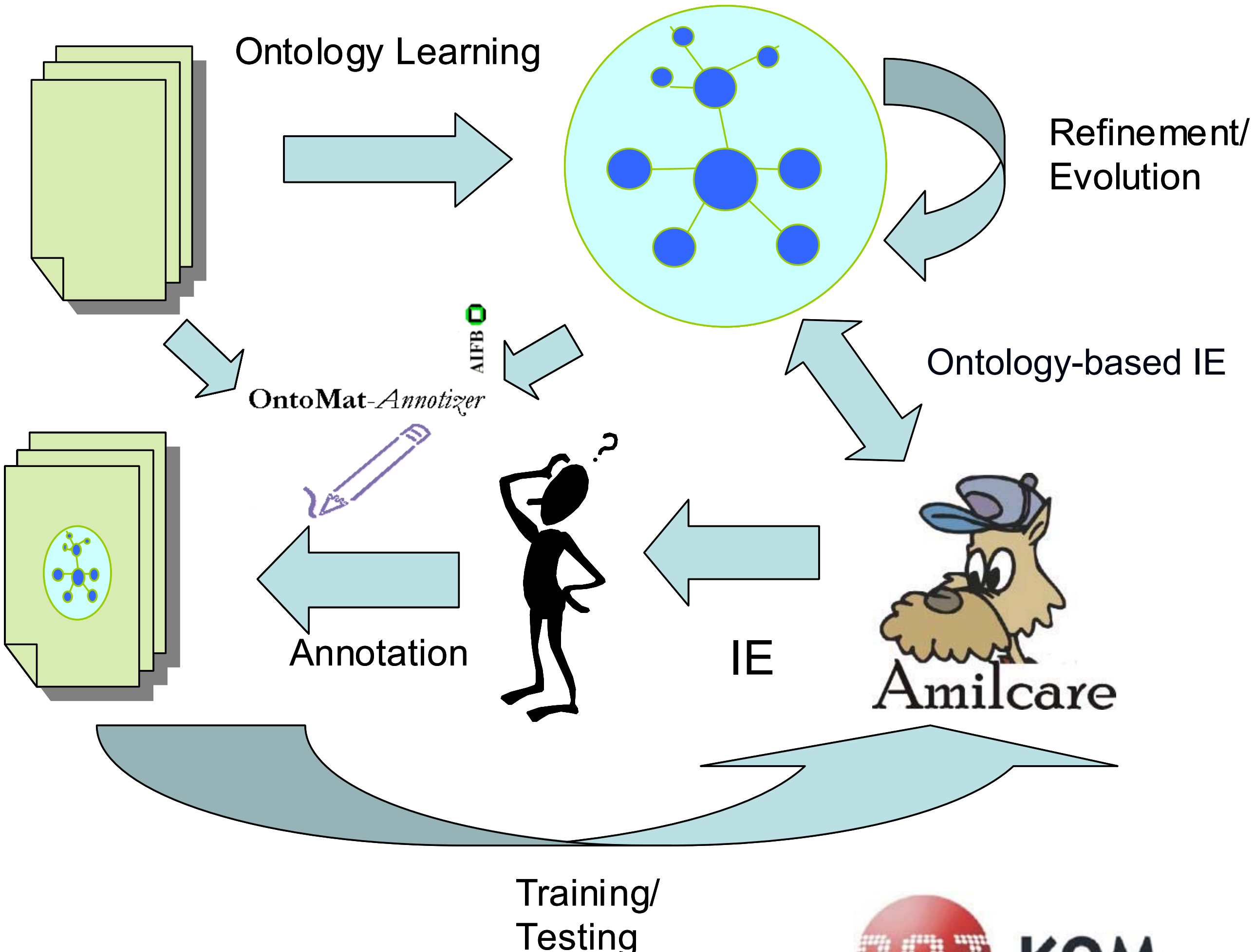
Results





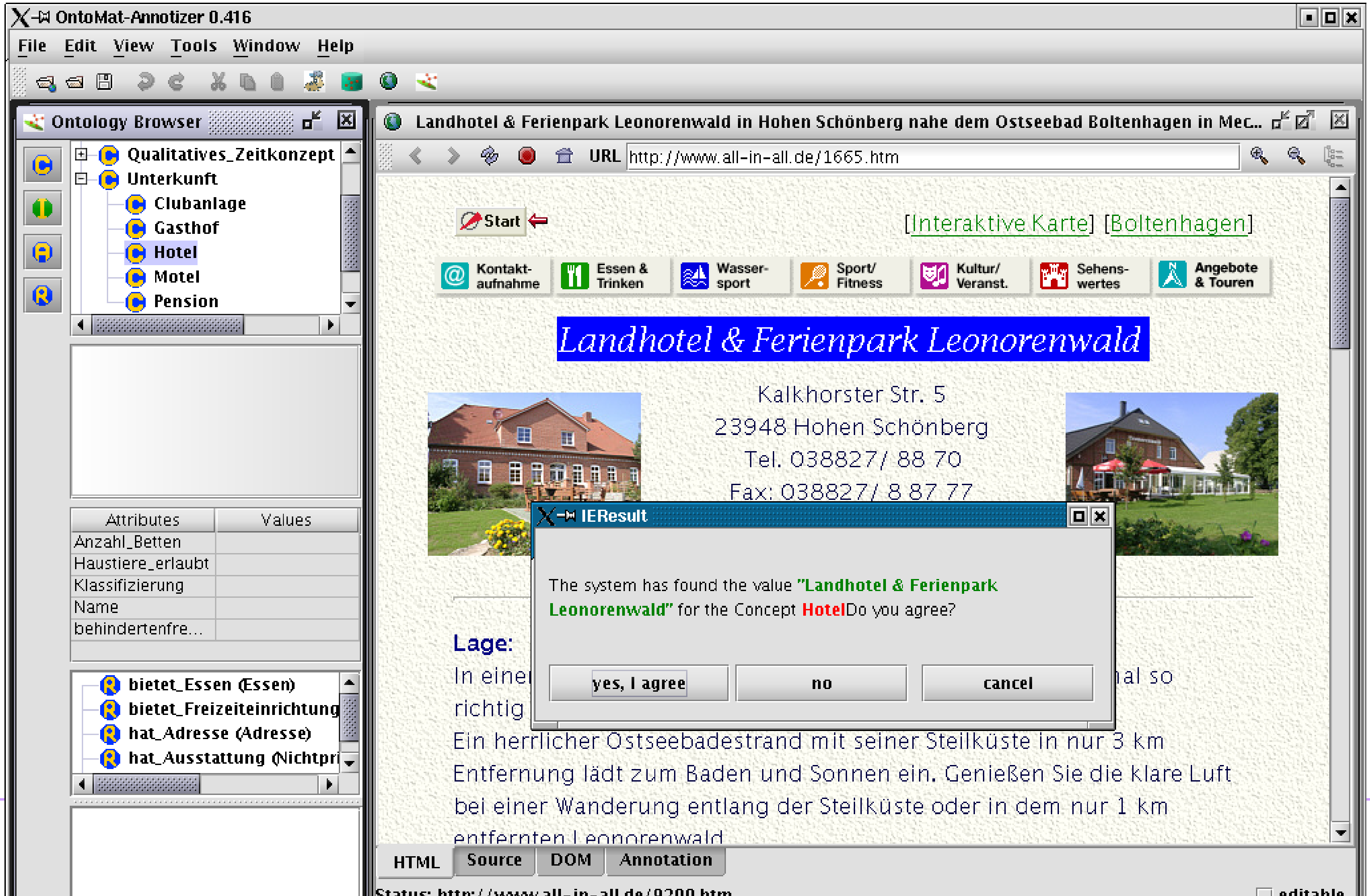
- **Without Ontology = Baseline**
  - **Purity = 62%, Inverse Purity = 61%**
- **With handmade Ontology**
  - **Purity = 60%, Inverse Purity = 57%**
- **Ontology improved by Ontology Learning**
  - **Purity = 67%, Inverse Purity = 64%**

Information Extraction (*)



(*) work done within the EU-IST funded DOT.KOM project.

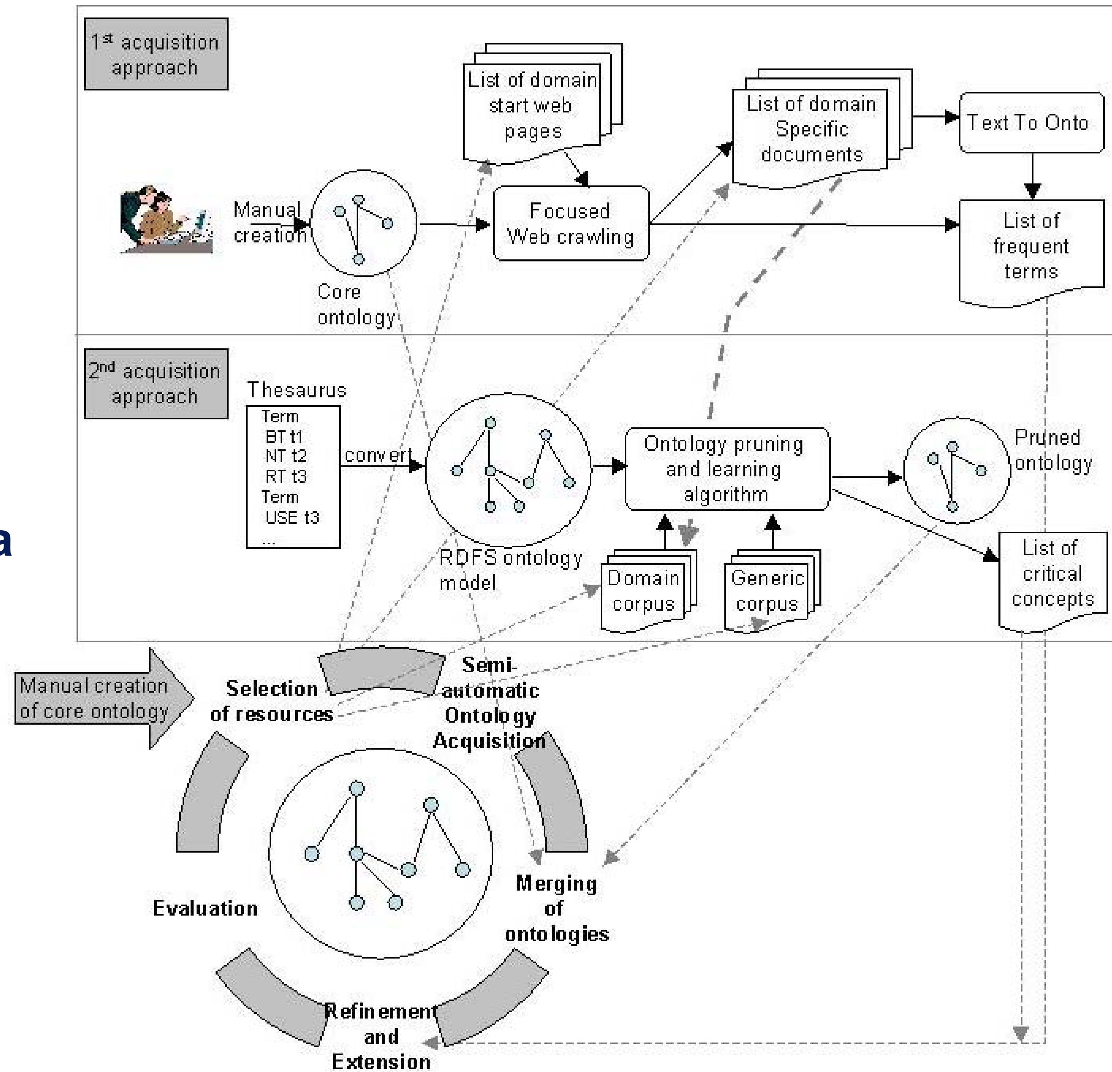
Ontology-based Information Extraction



Document Search Application

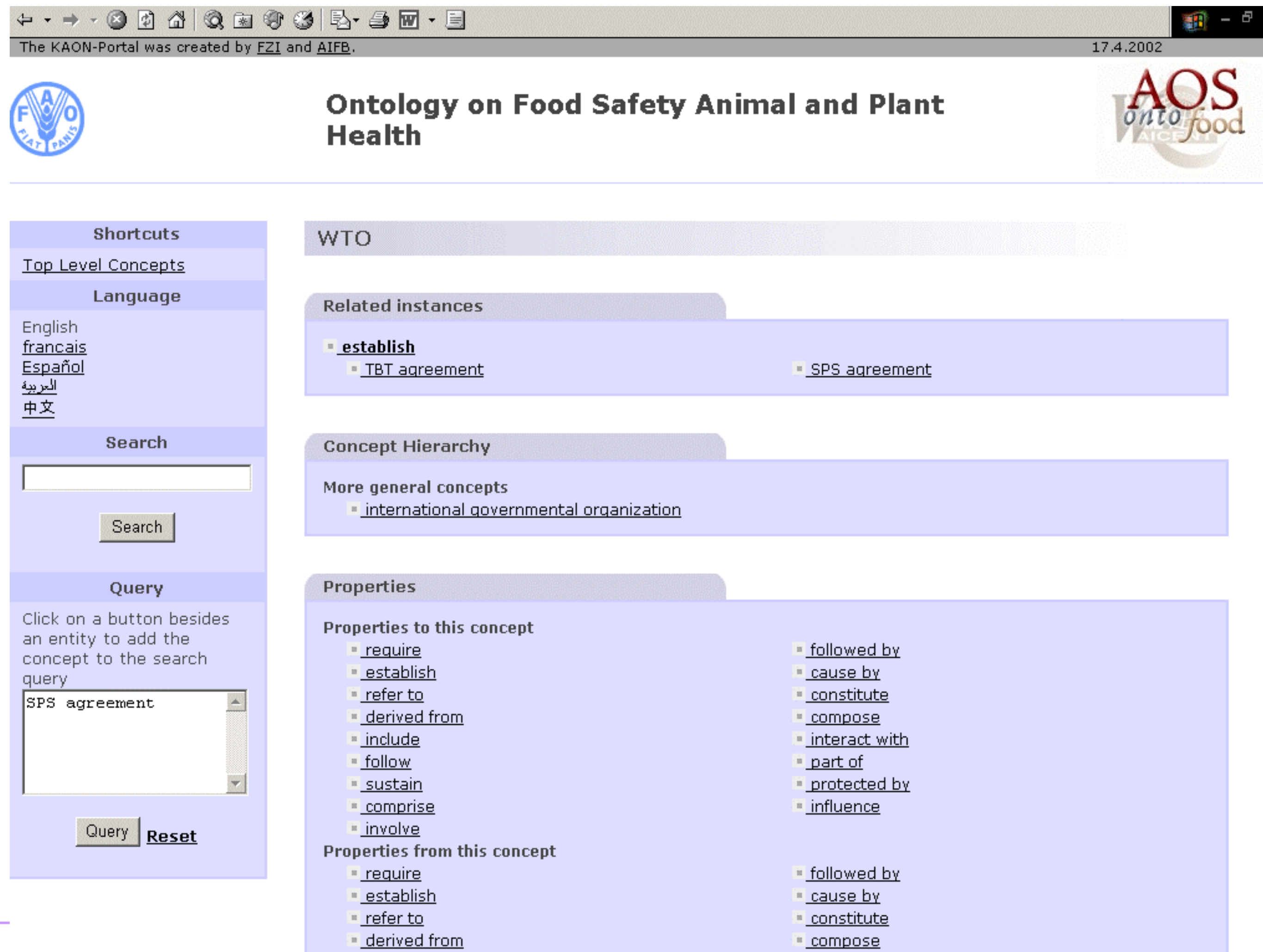
- The Food and Agricultural Organization (FAO) within United Nations is providing means for information dissemination.

- On the basis of the thesaurus AGROVOC a domain specific ontology (food safety animal and plant health) has been generated using pruning.



United Nations FAO Application

- Query expansion, ontology-based retrieval of documents
- Exploit extracted semantic relations for guiding the user in the search.



Agenda

- Introduction & Motivation
- Ontology Learning Framework & Techniques
- Text-To-Onto Tool-Environment
- Applications
- Conclusion

Conclusion

- Ontologies are central for realizing the vision of semantics-based processing of information.
- Ontology learning is a promising step towards approaching the knowledge acquisition bottleneck.
- In this presentation a balanced cooperative approach has been presented.

Some Comments for MEANING

- Knowledge representation issue: How far do you go with semantics?
- Standards issue: The MEANING repository should be somehow aligned with existing standards to make the resources more widely usable.
- Tool issue: To make algorithms usable they have to be integrated into a tool environment and a common framework.

Thank you for your attention!

A. Maedche

Forschungszentrum Informatik an der Universität Karlsruhe

Research Group WIM

http://www.fzi.de/wim

Results

filter	Corpus1 „Human Language Technology“	Corpus2 „EoI-Knowledge-Technologies“
[(NNS)(NN)]+	**88% (*)**	**80%**
	(1230,233,27) (**)	**(1079,202,40)**
	86%	**76%**
(RB)*(JJ)*[(NN)(NNS)]+		
	(1362,362,47)	**(1243,361,85)**
[(RB)(JJ)(NN)]*(IN)?	**64%**	**64%**
[(RB)(JJ)(NN)]*		
[(NN)(NNS)]	**(1511,511,181)**	**(1362,478,171)**

(*) Of precision

(**) (number of all extracted terms, number of multiword terms, number of incorectly extracted multiword terms)

Preprocessing

build a bag of words model

docid	term1	term2	term3	...
doc1	0	0	1	
doc2	2	3	1	
doc3	10	0	0	
doc4	2	23	0	
...				

extract word counts (term frequencies)
remove stopwords
pruning: drop words with less than 30 occurrences
weighting of document vectors with tfidf

(term frequency - inverted document frequency)

$$tfidf(d,t) = tf(d,t) * \log\left(\frac{|D|}{df(t)}\right)$$

$|D|$ no. of documents d
$df(t)$ no. of documents d which contain term t