

# Ontology Learning: Framework, Techniques and a Software Environment

MEANING WS Presentation,  
San Sebastian

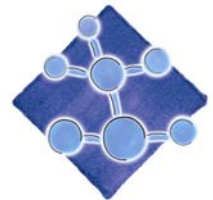


**Alexander Maedche**

Forschungszentrum Informatik an der Universität Karlsruhe

Forschungsbereich Wissensmanagement (WIM)

<http://www.fzi.de/wim>



# Agenda

- 
- A vertical strip on the left side of the slide shows a close-up of a calendar page. The numbers 7, 8, 9, and 10 are visible, along with a metal ring binding the page. The lighting is warm and golden.
- **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.**



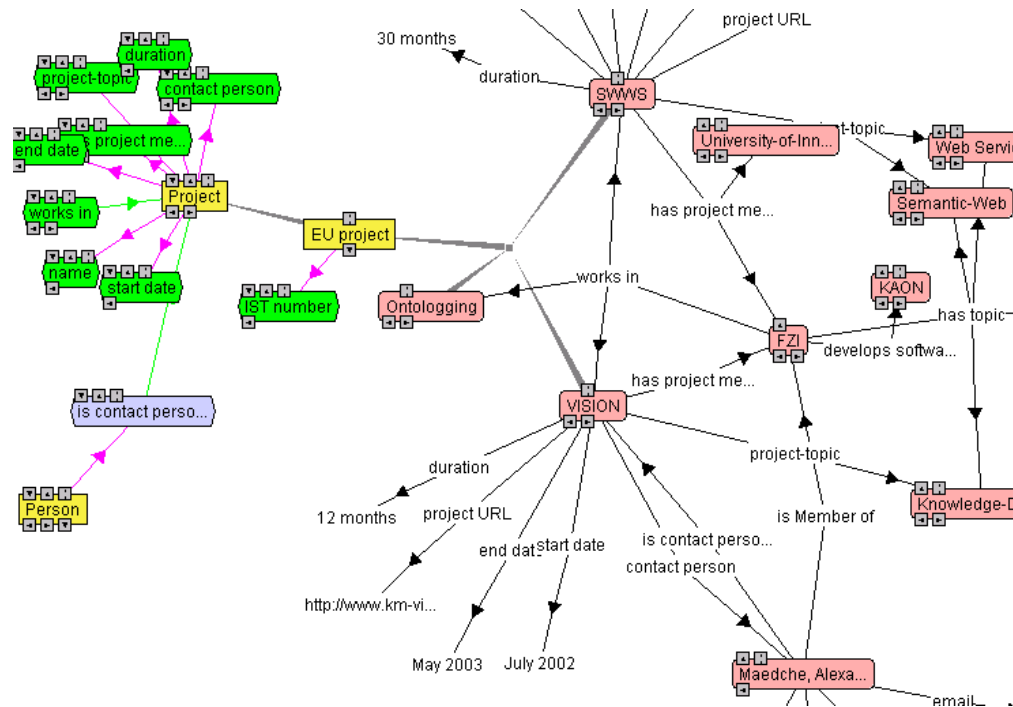
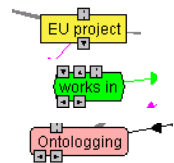
# Agenda

- 
- A vertical strip on the left side of the slide shows a close-up of a calendar page. The numbers 7, 8, 9, and 10 are visible, along with a metal ring binding the pages. The lighting is warm and golden.
- **Introduction & Motivation**
  - **Ontology Learning Framework & Techniques**
  - **Text-To-Onto Tool-Environment**
  - **Applications**
  - **Conclusion**

# Ontologies

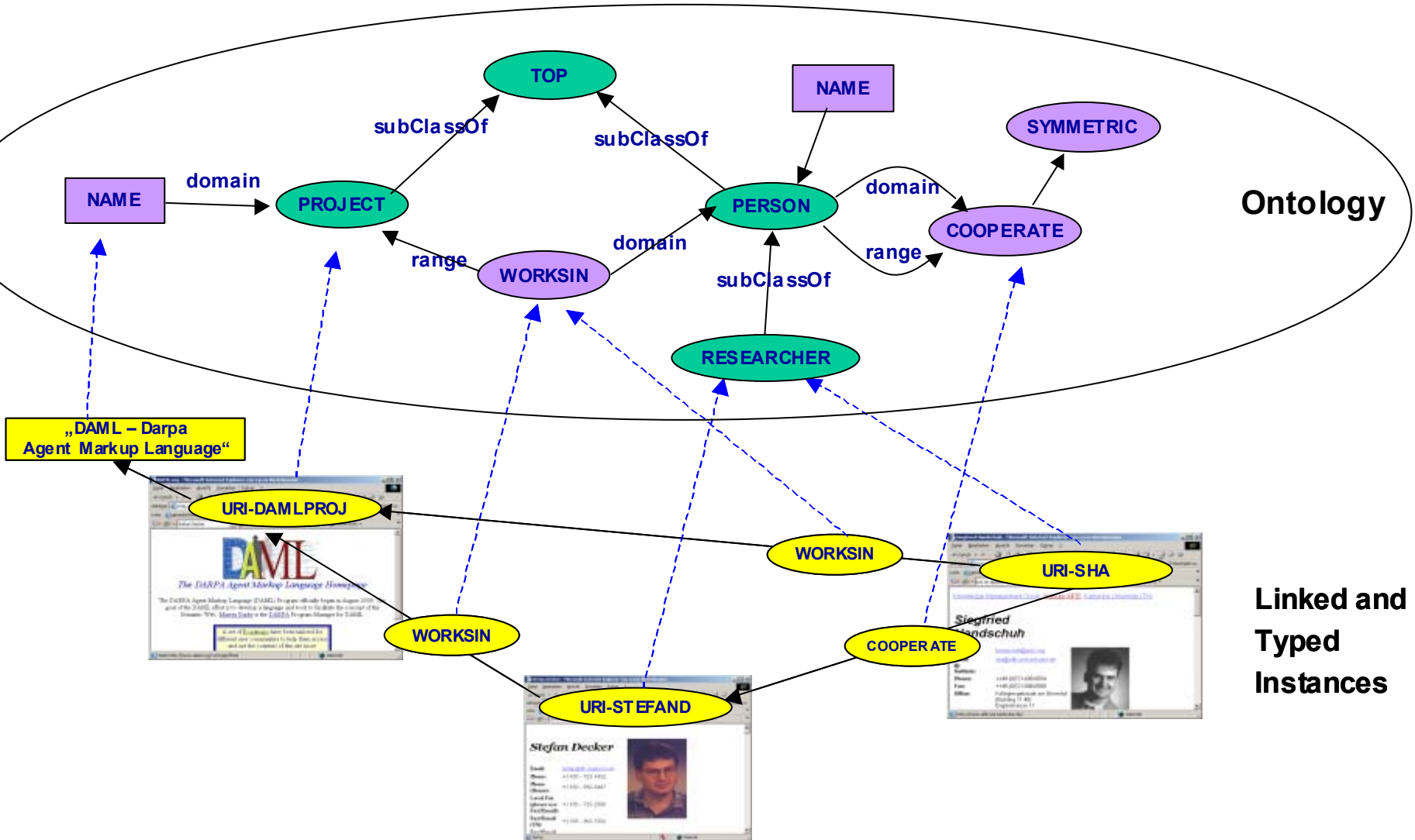
- Expressive conceptual models, no strict separation between schema and instance.
- OI-model (ontology-instance model) – elementary information container, contains ontology and instance data:

- concepts
- relations
- instances
- relation instances



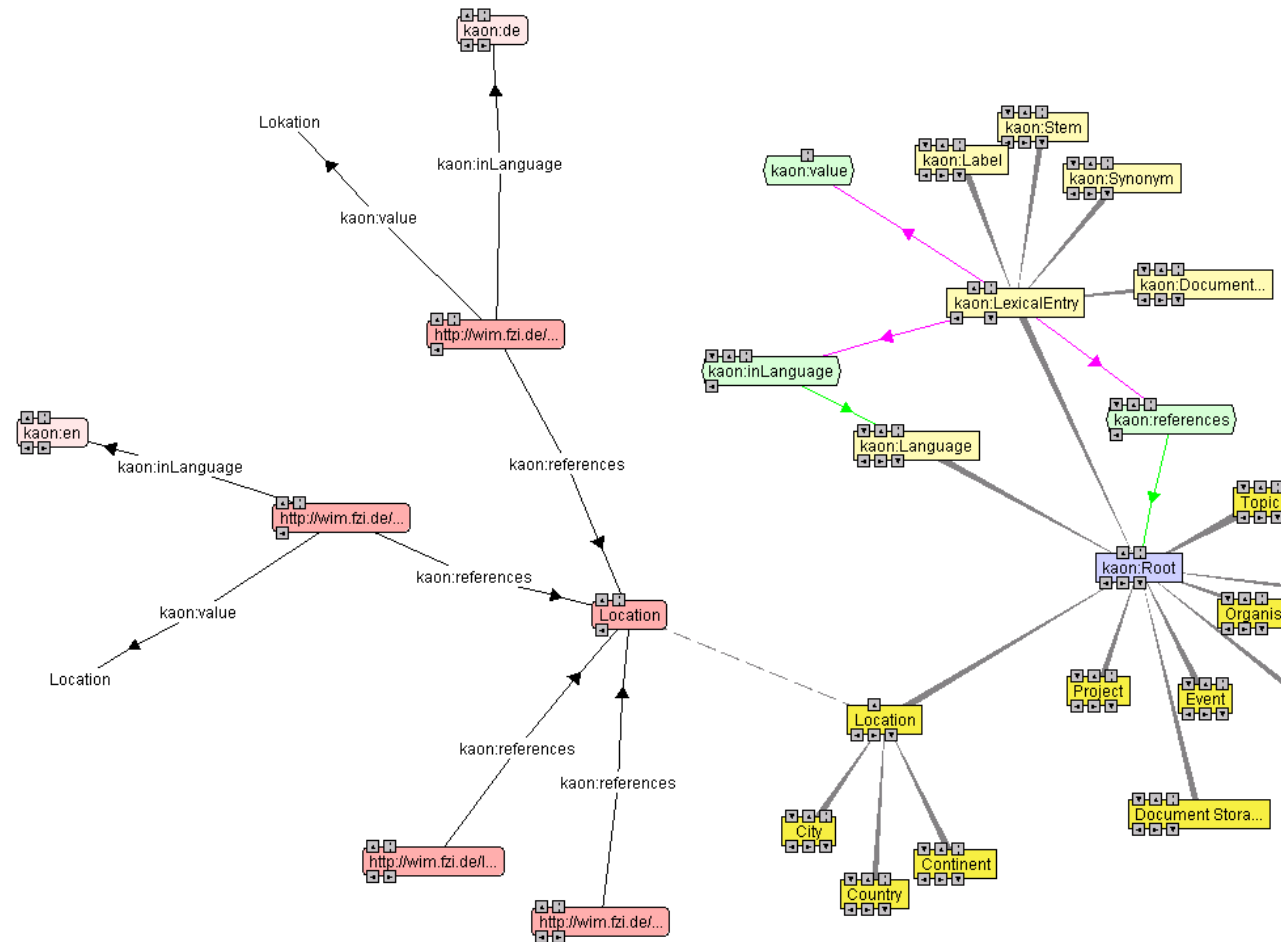
- Extensions of W3C's RDF-Schema, along the same lines of W3C OWL.
- Builds on an expressive hybrid knowledge representation mechanism, inspired by Description Logics paradigm, but executed using deductive database techniques.

# 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.

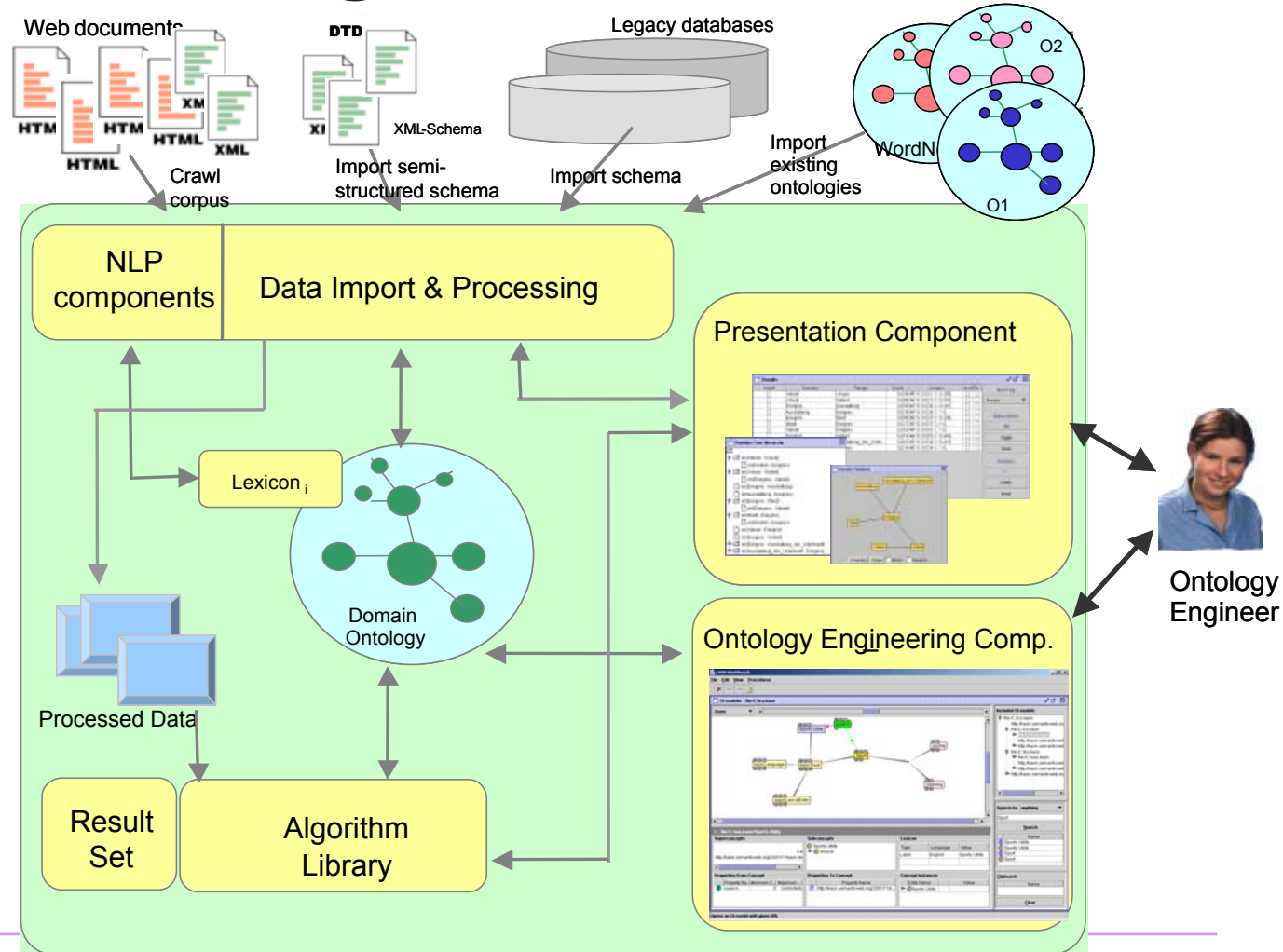






# Ontology Learning Framework

- 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:

$$\{\text{term}(\text{verb}_1, \text{freq}), \dots, (\text{verb}_n, \text{freq}), (\text{noun}_1, \text{freq}), \dots, (\text{noun}_t, \text{freq})\}$$

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

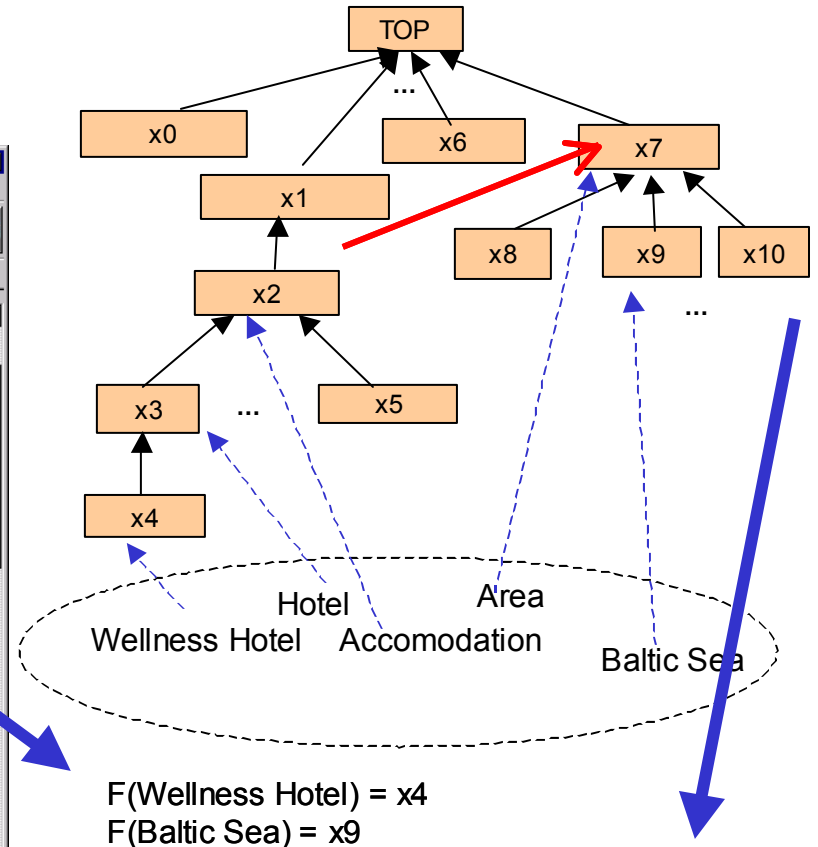
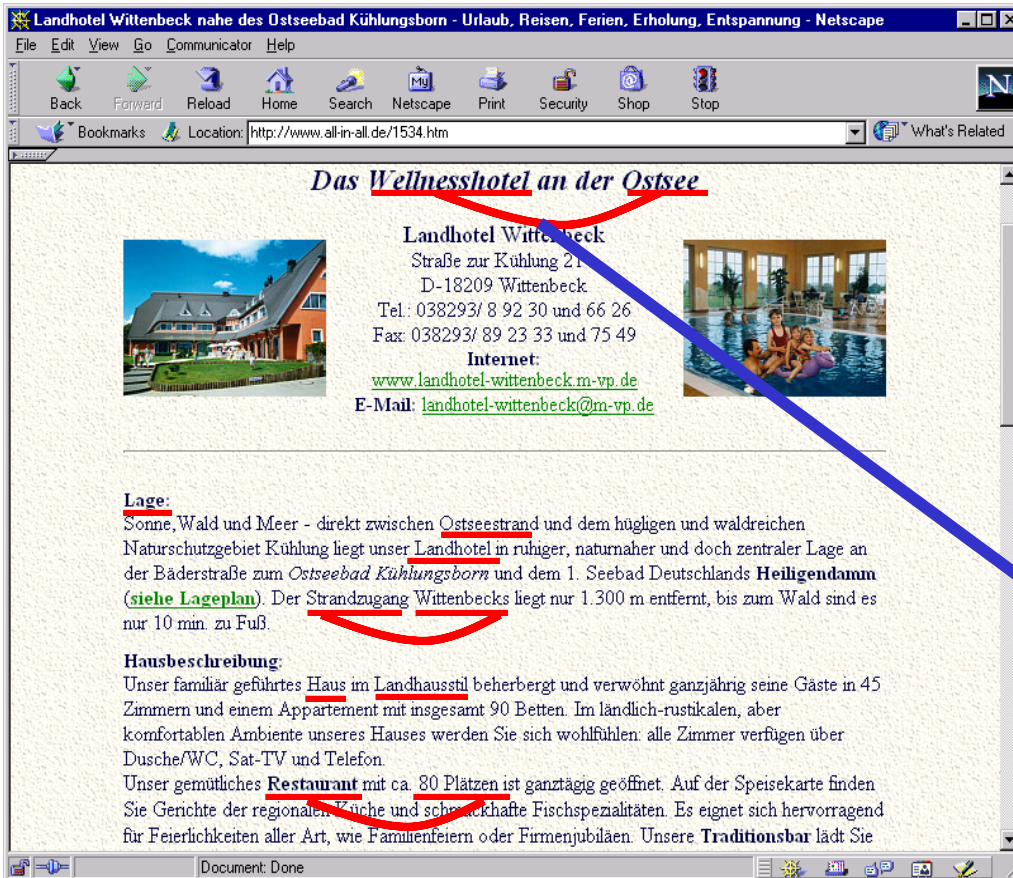
## Concept Hierarchy Extraction

- 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

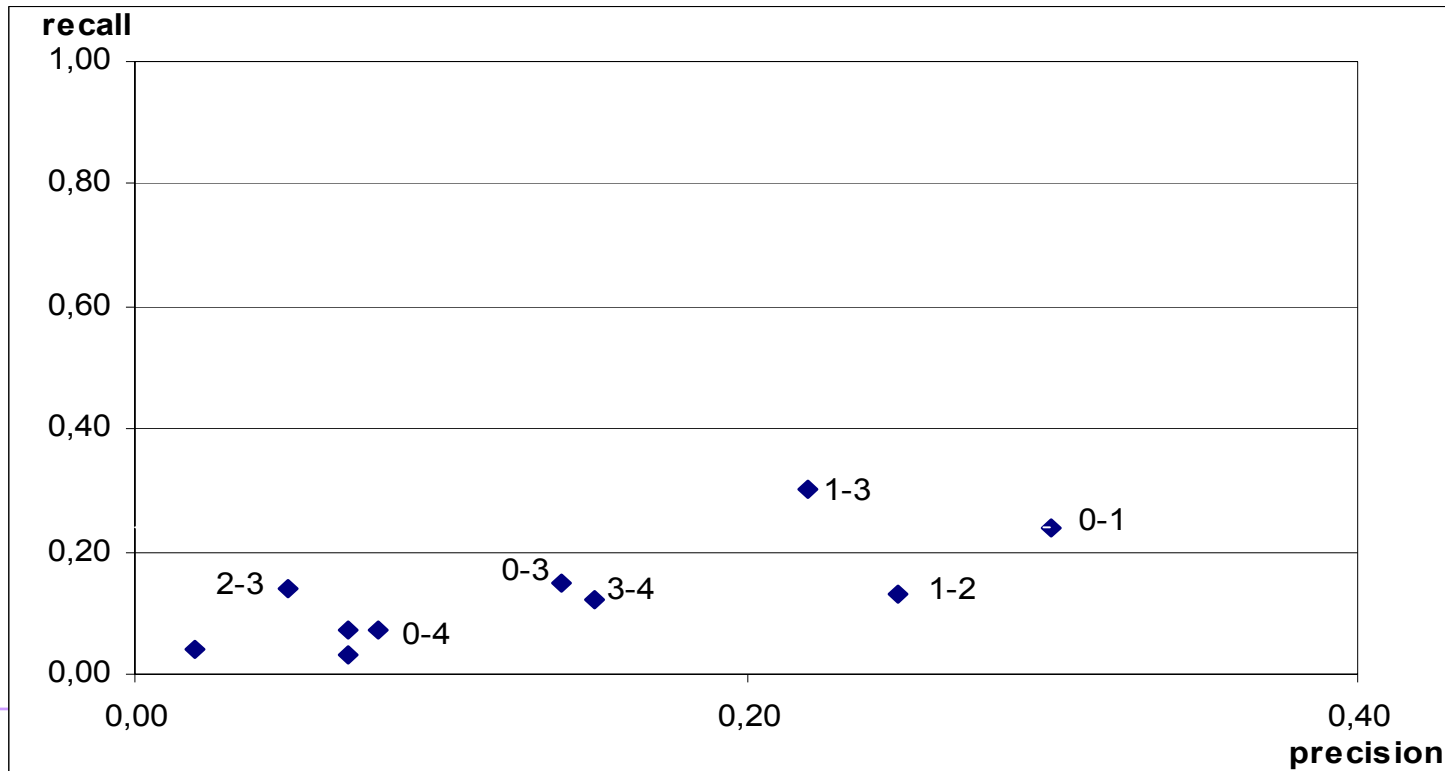
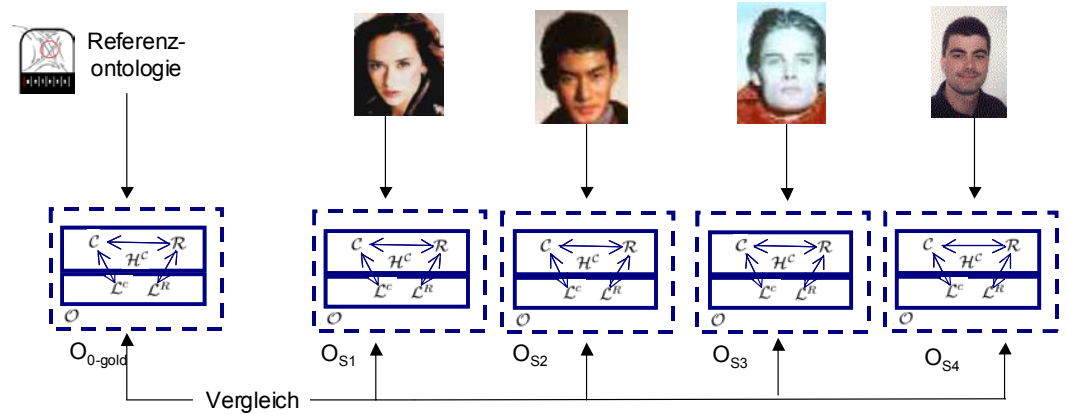


**Concept pair (ling. transaction)**  
 (x4,x9) bzw. (F(Wellness Hotel), F(Baltic Sea))

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



# Evaluation



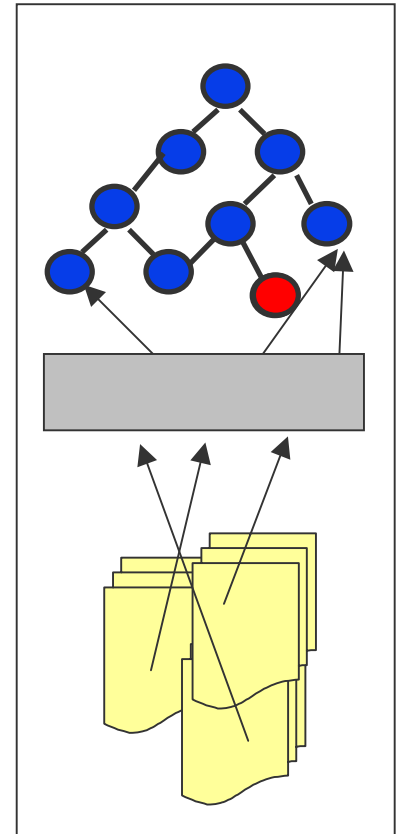
---

# 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 occurrences)**
    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

- 
- A vertical strip on the left side of the slide shows a close-up of a calendar page. The numbers 7, 8, 9, and 10 are visible, along with a metal ring binding the page. The lighting is warm and golden.
- **Introduction & Motivation**
  - **Ontology Learning Framework & Techniques**
  - **Text-To-Onto Tool-Environment**
  - **Applications**
  - **Conclusion**

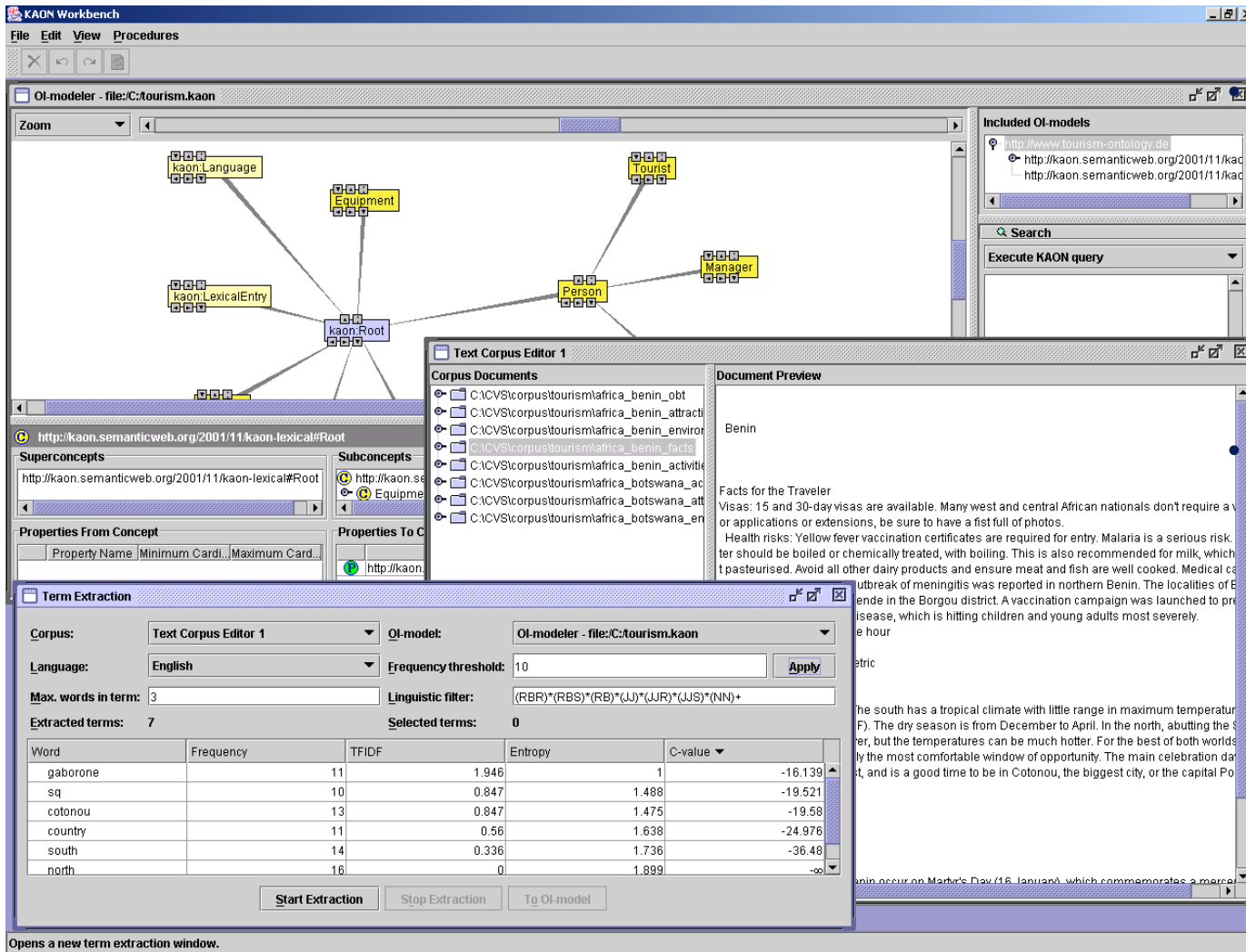
---

# 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



The screenshot displays the KAON Workbench interface. The main window shows an ontology model with a central 'kaon\_Root' node connected to 'kaon\_Language', 'kaon\_LexicalEntry', 'Equipment', 'Person', 'Tourist', and 'Manager'. A 'Text Corpus Editor 1' window is open, showing a list of corpus documents and a preview of a document about Benin. A 'Term Extraction' window is also open, showing settings for corpus, language, and frequency threshold, along with a table of extracted terms.

Word	Frequency	TFIDF	Entropy	C-value
gaborone	11	1.946	1	-16.139
sq	10	0.847	1.488	-19.521
cotonou	13	0.847	1.475	-19.58
country	11	0.56	1.638	-24.976
south	14	0.336	1.736	-36.48
north	16	0	1.899	-99

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

**Term Extraction**

Corpus:  OI-model:

Language:  Frequency threshold:

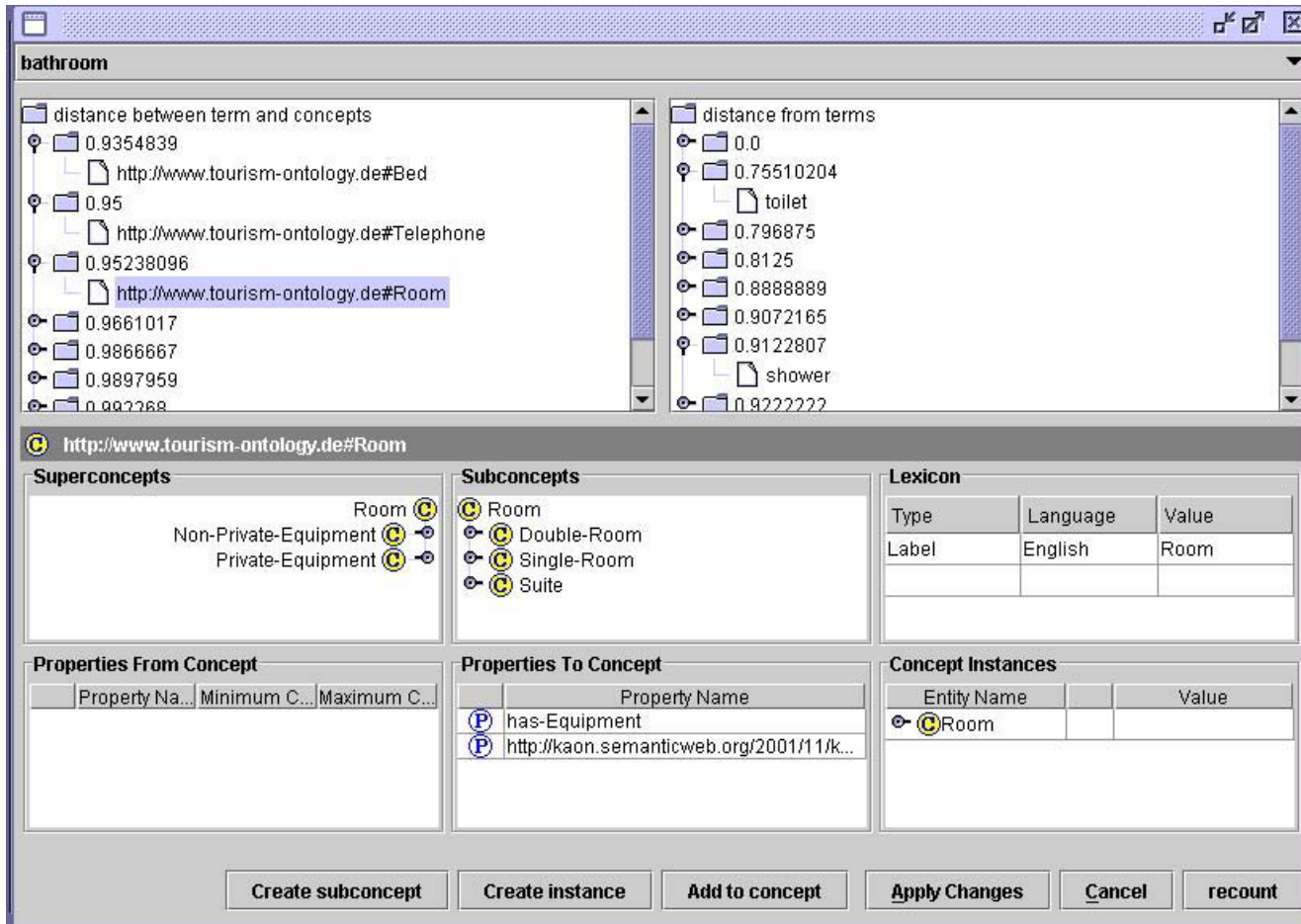
Max. words in term:  Linguistic filter:

Extracted terms: 971 Selected terms: 0

Word	Frequency	TFIDF	Entropy	C-value
knowledge managemen...	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
eurnean union	40	3.743	0.574	27.726

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

# Multi-Word Term Meaning



The screenshot shows a software interface for ontology management. The main window is titled "bathroom" and displays a hierarchical tree structure of terms. The left pane shows a list of terms with their distances from a root concept, and the right pane shows a similar list for a specific term. Below the tree, there are several panels for managing the selected concept, "Room".

**distance between term and concepts**

- 0.9354839
  - http://www.tourism-ontology.de#Bed
- 0.95
  - http://www.tourism-ontology.de#Telephone
- 0.95238096
  - http://www.tourism-ontology.de#Room
- 0.9661017
- 0.9866667
- 0.9897959
- 0.992268

**distance from terms**

- 0.0
  - 0.75510204
    - toilet
- 0.796875
- 0.8125
- 0.8888889
- 0.9072165
- 0.9122807
  - shower
- 0.9222222

**Superconcepts**

- Room
- Non-Private-Equipment
- Private-Equipment

**Subconcepts**

- Room
- Double-Room
- Single-Room
- Suite

**Lexicon**

Type	Language	Value
Label	English	Room

**Properties From Concept**

Property Na...	Minimum C...	Maximum C...
----------------	--------------	--------------

**Properties To Concept**

Property Name
has-Equipment
http://kaon.semanticweb.org/2001/11/k...

**Concept Instances**

Entity Name	Value
Room	

Buttons at the bottom: Create subconcept, Create instance, Add to concept, Apply Changes, Cancel, recount

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



# Concept Relation Extraction

**Relations Extraction**

Corpus:  OI-model:

Language:

Apply Text Patterns  Apply Association Rules

Minimum Support:  Minimum Confidence:

Apply Hierarchy Reuse  Apply Hierarchy Reuse OI-model for Hierarchy Reuse:

Premise ▼	Conclusion	Conclusion Freq...	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.

**Relations Extraction**

Corpus: Text Corpus Editor 1    OI-model: OI-modeler - file:C:/share/lonelyplanet/tourism.kaon

Language: English

Apply Text Patterns     Apply Association Rules

Minimum Support: 0    Minimum Confidence: 0

Apply Hierarchy Reuse  Apply Hierarchy Reuse

Premise	Conclusion	Conclusion Fre...
Travel-Agency	Tourist	72
Event	Festival	62
national galleri	Museum	72
memorial dai	Independence ...	19
art galleri	Museum	72
indian ocean	Country	77
constitution dai	Public Holiday	25
natural history ...	Museum	72
christmas dai	Public Holiday	25
royal palac	Museum	72
memorial dai	Public Holiday	25
national dai	Festival	62
cable car	City	82
Train-Station	City	82
Festival	Event	59
Zoo	City	82
Balcony	City	82
Motel	City	82
public transport	City	82
Concert	Festival	62

**Start Extraction**    **Stop Extraction**

**Relations explorer**

Premise	Conclusion
event	festival
memorial dai	Independence Day
indian ocean	Country
royal palac	Museum
Train-Station	City
Festival	Event
public transport	City
Museum	City
town hall	City
marine park	Coast
Theatre	Festival
royal palac	City

Verb	P-Count	C-Count
held	3	101
see	5	34
celebr	3	72
come	2	29
take	1	97
go	1	25
run	4	25
i held	1	64
includ	8	49
get	4	44

Property name:

**Verb Extraction**

Corpus: Text Corpus Editor 1    OI-model: OI-modeler - file:C:/share/lonelyplanet/tourism.kaon

Language: English    Frequency threshold: 10

Extracted verbs: 839    Selected verbs: 0

Word	Frequency	TFIDF	Entropy
get	759		0.675
includ	729		0.675
take	709		0.711
make	616		0.809
see	610		0.809
go	477		0.693
is in	409		0.769
is to	392		0.85
built	377		0.894
come	368		0.894
find	311		0.872
top	310		1.012
known	297		0.894
don	286		0.987
visit	267		1.012
run	258		1.037
offer	238		1.063
contain	230		1.266
is built	223		1.204

**Start Extraction**    **Stop Extraction**    **Add to OI model**

# Ontology Pruning/Reuse

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

KAON Workbench

File Edit View Procedures

Pruner

Corpus: Text Corpus Editor 1 Ol-model: Ol-modeler - file:C:/tourism.kaon

Language: English Cumulative frequency threshold: 2 Apply

Word	Prune	Frequency	Cumulative Frequency
countri	<input type="checkbox"/>	12	12
region	<input type="checkbox"/>	9	9
museum	<input type="checkbox"/>	8	8
tourist	<input type="checkbox"/>	5	5
camp	<input type="checkbox"/>	5	5
citi	<input type="checkbox"/>	2	2
hotel	<input type="checkbox"/>	2	2
event	<input type="checkbox"/>	2	4
coast	<input type="checkbox"/>	2	2
organis	<input type="checkbox"/>	2	17
restaur	<input type="checkbox"/>	2	2
festiv	<input type="checkbox"/>	2	2
locat	<input type="checkbox"/>	1	26
person	<input type="checkbox"/>	1	6
bar	<input checked="" type="checkbox"/>	1	1
suit	<input checked="" type="checkbox"/>	1	1
Hotel-Guest	<input checked="" type="checkbox"/>	0	0
archaeological museum	<input checked="" type="checkbox"/>	0	0
SwimmingPool	<input checked="" type="checkbox"/>	0	0
cable car	<input checked="" type="checkbox"/>	0	0
art museum	<input checked="" type="checkbox"/>	0	0
Train-Station	<input checked="" type="checkbox"/>	0	0
christmas dai	<input checked="" type="checkbox"/>	0	0
southern coast	<input checked="" type="checkbox"/>	0	0
music festiv	<input checked="" type="checkbox"/>	0	0
tourist offic	<input checked="" type="checkbox"/>	0	0
Sauna	<input checked="" type="checkbox"/>	0	0
Zoo	<input checked="" type="checkbox"/>	0	0
Double-Room	<input checked="" type="checkbox"/>	0	0
Commercial-Organization	<input type="checkbox"/>	0	7
national galleri	<input checked="" type="checkbox"/>	0	0
Hair Drier	<input checked="" type="checkbox"/>	0	0
post offic	<input checked="" type="checkbox"/>	0	0
Non-Private-Equipment	<input type="checkbox"/>	0	4

Prepare Pruning List Delete Pruned Concepts Stop Pruner

Opens a new pruner window.

# Agenda

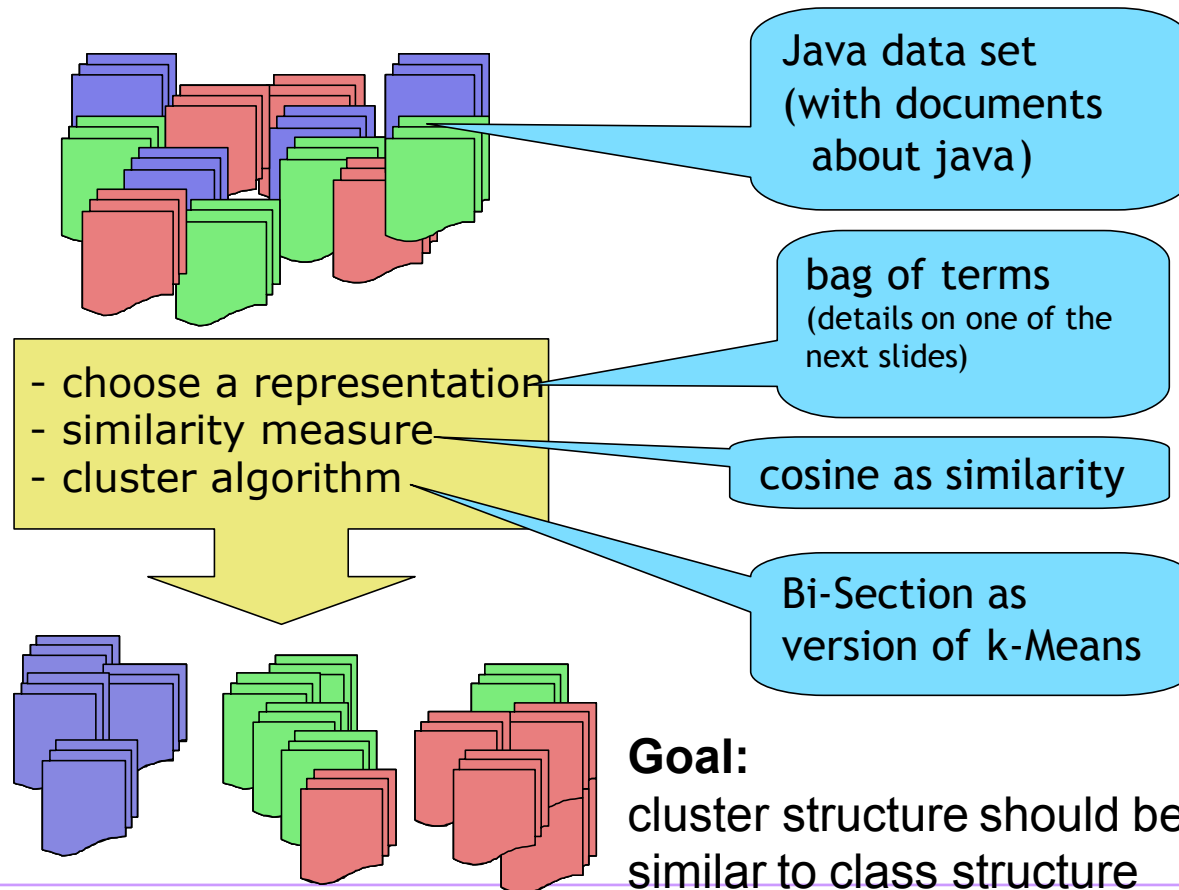
- 
- A vertical strip on the left side of the slide shows a close-up of a calendar page. The numbers 7, 8, 9, and 10 are visible, along with a metal ring binding the pages. The lighting is warm and golden.
- **Introduction & Motivation**
  - **Ontology Learning Framework & Techniques**
  - **Text-To-Onto 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(\*)



# Background Knowledge

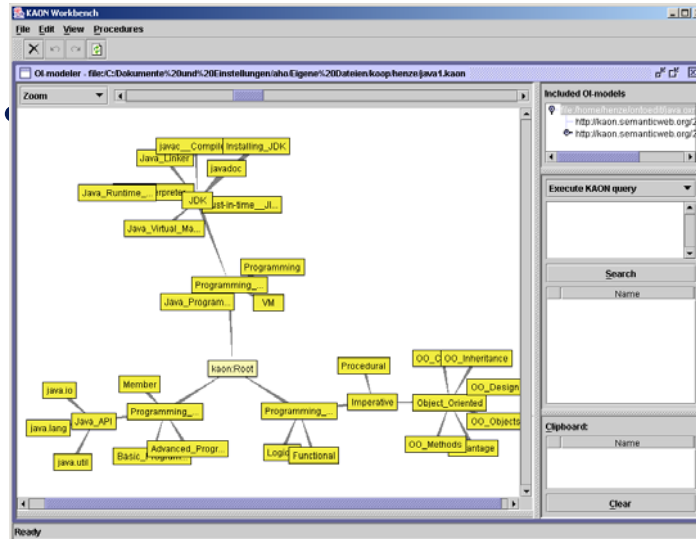
## Bag of words model:

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

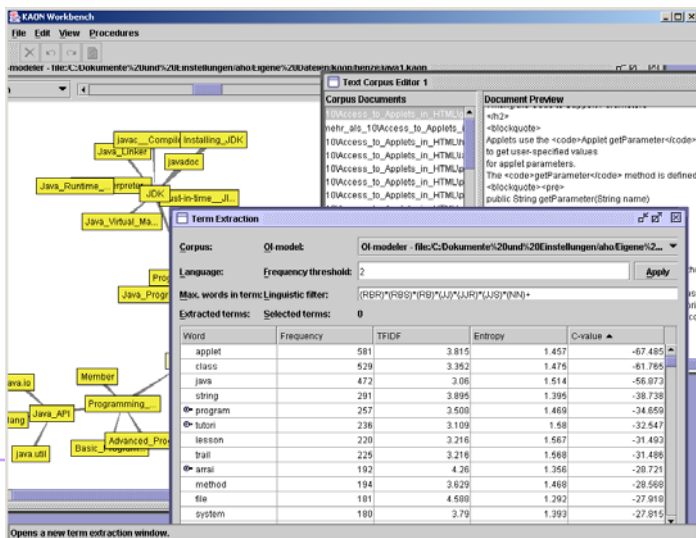
## Bag of concept model (term and concept vector):

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



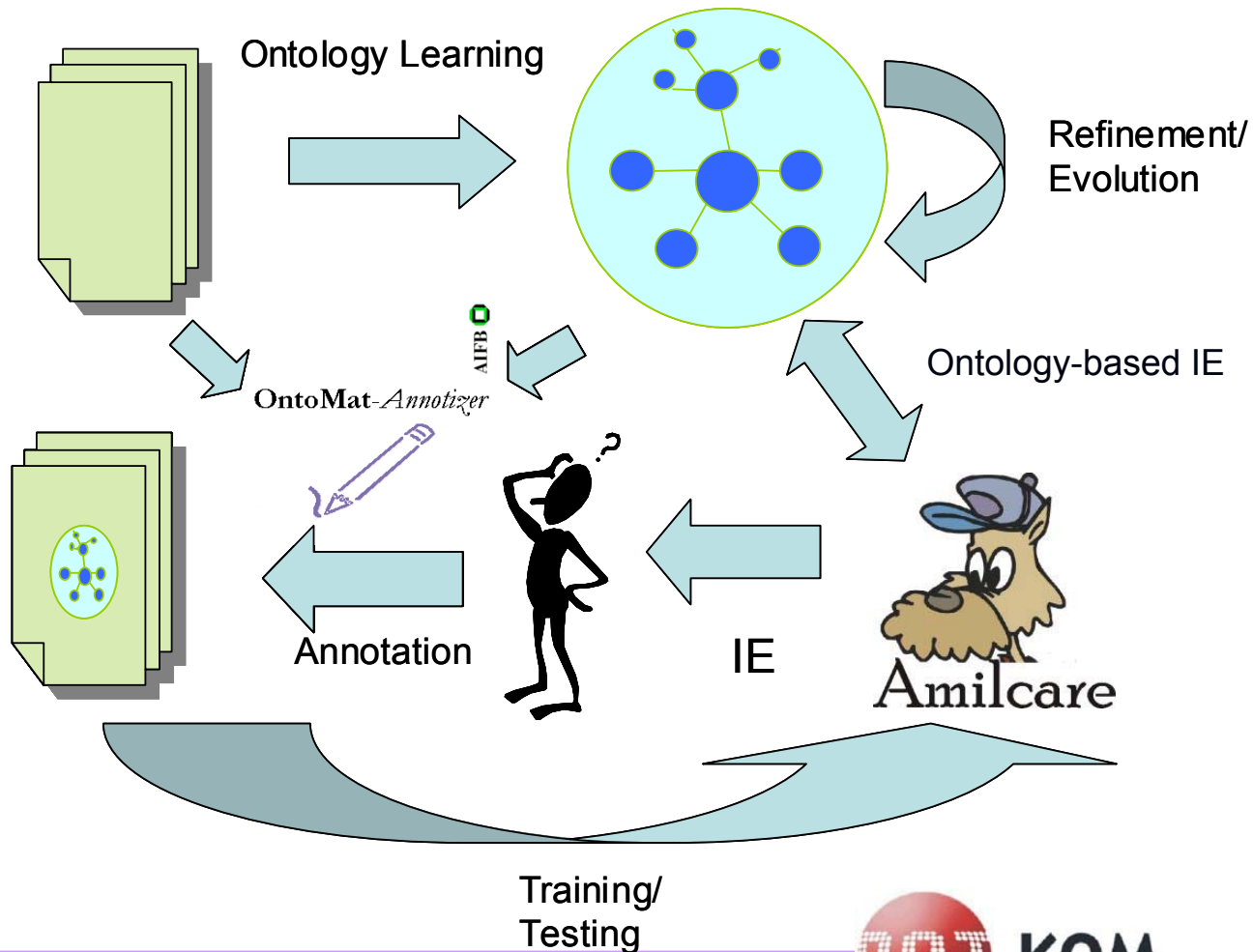
- Without Ontology = Baseline
  - Purity = 62%, Inverse Purity = 61%
- With handmade Ontology
  - Purity = 60%, Inverse Purity = 57%



- Ontology improved by Ontology Learning
  - Purity = 67%, Inverse Purity = 64%



# Information Extraction (\*)



# Ontology-based Information Extraction

OntoMat-Annotizer 0.416

File Edit View Tools Window Help

Ontology Browser

- Qualitatives\_Zeitkonzept
- Unterkunft
  - Clubanlage
  - Gasthof
  - Hotel
  - Motel
  - Pension

Attributes Values

Anzahl_Betten	
Hausiere_erlaubt	
Klassifizierung	
Name	
behindertenfre...	

- bietet\_Essen (Essen)
- bietet\_Freizeiteinrichtung
- hat\_Adresse (Adresse)
- hat\_Ausstattung (Nichtpr)

Landhotel & Ferienpark Leonorenwald in Hohen Schönberg nahe dem Ostseebad Boltenhagen in Mec...

URL: http://www.all-in-all.de/1665.htm

Start [Interaktive Karte] [Boltenhagen]

Kontakt-aufnahme Essen & Trinken Wassersport Sport/Fitness Kultur/Veranst. Sehenswertes Angebote & Touren

**Landhotel & Ferienpark Leonorenwald**

Kalkhorster Str. 5  
23948 Hohen Schönberg  
Tel. 038827/ 88 70  
Fax: 038827/ 8 87 77

Lage: In einer richtigen Lage. Ein herrlicher Ostseebadestrand mit seiner Steilküste in nur 3 km Entfernung lädt zum Baden und Sonnen ein. Genießen Sie die klare Luft bei einer Wanderung entlang der Steilküste oder in dem nur 1 km entfernten Leonorenwald.

HTML Source DOM Annotation

Status: http://www.all-in-all.de/0200.htm

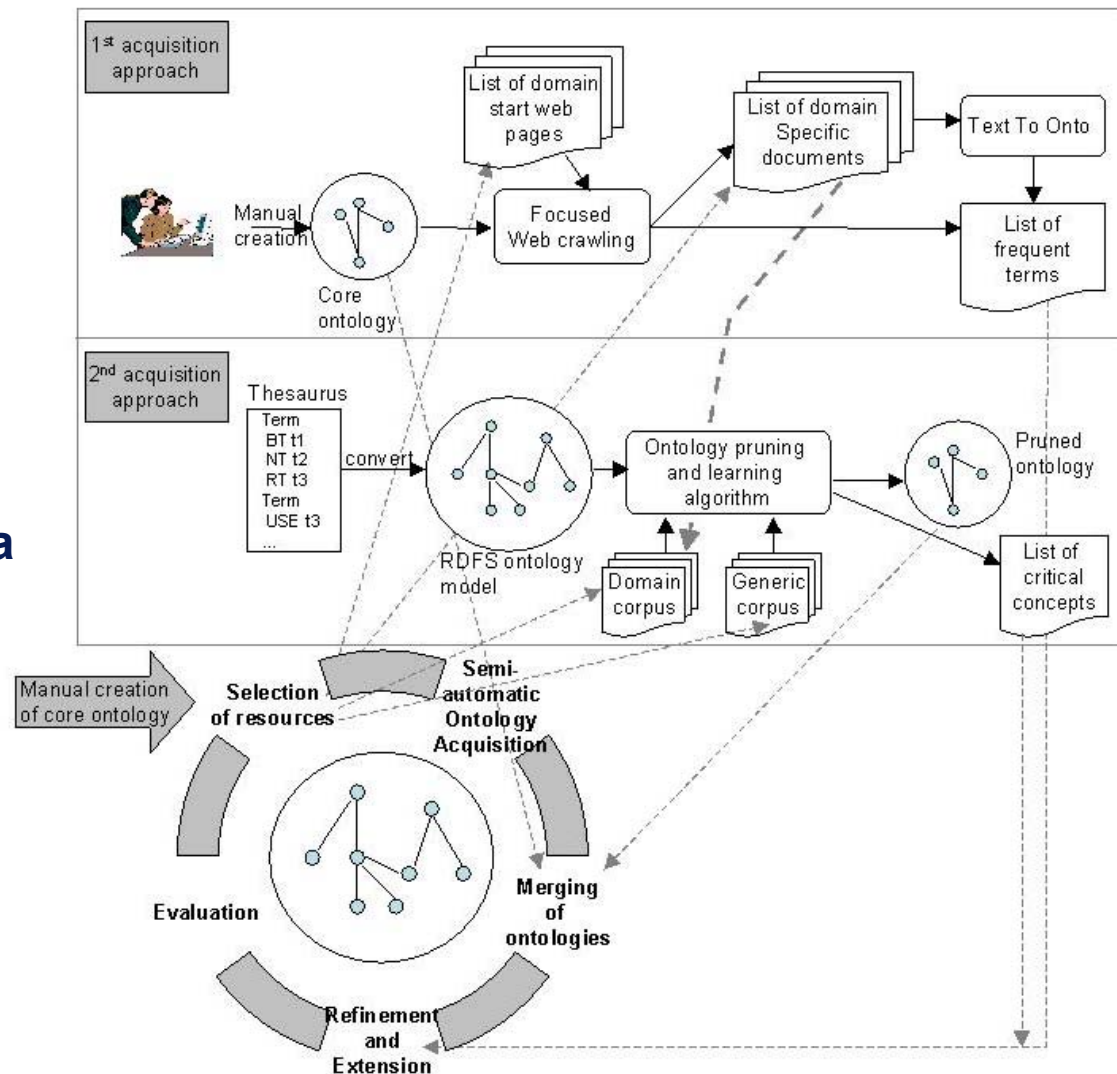
IEResult

The system has found the value "Landhotel & Ferienpark Leonorenwald" for the Concept Hotel. Do you agree?

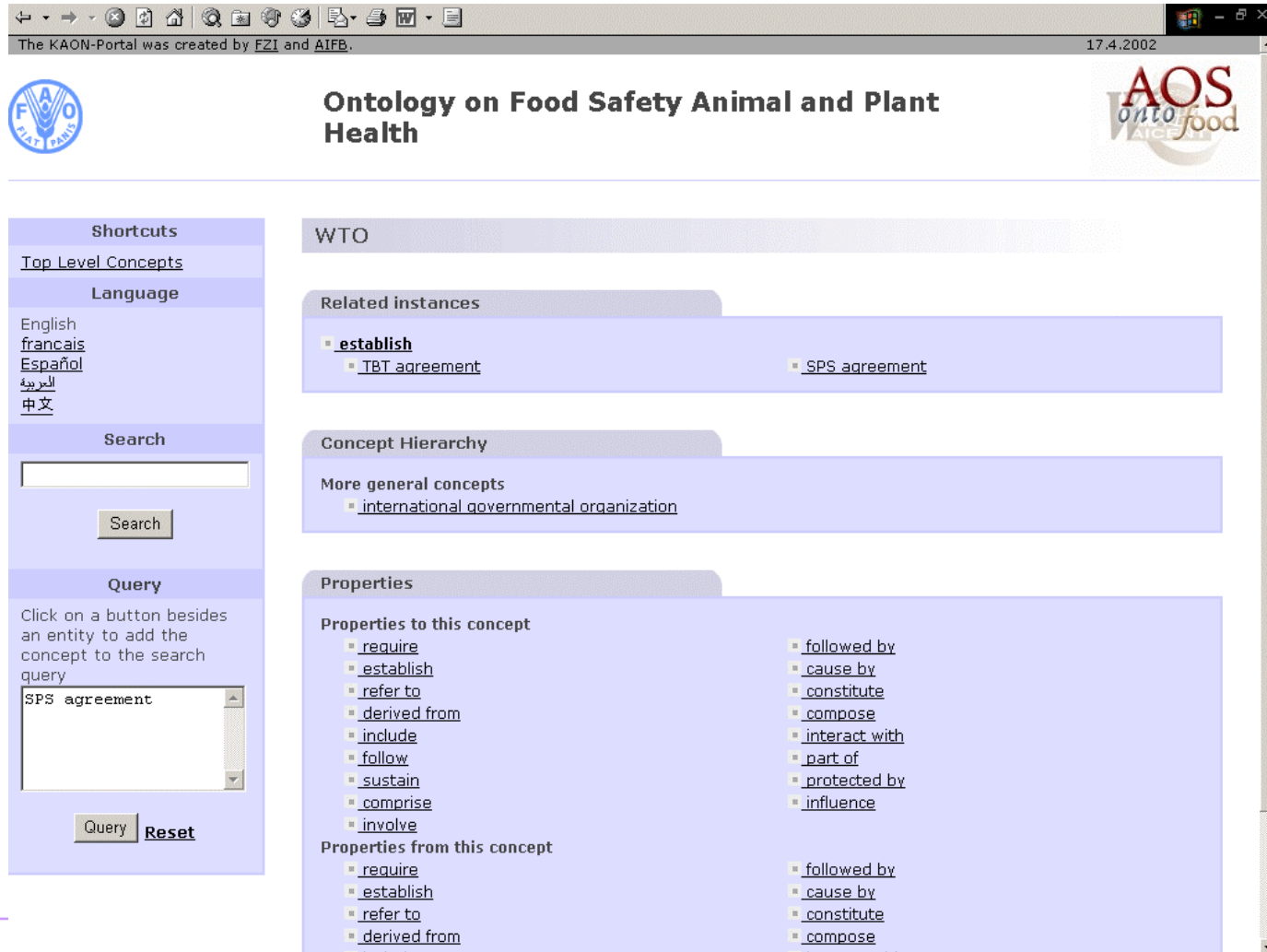
yes, I agree no cancel

# 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



The screenshot shows a web browser window displaying the 'Ontology on Food Safety Animal and Plant Health' application. The browser's address bar shows the URL 'http://www.fao.org/ontologies/food-safety/ontology/'. The page header includes the FAO logo and the text 'The KAON-Portal was created by FZI and AIFB.' and '17.4.2002'. The main content area is titled 'WTO' and features several sections: 'Related instances' with links to 'establish', 'TBT agreement', and 'SPS agreement'; 'Concept Hierarchy' with a link to 'international governmental organization'; and 'Properties' divided into 'Properties to this concept' and 'Properties from this concept', both listing various semantic relationships like 'require', 'establish', 'refer to', 'derived from', 'include', 'follow', 'sustain', 'comprise', 'involve', 'followed by', 'cause by', 'constitute', 'compose', 'interact with', 'part of', 'protected by', and 'influence'. A sidebar on the left contains 'Shortcuts', 'Language' options (English, français, Español, العربية, 中文), a search box, and a 'Query' section with a dropdown menu showing 'SPS agreement' and 'Query Reset' buttons.

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

# Agenda

- 
- A vertical strip on the left side of the slide shows a close-up of a calendar page. The numbers 7, 8, 9, and 10 are visible, along with a metal ring binding the page. The lighting is warm and golden.
- **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	Corpus2
	<b>„Human Language Technology“</b>	<b>„Eol-Knowledge-Technologies“</b>
<b>[(NNS)(NN)]+</b>	<b>88% (*)</b>	<b>80%</b>
	<b>(1230,233,27) (**)</b>	<b>(1079,202,40)</b>
	<b>86%</b>	<b>76%</b>
<b>(RB)*(JJ)*[(NN)(NNS)]+</b>		
	<b>(1362,362,47)</b>	<b>(1243,361,85)</b>
<b>[(RB)(JJ)(NN)]*(IN)?</b>	<b>64%</b>	<b>64%</b>
<b>[(RB)(JJ)(NN)]*</b>		
<b>[(NN)(NNS)]</b>	<b>(1511,511,181)</b>	<b>(1362,478,171)</b>

(\*) 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  $t$