Convergence of classical search and topic maps – evidences from a practical case in the chemical industry

Dr. Stefan Smolnik
Information Systems 2, European Business School (EBS)

TMRA 2007

11. October 2007

European ☰ Business School
International University · Schloss Reichartshausen

Agenda

1. Motivation and challenges

2. Classical search – text mining

3. Semantic technologies – Topic Maps

4. Evaluations and lessons learned

Agenda

1. Motivation and challenges
2. Classical search – text mining
3. Semantic technologies – Topic Maps
4. Evaluations and lessons learned

European ☰ Business School
International University · Schloss Reichartshausen

Motivation and challenges

> Focus: Information Center within the research group of a chemical firm

> Overall responsibility: provision of information services and information systems

> Main tasks: making inquiries as well as building and running information systems

> Typical application scenarios: patent analyses, competitive analyses, product developments, opportunity discoveries, news services

> Problems: very heterogeneous information and distributed information sources

Organisation of Metadata	Technology		Extend of Content	
			Full Text	Text + Metadata
			Degree of Structure weak	semi
No Metadata	Full Text Search	Statistical Approaches	M: ●	
			R: ○ Q: ◉ - ● P: ● I: ○	R: ○ Q: ○ - ● P: ● I: ○
		Linguistic Approaches	M: ◉	
			R: ◉ Q: ◉ P: ◉ - ● I: ○	
Flatly or Hierarchically Organisation of Metadata	Structured Search		M: ●	
			R: ● Q: ◉ P: ● I: ○	R: ◉ Q: ● P: ● I: ○
	Taxonomies & Thesaurus		M: ●	
			R: ● Q: ◉ - ● P: ● I: ○	R: ◉ Q: ◉ - ● P: ● I: ○
Network-like Organisation of Metadata	Topic Maps		M: ○	
			R: ● Q: ● P: ● I: ●	
	Ontologies		M: ◉	
			R: ● Q: ● P: ● I: ●	

M - Maturity   ● high / ◉ average / ○ low
R - Requirements for and Efforts of Implementation   ● high / ◉ averaqe / ○ low
Q - Quality of Search Results   ● high / ◉ average / ○ low
P - Search/Retrieval Performance   ● high / ◉ average / ○ low
I - Heterogeneous Information Repositories' Integration Capability   ● high / ◉ average / ○ low

Agenda

1. Motivation and challenges
2. Classical search – text mining
3. Semantic technologies – Topic Maps
4. Evaluations and lessons learned

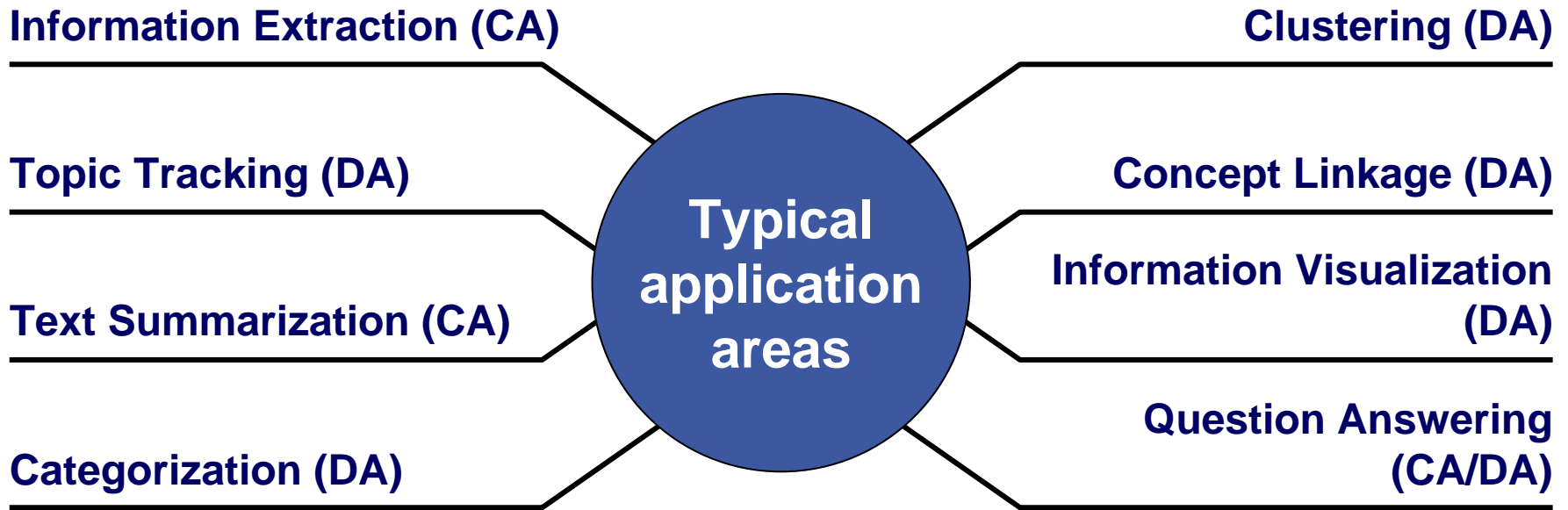# Classical search – text mining: Overview and foundations

Information retrieval perspective: methods for text summarization and information extraction (which improve the information retrieval)

Data mining perspective: exploration of (to be interpreted) data, that could be gained from texts/documents

Methodical perspective: methods for the (semi-)automatic analysis of huge document sets

Knowledge-oriented perspective: (automatic) text-based data analysis for the exploration of „unknown" information

Classical search – text mining: Typical application areas

Information Extraction (CA)

Topic Tracking (DA)

Text Summarization (CA)

Categorization (DA)

Typical application areas

Clustering (DA)

Concept Linkage (DA)

Information Visualization (DA)

Question Answering (CA/DA)

DA – document-based applications
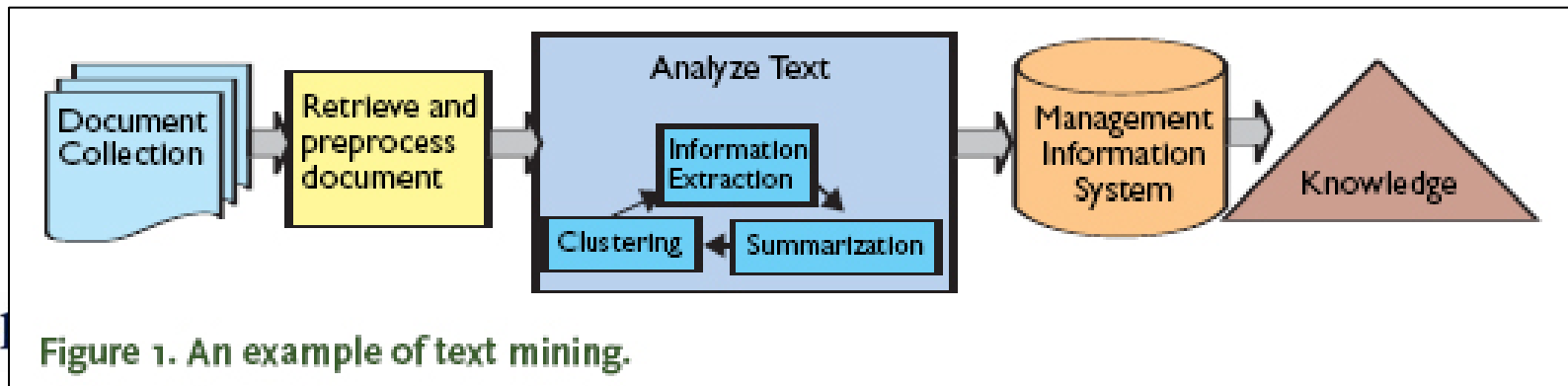CA – content-based applications

Figure 1. An example of text mining.

9

Classical search – text mining: Possible application areas for the chemical firm

Application areas	Information Extraction	Topic Tracking	Text Summarization	Categorization	Clustering	Concept Linkage	Information Visualization	Question Answering
Patent analysis	X				X	X	X	
Competitive analysis		X	X					
Product development	X				X	X		
Opportunity Discovery	X				X	X	X	
News service	X	X	X	X	X	X		

Wirtschaftsinformatik 2 / Information Systems 2

Benefits/pros

- **Benefits according to the application areas**
- **Determination of word meanings**
- **Nearly independent of the document/text length**
- **Interpretation of the word structures**
- **Understanding of flexion forms and modifications of words**
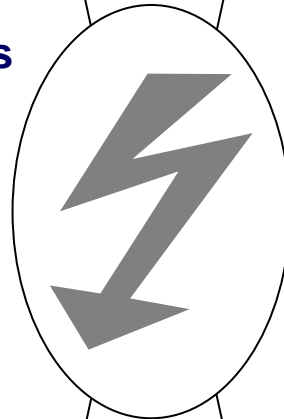
Limitations/cons

- **Source language: high degree of rule-based word structures needed**
- **Language dependent**
- **Without using a thesaurus only moderate results**
- **Problems with ambiguity**
- **Moderate relevance of (automatically) identified relations**
- **Only limited identification of relations across distributed information repositories**
- **Complete automation still unrealistic**

Agenda

1. Motivation and challenges

2. Classical search – text mining

3. Semantic technologies – Topic Maps

4. Evaluations and lessons learned

Delimitation, analysis, and modeling of the considered knowledge domain

In the case of non-stable or immature knowledge domains: modifications of the topic maps required during run time

Requirements and efforts

Relevant information needs to be provided, indexed, and pre-processed as well as linked with the topic maps

New information needs to be preprocessed during run time (e.g. indexing or annotation are required)
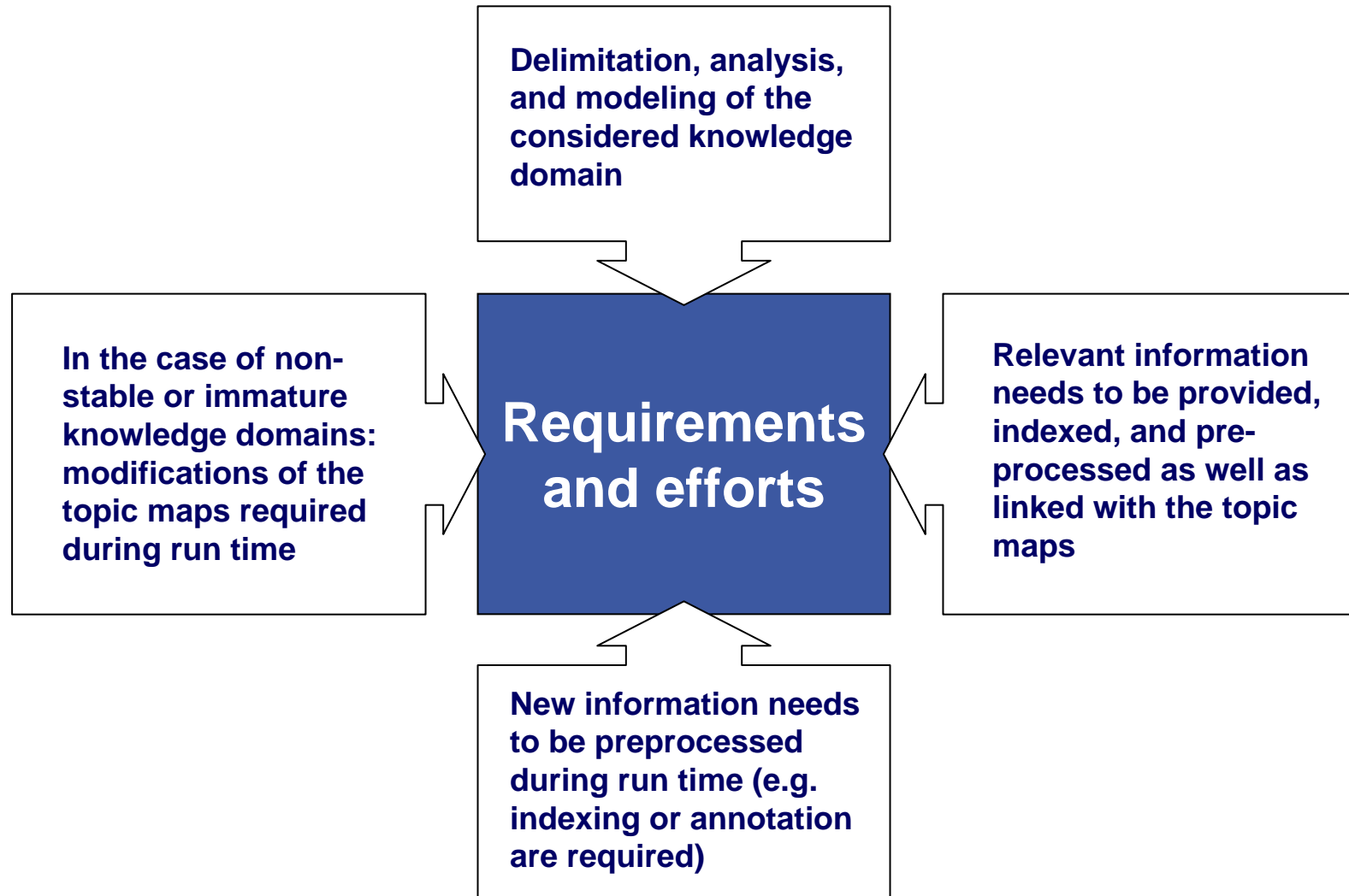
Semantic technologies – Topic Maps: Benefits/pros

Improved accessibility of heterogeneous information repositories

Improved retrieval of information

Discovery of hidden relations between information

Contextualization of themes/information

Benefits/ pros

Support of human thinking

Explorative („fuzzy") search and search in contexts

Description/communication of the shared understanding within organizations

Easier information maintenance and more consistent information repositories

Semantic technologies – Topic Maps: Limitations/cons

**Limitations/
cons**

Efforts for the modeling, generation/creation, maintenance, and operation of the topic maps

Complete and appropriate modeling of the knowledge domain as a topic map is critical for the users' benefits (**"Everything that should be found has to be modeled in advance."**)

Only a few standardized or established knowledge models available (particularly for the chemical industry); therefore, such models need to be **developed in-house**

Inconsistent perspectives and understandings of the Topic Maps' concepts and terms make a positive evaluation for the organizational deployment difficult

Agenda

1. Motivation and challenges

2. Classical search – text mining

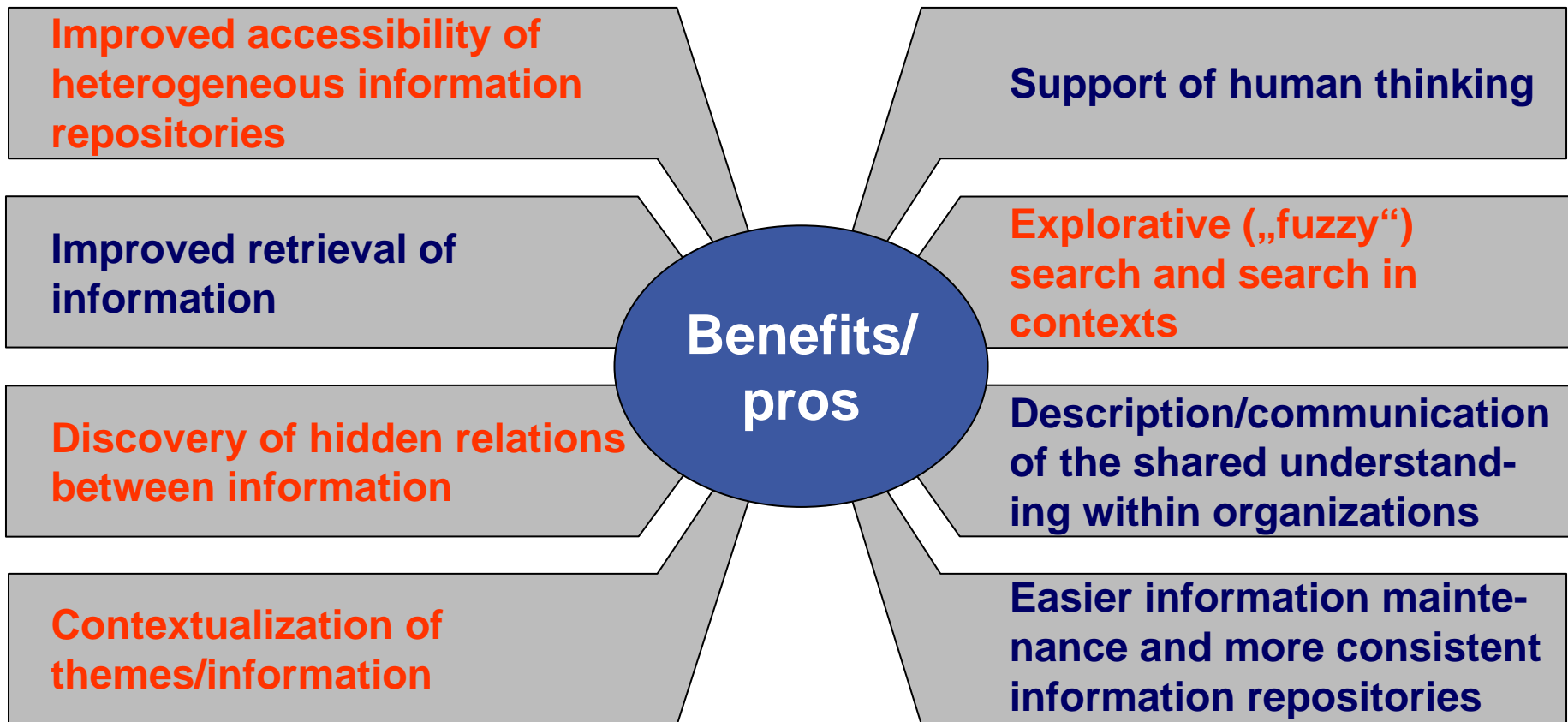3. Semantic technologies – Topic Maps

4. Evaluations and lessons learned
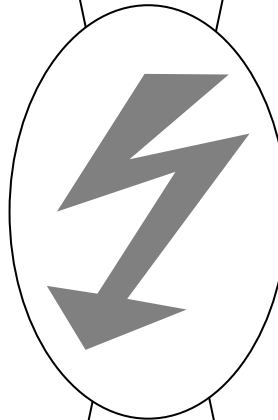
Evaluations – summary

Text mining

- **Automatic document processing**
- **Based on standard models**
- **"Concept Linkage"**
- **Dependent on concepts and relations defined in thesauruses and rule frameworks (linguistic methods)**
- **Efforts for both information systems and users**
- **Efforts for the document preprocessing**
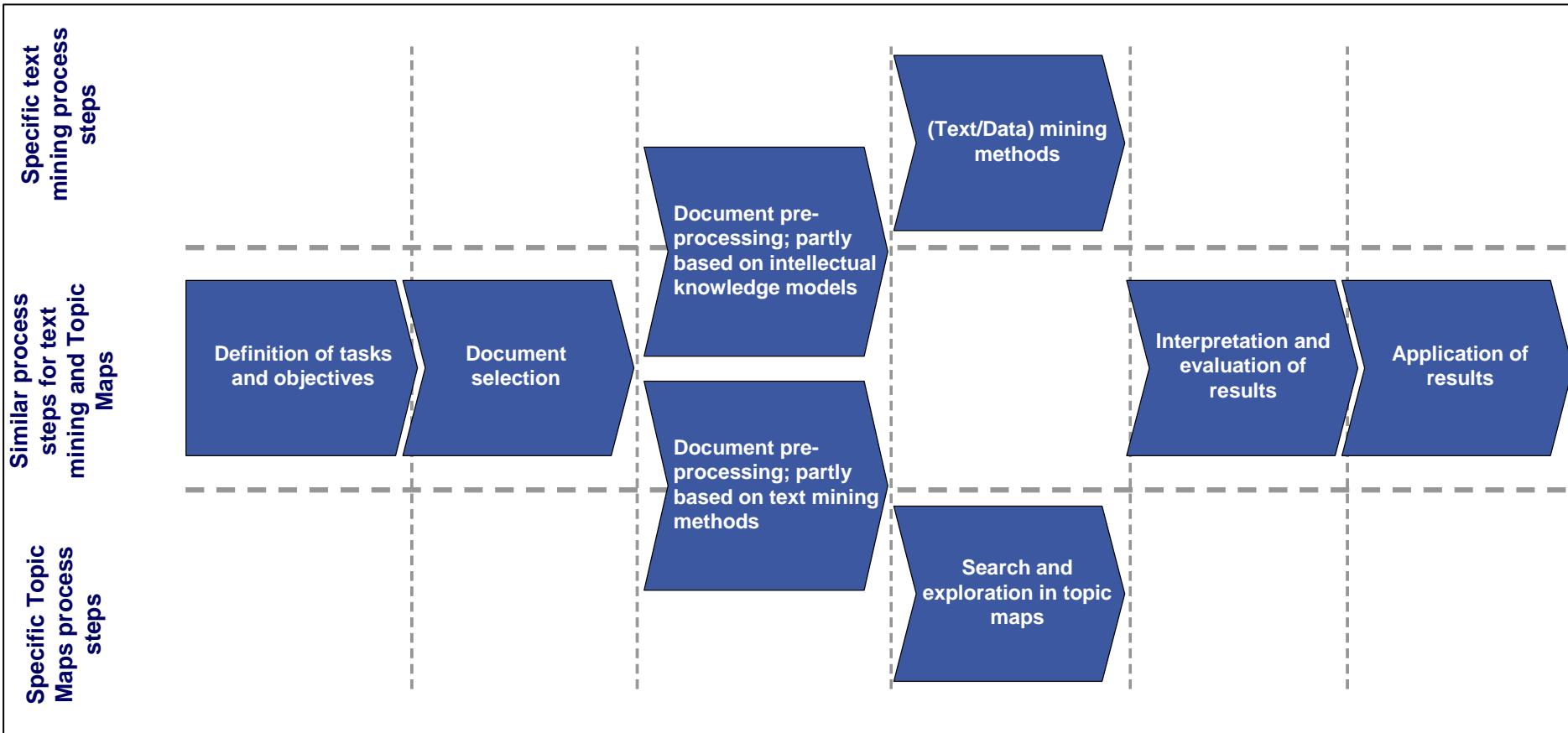- **No real integration of heterogeneous information repositories**

Topic Maps

- **Manual/semi-automatic knowledge modeling**
- **Flexible and according to specific requirements conducted knowledge modeling**
- **Topic associations**
- **Efforts for the knowledge modeling**
- **Efforts for the document preprocessing**
- **Semantic integration of heterogeneous information repositories**

Lessons learned (1/2): Information access/structuring process

Specific text mining process steps

Similar process steps for text mining and Topic Maps

Specific Topic Maps process steps

Definition of tasks and objectives

Document selection

Document pre-processing; partly based on intellectual knowledge models

Document pre-processing; partly based on text mining methods

(Text/Data) mining methods

Search and exploration in topic maps

Interpretation and evaluation of results

Application of results

European ≡ Business School
International University · Schloss Reichartshausen

Wirtschaftsinformatik 2 / Information Systems 2

Lessons learned (2/2):
Summary

„In a nutshell": automation vs. modeling
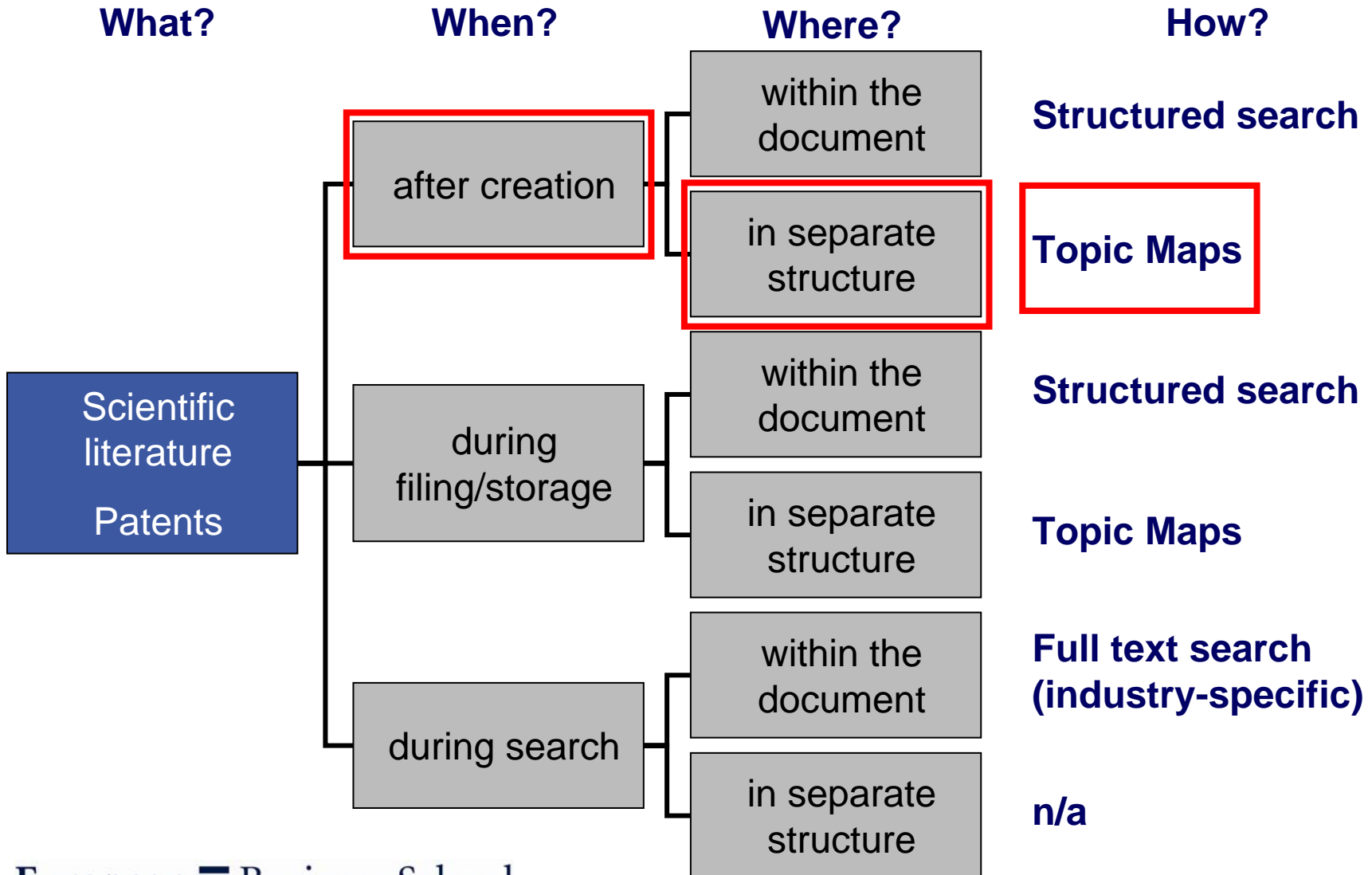but: text mining and Topic Maps applications converge

Text mining uses semantic technologies and vice versa (particularly, with respect to the document preprocessing)

Intellectual modeling takes place during the text mining process as well

Practical experiences in a typical environment is necessary for gaining further insights and results

Recommendations

What?	When?	Where?	How?
	after creation	within the document	**Structured search**
		in separate structure	**Topic Maps**
Scientific literature **Patents**	during filing/storage	within the document	**Structured search**
		in separate structure	**Topic Maps**
	during search	within the document	**Full text search (industry-specific)**
		in separate structure	**n/a**

European ≡ Business School
International University · Schloss Reichartshausen

Questions & discussion

Dr. Stefan Smolnik

Chair of Information Systems 2
European Business School (EBS)
International University Schloss Reichartshausen
Rheingaustr. 1
65375 Oestrich-Winkel
Phone: +49-6723-991-246
Fax: +49-6723-991-259
E-Mail: stefan.smolnik@ebs.edu
WWW: http://www.ebs.edu/is2

European ≡ Business School
International University · Schloss Reichartshausen