



seit 1558

Friedrich-Schiller-Universität Jena

Institut für Informatik

# Dokumentindexierung als Basis zur semantischen Dokument-Annotation

Heiko Peter

Harald Sack

Clemens Beckstein

## XML-Tage 2006

Berlin

25.-27. September 2006

# Motivation (I)

Ein alltägliches Problem ist das Auffinden von Information:



*Lösung:* Es ist eine **semantische Annotation** der informationstragenden Dokumente erforderlich.

# Motivation (II)

Dokument



semantische Annotation

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE rdf:RDF [
  <!ENTITY rdf 'http://www.w3.org/1999/02/22-rdf-syntax-ns#'>
  <!ENTITY rdfs 'http://www.w3.org/2000/01/rdf-schema#'>
  <!ENTITY owl 'http://www.w3.org/2002/07/owl#'>
  <!ENTITY swrc 'http://swrc.ontoware.org/ontology#'>
  <!ENTITY xsd 'http://www.w3.org/2001/XMLSchema#'>
]>
<rdf:RDF
  xml:base="http://www.aifb.uni-karlsruhe.de/Publikationen/viewPublikatio
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:swrc="http://swrc.ontoware.org/ontology#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#">
  <owl:Ontology rdf:about="">
    <rdfs:comment>Instance data for publication "Proceedings of the First
    Workshop on Ontology Learning OL'2000, Berlin, Germany, August 25,
    2000."</rdfs:comment>
```

*Lösung 1:* manuell (z.B. durch den Dokumentautor)

Problem: zu zeitaufwendig, zu teuer

*Lösung 2:* Einsatz von Data-Mining-Techniken

Problem: domänenabhängig, unzuverlässiges Ergebnis



Beide Ansätze sind allein unbefriedigend

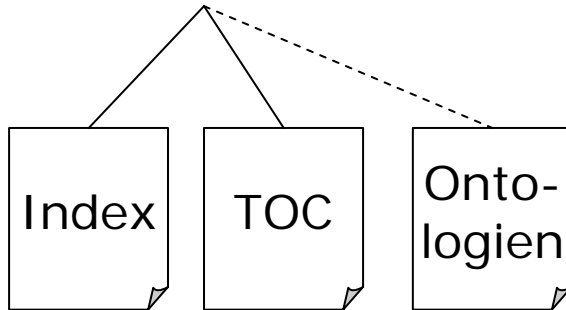
# Motivation (III)

Dokument



semantische Annotation

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE rdf:RDF [
  <!ENTITY rdf 'http://www.w3.org/1999/02/22-rdf-syntax-ns#'>
  <!ENTITY rdfs 'http://www.w3.org/2000/01/rdf-schema#'>
  <!ENTITY owl 'http://www.w3.org/2002/07/owl#'>
  <!ENTITY swrc 'http://swrc.ontoware.org/ontology#'>
  <!ENTITY xsd 'http://www.w3.org/2001/XMLSchema#'>
]>
<rdf:RDF
  xml:base="http://www.aifb.uni-karlsruhe.de/Publikationen/viewPublikatio
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:swrc="http://swrc.ontoware.org/ontology#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  <owl:Ontology rdf:about="">
    <rdfs:comment>Instance data for publication "Proceedings of the First
    Workshop on Ontology Learning OL'2000, Berlin, Germany, August 25,
    2000."</rdfs:comment>
```



Index  
TOC (Dokumentstruktur)  
externes konzeptuelles Wissen (Ontologie)

Basis zur semantischen  
Dokumentannotation

# Motivation (IV)



Problem: Informationen dieser Dokument-Annotationen nicht explizit

Beispiel:

Nagetier, 1

Biber, 10, 11 → Biber gehören zur Ordnung der Nagetiere.

Gebiss → Das Gebiss ist ein morphologisches Kennzeichen der Nagetiere.

Schneidezahn, 4 → Der Schneidezahn ist ein wichtiges Merkmal im Gebiss eines Nagetiers.

Zahnwechsel, 5

Hamster, 2 - 4 → Hamster gehören zur Ordnung der Nagetiere.

*Siehe auch* Feldmaus → Feldmäuse sind verwandt mit den Nagetieren.



Informationen dieser Annotationen müssen in eine computerverarbeitbare Form gebracht werden.

# Motivation (V)



- Die Verwendung dieser vorhandenen Hilfsmittel erfordert:

- Formalisierung des Indexes
- Formalisierung der Dokumentstruktur
- Repräsentation der beiden Strukturen



Index-Ontologie



Index-Graph

- Dies bildet die Basis für:

- semantische Annotation von Dokumenten
- neue Anwendungen

# Agenda

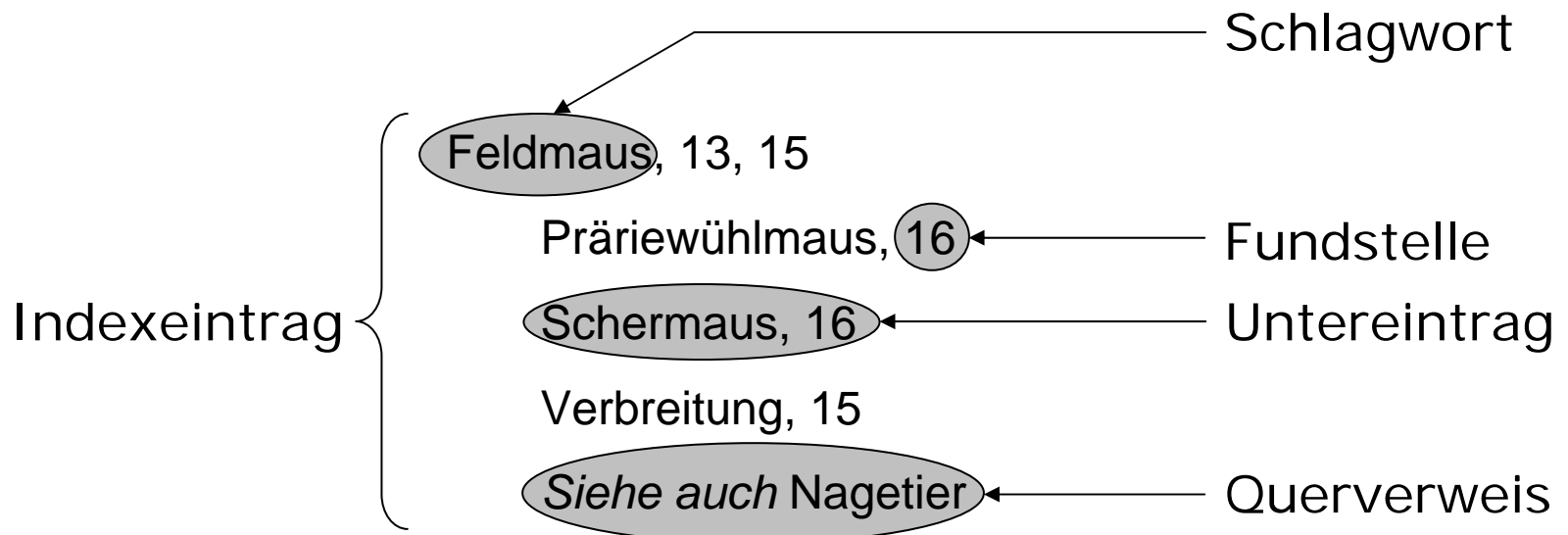


- Der Index
- Die Dokumentstruktur
- Die Index-Ontologie
- Der Index-Graph
- Anwendungen
- Zusammenfassung & Ausblick

# Syntaktischer Aufbau eines Indexes



- Ein Index ist eine systematische Anordnung von Einträgen.
- *Beispiel:*



# Begriffe und Namen (I)



Grundlage zur Beschreibung der Semantik eines Indexes sind:

**Gegenstand:** ist alles, worüber eine Aussage gemacht werden kann.

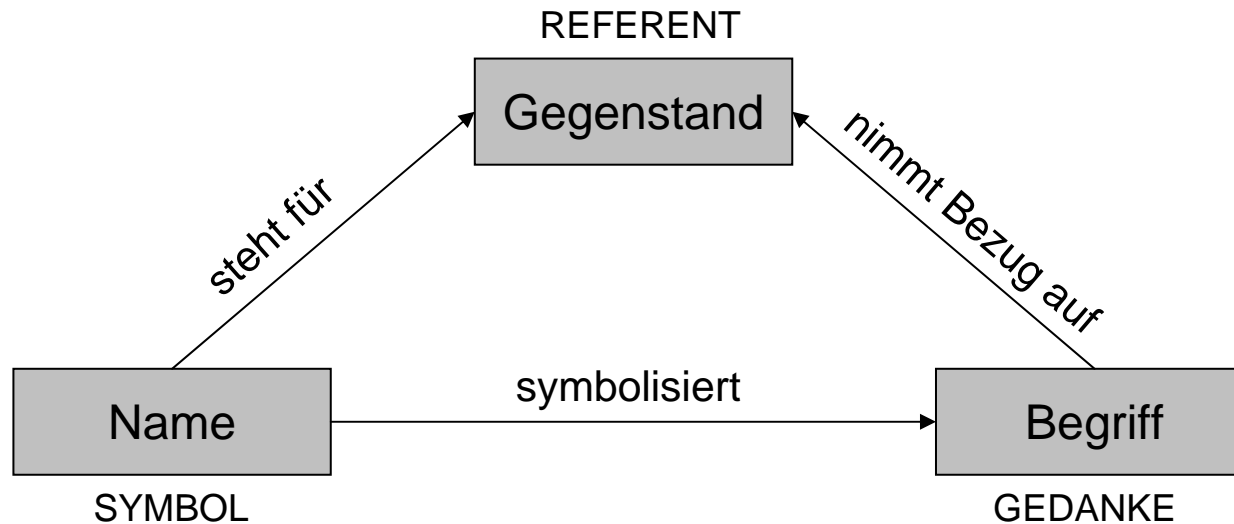
**Begriff:** ist die Gesamtheit aller wesentlichen Aussagen, die über einen Gegenstand gemacht werden.

**Name:** ist ein Symbol, das für einen Begriff steht.

# Begriffe und Namen (II)



- Gegenstand, Begriff und Name entsprechen den drei Elementen des Semiotischen Dreiecks:



# Begriffsbeziehungen



- Begriffe können in semantischer Beziehung zueinander stehen:
  - Hypernymie / Hyponymie
  - Meronymie / Holonymie
  
- Es bestehen zudem Beziehungen zwischen Namen und Begriffen:
  - Synonymie
  - Homonymie

# Die Semantik eines Indexes: Schlagwörter



- Schlagwörter sind Namen für im Dokument enthaltene Begriffe.
  
- Eigenschaften eines Schlagwortes:
  - Einzelwort oder Wortgruppe
  - vorhersagbar
  
- Man unterscheidet zwischen:
  - **Titel-Stichwörtern:** Sie entstammen dem Titel des Dokuments.
  - **Text-Stichwörtern:** Sie entstammen dem Dokumenttext.

# Die Semantik eines Indexes: Untereinträge & Querverweise



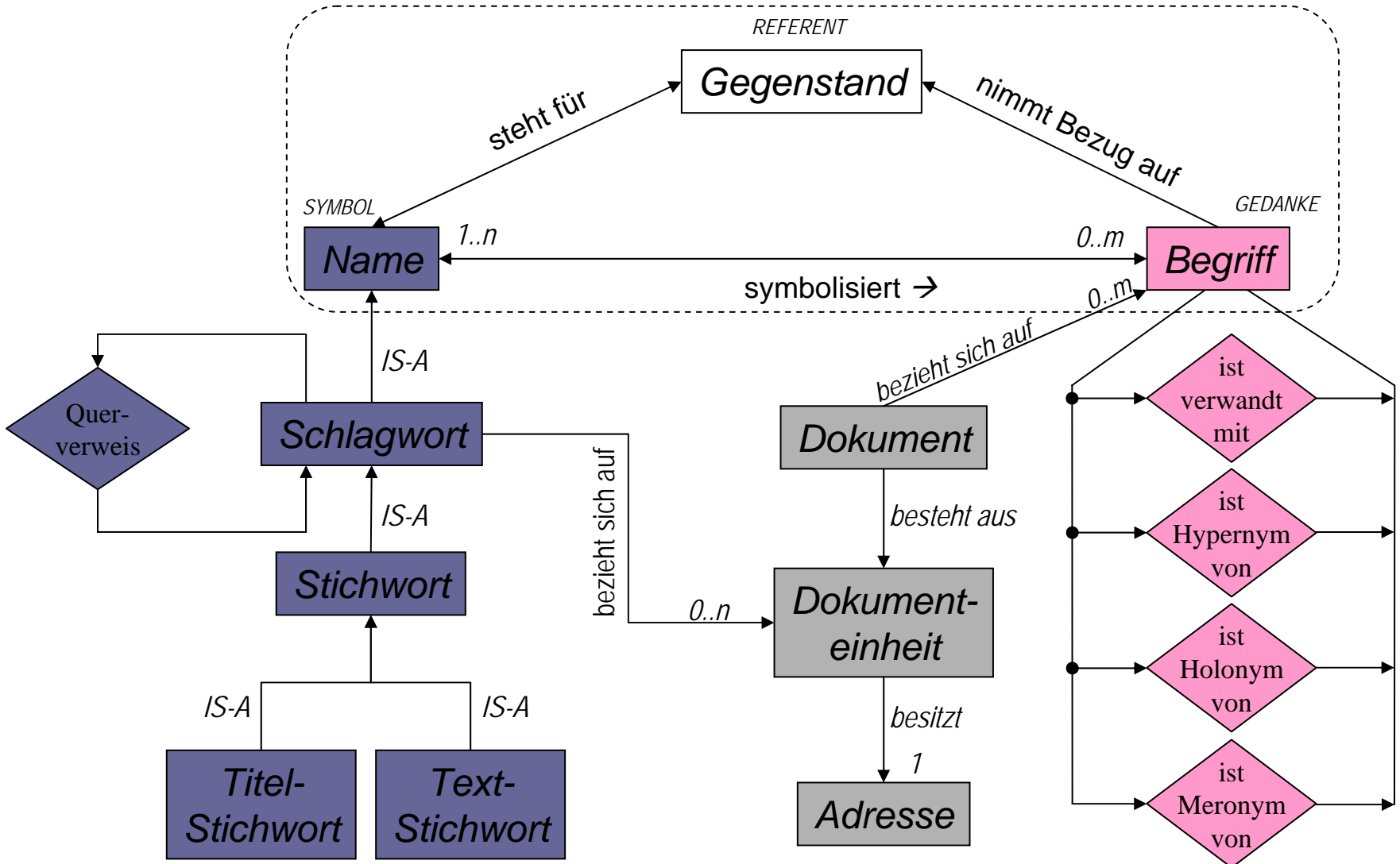
- Untereinträge und Querverweise etablieren semantische Beziehungen zwischen Begriffen.
- Ein *Unterschlagwort* zeigt eine semantische Beziehung zum übergeordneten Schlagwort an.
- Bei *Querverweisen* unterscheidet man zwischen:
  - **Siehe Verweisen**: für synonyme Beziehungen
  - **Siehe auch Verweise**: für jede andere Art von semantischer Beziehung

# Die Dokumentstruktur



- **Dokument** = total geordneter String von kleinsten, adressierbaren **Dokumenteinheiten**
- Mit Hilfe von Tags können die kleinsten Dokumenteinheiten zu größeren gruppiert werden.
- Die **Dokumentstruktur** erfasst diese part-of Beziehung zwischen den Einheiten.
- Das TOC spiegelt einen Teil dieser Struktur wider.
- **Fundstelle** = Adresse einer Dokumenteinheit

# Die Index-Ontologie (I)



# Index-Ontologie (II)



- allgemeines Wissen über Indexelemente + deren Beziehungen zueinander
  - allgemeines Wissen über Dokumentstruktur
- } Index-Ontologie
- 
- Umsetzung der Index-Ontologie in OWL-DL
    - XML-basierte Wissensrepräsentationssprache
- ↩ allgemeines Indexwissen in computerverarbeitbarer Form beschrieben
- Austauschbarkeit, Übertragbarkeit durch W3C-Standardisierung gegeben
  - frei wählbare Kardinalitäten in OWL-DL möglich

# Der Index-Graph (I)



- Es ist weiterhin eine Repräsentation des impliziten Wissens eines konkreten Dokuments notwendig.
- Repräsentation muss zudem kompatibel zur Index-Ontologie sein.
- Ein konkreter Index kann zusammen mit der Dokumentstruktur als 2-schichtiger Graph repräsentiert werden: **Index-Graph**.
- Der Index-Graph besteht aus:
  - dem **Begriffs-Graphen**: repräsentiert die konkreten Indexelemente und deren Beziehungen.
  - dem **Dokument-Graphen**: repräsentiert die inhärente hierarchische Struktur des Dokuments.

# Der Index-Graph (II)



## Beispiel: Begriffs-Graph

Nagetier, 1

Biber, 10, 11

Gebiss

Schneidezahn, 4

Zahnwechsel, 5

Hamster, 2 - 4

*Siehe auch* Feldmaus

Feldmaus, 13, 15

Präriewühlmaus, 16

Scherm Maus, 16

Verbreitung, 15

*Siehe auch* Nagetier

# Der Index-Graph (II)



## Beispiel: Begriffs-Graph

Nagetier, 1

Biber, 10, 11

Gebiss

Schneidezahn, 4

Zahnwechsel, 5

Hamster, 2 - 4

*Siehe auch* Feldmaus

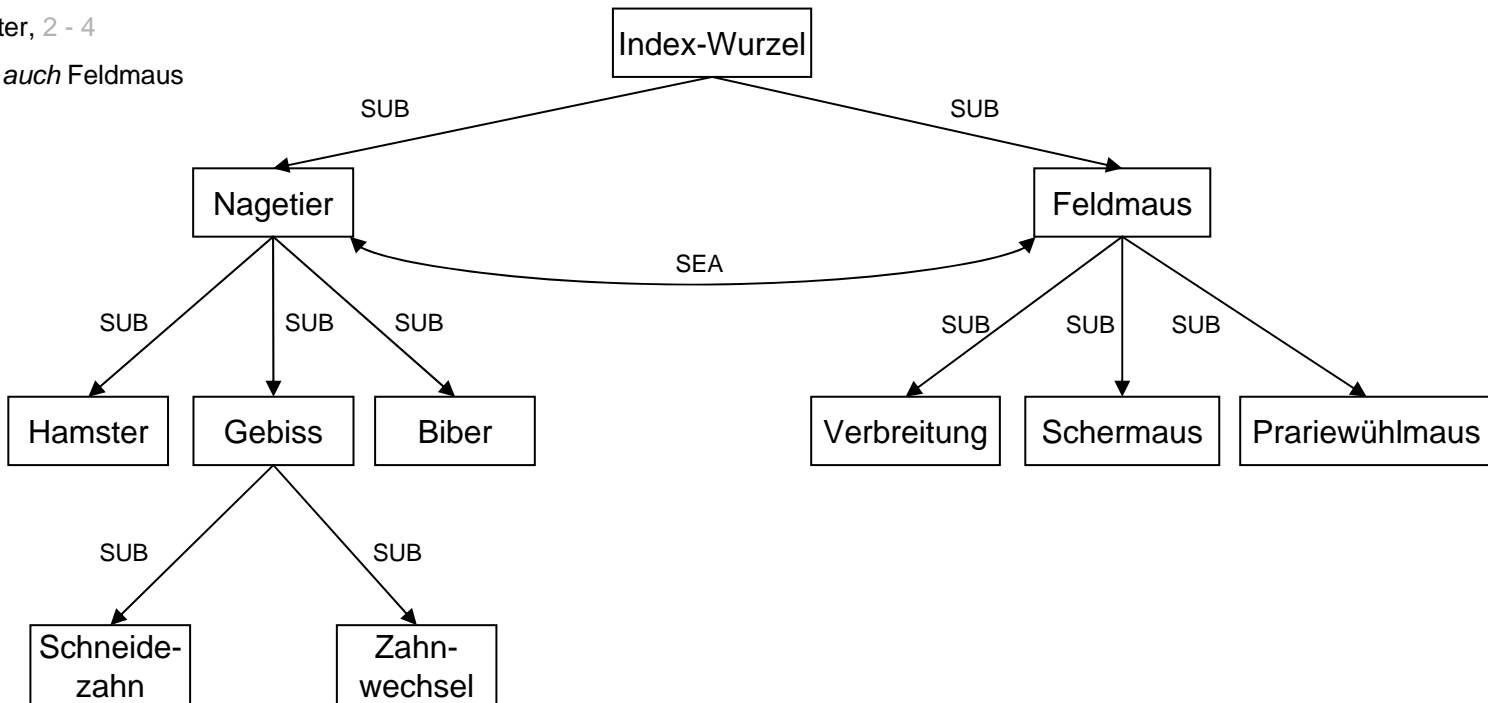
Feldmaus, 13, 15

Präriewühlmaus, 16

Schermaus, 16

Verbreitung, 15

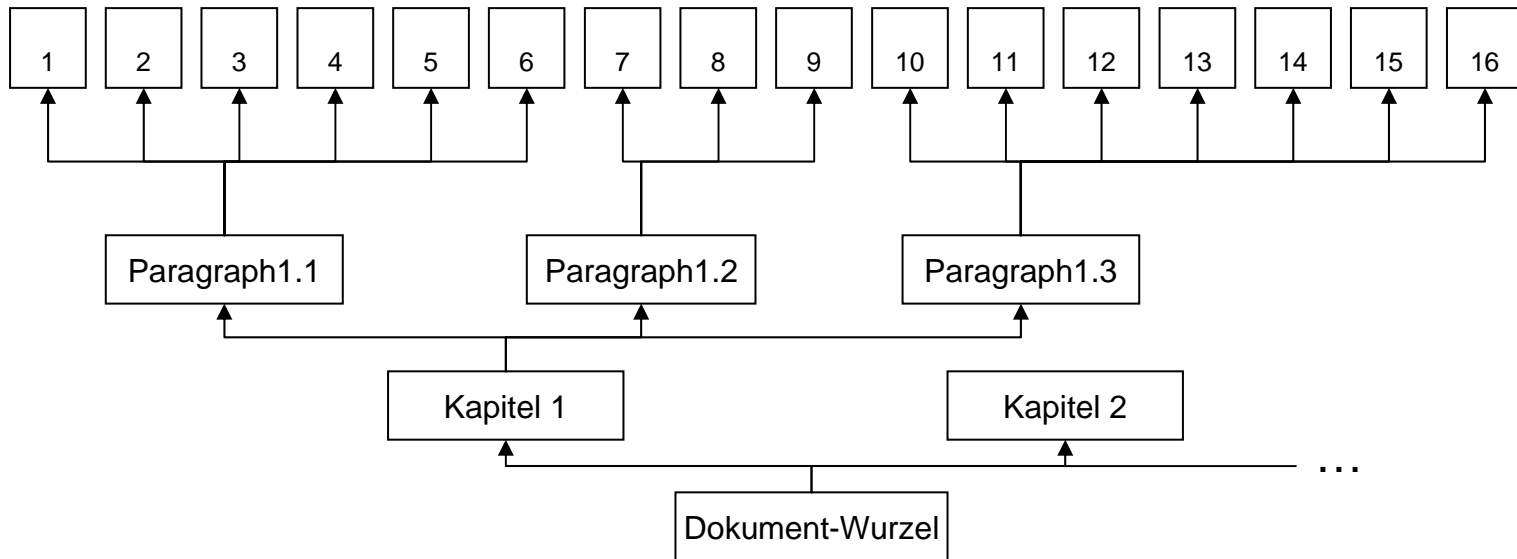
*Siehe auch* Nagetier



# Der Index-Graph (III)



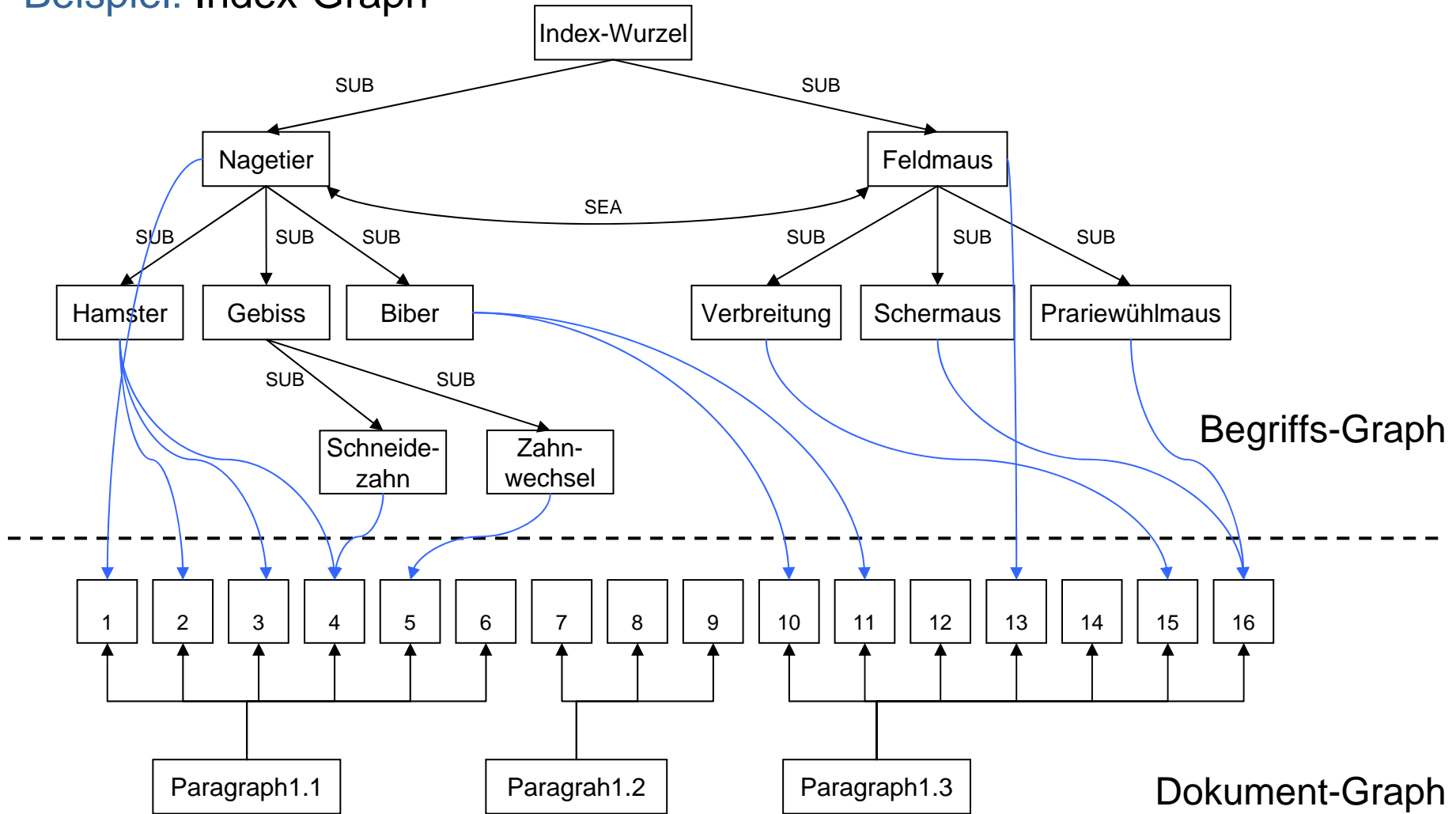
## Beispiel: Dokument-Graph



# Der Index-Graph (IV)



## Beispiel: Index-Graph



# Anreicherung der Indexdaten



- *Ergebnis:* Index-Ontologie und Index-Graph liefern eine computerverarbeitbare Annotation.
- Mit Hilfe von externem Wissen kann die Annotation semantisch angereichert werden.
- Externe Wissensquellen können sein:
  - Wiktionary
  - lexikalische Ressourcen (z.B. WordNet)
  - Domänenontologien
- Externe Wissensquellen liefern:
  - explizite Beschreibung der auftretenden Begriffe
  - exaktere oder neue semantische Beziehungen zwischen den Begriffen

# Verwendung der Index-Annotation



- Index-Ontologie + Indexgraph + externes Wissen  
➔ semantische Annotation
  
- Diese kann verschiedenartig eingesetzt werden:
  - im Autorensystem bei der Generierung eines verbesserten Indexes:  
SMARTINDEXER
  
  - Dokumentnavigation
  
  - Indexvisualisierung

# SMARTINDEXER



seit 1558

Akzeptanz / Ablehnung / Modifikation  
des vorgeschlagenen Indexeintrags

Indexeintragsvorschlag

Autor

Textverarbeitungs-  
programm

Dokument  
(z.B. \*.doc, \*.tex)

Index-  
prozessor

Dokumentindex  
(z.B. \*.idx)

1: potentieller Eintrag

6: Indexbefehle

2: existierender Index

SMARTINDEXER

Index-  
Generator

3: vorver-  
arbeiteter  
Indexein-  
trag

5: gefiltertes  
lexikalisches  
Umfeld

Ontologie-  
Prozessor

4: lexikal. Umfeld  
des Eintrag

Index-  
Ontologie

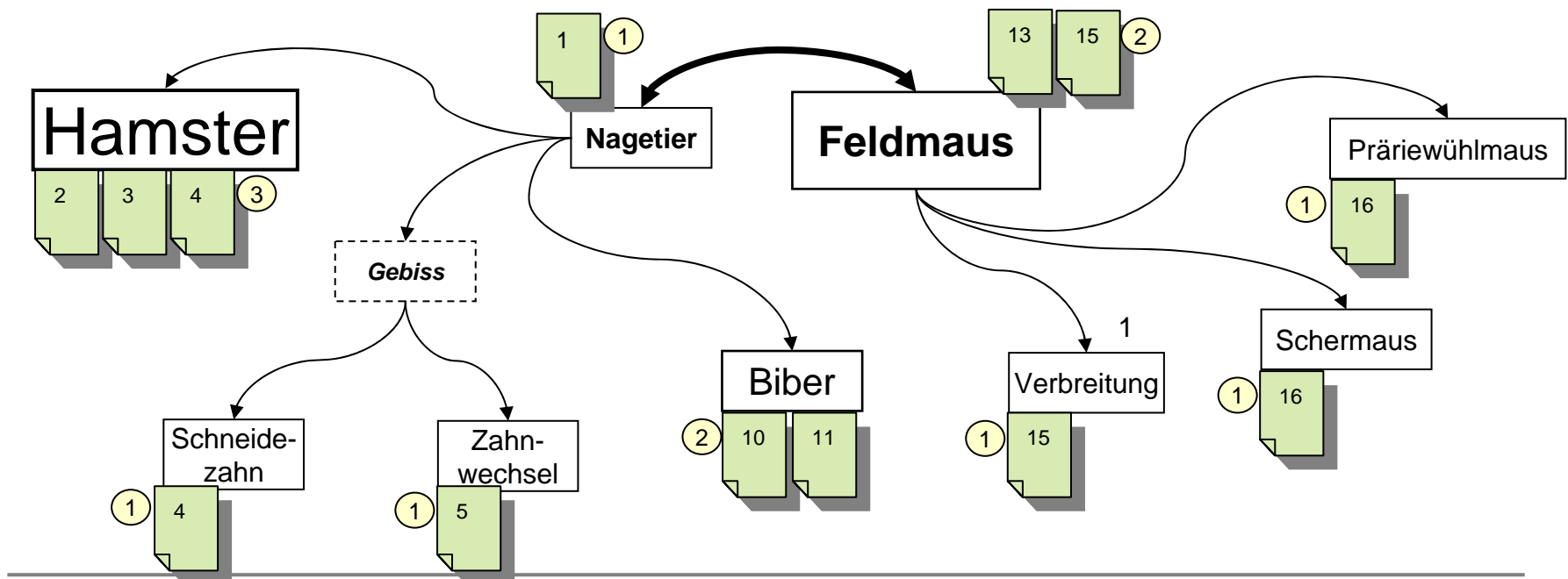
Domänen-  
Ontologien

WordNet

Glossary - Help  
SEARCH DISPLAY OPTIONS: [Select option to change] Change  
Enter a word to search for:  Search WordNet  
KEY: "S." = Show Synset (semantic) relations, "W." = Show Word (lexical) relations  
**Noun**  
• S. (n) mouse (any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails)  
• S. (n) shiner, black eye, mouse (a swollen bruise caused by a blow to the eye)  
• S. (n) mouse (person who is quiet or timid)  
• S. (n) mouse, computer mouse (a hand-operated electronic device that controls the coordinates of a cursor on your computer screen as you move

# Index-Visualisierung

- Begriffs-Graph ermöglicht eine alternative Visualisierung des herkömmlichen Indexes.
- Die Anzahl der Fundstellen eines Schlagwortes kann Rückschlüsse auf seine Relevanz im Dokument geben.



- Der Begriffs-Graph stellt Informationen zur Navigation durch das Dokument bereit.
- Pfadlänge zwischen zwei Knoten  $\longrightarrow$  semantische Distanz der Begriffe
- Aufgabe 1:
  - *Gegeben*: Begriff
  - *Gesucht*: Menge von weiteren relevanten Begriffen
- Aufgabe 2:
  - *Gegeben*: eine Menge von Begriffen, die man verstehen möchte
  - *Gesucht*: minimale Anordnung von Dokumenteinheiten, die Grundlage für das Verstehen der Begriffsmenge sind

- Das Generieren von semantischer Annotation ist kompliziert.
- Index und TOC eines Dokuments im Zusammenspiel mit externem Wissen stellen nützliche Informationen zur Annotation bereit.
- Dies erfordert:
  - allgemeines konzeptuelles Wissen über die Struktur des Dokuments,
  - allgemeines konzeptuelles Wissen über das Indizieren
  - Eine geeignete Repräsentation des Dokumentwissens
- Die resultierende Annotation kann verschiedenartig eingesetzt werden.
  
- Wie können andere Informationsquellen für die Annotation herangezogen werden?
- Wie können die Annotationen weiter angewandt werden?



seit 1558

# Vielen Dank für Ihre Aufmerksamkeit!