Text Mining in der Wettbewerberanalyse: Konvertierung von Text-archiven in XML-Dokumente

6. Konferenz der SAS Anwender in Forschung und Entwicklung (KSFE 2002)

Myra Spiliopoulou und Karsten Winkler
Handelshochschule Leipzig

Agenda

1. Semantische Auszeichnung fachspezifischer Textarchive in der Wettbewerberanalyse
2. DIAsDEM-Vorgehensmodell im Überblick
3. Prozeß der Wissensentdeckung zur semantischen Textauszeichnung
4. Fallstudie I: Handelsregistereinträge
5. Fallstudie II: Ad-hoc-Mitteilungen
6. Ihre Fragen und Diskussion

HHL. . .
Leipzig Graduate School of Management

Agenda

1. **Semantische Auszeichnung fachspezifischer Textarchive in der Wettbewerberanalyse**
2. DIAsDEM-Vorgehensmodell im Überblick
3. Prozeß der Wissensentdeckung zur semantischen Textauszeichnung
4. Fallstudie I: Handelsregistereinträge
5. Fallstudie II: Ad-hoc-Mitteilungen
6. Ihre Fragen und Diskussion

HHL. . .
Leipzig Graduate School of Management

Wettbewerberanalyse

- **Engl.: Competitive Intelligence (CI)**
- **"Competitive analysis is a systematic program for gathering and analyzing information about your competitors' activities and general business trends to further your own company's goals" (Kahaner)**
- **Ziele der Wettbewerberanalyse:**
  - **Antizipation von Wettbewerberaktivitäten oder relevanten Marktveränderungen**
  - **Sammlung von Wissen über potentiell relevante Technologien, Produkte und Vorschriften**

HHL. . .
LEIPZIG GRADUATE SCHOOL OF MANAGEMENT

Wettbewerberanalyse: Datenbasis

Patente

Ad-hoc-Mitteilungen

Aktienkurse

Meinungs-portale

Presse-mitteilungen

Handels-register

Branchen-verbände

Produkt-kataloge

Jahresab-schluß (XBRL)

(...)

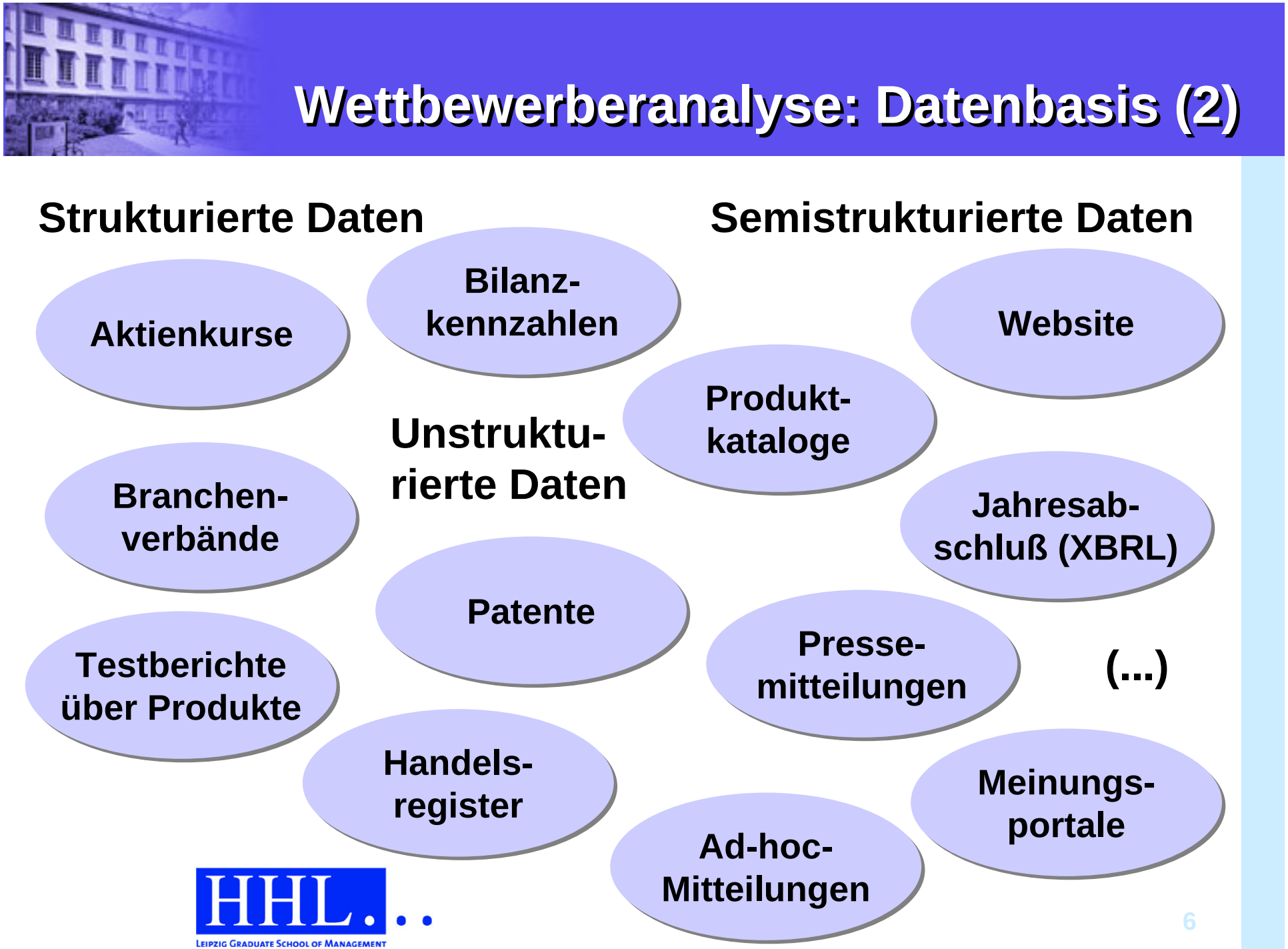
Bilanz-kennzahlen

Testberichte über Produkte

Website

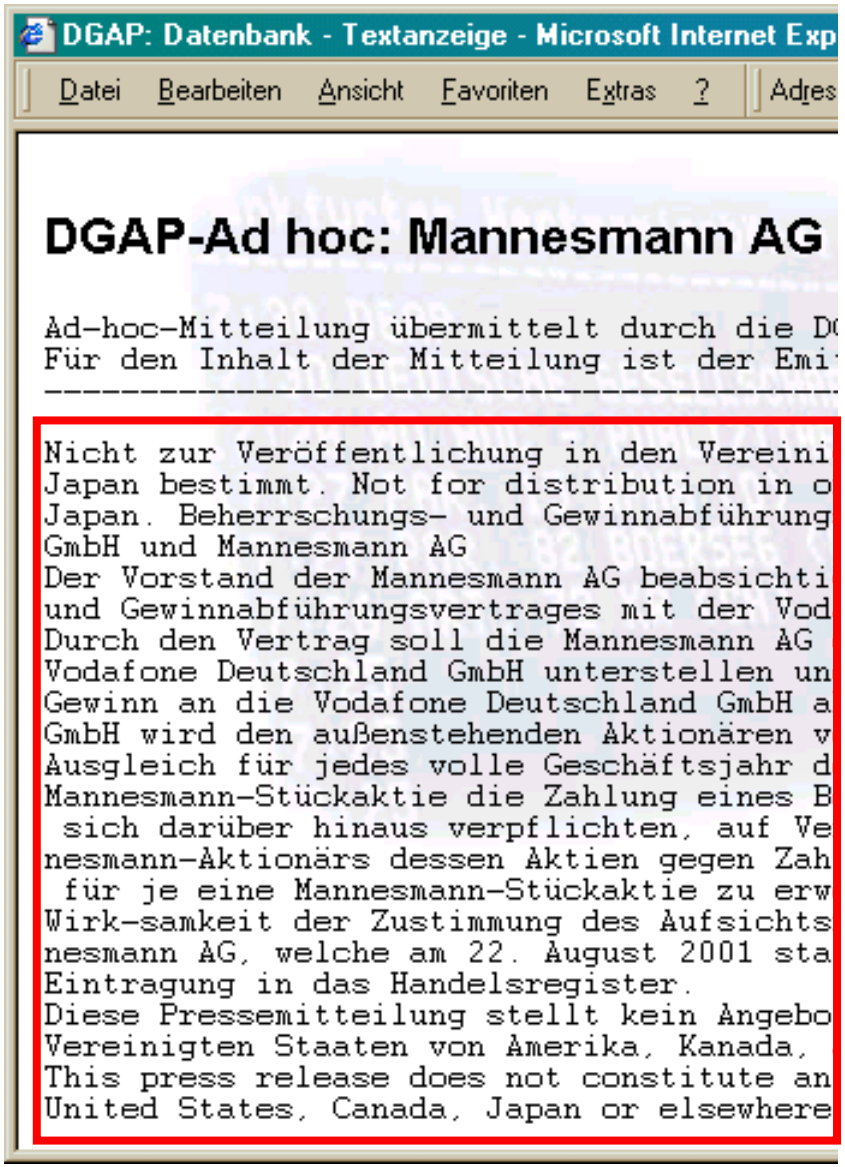
HHL. . .
Leipzig Graduate School of Management

Wettbewerberanalyse: Datenbasis (2)



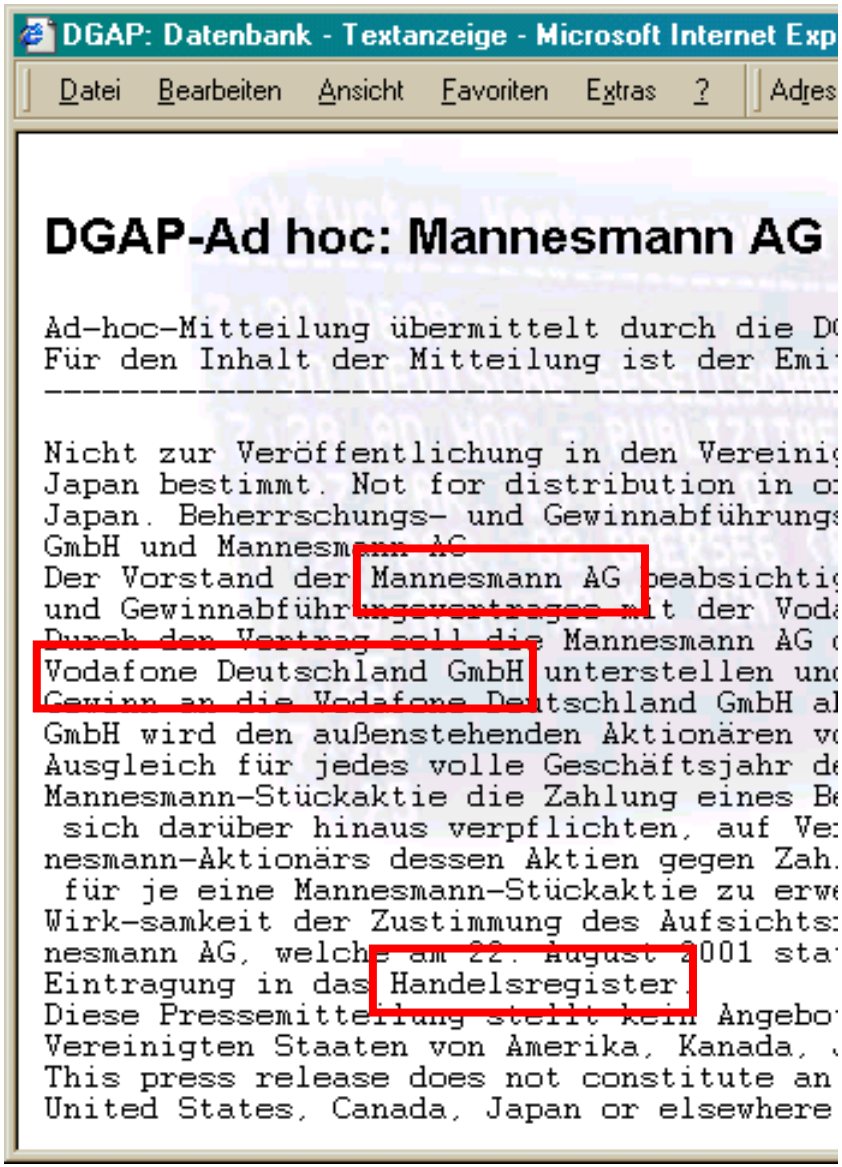
HHL. . .
Leipzig Graduate School of Management

Text Mining: Fokus Dokument



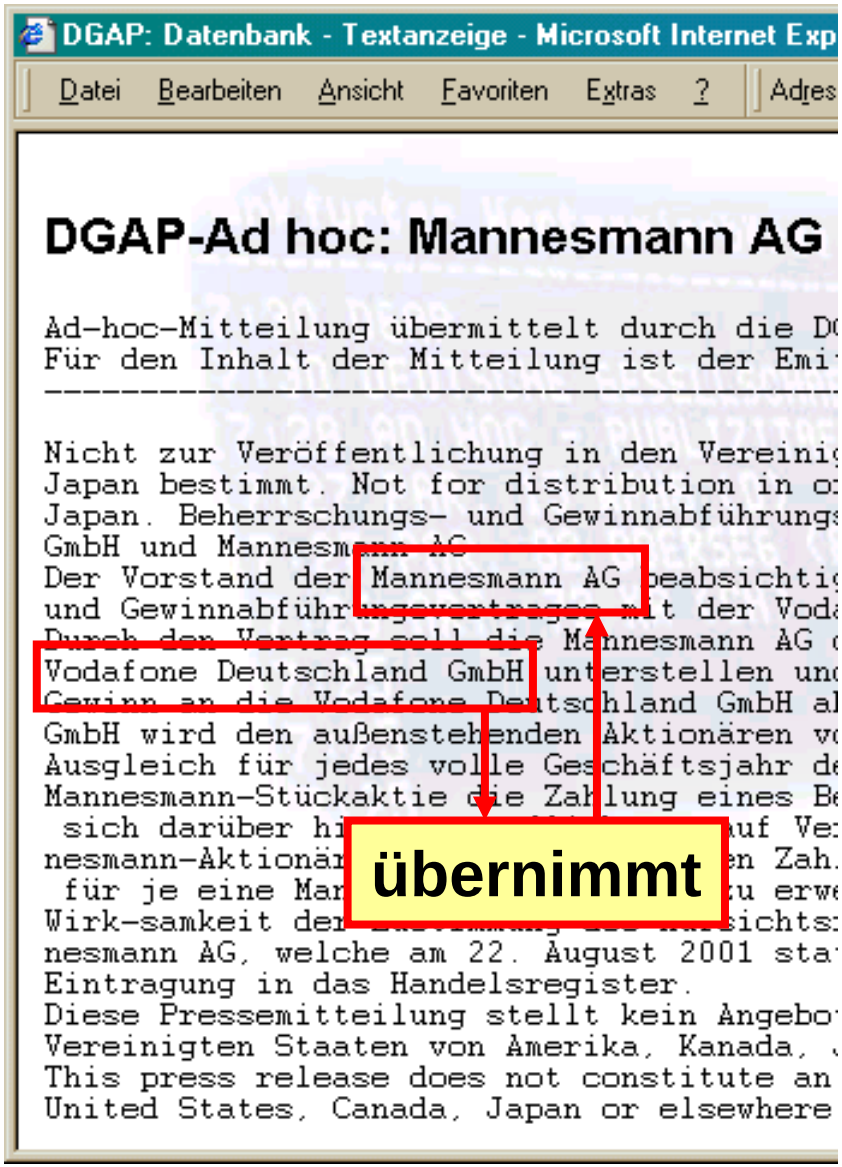
- **Clustering von Dokumenten nach Inhalt**
- **Klassifizierung von Dokumenten**
- **Entdeckung von Themen in Dokumenten**
- **Zusammenfassung von Dokumenten**

Text Mining: Fokus Terme



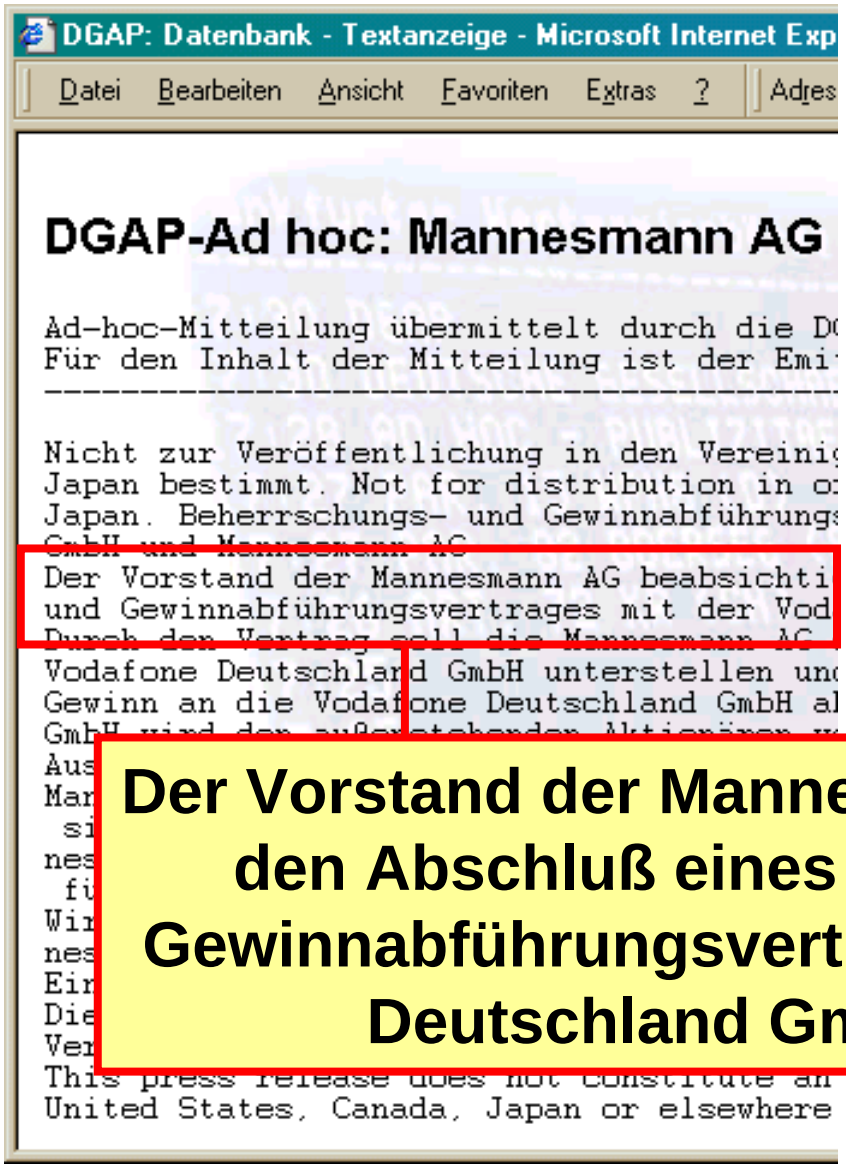
- **Extraktion wichtiger benannter Entitäten**
- **Entdeckung von Ontologien oder Thesauri**
- **Computerlinguistik: Grammatische Auszeichnung von Termen**

Text Mining: Fokus Textelemente



- **Entdeckung von Beziehungen zwischen benannten Entitäten (Strukturierung)**
- **Ableitung einer XML DTD und semantische Textauszeichnung**

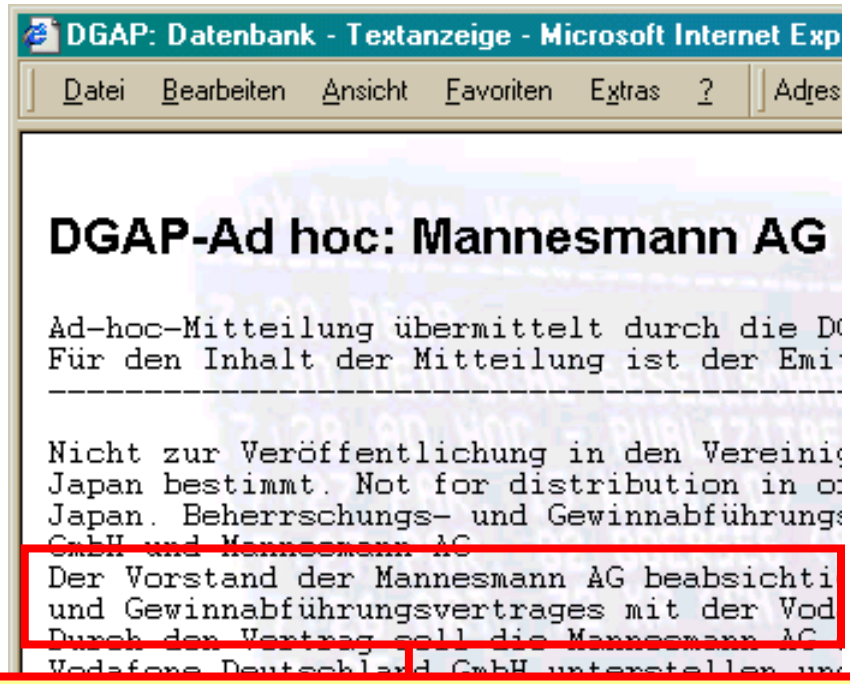
Text Mining: Fokus Textelemente (2)



- Entdeckung von Beziehungen zwischen benannten Entitäten (Strukturierung)
- **Ableitung einer XML DTD und semantische Textauszeichnung**

Der Vorstand der Mannesmann AG beabsichtigt den Abschluß eines Beherrschungs- und Gewinnabführungsvertrages mit der Vodafone Deutschland GmbH, Düsseldorf.

Semantische Textauszeichnung



- Entdeckung von Beziehungen zwischen benannten Entitäten (Strukturierung)
- **Ableitung einer XML DTD und semantische Textauszeichnung**

<AbschlußBeherrschungsvertrag Unternehmen= "Mannesmann AG && Vodafone Deutschland GmbH">
Der Vorstand der Mannesmann AG beabsichtigt den Abschluß eines Beherrschungs- und Gewinnabführungsvertrages mit der Vodafone Deutschland GmbH, Düsseldorf. </AbschlußBeherrschungsvertrag>

Agenda

1. Semantische Auszeichnung fachspezifischer Textarchive in der Wettbewerberanalyse
2. **DIAsDEM-Vorgehensmodell im Überblick**
3. Prozeß der Wissensentdeckung zur semantischen Textauszeichnung
4. Fallstudie I: Handelsregistereinträge
5. Fallstudie II: Ad-hoc-Mitteilungen
6. Ihre Fragen und Diskussion

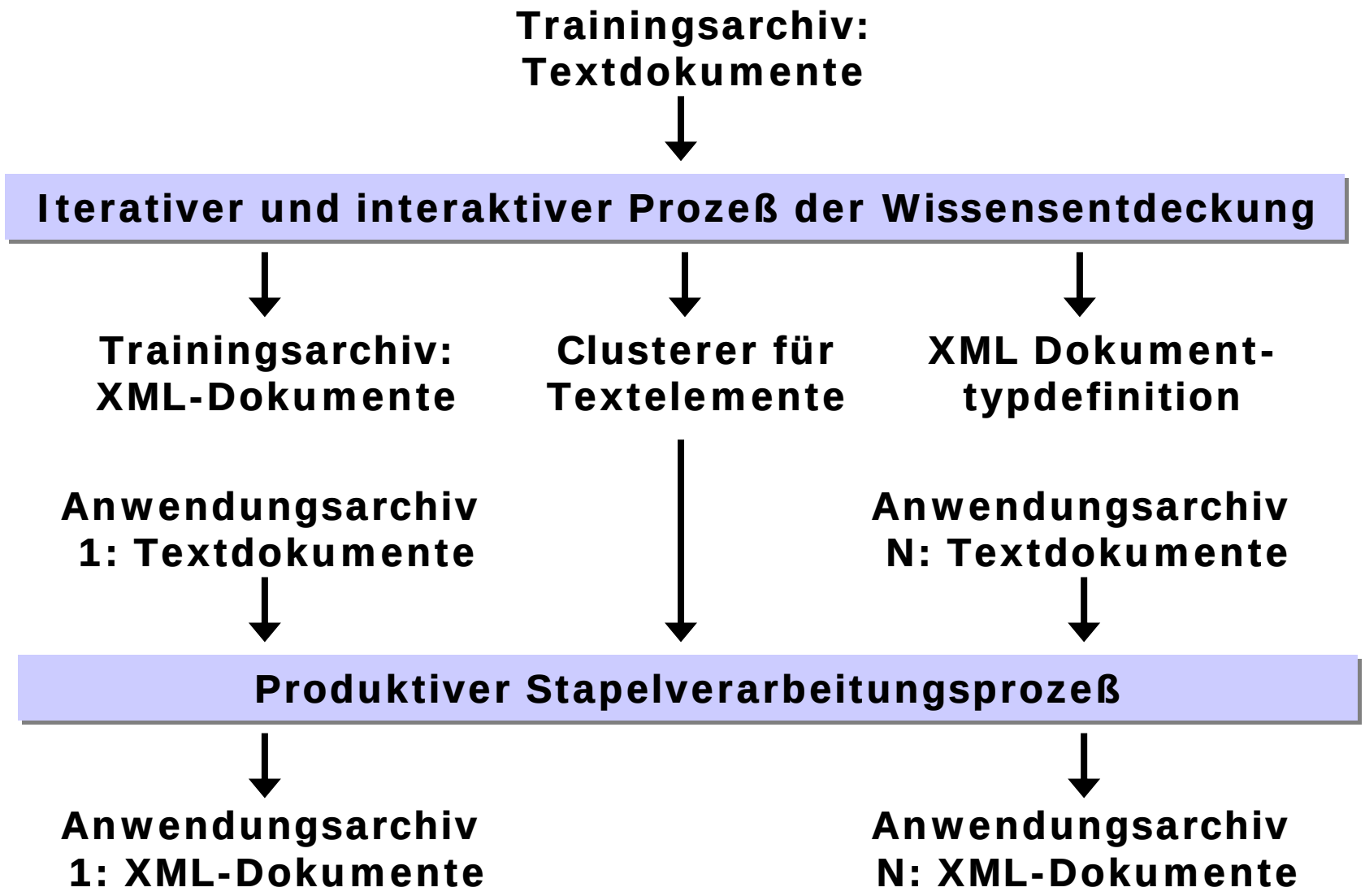
HHL. . .
Leipzig Graduate School of Management

DIAsDEM-Vorgehensmodell

- **Ziel 1: Semantische Auszeichnung großer, anwendungsspezifischer Textarchive mit der Textauszeichnungssprache XML**
  - **Überführung des Textarchivs in ein Archiv semistrukturierter XML-Dokumente**
- **Ziel 2: Ableitung einer möglichst strukturierten XML Dokumenttypdefinition für das Archiv**
  - **Nutzung einer XML-Anfragesprache für inhalts- und strukturbasierte Suche in XML-Archiven**
- **Methodik: Wissensentdeckung in Datenbanken**

HHL. . .
Leipzig Graduate School of Management

DIAsDEM-Vorgehensmodell (2)



Agenda

1. Semantische Auszeichnung fachspezifischer Textarchive in der Wettbewerberanalyse
2. DIAsDEM-Vorgehensmodell im Überblick
3. **Prozeß der Wissensentdeckung zur semantischen Textauszeichnung**
4. Fallstudie I: Handelsregistereinträge
5. Fallstudie II: Ad-hoc-Mitteilungen
6. Ihre Fragen und Diskussion

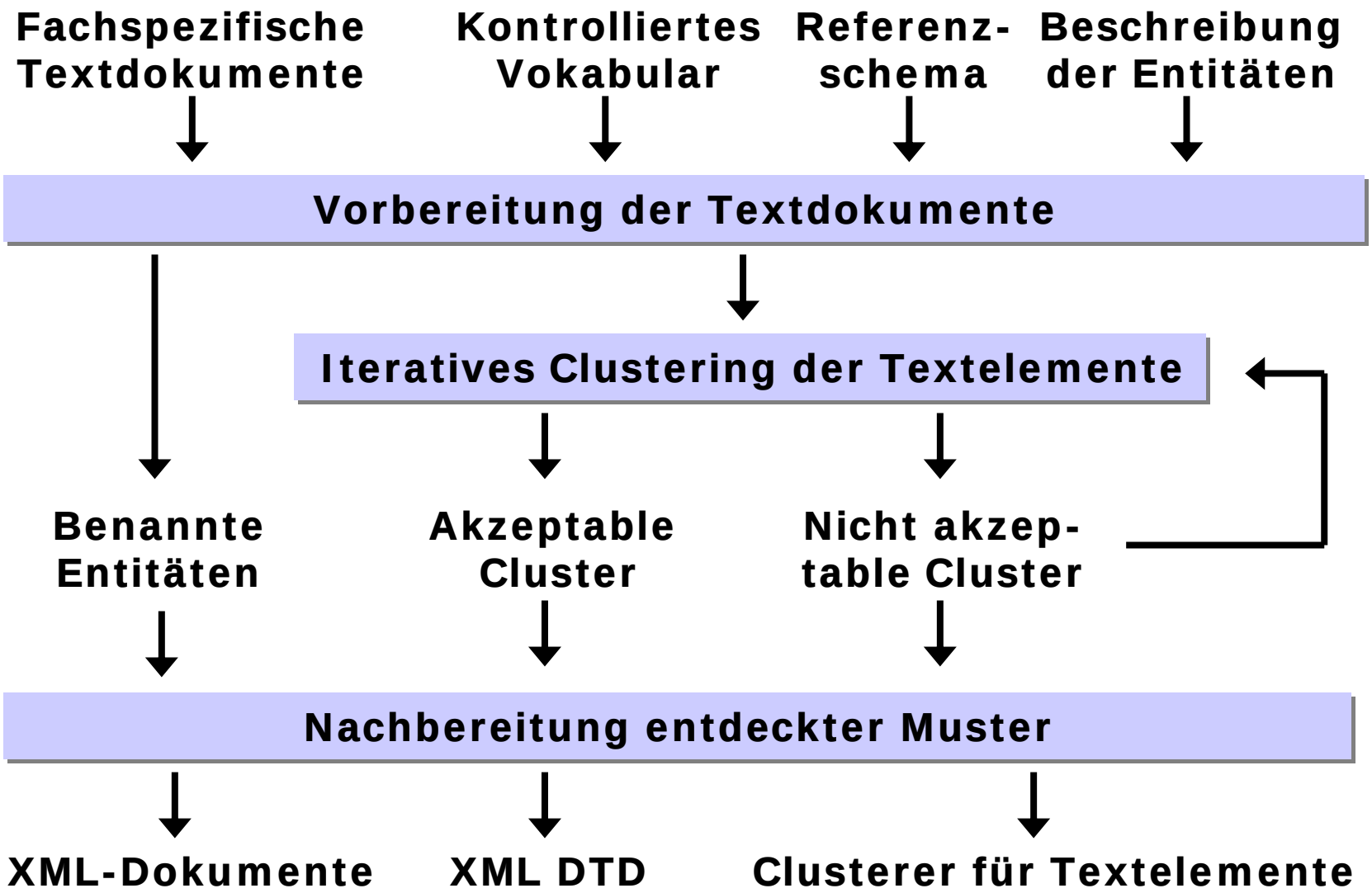
HHL. . .
Leipzig Graduate School of Management

Prozeß der Wissensentdeckung

1. Inhaltsbasierte Gruppierung struktureller Textelemente (z.B. Sätze oder Absätze) mit Clustering-Algorithmus
2. Inhaltliche Charakterisierung der entdeckten Gruppen von Textelementen: Benennung qualitativ hochwertiger Cluster
3. Identifizierung und Benennung benannter Entitäten (z.B. Personen) in Textelementen
4. Zusammenführung von Cluster-Bezeichnern und Bezeichnern benannter Entitäten zu einer XML Dokumenttypdefinition

HHL. . .
Leipzig Graduate School of Management

Prozeß der Wissensentdeckung (2)

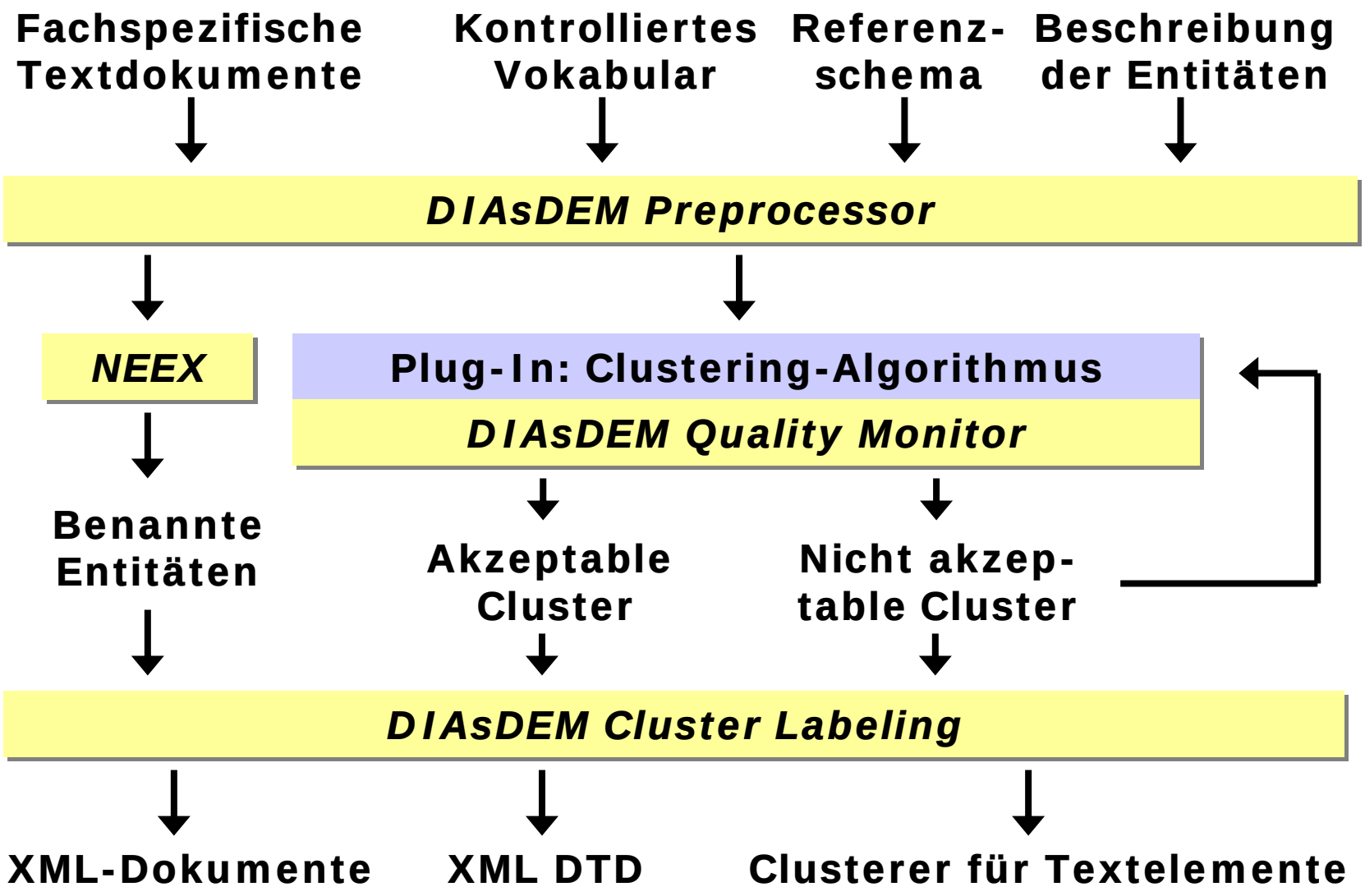


Iteratives Clustering

1. Festlegung der Parameter des Clustering-Algorithmus (Ähnlichkeit?)
2. Gruppierung der Vektoren entsprechend des gewählten Ähnlichkeitsmaßes
3. Evaluation der Qualität entdeckter Cluster mit Qualitätskriterien (Interessantheit?)
4. Ausblendung aller Vektoren in "akzeptablen" Clustern für spätere semantische Benennung
5. Wiederholung des Gruppierungsvorgangs mit Vektoren in "nicht akzeptablen" Clustern, sofern die Stop-Bedingung nicht erfüllt ist

HHL. . .
Leipzig Graduate School of Management

Die *DIAsDEM-Workbench*



Agenda

1. Semantische Auszeichnung fachspezifischer Textarchive in der Wettbewerberanalyse
2. DIAsDEM-Vorgehensmodell im Überblick
3. Prozeß der Wissensentdeckung zur semantischen Textauszeichnung
4. **Fallstudie I: Handelsregistereinträge**
5. Fallstudie II: Ad-hoc-Mitteilungen
6. Ihre Fragen und Diskussion

HHL. . .
Leipzig Graduate School of Management

Fallstudie I: Handelsregistereintrag

Daniel Spiel-Center GmbH Potsdamer Str. 94, 14513 Teltow	HRB 12576 06.05.99

Der Betrieb von Spielhallen in Teltow und das Aufstellen von Geldspiel- und Unterhaltungsautomaten. Stammkapital: 25.000 EUR. Gesellschaft mit beschränkter Haftung. Der Gesellschaftsvertrag ist am 12. November 1998 abgeschlossen und am 19. April 1999 abgeändert. (...) Pawel Balski, 14.04.1965, Berlin, ist zum Geschäftsführer bestellt. Er vertritt die Gesellschaft stets einzeln und (...)

HHL. . .
Leipzig Graduate School of Management

Fallstudie I: Vorbereitung

1. Zerlegung der Texte in Textelemente und Terme

```
(...) 1999 abgeändert | Pawel Balski , 14.04.1965 , Berlin,
ist zum Geschäftsführer bestellt . | Er vertritt die (...)
```

2. Ausblendung identifizierter benannter Entitäten

```
PERSON , DATE , PLACE , ist zum Geschäftsführer bestellt .
```

3. Ermittlung der Wortstämme sämtlicher Terme

```
PERSON , DATE , PLACE , sein zu Geschäftsführer bestellen .
```

4. Auswahl von Deskriptoren aus dem Thesaurus

```
(Bestellung, ..., Geschäftsführer, Gründung, Gesellschaft)
```

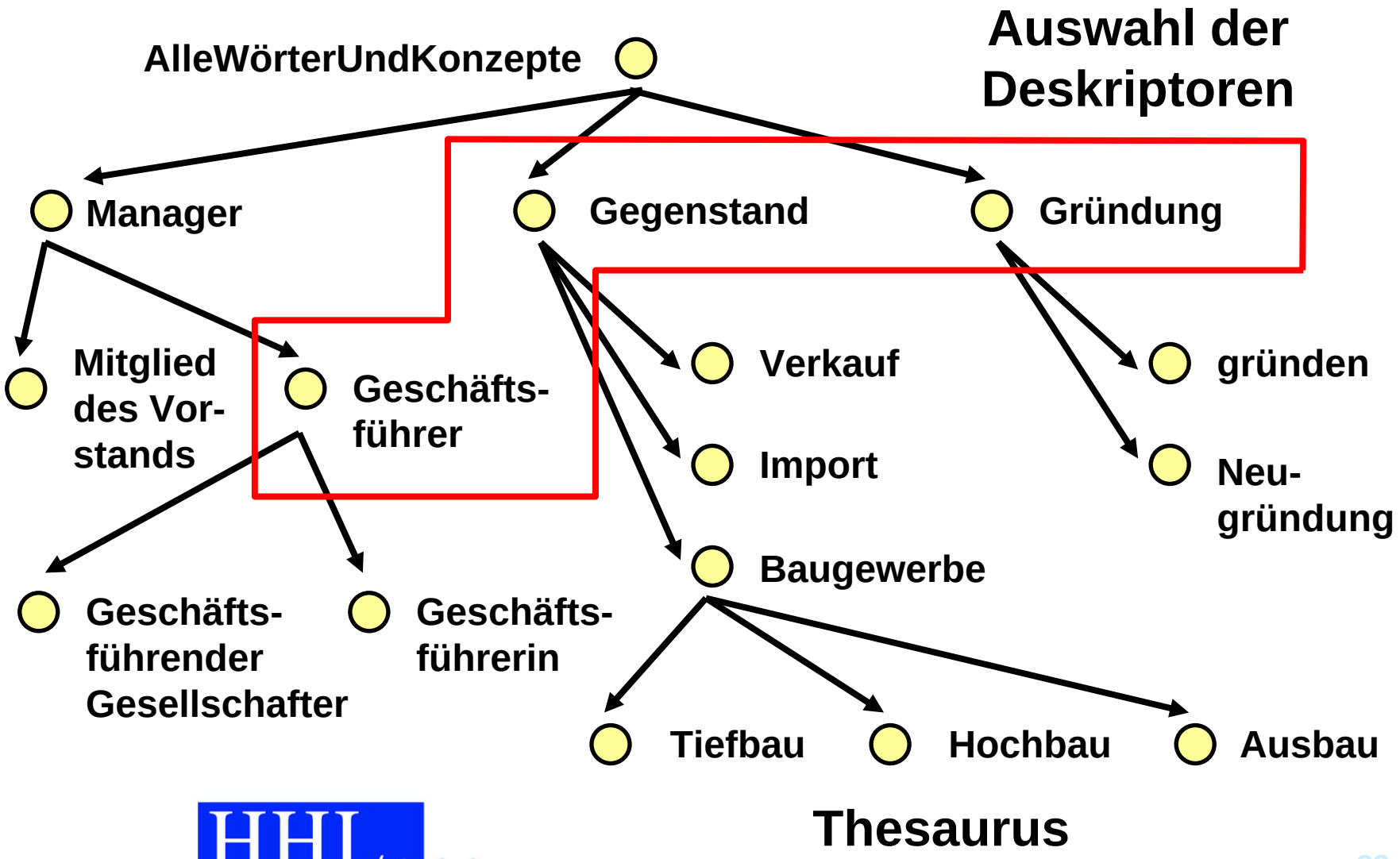
5. Abbildung der Textelemente auf Boolesche Vektoren

```
(1,          ..., 1,                0,       0          )
```

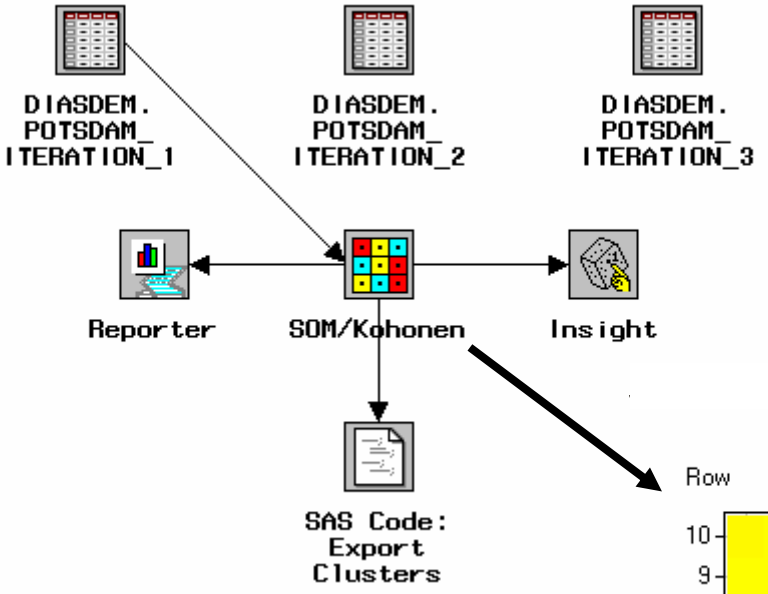
6. Gewichtung der Textelementvektoren mit TFxIDF

```
(1.4283,     ..., 0.9010,           0,       0          )
```

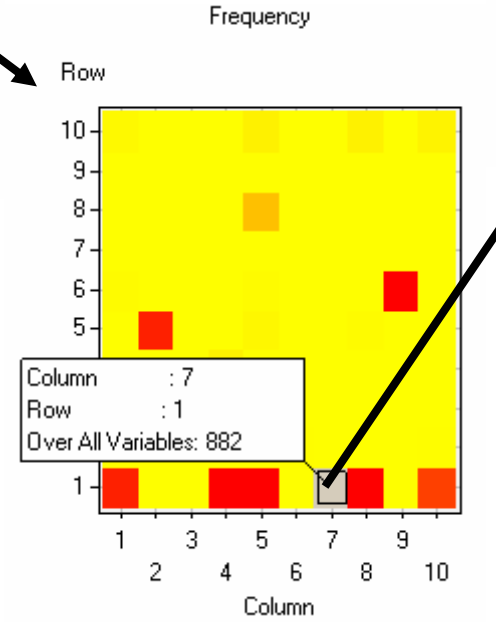
Fallstudie I: Vektorraum

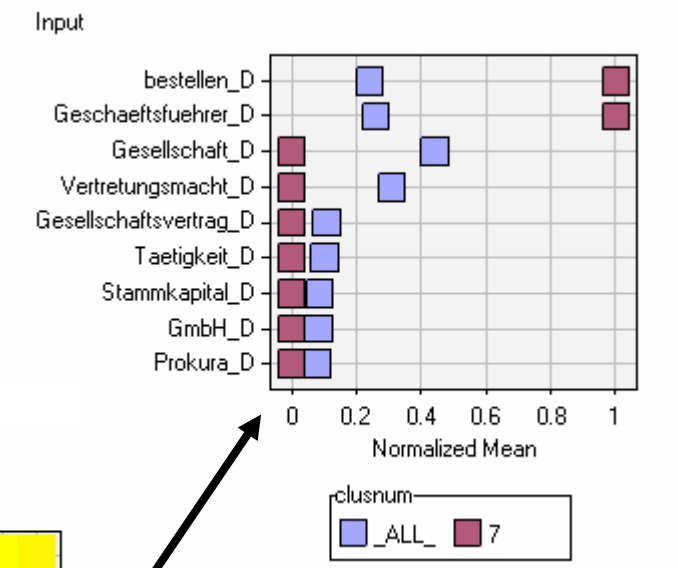


Fallstudie I: Data Mining mit SAS/EM



Data-Mining-Diagramm





Visualisierung des Cluster 7

Ergebnis der 1. Iteration

HHL Leipzig Graduate School of Management

Fallstudie I: XML-Dokument

Daniel Spiel-Center GmbH Potsdamer Str. 94, 14513 Teltow	HRB 12576 06.05.99

```
<Handelsregistereintrag> <Gegenstand>
Der Betrieb von Spielhallen in Teltow
und das Aufstellen von Geldspiel- und
Unterhaltungsautomaten. </> (...)
<GeschäftsführerBestellen Person="Balski,
Pawel, 14.04.1965, Berlin, null"> Pawel
Balski, 14.04.1965, Berlin, ist zum
Geschäftsführer bestellt. </>
(...) </Handelsregistereintrag>
```

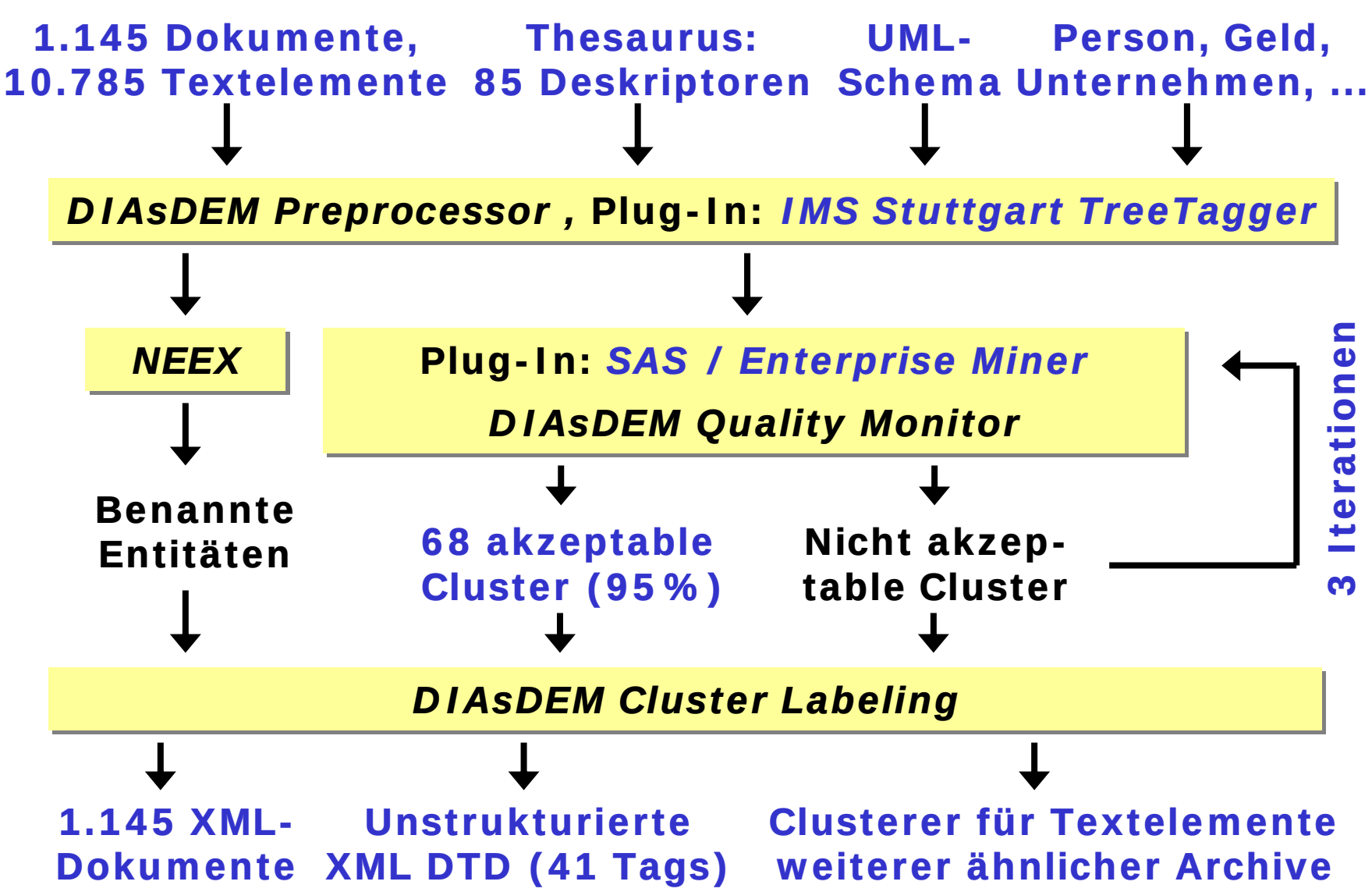
HHL. . .
Leipzig Graduate School of Management

Fallstudie I: XML DTD

```
<?xml version="1.0" encoding="ISO-8859-1"?>

<!ELEMENT Handelsregistereintrag ( #PCDATA |
Gegenstand | (...) |  Stammkapital | Verän-
derungSitz | PersönlichHaftendeGesellschafter
| GeschäftsführerBestellen | AufteilungGrund-
kapital | (...) | GründungGesellschaft )* >

<!ELEMENT Gegenstand (#PCDATA)> (...)
<!ELEMENT Stammkapital (#PCDATA)> (...)
<!ELEMENT GründungGesellschaft (#PCDATA)>

(...) <!ATTLIST GeschäftsführerBestellen
Person CDATA #IMPLIED> (...)
```

HHL. . .
LEIPZIG GRADUATE SCHOOL OF MANAGEMENT

Fallstudie I: Zusammenfassung



Fallstudie I: Evaluation

- **Evaluation im Hinblick auf Auszeichnungsfehler:**
  - **Fehlertyp I: Ein XML-Tag reflektiert nicht den genauen Inhalt eines Textelements**
  - **Fehlertyp II: Ein Textelement ist nicht annotiert, obwohl es ein Konzept enthält, das Teil der abgeleiteten XML DTD ist**
- **Manuelle Überprüfung von 5% der Texteinheiten:**
  - **Fehlertyp I: 1,5% in der Stichprobe**
  - **Fehlertyp II: 1,8% in der Stichprobe**
  - **Gesamtfehler mit 0,95-Konfidenz: [2,35%; 4,24%]**

HHL. . .
LEIPZIG GRADUATE SCHOOL OF MANAGEMENT

Agenda

1. Semantische Auszeichnung fachspezifischer Textarchive in der Wettbewerberanalyse
2. DIAsDEM-Vorgehensmodell im Überblick
3. Prozeß der Wissensentdeckung zur semantischen Textauszeichnung
4. Fallstudie I: Handelsregistereinträge
5. **Fallstudie II: Ad-hoc-Mitteilungen**
6. Ihre Fragen und Diskussion

HHL. . .
Leipzig Graduate School of Management

Fallstudie II: XML-Dokument

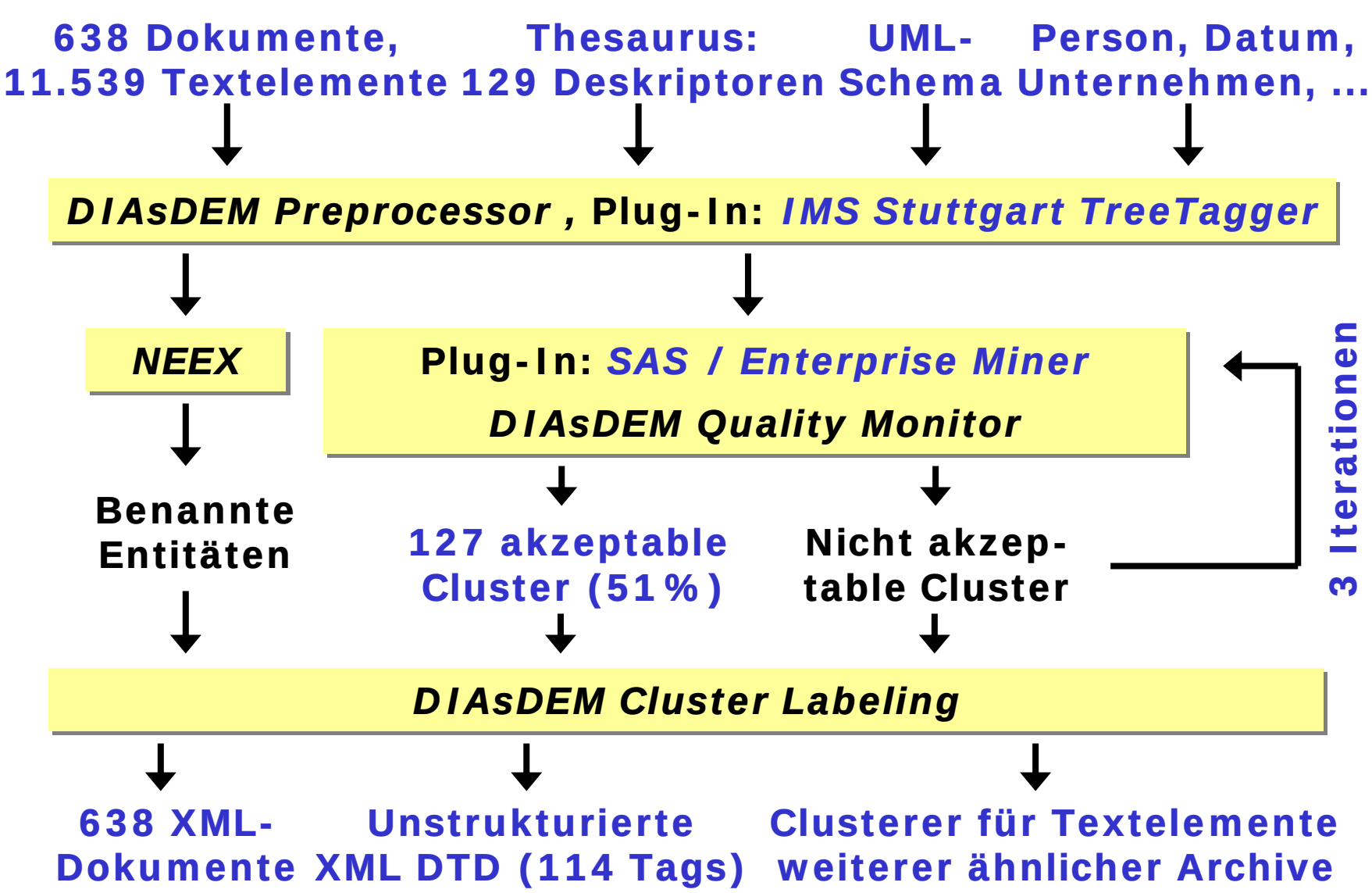
```
<AdHocMitteilung>
<NiederlegungMandatAufsichtsrat Company=
"WESTGRUND AG" Person="Rurack; Klaus" Date=
"04.01.2000> Der Vorstand der WESTGRUND AG teilt
gem. § 15 WpHG mit, dass das Mitglied des
Aufsichtsrats der WESTGRUND AG, Herr Klaus
Rurack (Remscheid), mit Wirkung zum 4.01.2000
sein Amt als Mitglied des Aufsichtsrats der
WESTGRUND AG niederlegt. </> <Hauptversammlung
Company="WESTGRUND AG"> Da die nächste
ordentliche Hauptversammlung der WESTGRUND AG
voraussichtlich im Mai 2000 stattfinden wird,
wird die WESTGRUND AG ein neues Aufsichtsrats-
mitglied durch Gerichtsbeschluss bestellen
lassen. </> (...) </AdHocMitteilung >
```

Fallstudie II: XML DTD

```
<?xml version="1.0" encoding="ISO-8859-1"?>

<!ELEMENT AdHocMitteilung ( #PCDATA |
PositivesQuartal | AuftragseingangImQuartal |
NeuesProdukt | AnstiegUmsatzImQuartal |
EntwicklungVonProdukten | ErwartungenUmsatz |
AkquisitionVonUnternehmen | NeuerKunde |
Grundkapital | AusbauVonBereichen | Planung |
Marktposition | Kapitalerhoehung | Vorstand |
ErfolgAmMarkt | NiederlegungMandatAufsichtsrat
| (...) | NeueTochtergesellschaft )* >

<!ELEMENT PositivesQuartal (#PCDATA)> (...)
<!ELEMENT NeueTochtergesellschaft (#PCDATA)>
```

HHL. . .
LEIPZIG GRADUATE SCHOOL OF MANAGEMENT

Fallstudie II: Zusammenfassung



Fallstudie II: Evaluation

- **Evaluation im Hinblick auf Auszeichnungsfehler:**
  - **Fehlertyp I: Ein XML-Tag reflektiert nicht den genauen Inhalt eines Textelements**
  - **Fehlertyp II: Ein Textelement ist nicht annotiert, obwohl es ein Konzept enthält, das Teil der abgeleiteten XML DTD ist**
- **Manuelle Überprüfung von 5% der Texteinheiten:**
  - **Fehlertyp I: 10,8% in der Stichprobe**
  - **Fehlertyp II: 6,4% in der Stichprobe**
  - **Gesamtfehler mit 0,95-Konfidenz: [16,29%; 18,11%]**

HHL. . .
Leipzig Graduate School of Management

Zusammenfassung und Ausblick

- ***DIAsDEM Workbench* + *SAS / Enterprise Miner* = Text Mining für die Strukturierung von Archiven**
- **Nächste Schritte im Forschungsprojekt DIAsDEM:**
  - **Evaluierung und Verbesserung des Vorgehens-modells in sprachlich komplexen Domänen**
  - **Zielorientierte Bewertung klassischer Clustering-Algorithmen und Ähnlichkeitsmetriken**
  - **Ableitung einer strukturierteren XML DTD**
  - **Integration von XML-Archiven mit relationalen Daten und Auswahl einer XML-Anfragesprache**

HHL. . .
LEIPZIG GRADUATE SCHOOL OF MANAGEMENT

Vielen Dank

an die DFG

und den SAS e-Intelligence Academic Club !

Fragen ?

Karsten Winkler
kwinkler@ebusiness.hhl.de
http://ebusiness.hhl.de

HHL Leipzig Graduate School of Management