PG-402 Wissensmanagement: Ontologiebasierte Wissensextraktion

WS2001/2002

Klaus Unterstein

Verlauf

- Begriffsklärung
- Ontologiebasierte Wissensextraktion
- Methoden
- Vor- und Nachteile im Vergleich
- Bewertung der Ansätze
- Praxis/Trends
- Schlußwort

Begriffsklärung

- Ontologie(n)
- (Wissens-) Extraktion
- Ontologiebasierte Wissensextraktion (OWE)

Ontologie(n)

- Definition
- Motivation
- Zweck
- Beschreibung
- Einsatz
- Bewertung

Ontologie(n) - Definition

Was ist eine Ontologie ?

Definition (Gruber):

„An ontology is a formal, explicit specification of a shared conceptualization.“ [1993]

Eine Ontologie beschreibt explizit eine formale, verteilte Konzeptualisierung eines bestimmten, uns interessierenden Bereichs.

Ontologie(n) – Motivation (1)
(allgemein)

Warum benutzen wir Ontologien ?

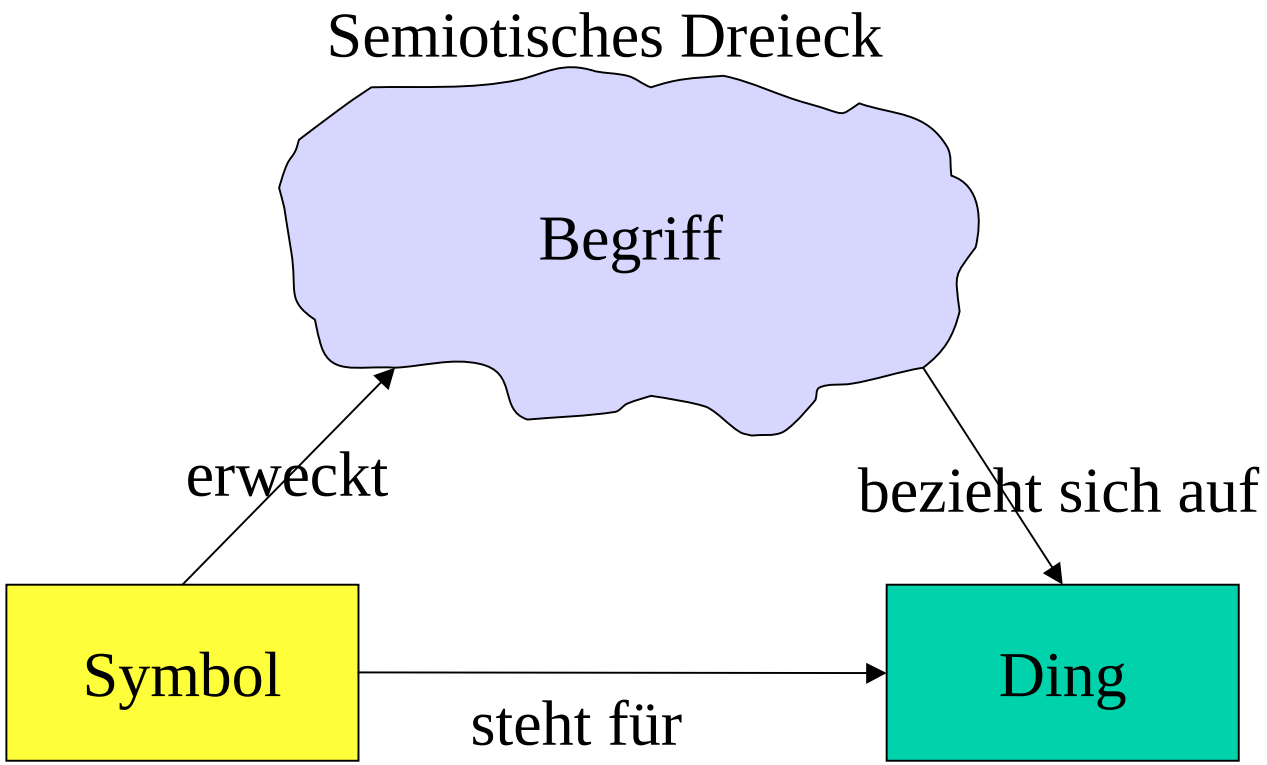
- Anzahl gespeicherter Informationsquellen wachsen
- Zugriff, Finden und Zusammenfassen von Informationen immer schwieriger
- Große Lücke zwischen Konzeptualisierung der Informationen und gespeicherte Form

Ontologie(n) - Motivation (2)
(spezieller Zweck)

Warum benutzen wir Ontologien ? (Fortsetzung)

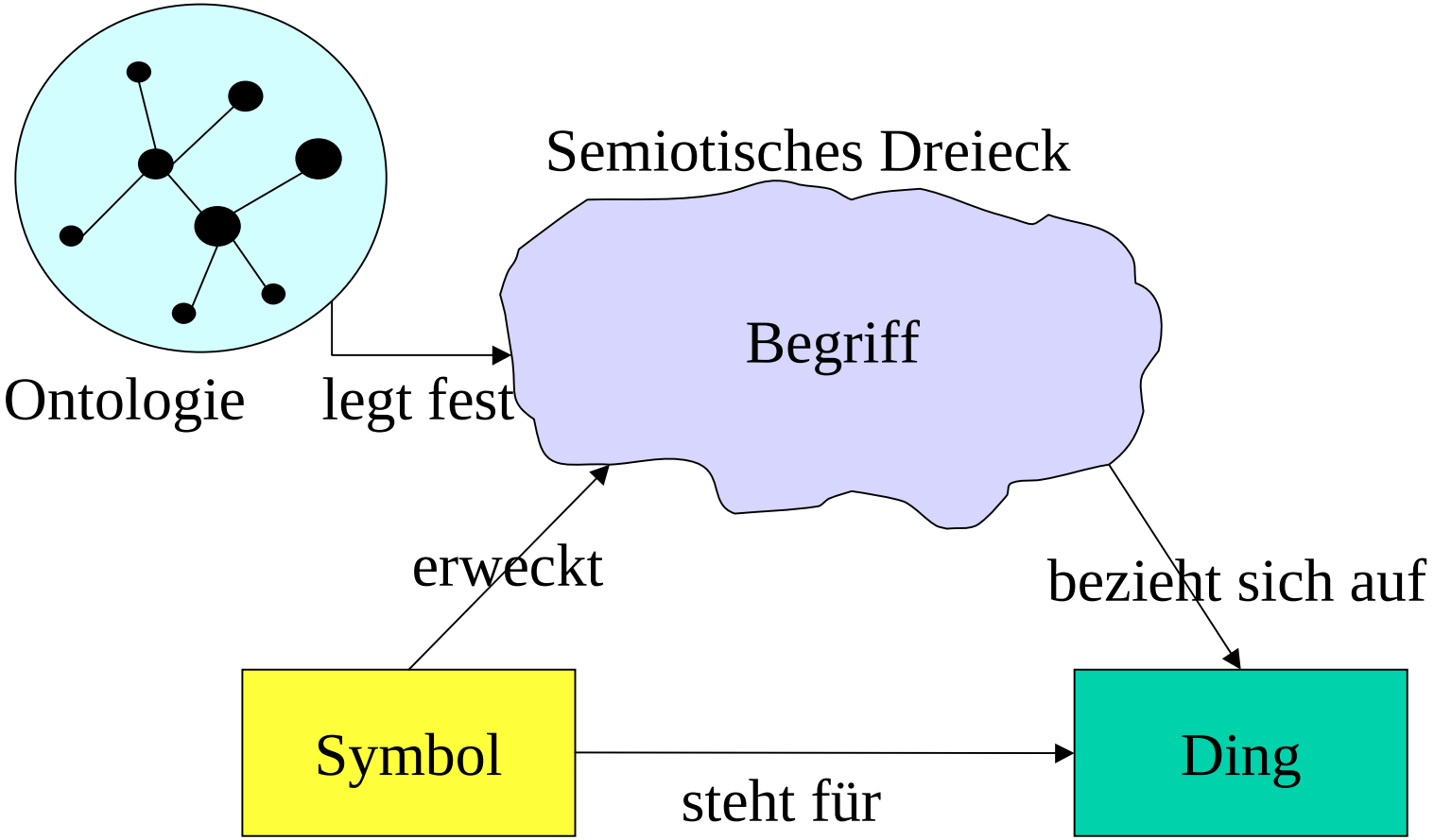
- Kommunikationshilfe zwischen Mensch und Maschine, was den Austausch von Semantik UND Syntax unterstützen soll
- Knowledge sharing und Wiederverwendung
- Zur Festlegung bestimmter Begriffe (Eindeutigkeit)
- Sie erzwingen eine wohldefinierte Semantik auf solche Konzeptualisierungen
- Sie sollen Hintergrund-Wissen zur Verfügung stellen, um die Leistung von Informations-Extraktions-Systemen zu erhöhen
- Formalisierung von implizit vorhandenem Wissen

Ontologie(n) - Motivation (3)



[Odgen, Richards, 1923]

Ontologie(n) - Motivation (4)



[nach: S. Staab, 2001]

Ontologie(n) – Beschreibung (1)

Eine Ontologie wird beschrieben durch:

- Eine Menge von Zeichenketten, die die lexikalen Einträge L für Konzepte und Relationen beschreiben
- Eine Menge von Konzepten C
- Eine Taxonomie von Konzepten (bei einigen Definitionen Heterarchie) H_C

Ontologie(n) – Beschreibung (2)

(Fortsetzung)

- Ein Satz an nicht-taxonomischen Relationen R (beschrieben durch ihre Domain)
- Relationen F und G, die Konzepte und Relationen verknüpfen
- Die Taxonomie der Relationen (bzw. Heterarchie H_R) (optional)
- Axiome A, die weitere Constraints der Ontologie beschreiben und es erlauben, implizite Fakten explizit zu machen (optional)

Ontologie(n) - Zweck

Ontologien beschreiben:

- Domain-relevante Konzepte
- Beziehungen zwischen den Konzepten
- Axiome für die Konzepte und Beziehungen

Ontologie(n) - Einsatz

Einsatz von Ontologien in Informations-Extraktions-Systemen zur:

- Integration von Informationen aus heterogenen Quellen
- Extraktion weiterer Fakten durch „Schliessen“ (Inferenz)
- Generierung verschiedener Ziel-Strukturen zur Informationsspeicherung
- Einfache Anpassung/Änderung während der Laufzeit

Ontologie(n) – Bewertung

Vorteile:

- Einfaches Prinzip
- Betrachtung relevanter Bereiche (Fokussierung)
- Vorteile durch Nutzung von Semantik und Hintergrundwissen
- Dynamische Entwicklung (siehe Such-Maschine)
- Semi-automatische Ansätze

Nachteile:

- (bisher) manuelle Erstellung
- Zeitliche Erstellung
- Problematik: Vollständigkeit vs. Minimalität

(Wissens-) Extraktion

- Definition
- Extraktion von Informationen
- Verschiedene Quellen (DB, WWW, Mail...)
- Verschiedene Datenformate (HTML, XML, unstrukturierter Text,...)
- Verschiedene Extraktions-Methoden (Anwendung abhängig vom Datenformat)

Wissensextraktion

Eine mögliche Definition:

Der Prozeß, in dem Information automatisch aus textuellen Dokumenten in eine zur Speicherung in Datenbanken geeignete Form generiert wird. [J. M. Lawler, 1998]

Ontologiebasierte Wissensextraktion:

Was ist ontologiebasierte Wissensextraktion (kurz: OWE) ?

Die Verwendung von Ontologien zur Unterstützung des Wissensextraktionsprozesses auf verschiedene Weisen.

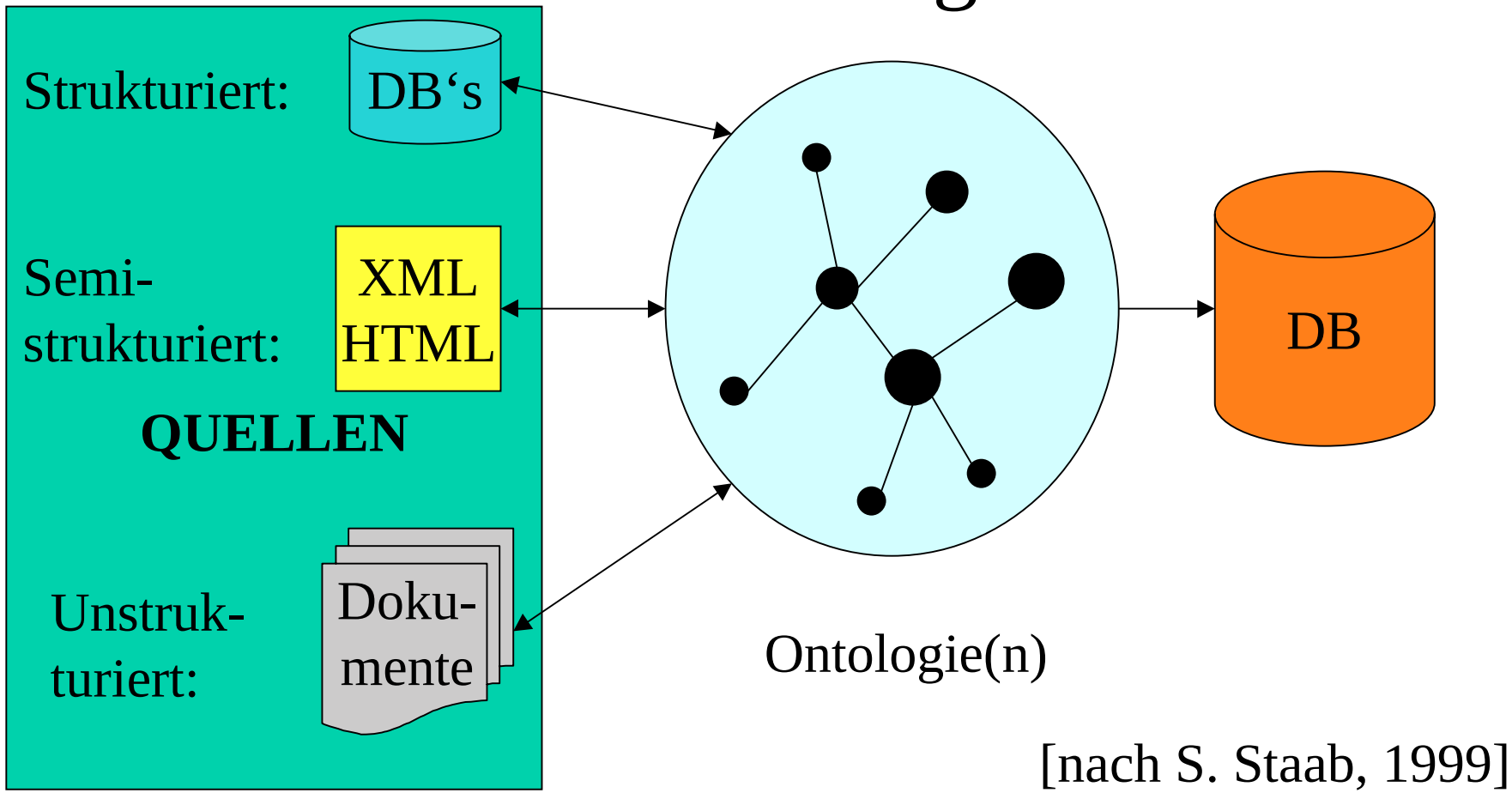
OWE

- Allgemein
- Informationsextraktion und Integration mittels Ontologien
- Klassifikationskriterien
- Verfahren

OWE - Allgemein

- Verwendung von Ontologien im Extraktionsprozeß
- Wahl der Ontologie abhängig vom Anwendungsbereich
- Flexible Extraktion abhängig von Ontologie
- Extraktionsprozeß liefert Informationen für die semantische Annotation der Texte
- Annotation liefert als Nebenprodukt die Klassifikation der Daten, die dadurch direkt integriert werden können

Informationsextraktion & Integration mittels Ontologien

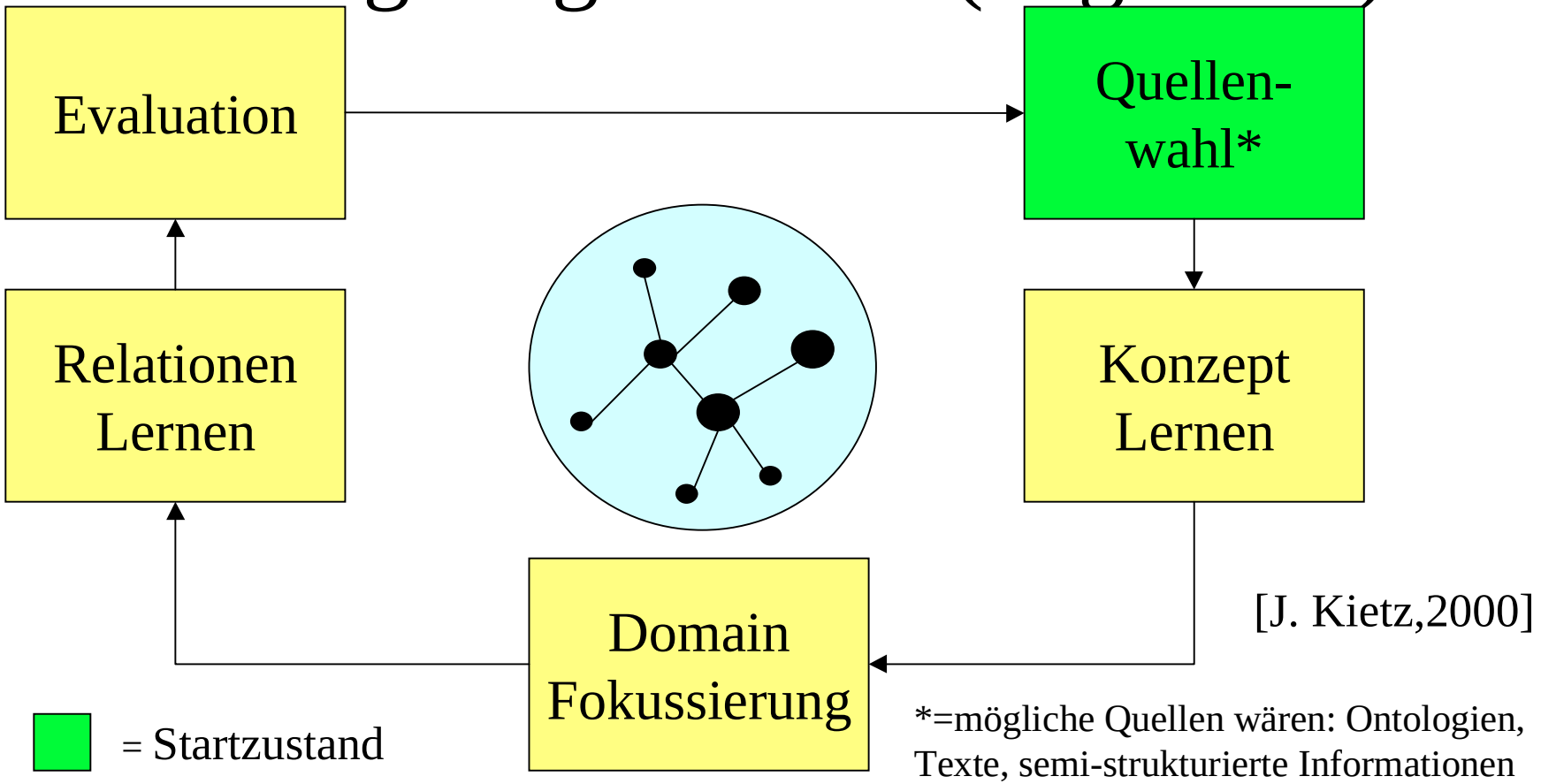


[nach S. Staab, 1999]

OWE - Klassifikationskriterien

- Autonomie-Grad
  - manuell (durchführbar, aber zeit-intensiv)
  - semi-automatisch (aktueller Stand)
  - automatisch (Zukunftsvision)
- Verwendete Methoden
- Verschiedene Verfahren
  - Bottom-up
  - Top-down
  - Merging & Mapping
- Eingabedaten (Strukturiertheit)
- Extraktion on-demand vs. Vorab-Extraktion

Semi-automatischer Ontologie-Aneignungs-Prozeß (allgemein)



[J. Kietz,2000]

*=mögliche Quellen wären: Ontologien, Texte, semi-strukturierte Informationen

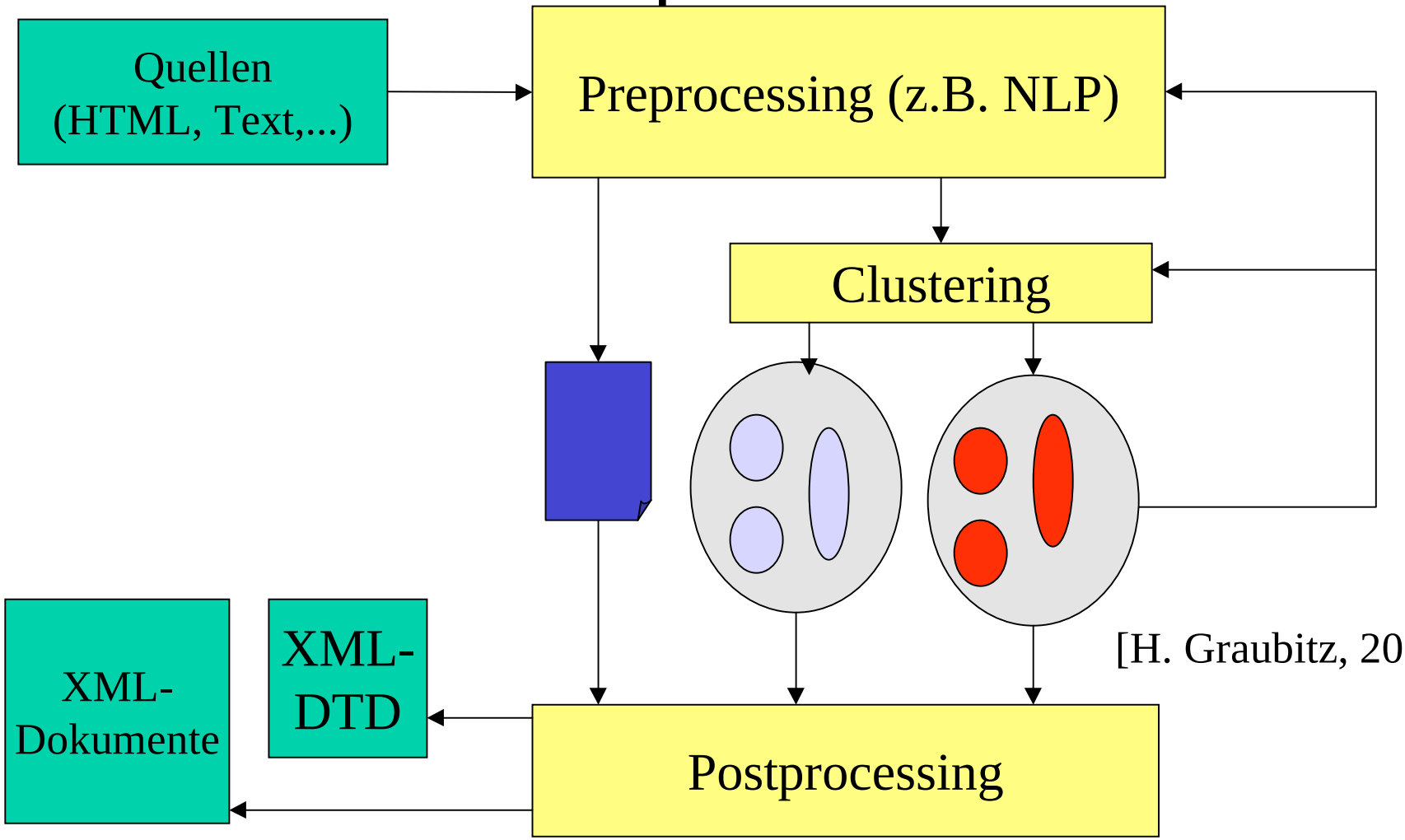
Methoden

- NLP: (z.B. SMES):
  - morphologische Analyse (Stamm)*
  - Semantik-Analyse
  - Erkennung benannter Entitäten*
  - Nutzung domain-spezifischer Informationen
- Text-Clustering:
  - Reduktion der Text-Dimension durch NLP
  - Clusterbildung (iterativ)
  - Klassifikation anhand der Cluster

Methoden

- Muster-Abgleich
- Induktive Verfahren
  - Erkennung/Klassifikation unbekannter Konzepte
  - Erkennung von Relationen zwischen Konzepten
- Inferenz (mit Description Logic)
- Statistik

KDT - Beispiel-Architektur



[H. Graubitz, 2001]

Vor- und Nachteile im Vergleich

- NLP
  + orientiert sich an Sprache, Lexika
  - viele Heuristiken, manuelle Regelerstellung
- Text-Clustering
  + iterative automatisierte Variante
  - Einschränkung auf eine Domain, Erklärbarkeit
- Muster-Abgleich
  + allgemein anwendbar
  - viele Heuristiken, manuelle Regelerstellung

Vor- und Nachteile im Vergleich

- Induktive Verfahren
  + Automatisierung
  - Erlernen der Regeln kompliziert
- Inferenz (mittels Description Logic)
  + Ableitung von weiteren Regeln durch Inferenz
  & unvollständige/fehlerhafte Daten sind nutzbar
  - verschiedene Standards
- Statistik
  + schnell, zuverlässig, bereits bekannt
  - manchmal absurde Ergebnisse, Verständlichkeit

Bewertung der Ansätze

Einzelne Anwendung einer Methode ist nicht optimal. Kombination mehrer Methoden, um die Stärken zu kombinieren und Nachteile einzelner Verfahren zu mildern.

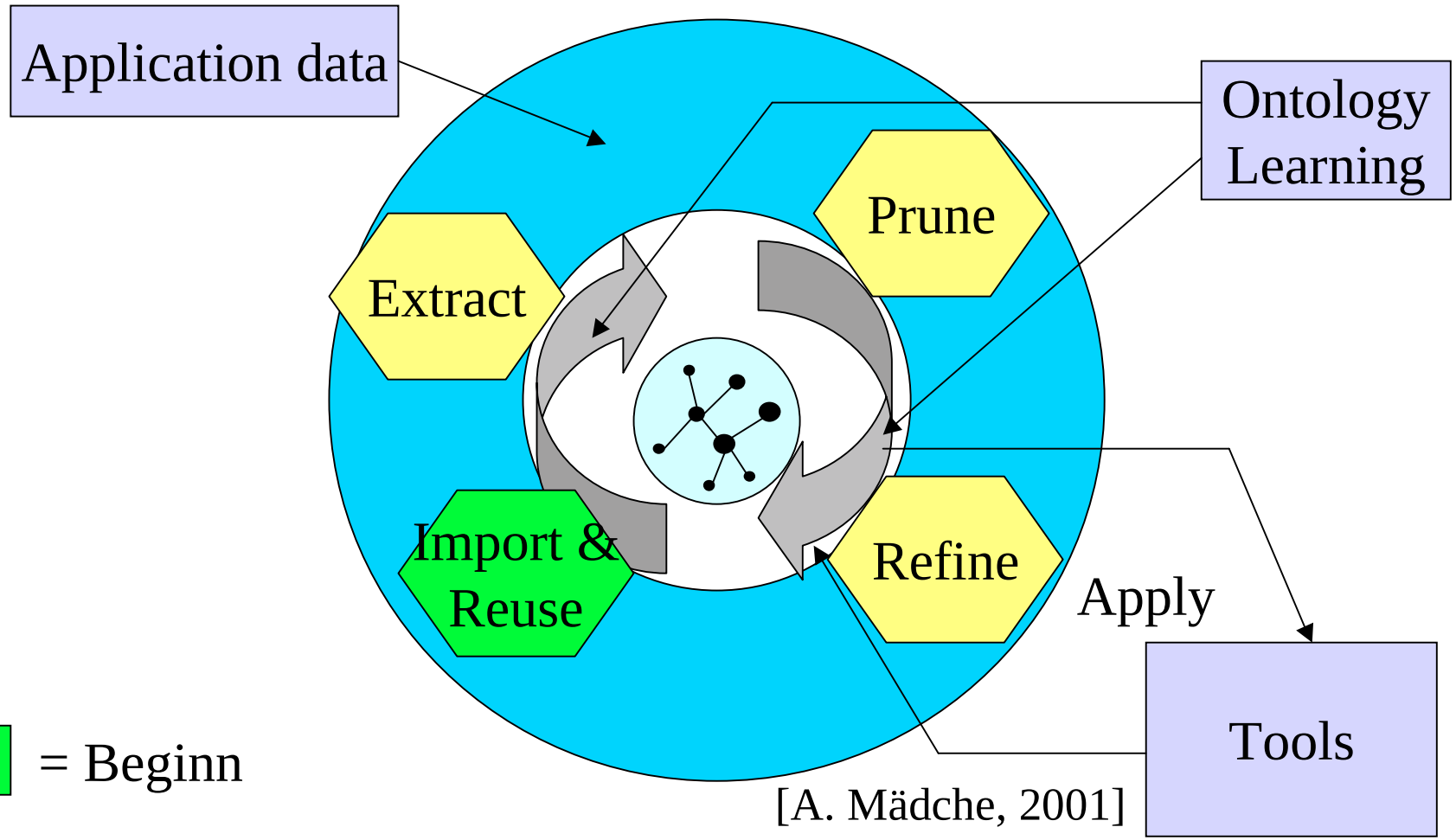
Kombination von Text-Clustering und NLP hat sich in einigen Situationen bewährt.

OWE - Verfahren

Prozeß-Schritte (allgemein):

- Import/Wiederverwendung/Konvertierung von Ontologien (optional)
- Extraktion von Daten (bottom-up; top-down)
- Pruning (Beschneidung)
- Refining (Veredelung)
- Verifikation/Evaluation

Ontologie-Lernen: Prozeß-Schritte



[A. Mädche, 2001]

OWE – Verfahren (Bottom-up)

Angefangen wird mit einem Datensatz, aus dem eine Ontologie erstellt wird, die die Daten strukturiert.

Genauer:

- Verwendung von zwei Text-Sammlungen (domain-spezifische vs. allgemeine)
- Statistische Erfassung (Wörter, Häufigkeit,...)

OWE – Verfahren (Bottom-up)

- Dimensionsreduktion (NLP, Stammbildung, ...)
- Erstellung eines domain-spezifischen Lexikons (Konzepte)
- Anwendung heuristischer Verfahren zur Relationserstellung (semantische Analyse)
- Pruning
- Refining

OWE – Verfahren (Top-down)

Anfangs hat man bereits eine allgemeine Ontologie, die dann im Verlauf durch bereichsbezogene Daten an den interessierenden Bereich angepaßt wird. (Domain-Fokussierung)

Genauer:

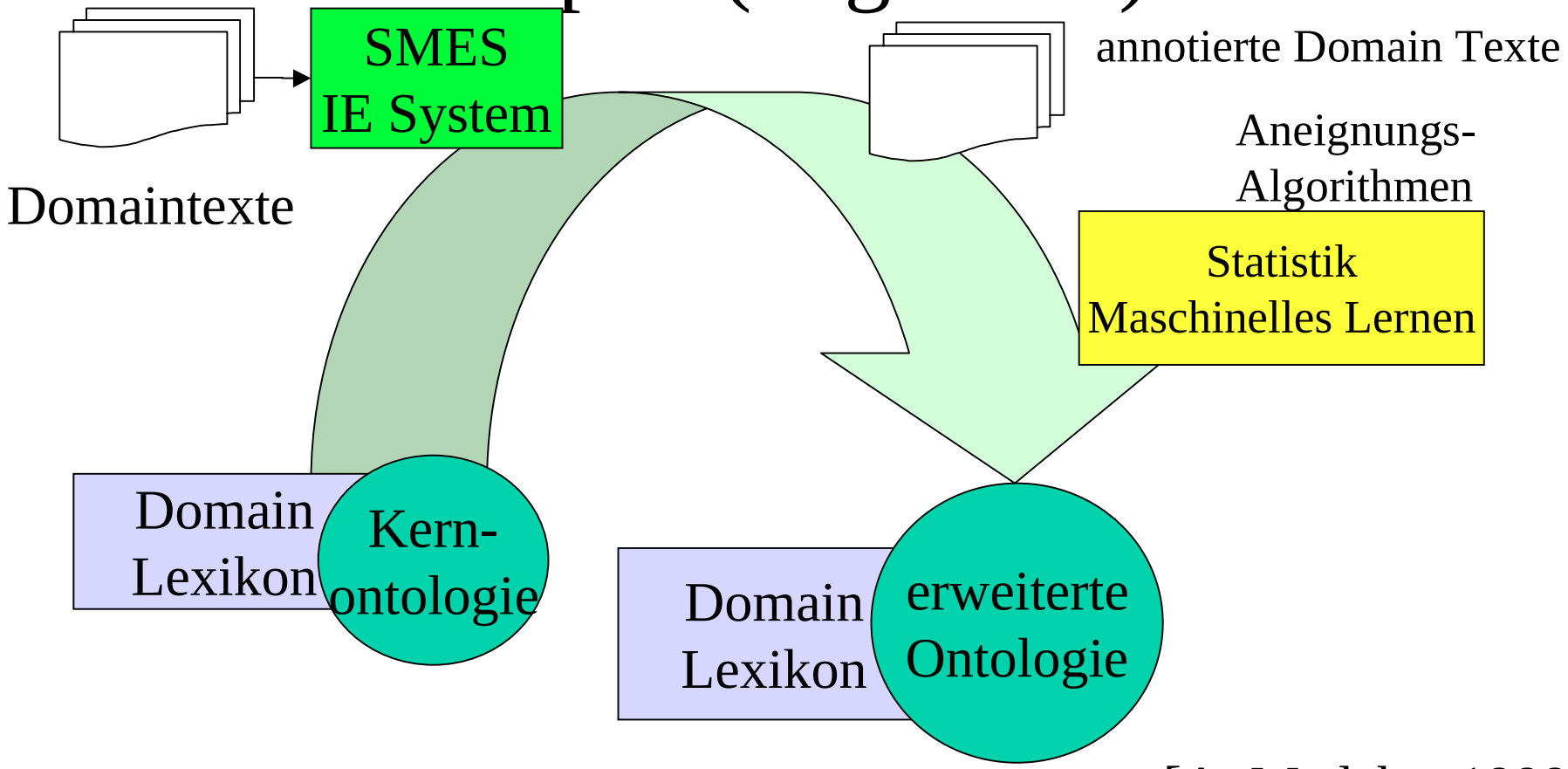
- Wahl einer (allgemeinen) Ontologie und domain-spezifischen Quellen (Import)

OWE – Verfahren (Top-down)

- Anwendung heuristischer Verfahren zur Konzept- und Relationsextraktion.
- Erweiterung der bestehenden Ontologie durch gefundene Konzepte und Relationen (Fokussierung)
- Pruning
- Refining

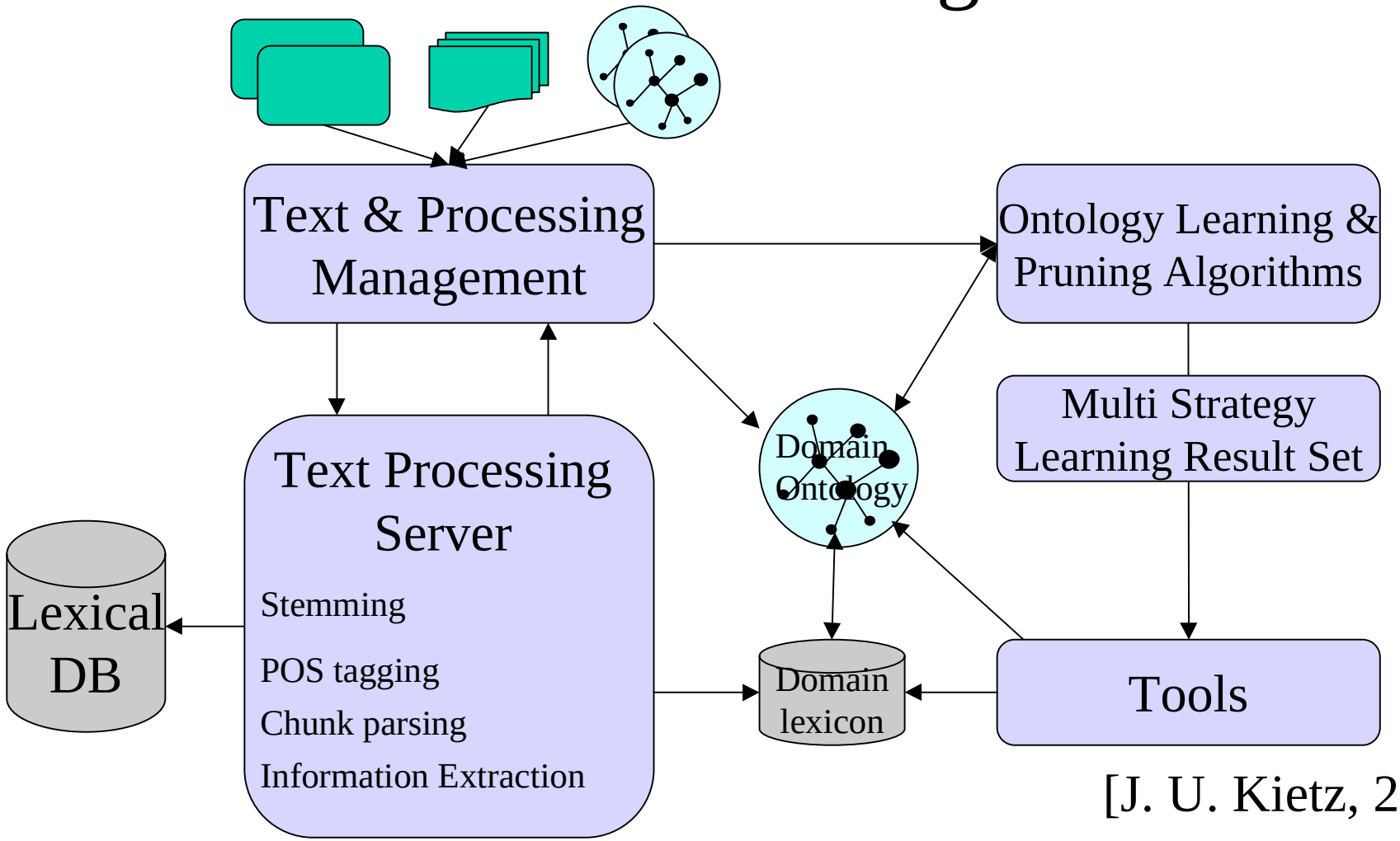
Wahl der Kern-Ontologie hat starke Auswirkungen

Wissensextraktion – „Top Down“-Beispiel (allgemein)



[A. Mädche, 1999]

Architektur für Ontologie – Lernen



[J. U. Kietz, 2000]

OWE - Verfahren

Merging:

Zusammenführung von Ontologien zur Konstruktion einer neuen Ontologie.

Mapping:

Erstellung von Regeln, die Entsprechungen aus den Ontologien zuordnen.

Praxis und Trends

- Vereinfachung in der Entwicklung fördert Verbreitung
- Verbesserung der Extraktionsfähigkeiten
- Automatisierung des kompletten Prozesses
- Steigende Integration und Verwendung von Ontologien in vielen Bereichen
- Semantic Web und Knowledge-Portale sind wichtige Gebiete

Schlußwort

- Hilfreiche Technik, die auf spezielle Bereiche zugeschnitten wird
- Anpassung an Aufgabenstellung durch Änderung der Ontologie
- Unterschiedliche Ansätze zur Extraktion
- Verschiedene Methoden aus vielen Bereichen (Maschinelles Lernen, Assoziationsregeln, Clustering,...). Profitiert aus Erfolgen aus jedem dieser Bereiche
- Mißbrauch
- Verkettung vieler Verfahren, Komplexität, Aufwand

Literaturangaben

[OBE98] D. W. Embley, D. M. Campbell, S. W. Liddle, R. D. Smith. *Ontology-Based Extraction and Structuring of Information from Data-Rich Unstructured Documents in CIKM'98.*

[OBI'99] A. Mädche, S. Staab, R. Studer. *Ontology-based Information Extraction and Integration in DGfS/CL'99.*

[SOAC] J.-U. Kietz, A. Mädche, R. Volz. *A Method for semi-automatic Ontology Acquisition from a corporate Intranet in EKAW'2000.*

[STDS] H. Graubitz, K. Winkler, M. Spiliopoulou. *Semantic Tagging of Domain-Specific Text Documents with DIAsDEM in DBFusion 2001*.

[OBTC] A. Hotho, S. Staab, A. Mädche.*Ontology-based Text-Clustering in IJCAI'2000.*

[LOSW] A. Mädche, S. Staab.*Learning Ontologies for the Semantic Web in ECML/PKDD2001*.

[DLOE] A. Todirascu. *Using Description Logics for Ontology Extraction in Ontology Learning 2000 at ECAI2000.*

☺ Danke! ☺