Leuchter, S., Reinert, F. & Schönbein, R. (2006). Gestaltung von semantischen Anwendungen: Repräsentation von ingenieurwissenschaftlicher Expertise. In: *Useware 2006. Nutzergerechte Gestaltung technischer Systeme. Tagung Düsseldorf, 10. und 11. Oktober 2006* (S. 21-30). Düsseldorf: VDI-Verlag (VDI-Berichte; 1946). http://www.safety-critical.de/doc/useware2006.pdf

Gestaltung von semantischen Anwendungen

Clusteranalyse zur ontologischen Repräsentation von ingenieurwissenschaftlicher Expertise

S. Leuchter, F. Reinert, R. Schönbein, Karlsruhe

Kurzfassung

Semantische Anwendungen verwenden für die in ihnen gespeicherten und verarbeiteten Informationen eine explizite Repräsentation der zugrunde liegenden Konzepte und Relationen in Form einer Ontologie. Im Bereich der Repräsentation ingenieur-wissenschaftlicher Expertise wurde eine solche Ontologie neu entwickelt. Für die Konstruktion der Ontologie wurde die empirische Methode *Card Sorting Task* und zur Auswertung eine Clusteranalyse benutzt. Die Softwarearchitektur und Interaktionskonzepte des resultierenden Prototypen ExperOnto zur Expertensuche in virtuellen Organisationen werden vorgestellt.

Abstract

Semantic applications use an explicit representation of grounding concepts and relationships of information, which is stored and processed. This representation is formalised as an ontology. A new ontology was developed for representing knowledge about scientific expertise in engineering domains. The empirical method “card sorting task” was applied for the construction of this ontology. Results were analysed with cluster analyses. The ontology was implemented in the software prototype ExperOnto. This web-based application supports users to find experts in virtual enterprises. Software architecture and interaction design of ExperOnto are presented.

1. Problembeschreibung

Semantische Anwendungen sind Softwaresysteme, die die in ihnen gespeicherten und verarbeiteten Informationen explizit repräsentieren. Die Repräsentation wird verwendet, um Interoperabilität mit externen Systemen herzustellen und „smarte“ Assistenzfunktionen mit Methoden der symbolischen Künstlichen Intelligenz zu realisieren. Es gibt eine breite Palette von semantischen Anwendungen, die Informationen über Expertiseträger verarbeiten.

Eine Einigung auf ein gemeinsames Format z.B. in Form eines Datenbankschemas im Sinne der Interoperabilität auf der syntaktischen Ebene ist technisch leicht zu erreichen. Eine allgemein akzeptierte Standardisierung der Inhalte ist schwieriger. Zum einen muss festgelegt werden, was für Sachverhalte repräsentiert werden sollen, zum anderen müssen gemeinsame verbindlich zu benutzende Konzepte und die Abbildung auf die Bezeichner, auf die die Konzepte abgebildet werden, bestimmt werden.

Ein Anwendungsbeispiel ist das Zusammenstellen von interdisziplinären Projektteams aus Fachexperten in großen, verteilten Organisationen und virtuellen Unternehmen. Für derartige Anwendungsszenarien müssen Profile von Mitarbeitern organisationsübergreifend gespeichert werden. Auf der semantischen Ebene muss Fachkompetenz in Form von Expertise erfasst, repräsentiert und verarbeitet werden. Expertise ist das spezialisierte Wissen eines Experten über ein bestimmtes Fachgebiet. Das schließt sowohl deklaratives („was“) als auch prozedurales Wissen („wie“) ein. Beispiele für deklaratives Wissen sind Fakten, Konzepte, Beziehungen zwischen Konzepten, Regeln und einschränkende Bedingungen. Beispiele für prozedurales Wissen sind die Fähigkeit zu Schlussfolgerungen, geeignete Vorgehensweisen auszuwählen und Problemlösungsprinzipien anzuwenden.

Im Bereich der Ingenieurwissenschaften sind die Problemstellungen, die durch Expertenwissen gelöst werden, anwendungsorientiert. Daher muss sich Expertise hier sowohl auf methodisches Wissen („Fachwissen“) als auch auf Wissen über die Anwendungsdomäne („Bereichswissen“) beziehen. Ontologien beinhalten Taxonomien, die aufgaben- und nutzerorientiert nebeneinander existieren [7] und die Konzepte und Relationen zu einem semantischen Netzwerk verknüpfen.

2. Ontologische Wissensrepräsentation

Ontologien werden in der Informatik benutzt, um Konzepte und ihre Semantik zu repräsentieren. Sie enthalten Begriffe, Eigenschaften, Beziehungen und Regeln (s. Tabelle 1). Die Konzepte werden repräsentiert durch *Begriffe* und ihre *Eigenschaften*. Die *Beziehungen* zwischen den Begriffen sind weitere Eigenschaften, die Konzepte hinter den Begriffen untereinander in Relation setzen. *Regeln* können in Ontologien benutzt werden, um die Repräsentation der Extension eines Sachverhaltes einzusparen und durch die Intension, also eine Anweisung, die den Aufbau des Sachverhaltes beschreibt, zu ersetzen.

Tabelle 1: Beispiele für ontologische Bereiche

Begriffe:	Person, Experte, Softwareentwickler
Eigenschaften:	Person/Name, Experte/Fachgebiet, Experte/Expertise
Beziehungen:	Experte „ist eine“ Person, Softwareentwickler „acts_as“ Programmierer
Regeln:	„Alle Personen, die Softwareentwickler sind, sind Experten mit der Eigenschaft Experte/Expertise=Informationstechnologie.“

Das Ziel beim Aufstellen einer Ontologie ist, das Wissen eines Anwendungsgebietes explizit herauszuarbeiten und zu repräsentieren, damit es automatisch verarbeitet werden kann. Auch implizites Wissen (z.B. Arbeitsabläufe) muss explizit modelliert werden. Daher enthalten Ontologien sowohl deklaratives Wissen, als auch prozedurales Wissen.
Formal ist eine Ontologie O ein Zeichensystem bestehend aus: O := (L, C, R, F, G, H, A) [8]. Mit

- **L** - Lexikon mit Begriffsdefinitionen:
  z.B. „Person“, „Experte“, „Softwareentwickler“
- **C** - Menge von Konzepten:
  abstrakte Konzepte und ihre Eigenschaften
- **R -** Menge von Relationen:
  stellen Beziehungen der Konzepte zueinander dar
- **F** - Abbildungsrelation von L→ C
  legt die Benennung der Konzepte fest
- **G** - Abbildungsrelation von L→ R
  legt die Benennung der Relationen fest
- **H** - Taxonomien zur Abbildung von Hierarchien,
  z.B. Softwareentwickler → Experte → Person
- **A** - Satz von Axiomen oder Regeln
  drücken mittels einer formalen Logik Beziehungen zwischen Konzepten und Relationen aus

2.1 Repräsentation ingenieurwissenschaftlicher Expertise

Für die Anwendung zur Expertensuche wurde ein Modell zur Repräsentation ingenieurwissenschaftlicher Expertise, in der Form einer Top-Level Ontologie, die leicht

erweitert bzw. angepasst werden kann, entwickelt. Die Ontologie besteht gegenwärtig aus 480 Konzepten und 190 Relationen.

In der Ontologie wird Wissen über Personen in ihrer Eigenschaft als Träger von Expertise repräsentiert. Dazu wurde ein geeignetes Modell der Akteure, Situationen, Abläufe und Fakten erstellt. Die Auswahl von Projektteammitgliedern stützt sich auf Informationen zu: Mitarbeit an Projekten, eingesetzten Werkzeugen, Methodiken, Technologien und Modellen sowie die räumliche und zeitliche Verfügbarkeit, die Zugehörigkeit zu und der Rang in einer Organisationseinheit. In dieser neuen Ontologie wurden dazu die folgenden Konzepte erfasst: (Arbeits-) Methodiken, Modelle, Werkzeuge, Projekte, Raum, Zeit, Verfügbarkeit (Konzept Kalender, Arbeitszeit etc.), Organisation, Person (Experte), Kontaktmöglichkeiten. Bild 1 zeigt einen Ausschnitt aus der Ontologie mit den wesentlichen Verknüpfungen zwischen den zentralen Konzepten zur Expertise.

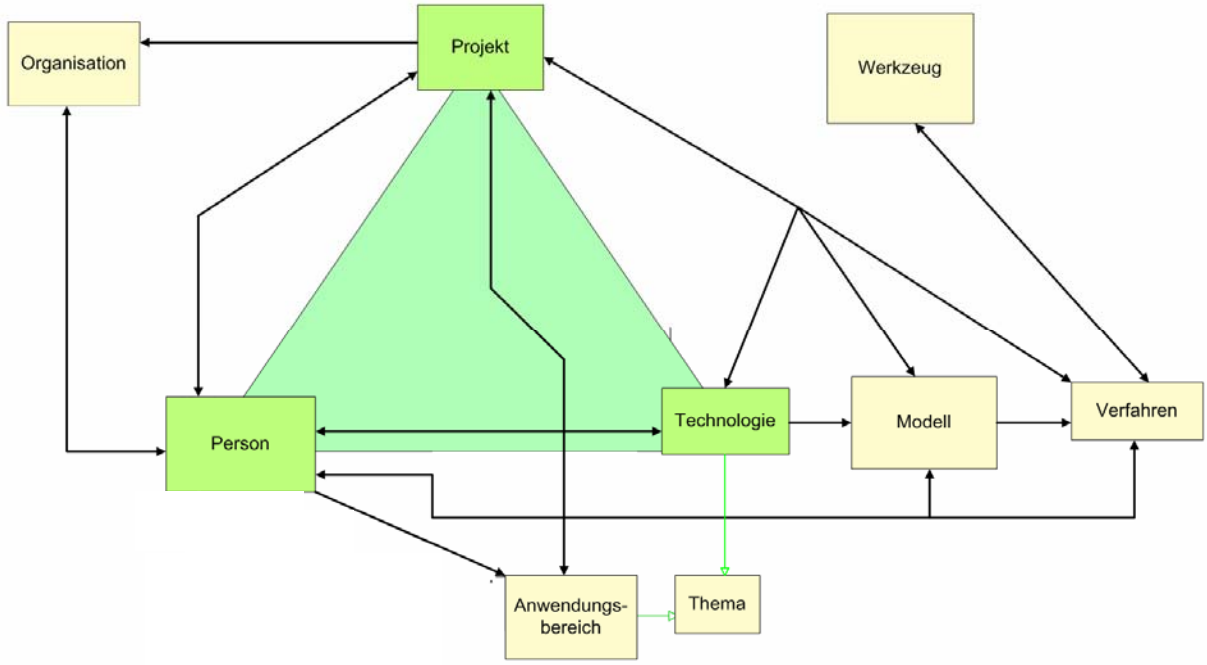

Bild 1: Ausschnitt aus der Ontologie zur Repräsentation ingenieurwissenschaftlicher Expertise

2.2 Kooperative Ontologieentwicklung mittels Clusteranalyse

Ein wesentlicher Bestandteil der Ontologie zur Expertensuche ist die Repräsentation von Expertise. Die dazu benötigten Konzepte zu Technologien und Anwendungsbereichen

wurden analytisch durch die Auswertung von Dokumenten und empirisch durch eine Befragung von Domänen-Experten erhoben.

Die einzelnen Themenbereiche Technologie, Modell, Verfahren etc. müssen weiter strukturiert werden. Dazu werden Gruppen gebildet. Die Begriffe innerhalb einer Gruppe sollen möglichst ähnlich und von ihrer Bedeutung verwandt sein. Zwischen den Gruppen sollen dagegen möglichst große Unterschiede sein. Es ist schwierig ein externes Ähnlichkeitsmaß für den Vergleich zwischen den Begriffen zu finden. Deshalb haben wir eine empirische Herangehensweise zur Strukturierung der Konzepte gewählt und an einem Sub-Set von 100 Begriffen aus informationstechnischen Feldern, die am Fraunhofer IITB bearbeitet werden, erprobt. Versuchspersonen sollten die Begriffe in max. 10 Gruppen aufteilen. Vorgaben über Sortierkriterien oder die Größe der Gruppen wurden nicht gemacht. Die resultierenden Gruppen brauchten nicht benannt zu werden. Zur praktischen Durchführung wurde das Werkzeug CardSort [6] verwendet. Bild 2 zeigt den Einsatz mit einem fiktiven Satz von Begriffen. Initial sind alle Begriffe im Bereich Cardset am linken Rand des Fensters auf Karten angezeigt. Im rechten Bereich gibt es Kartenstapel, die unter den roten Oberkarten angelegt werden können. Im Verlauf können die Karten mit der Maus überall zwischen Kartenstapeln und dem Bereich Cardset verschoben werden. Am Ende kann ein Protokoll mit den endgültig erzeugten Stapeln ausgegeben werden.

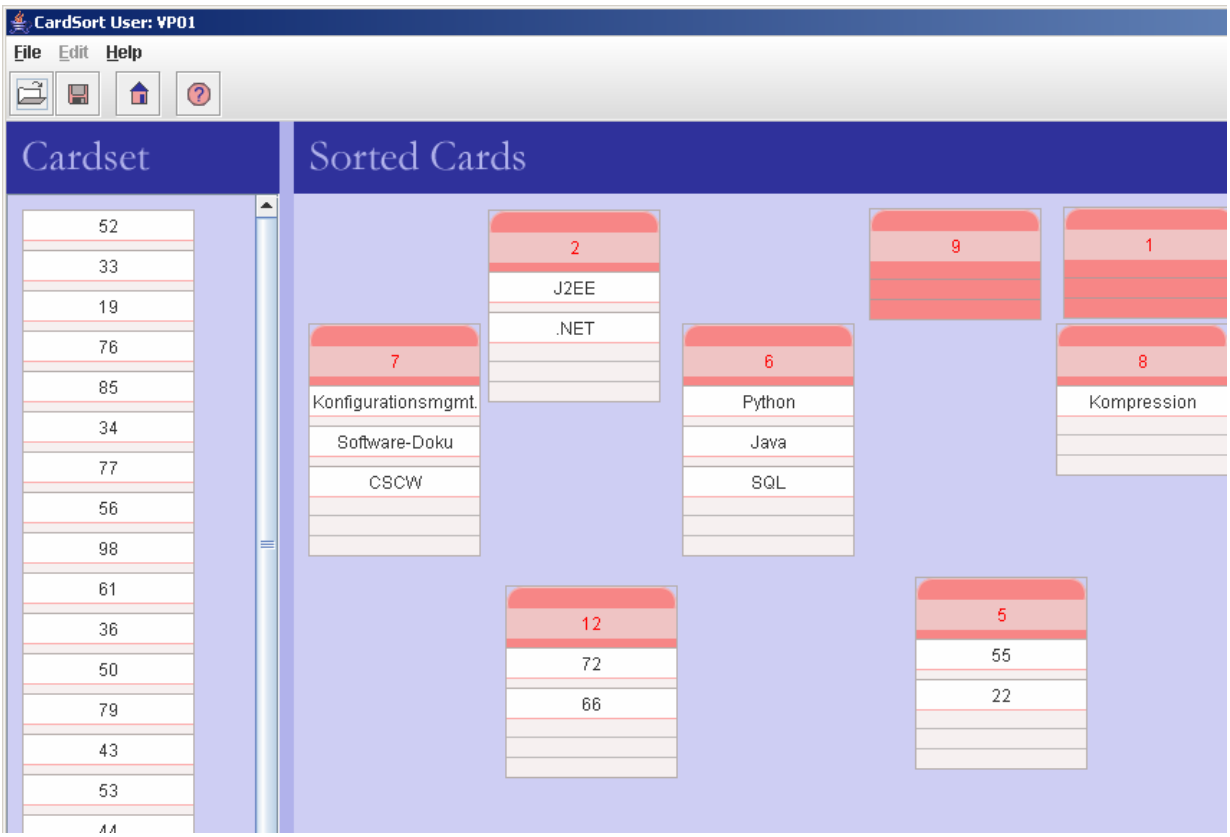

Bild 2: CardSort im Einsatz zur Clusterung der Konzeptbegriffe

Tabelle 2: Distanzmatrix zur Clusteranalyse

d(x,y)	SQL	J2EE	Ontologie
SQL	1	0,5	0,2
J2EE	0,5	1	0
Ontologie	0,2	0	1

Um einen „Mittelwert“ der Clusterung zu berechnen, muss zuerst die Distanzmatrix aufgestellt werden. Hier wird für alle zu clusternden Konzepte aufgetragen, in welchem Anteil aller Fälle zwei Konzepte im selben Cluster waren. Im Beispiel in Tabelle 2 waren „J2EE“ und „SQL“ bei 50% aller Versuchspersonen im selben Cluster, während „Ontologie“ und „SQL“ nur bei 20% der Versuchspersonen im selben Cluster waren.

Um aus der Distanzmatrix Gruppen zu bilden, gibt es mehrere Verfahren aus der Clusteranalyse, die verwendet werden können. Wir verwenden *agglomerative hierarchical Clustering:*

1. Initial wird jedes Konzept einem eigenen Cluster zugeordnet.
2. Alle paar-weisen Distanzen zwischen den Clustern werden berechnet (s.u.).
3. Das Cluster-Paar mit der geringsten Distanz wird zu einem neuen Cluster zusammengeführt.
4. Weiter bei Schritt 2 oder
   Abbruch, wenn die gewünschte Anzahl der Cluster erreicht ist oder eine maximal tolerierbare Distanz zwischen den verbleibenden ähnlichsten Clustern erreicht ist.

Zur Berechnung der paar-weisen Distanzen zwischen den Clustern wird die Distanzmatrix (s.o.) benutzt. Es gibt drei unterschiedliche Maße für die paarweise Distanz zwischen zwei Clustern:

- *Average Linkage*: Die Distanz zwischen zwei Clustern ist der Durchschnitt der Distanzen zwischen allen Punkten in den Clustern.
- *Single Linkage*: Die Distanz zwischen zwei Clustern ist die kleinste Distanz zwischen Elementen der beiden Cluster.
- *Complete Linkage*: Die Distanz zwischen zwei Clustern ist die größte Distanz zwischen Elementen der beiden Cluster.

Zur Ableitung der Taxonomie von Technologien haben wir das Distanzmaß *complete linkage* gewählt, weil so Cluster nach „worst case“ Ähnlichkeit gebildet werden. Das Werkzeug EZCalc [1] von IBM liefert *Tree Plots* zur Clusteranalyse (Prinzip s. Bild 3). Mit zwei verschiebbaren Balken können interaktiv die Grenzen für maximale und minimale Ähnlichkeit verschoben werden. Zur Berechnung des Distanzmaßes können *average*, *single* und *complete Linkage* verwendet werden.

Mit der Methode der Clusteranalyse konnten wir eine „durchschnittliche“ Taxonomie erstellen. Eine genauere Analyse muss jedoch noch erfolgen, um zu ermitteln, ob es mehrere in sich homogene, aber gegeneinander unterschiedlich clusternde Gruppen gibt. Fiktive Beispiele für solche Gruppen können „technische Informatiker“ vs. „Medieninformatiker“ und „um 1980 ausgebildete“ vs. „2005 ausgebildete Informatiker“ sein. Wenn es mehrere solche Gruppen gibt, müssten Sichten in der Ontologie verankert werden, die einen gruppenspezifischen Zugang in Form eigener Taxonomien ermöglichen.

Eine Strukturierung von Konzepten kann auch in Workshops mit Vertretern der Zielgruppe erfolgen. Dieses Vorgehen bietet den Vorteil, dass Benutzer durch die Einbeziehung in den Entstehensprozess „abgeholt“ werden und ein gruppengemeinsames Verständnis für die entwickelte Struktur entwickeln. Wenn das Ziel jedoch ist, eine Ontologie für eine sehr große Benutzergruppe zu entwickeln, bietet die objektivierende Methode der statistischen Auswertung mit der Clusteranalyse den Vorteil, dass eine akzeptable Lösung vorgegeben wird, an die sich die Benutzer zwar anpassen müssen, die aber nicht allzu weit von deren individuellen mentalen Modellen entfernt ist.

3. Anwendung in prototypischer Implementierung

Die Repräsentation von ingenieurwissenschaftlicher Expertise wurde im System „ExperOnto“, einer Entwicklung des Fraunhofer IITB, implementiert. ExperOnto ist eine J2EE-Anwendung. Bild 4 zeigt den prinzipiellen Aufbau auf der Basis der Entwicklungsplattform WebGenesis [2]. Die Wissensrepräsentation erfolgt in OWL [3] mit dem Werkzeug Protege [5].

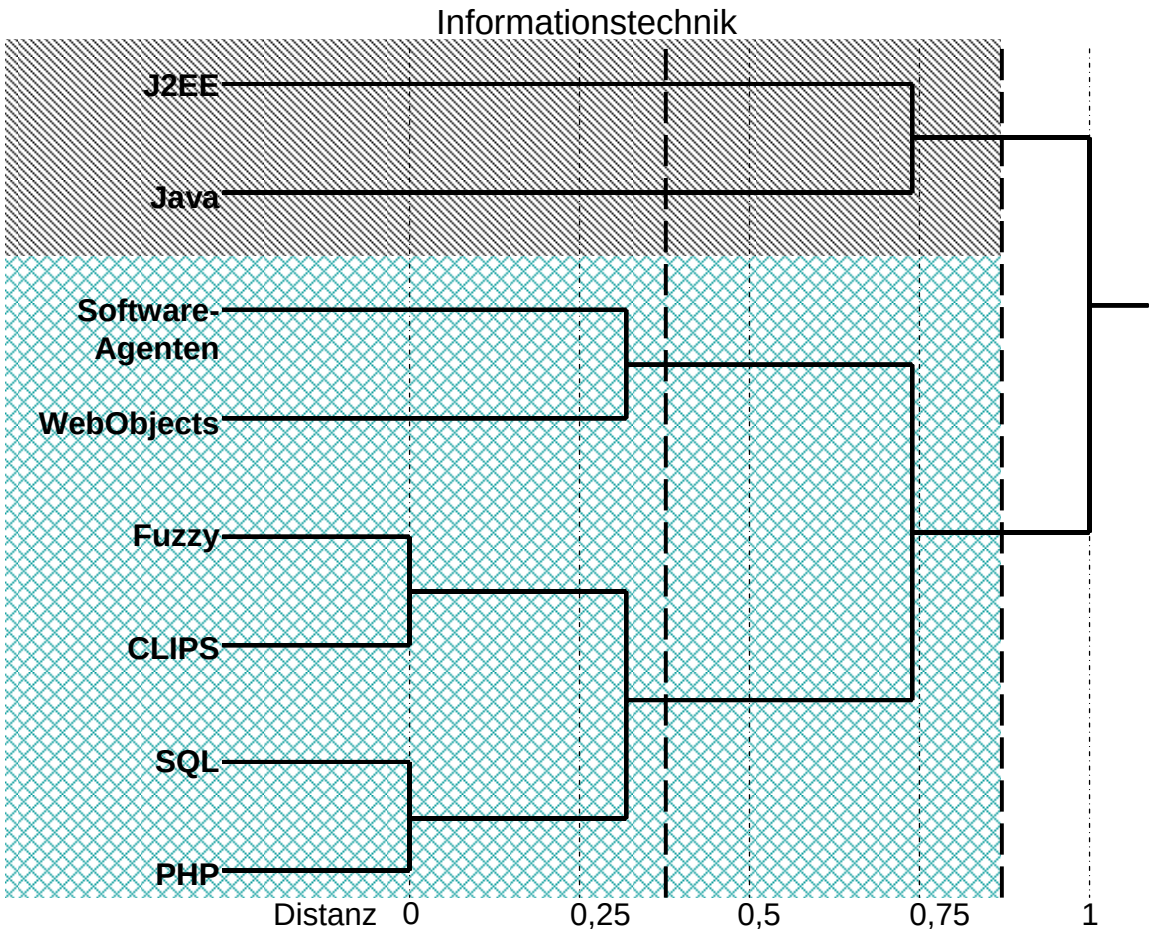

Bild 3: Ergebnis der Clusteranalyse: Zwei Cluster im Matrix Tree Plot

3.1 Grad der Expertise

Mit den Konzepten der Ontologie werden einzelne Personen mit Technologien verknüpft. Der Grad der Expertise ist damit aber noch nicht erfasst. Dazu wurden verschiedene Heuristiken entwickelt, die als Fuzzy-Funktionen realisiert sind. Sie basieren auf einem Modell, das Projekte und fachspezifische Veröffentlichungen berücksichtigt und Personen zuordnet. Projekte und Veröffentlichungen wiederum beziehen sich auf Anwendungsgebiete, Technologien, Methoden, Verfahren und Modelle. Die Heuristiken nutzen die Anzahl und Dauer der zu einer Anfrage passenden Projekte. Auch die Aktualität der fachspezifischen Erfahrungen, also wie lange Projekte bzw. Veröffentlichungen zurückliegen, wird in der Berechnung berücksichtigt.

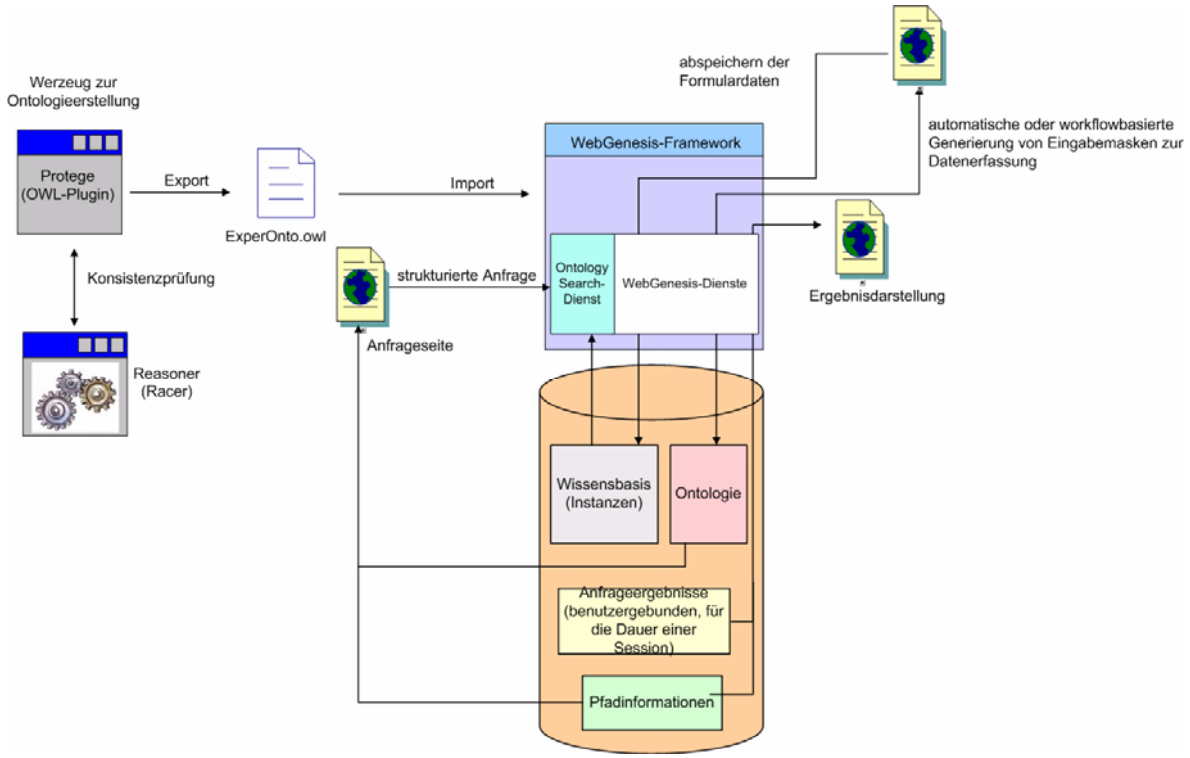

Bild 4: Prinzipieller Aufbau des ExperOnto-Prototypen

3.2 Anfragen

In ExperOnto können administrativ festgelegte „Super-User“ über eine webgestützte Benutzungsschnittstelle generische Anfragen anhand der Relationen und Konzepte in der Expertenontologie erstellen. Die generischen Anfragen werden gespeichert und stehen allen Benutzern zur Verfügung. Wenn sie ausgeführt werden, werden generalisierte durch spezialisierte Konzepte bzw. ihre Instanzen ersetzt. Um die Anfrage auszuführen, wählt der Benutzer aus einer Baumansicht die entsprechenden Elemente der Ontologie aus.

Die von Super-Usern vorgefertigten Suchanfragen werden in ExperOnto aus zwei Gründen angeboten: Eine genaue Kenntnis der Ontologie ist erforderlich, um den Informationsbedarf effektiv in einer formalen Anfrage abzubilden. Außerdem würde der freie Zugriff auf alle repräsentierten Informationen eine möglicherweise datenschutz- oder arbeitsrechtlich unzulässige Verknüpfung von Einzelinformationen zu einer verdichteten Aussage wie bei der Rasterfahndung ermöglichen.

4. Ausblick

Zur Optimierung wird das System gegenwärtig um eine resolutionsartige Verarbeitung erweitert.

Die vorgestellte Repräsentation von ingenieurwissenschaftlicher Expertise ist eine Möglichkeit, Experten als Expertiseträger und Anbieter „nicht-automatisierter Dienste“ in ihrem Fachbereich darzustellen. Diese Repräsentation wird in Netzwerken semantisch interoperabler Systeme benötigt, um Informations- und Diensträume zu integrieren. Semantic Web und Semantic Grid sind Ausprägungen solcher Netzwerke [4, 7].

Durch eine zu den Expertise-Modellen kompatible semantische Beschreibung von automatischen Diensten („Services“) ist darüber hinaus eine aufgabenbezogene Vergleichbarkeit von automatischen und „Experten-Diensten“ aufbauend auf einem integrierten Kosten-Nutzen-Modell möglich. Diese Eigenschaft ermöglicht die „on demand“ Formierung virtueller Organisationen.

5. Literatur

[1] Dong, J., Martin, S., & Waldo, P. (2000): *A User Input and Analysis Tool for Information Architecture*. In *Extended Abstracts of the ACM Conference on Human Factors in Computing System – CHI 2001*. Online-Dokument http://www.stcsig.org/usability/topics/articles/EZSortPaper.pdf (letzter Zugriff: 24.07.2006).

[2] IITB (2006): *WebGenesis. Ein System für Content-, Wissens-, und CommunityManagement*. Online-Dokument http://www.iitb.fraunhofer.de/servlet/is/279/WebGenesis-Produktblatt.pdf (letzter Zugriff: 24.07.2006).

[3] McGuinness, D.L. & van Harmelen, F. (Hrsg.) (2004): *OWL Web Ontology Language Overview. W3C Recommendation 10 February 2004*. Online-Dokument http://www.w3.org/TR/2004/REC-owl-features-20040210/ (letzter Zugriff 24.07.2006).

[4] De Roure, D., Jennings, N.R. & Shadbolt, N.R. (2005): The Semantic Grid: Past, Present, and Future. *Proceedings of the IEEE, 93*, (3), 669- 681.

[5] Noy, N.F., Sintek, M., Decker, S., Crubezy, M., Fergerson, R.W. & Musen, M.A. (2001): Creating Semantic Web Contents with Protege-2000. *IEEE Intelligent Systems, 16* (2), 60-71.

[6] Schilb, S. (2003): *Konzeption und Implementierung eines Card Sorting Tools*. Diplomarbeit, FH Kaiserslautern.

[7] Schönbein, R. (2006): *Agenten- und ontologiebasierte Software-Architektur zur interaktiven Bildauswertung* (Dissertation, Universität Karlsruhe, Fakultät für

Informatik). Karlsruhe: Universitätsverlag. Online-Dokument
http://www.uvka.de/univerlag/volltexte/2006/99/ (letzter Zugriff 24.07.2006).

[8] Staab, S. & Studer, R. (2004): *Handbook on Ontologies*. Heidelberg: Springer Verlag.