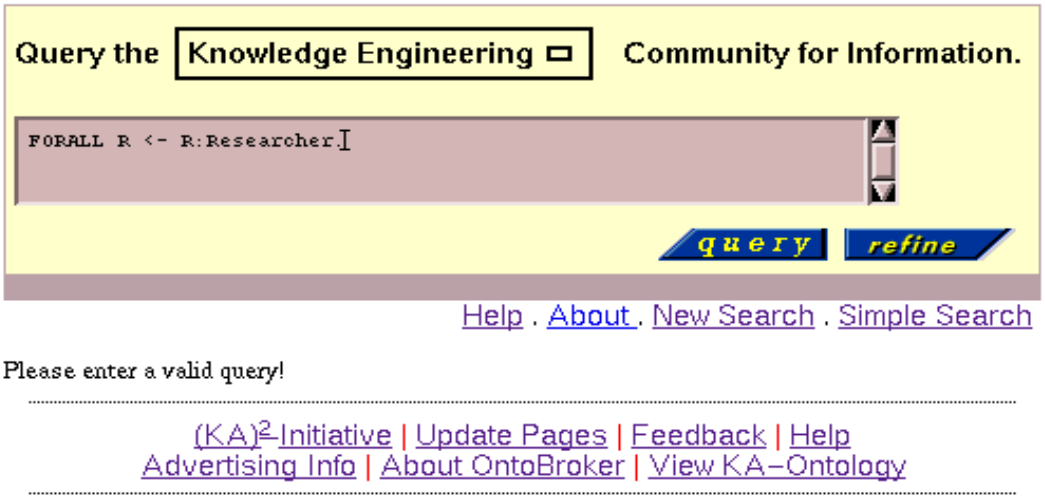


**Fig. 2** Query interface of the Ontobroker.

search metaphor of SHOE to the capability to express complex inferences using the knowledge as it is provided by the web. The ontological formalism used by SHOE is rather limited in regard to this purpose.

Currently, there is clear trend for content descriptions of web documents. Some browser provider would like to use content descriptions for improved presentations (cf. Apple' s Meta-Content Format) and several automatic search services would like to exclude documents from retrieval. Also, several brokering services use complex indexes to guide the search process. Using such keyword indexes for describing the content of a page is a clear step in our direction. Ontobroker provides two main surpluses: it allow to access pieces of a HTML page and does not tread a HTML document as an atomic unit and it provides ontologies that are a very rich representation formalism.

## 4 References

[Gru93] T. R. Gruber: A Translation Approach to Portable Ontology Specifications, *Knowledge Acquisition*, 5(2), 1993.

[KLW95] M. Kifer, G. Lausen, and J. Wu: Logical Foundations of Object-Oriented and Frame-Based Languages, *Journal of the ACM*, vol 42, 1995.

[LSR96] S. Luke, L. Spector, and D. Rager: Ontology-Based Knowledge Discovery on the World-Wide Web. Proceedings of the *Workshop on Internet-based Information Systems* at the *AAAI-96*, Portland, Oregon, 1996.

[LSR+97] S. Luke, L. Spector, D. Rager, and J. Hendler: Ontology-based Web Agents. In P*roceedings of First International Conference on Autonomous Agents*, 1997.

[LT84] J. W. Lloyd and R. W: Topor: Making Prolog more Expressive, *Journal of Logic Programming*, 3:225-—240, 1984.

[Mau97] M. L. Mauldin: Lycos: Design Choices in an Internet Search Engine, *IEEE Expert*, January-February 1997. http://www.lycos.com.

[SeE97] E. Selberg and O. Etzioni: The MetaCrawler Architecture for Resource Aggregation on the Web, *IEEE Expert*, January-February 1997. http://www.metacrawler.com.

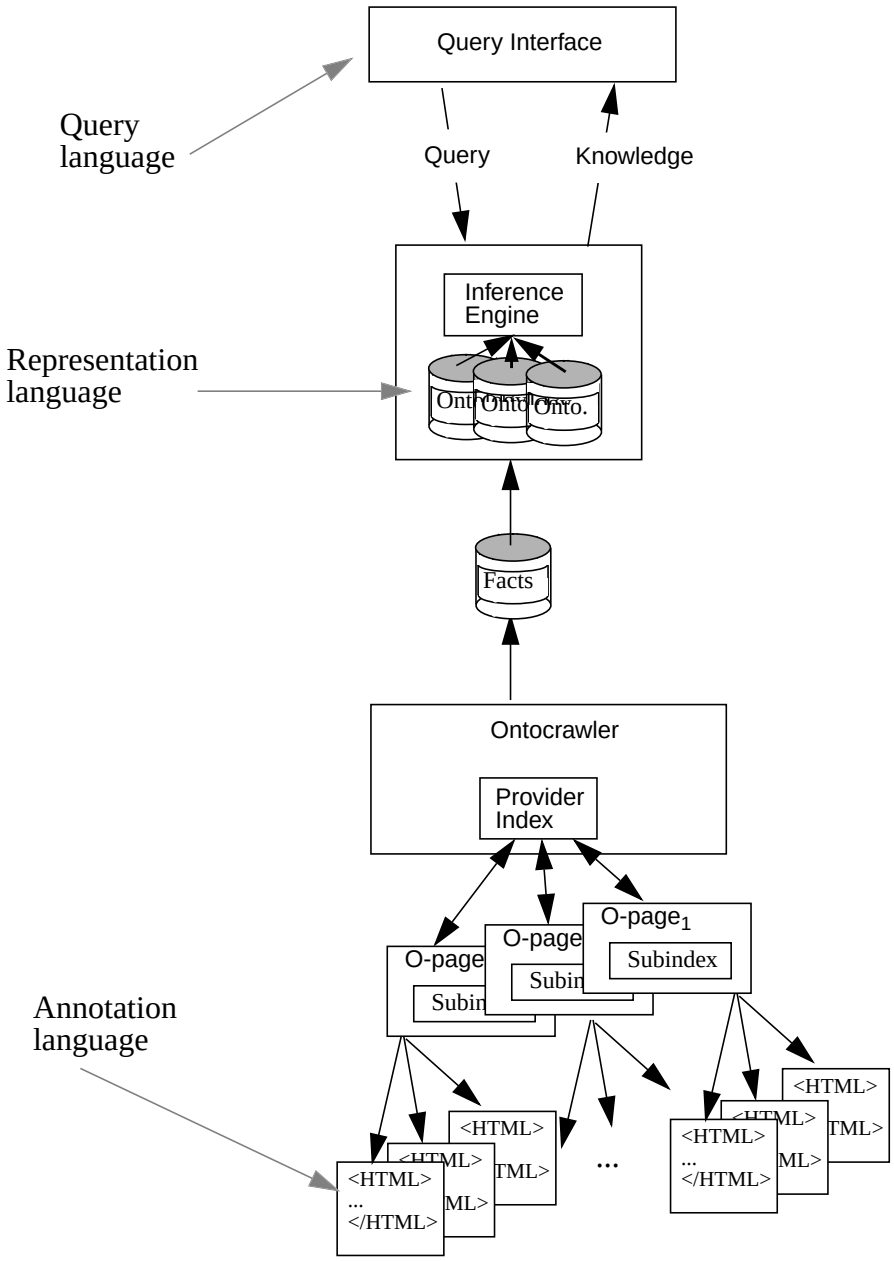


**Fig. 1** The architecture of the Ontobroker

may not know the ontological terms that he must use in a query and the web crawler may miss knowledge chunks because it cannot parse the entire WWW. In SHOE, ontologies are proposed as gradual improvement of the competence of global search engines in the WWW. If the user knows for some reasons parts of the ontology (like he has to know the right key words) and if the search engines knows for some reasons the appropriate URLs (for example, by executing keyword search on ontological terms) it can be used for a semantically guided search through the web. We present a much more restricted approach because our approach is suitable for homogeneous intranets and subnets of the WWW created by a community that agree upon a common ontology. As a consequence we can provide the entire ontology used for annotation to the questioner and we can deliver complete answers. Finally, we extend the

joint ontology.

## 2 The Ontobroker

A couple of tools is necessary to realise KBWs through the use of ontologies. The general architecture of the ontology-based brokering service *Ontobroker* is shown in Figure 1. It consists of three main elements: a query interface, an inference engine, and a webcrawler (called Ontocrawler). Each of these elements is accompanied with a formalization language: the query language for formulating queries, the representation language for specifying ontologies, and the annotation language for annotating web documents with ontological information.

**Query interface.** The broker has to communicate with clients who ask for some knowledge using web browsers like Netscape and Explorer. The query interface of Ontobroker is realised by a couple of active HTML pages and cgi-scripts that are executed by the browser of the client. The query language is a subset of the representation language customized for formulating queries.

**Inference engine.** The inference engine receives the query of a client and uses two information sources for deriving an answer: It uses the ontology chosen by the clients and it uses the facts that were found by the Ontocrawler in the WWW. The basic inference mechanism of the inference engine is the derivation of a minimal model of a set of Horn clauses. However, the language for representing ontologies is syntactical enriched. First, ideas of [LT84] were used to get rid of some of the limitation of Horn logic without requiring a new inference mechanism. Second, languages with richer epistemological primitives than predicate logic are provided. Currently Frame logic [KLW95] is used as representation languages for ontologies. It incooperates objects, relations, attributes, classes, and is-subclass-of and is-element-of relationships within a first-order semantic framework. To improve the accessibility of our service we are currently realizing translators for KIF, Ontolingua and LOOM.

**Ontocrawler.** First, Ontocrawler searches through a fragment of the WWW that makes use of one of the ontologies and collects these knowledge fragments. Second, it realises a wrapper that translated annotated web documents into facts formulated in the representation language. Neither the inference engine nor the query client have to be aware of the syntactical way, the facts are represented in the web. The Ontocrawler provides this abstraction mechanism. Only a knowledge *provider* has to use the annotation language. Each provider of an ontological annotated knowledge chunk has to register a provider index and he has to use the annotation language and an ontology of the Ontobroker to annotate his bits of knowledge.

## 3 Related Work

The idea of using ontologies to annotate information in the WWW is part of the SHOE-approach [LSR96], [LSR+97]. HTML pages are annotated via ontologies to support information retrieval based on semantic information. However, there is a main differences in the underlying philosophy. Providers of information in SHOE can introduce arbitrary extensions of ontologies and no central provider index is defined. As a consequence, the client

# Ontobroker:
# Transforming the WWW into a Knowledge Base

(Extended Abstract, 1515 words)

## 1 Introduction

The World Wide Web (WWW) could be seen as the largest knowledge-based system that has ever existed. It contains huge amounts of knowledge about any subject one can think of. HTML documents enriched by multimedia provide knowledge in different representations (i.e., text, graphics, moving pictures, video, sound, virtual reality, etc.). Hypertext links between web documents represent relationships between different knowledge entities. Based on the HTML standard, browsers are available that present the material to humans. Browsers can use the HTML-links to browse through distributed information and knowledge units. However, taking *the metaphor of a knowledge base* as a way to look at the WWW brings the big bottleneck of the web into mind. Its support in automated inference is very limited. Deriving new knowledge from existing knowledge is hardly supported. Actually, the main inference services the web provides are keyword-based search facilities realized by different search engines, web crawlers, web indices, man-made web catalogues etc. (see [Mau97], [SeE97]). Given a keyword, such an engine collects a set of knowledge chunks from the web that use this keyword. This limited inference access to existing knowledge stems from the fact that there exist only two main types of standardization for knowledge representation in the web. The HTML standard is used to present knowledge in a (browser and) human-readable way and to define links between different knowledge units and mainly the English language is used to represent the knowledge units.

Deriving automatically semantic information from sentences in natural language is still an unsolved problem. Inference by keyword search may deliver some results but it also results in a lot of unrelated information and at the same time it may miss a lot of important information. [LSR96] and [LSR+97] propose *ontologies* to improve the automatic inference support of the knowledge base WWW. An ontology provides “an explicit specification of a conceptualization“ [Gru93]. Ontologies are discussed in the literature as means to support knowledge sharing and reuse. This approach to reuse is based on the assumption that if a modelling schema - i.e. an ontology - is explicitly specified and agreed upon by a number of agents, it is then possible for them to share and reuse knowledge. Standardizing the syntactical way in which semantic information is presented allows the automatic derivation of semantic information via syntactical manipulation and creates what we will call a *knowledge-based WWW (KBW)*.

Clearly, we cannot expect that ontologies will be used by any web user and even if everybody would use ontologies to annotate his web pages it will be hardly ever possible to negotiate on a worldwide-used standard for representing knowledge about all possible subjects. Therefore, we use the *metaphor of a newsgroup* in to define the role of such an ontology. It is used by a group of people that share a common subject and a related point of view on this subject. Thus it allows them to annotate their knowledge to enable automatic inference based on the shared ontology. We coined the term *Ontogroup* to refer to such a group of web users that agree on a