On-To-Knowledge: Ontology-based Tools for Knowledge Management

Dieter Fensel, Frank van Harmelen, Michel Klein, Hans Akkermans
Free University Amsterdam VUA, Division of Mathematics and Informatics
De Boelelaan 1081a, NL-1081 HV Amsterdam, The Netherlands
http://www.ontoknowledge.org ; contact: dieter@cs.vu.nl
and
Jeen Broekstra, Christiaan Fluit, Jos van der Meer, *AIdministrator, The Netherlands*
Hans-Peter Schnurr, Rudi Studer, *AIFB, University of Karlsruhe, Germany*
John Hughes, Uwe Krohn, John Davies, *BT, Ipswich, UK*
Robert Engels, Bernt Bremdal, *CognIT, Oslo, Norway*
Fredrik Ygge, *Enersearch AB, Gothenburg, Sweden*
Thorsten Lau, Bernd Novotny, Ulrich Reimer, *Swiss Life, Zürich, Switzerland*
Ian Horrocks, *University of Manchester, UK*

Executive Summary. *On-To-Knowledge*, the European EU-IST project No. 10132, builds an ontology-based tool environment to speed up knowledge management, dealing with the large numbers of heterogeneous, distributed, and semi-structured documents typically found in large company intranets and the World-Wide Web. Results aimed for by the project are: (1) a toolset for semantic information processing and user access; (2) OIL, an ontology-based inference layer on top of the World-Wide Web; (3) an associated methodology; (4) validation by industrial case studies. This paper gives an overview of the *On-To-Knowledge* approach to knowledge management.

1. Heterogeneous Information Resources Need Semantic Access

Support for information and knowledge exchange is a key issue in the Information Society. The exponential growth of online information on intranets and the Web leads to information overload. To cut down on the time wasted in searching and browsing, and reduce associated user frustration, much more selective user access is needed. This is possible by automatic meaning-directed or semantic information processing of online documents. The European IST project *On-To-Knowledge* builds tools that achieve this.

Ontologies [1,2] are the key technology used to describe the semantics of information exchange. Defined as "specifications of a *shared conceptualization* of a particular domain", they provide a shared and common understanding of a domain that can be communicated across people and application systems, and thus facilitate knowledge sharing and reuse. Ontologies will play a key role in growth areas such as knowledge management [3,4] and electronic commerce. In the US, funding agencies have recognized the importance of these issues by setting up the DAML program (http://www.darpa.mil/iso/ABC/BAA0007PIP.htm) that aims at machine-processable semantics of information resources accessible to agents.

The World-Wide Web (WWW) has drastically boosted the availability of electronic information. Already the present first generation of the web has changed our daily practice, and these changes will become even more significant in the near future. However, the Web itself has to change if it is to reach the next level of user service [5]. Currently, the Web is

an incredibly large, (mostly) static information source. The main burden in information access, extraction, and interpretation still rests with the human user. Document management systems now on the market have severe weaknesses:

- *Searching information*: Existing keyword-based search also retrieves irrelevant information that uses a certain term in a different meaning, and misses information when different terms with the same meaning about the desired content are used.
- *Extracting information*: Currently, human browsing and reading is required to extract relevant information from information sources since automatic agents do not possess the commonsense knowledge required to extract such information from textual representations, and they fail to integrate information spread over different sources.
- *Maintaining* weakly structured text sources is a difficult and time-consuming activity when such sources become large. Keeping such collections consistent, correct, and up-to-date requires mechanized representations of semantics that help to detect anomalies.
- *Automatic document generation* would enable adaptive websites that are dynamically reconfigured according to user profiles or other aspects of relevance. Generation of semi-structured information presentations from semi-structured data requires a machine-accessible representation of the semantics of these information sources.

Tim Berners-Lee coined the vision of a *Semantic Web* that provides much more automated services based on machine-processable semantics of data, and on heuristics that make use of these metadata. The explicit representation of the semantics of data accompanied with domain theories (i.e., ontologies) will enable a Knowledge Web that provides a qualitatively new level of service. It weaves together a net linking incredibly large parts of human knowledge and complements it with machine processability. Various automated services will support the human user in achieving goals via accessing and providing information present in a machine-understandable form. This process will ultimately lead to a highly knowledgeable world-wide system with specialized reasoning services that support us in many aspects of our daily life, becoming as central as access to electric power.

The competitiveness of companies depends heavily on how they exploit their corporate knowledge and memory. Most information in modern electronic media is mixed-media and rather weakly structured. This holds for the Internet but also for large company intranets. Finding and maintaining information is a hard problem in weakly structured representation media. Increasingly, companies realize that their intranets are valuable repositories of corporate knowledge. But with the now rapidly increasing volumes of information, turning this into useful knowledge has become a major problem. *Knowledge Management* is about leveraging corporate knowledge for greater productivity, value, and competitiveness [3,4]. Due to Internet-enhanced globalization, many organizations are increasingly geographically dispersed and organized around virtual teams. Such organizations need knowledge management and organizational memory tools that encourage users and foster collaboration while capturing, representing and interpreting corporate knowledge resources and their meaning. *On-To-Knowledge* provides the tools to speed up knowledge management in large distributed organizations, by applying ontologies to electronic information as a basis for semantic information processing and fast, meaning-directed user access.

2. Tool Environment for Ontology-based Knowledge Management

The *On-To-Knowledge* architecture and all its major components are shown in Figure 1. To illustrate these components and their interactions, we present a simple querying

scenario, where a user poses a query to the system that must be answered on the basis of a set of weakly structured data sources in a repository.

Two example scenarios in our case studies are:

- Querying a skills description repository on the Swiss Life intranet, where this repository is filled with a large variety of weakly structured documents (CV's, project descriptions, course descriptions, etc.).
- Locating the material that is required to answer a query at one of the helpdesks operated by BT. Again, much of the relevant material is very heterogenous in nature: technical specifications of devices, previous fault reports, customer data, etc.

The sequence of numbers in Figure 1 indicates the steps that must be taken in order to perform any of the above queries. Each of the components in Figure 1 is based on existing solutions already provided *On-To-Knowledge* consortium partners [6-10]. The red labels indicate which partner is providing crucial technology to each of the steps of the scenario.

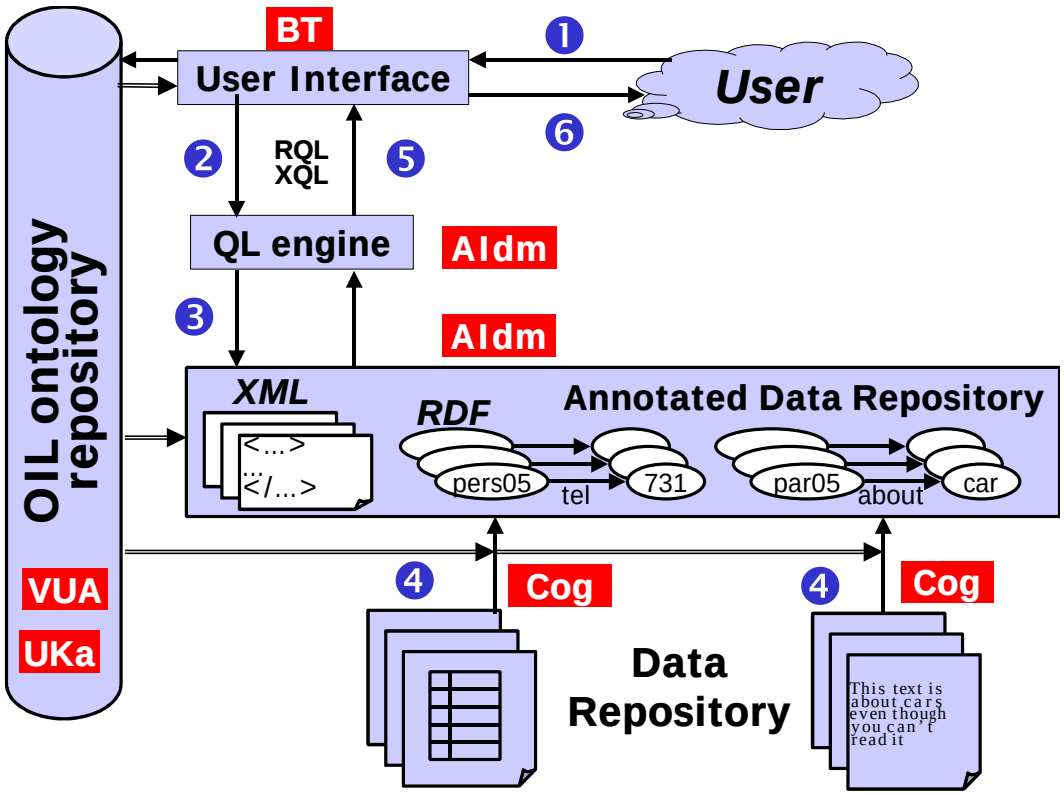

Figure 1. The On-To-Knowledge toolset for ontology-based knowledge management.

Step [1]. The system interacts with a user in order to elicit a specific query to be answered. Both the interaction with the user and the resulting query are entirely in terms of a domain-specific ontology, expressed in the OIL language developed within the consortium (see Section 3). The required ontologies are constructed using tools such as *OntoEdit* (http://ontoserver.aifb.uni-karlsruhe.de/ontoedit/), developed by the University of Karlsruhe [10]. Such an ontology-based user interaction has as main advantage that the user is shielded from any document-specific representation details, and can instead communicate in meaningful domain-specific terms. Furthermore, it makes the system much more robust against changes and variability in the formats of the underlying documents.

Step [2]. The user interaction results in a specific query to be answered by the data repository layer. We rely on the *Resource Description Framework* (RDF, http://www.w3c.org/Metadata/) currently being developed by the World-Wide Web consortium (W3C), to structure the data repository and to express queries over this

repository. The required translation from OIL-based user interaction to RDF-based queries is feasible because OIL is itself definable in terms of RDF-Schema definitions.

Step [3]. The consortium is developing an RDF query engine to efficiently process queries over medium-size data-repositories (with up to a million RDF triples in the repository). The University of Karlsruhe's *SilRI engine* (http://www.aifb.uni-karlsruhe.de/~sde/rdf/) is a starting point for such an engine. Besides RDF, XML may also be used to represent part of the semantically annotated data in the repository. In that case, an XML query language like XQL or XML-QL forms the basis for an XML-based query engine.

Step [4]. Of course, the above steps all assume that the data repository is filled with data that is annotated with sufficiently rich semantic information. Furthermore, the annotations must be related to the ontological vocabulary that was the basis for the original user query. Different technologies will be exploited to provide these annotations, depending on whether we are dealing with weakly structured data sources, or data sources that consist of free text only. In the first case, we will use wrapper technology in the vein of Jedi or W4. In the second case, the *Corporum* technology from CognIT ([9], http://www.cognit.no/) is the main platform for concept extraction from free text. Other tools will be based on automated summarization technology as developed for *ProSum* by BT [7,8].

Steps [5,6]. After the RDF query has been executed over the data repository, the resulting information is communicated to the user. Again, this must be done using an ontology-based vocabulary. Furthermore, powerful graphical visualizations of query results in the context of large data sets are developed. Examples of such visualizations are the semantic sitemaps produced by the *WebMaster* tool of AIdministrator [6] (for some results see Section 4).

3. OIL: Inference Layer for the Semantic World-Wide Web

The technical backbone of *On-To-Knowledge* is the use of ontologies for the tasks of meaningful information access, integration, and mediation. A major result from the *On-To-Knowledge* project is OIL (the ***O**ntology-based **I**nference **L**ayer*) [11]. OIL is a representation and inference layer on top of the Web that is based on ontologies. It unifies three important aspects provided by different communities: (i) formal semantics and efficient reasoning support as provided by Description Logics, (ii) epistemologically rich modelling primitives as provided by the Frame community, and (iii) a standard proposal for syntactical exchange notations as provided by the Web community.

- **Description Logics (DL).** DLs describe knowledge in terms of concepts and role restrictions that are used to automatically derive classification taxonomies. The main effort of the research in knowledge representation is in providing theories and systems for expressing structured knowledge and for accessing and reasoning with it in a principled way. OIL inherits from DL its *formal semantics* and the *efficient reasoning support* developed for these languages. In OIL, *subsumption* is decidable and with the developed *FaCT* engine we provide an efficient reasoner for this.
- **Frame-based systems.** The central modelling primitives of predicate logic are predicates. Frame-based and object-oriented approaches take a different point of view. Their central modelling primitives are classes (frames) with certain properties called attributes. Many other refinements of these constructs have been developed leading to the success of this modelling paradigm. Therefore, OIL incorporates the *essential modelling primitives* of frame-based systems into its language. OIL is based on the notion of a concept and the definition of its superclasses and attributes. Relations can also be defined not as an attribute of a class but as an independent entity having a certain domain and range. Like classes, relations can be arranged in a hierarchy.

- **Web standards: XML and RDF.** Modelling primitives and their semantics are one aspect of an ontology-based exchange language. Given the dominance and importance of the WWW, the syntax of such a language must be formulated using existing web standards for information representation. As already proven with XOL (http://www.ai.sri.com/pkarp/xol/), XML can be used as a serial syntax definition language for ontology-based information exchange. OIL is closely related to XOL and can be seen as an extension of it. Another candidate for a web-based syntax for OIL is RDF and RDFS. The RDF framework for the encoding, exchange, and reuse of structured metadata provides a means for adding semantics to a document without making any assumptions about the structure of the document. RDF schemas (RDFS) provide a basic type schema for RDF. Objects, Classes, and Properties can be described. In relation to ontologies, RDF provides two important contributions: a standardized syntax for writing ontologies, and a standard set of modelling primitives like instance-of and subclass-of relationships. Therefore, OIL offers two syntactical variants: one based on XML schema and one based on RDF schema.

Why not Ontolingua? Ontolingua (http://ontolingua.stanford.edu/) is an existing proposal for an ontology language. It has been designed to support the design and specification of ontologies with a clear logical semantics based on KIF. Ontolingua extends KIF with additional syntax to capture intuitive bundling of axioms into definitional forms with ontological significance; plus a Frame Ontology to define object-oriented and frame-language terms. The problem with Ontolingua is its high expressive power provided without any means to control it. Not surprisingly, no reasoning support has been provided for Ontolingua. OIL takes the opposite approach. We start with a very simple and limited core language. The web has proven that restriction of initial complexity and controlled extension when required is a very successful strategy. OIL takes this lesson to heart.

The use of OIL is currently evaluated in two running IST projects *On-To-Knowledge* and Ibrow (http://www.swi.psy.uva.nl/projects/ibrow/home.html). In *On-To-Knowledge*, OIL will be extended to a full-fledged environment for knowledge management in large intranets and websites. Unstructured and semi-structured data will be automatically annotated, and agent-based user interface techniques and visualization tools will help the user to navigate and query the information space. Here, *On-To-Knowledge* continues a line of research that was set up with SHOE and Ontobroker [5]: using ontologies to model and annotate the semantics of information resources in a machine-processable manner.

4. Business Applications in Semantic Information Access

Industry case studies. *On-To-Knowledge* is carrying out three industrial case studies to evaluate the tool environment for ontology-based knowledge management (Section 2) and the associated web inference layer OIL (Section 3). These case studies are chosen such that they ensure a broad coverage, involving three different industry sectors (insurance, telecom, energy) in three different countries, and facing different knowledge management problems.

Swiss Life: organizational memory. Swiss Life's vision is to build an *organizational memory* with an intranet-based portal. Three case studies explore the problem space:

1. A skills database contains a large variety of structured and unstructured documents like CVs, recruitment profiles, course and project descriptions. Today these documents do not exist or are not integrated into a single repository. Furthermore, there is no common vocabulary (i.e. ontology) that guarantees a unified usage and understanding of the documents, resulting in insufficient retrieval results.

2. Information about an insurance product comprises documents for sales persons, for training purposes, about performing office tasks, etc. This information is created in different places, in different formats and often not distributed to the right places.
3. The IAS document ("International Accounting Standards") is part of the global Swiss Life Intranet, called "GroupNet". The document's 1000 web pages make it very hard to find relevant passages, even though there is a division into chapters and sections.

BT: call centres. Call Centres are an increasingly important mechanism for customer contact in many industries. Every transaction should emphasize the uniqueness of both the customer and the customer service person. To do this one needs effective knowledge management [7, 8]. This includes knowledge about the customer but also knowledge about the customer service person, so that the customer is directed to the right person to answer their query. This knowledge must also be used in a meaningful and timely way. One or more of BT's own Call Centres will be targeted to identify opportunities for effective knowledge management using the *On-To-Knowledge* tools. More specifically, call centre agents tend to use a variety of electronic sources for information when interacting with customers, including their own specialized systems, customer databases, the organization's intranet and, perhaps most importantly, case bases of best practice. The *On-To-Knowledge* techniques provide an intuitive front-end to these heterogeneous information sources, to ensure that the performance of the best agents is transferred to the others.

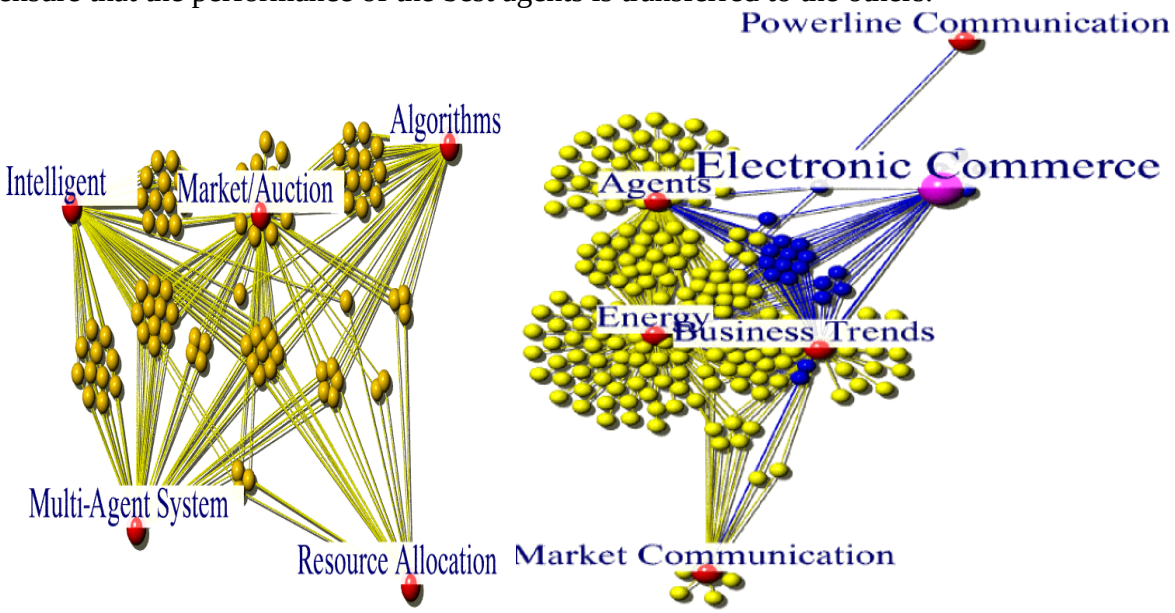

Figure 2. Automatically generated semantic structure maps of the EnerSearch website.

EnerSearch: virtual enterprise. EnerSearch is a virtual organization researching new IT-based business strategies and customer services in deregulated energy markets (e.g., [12], see further http://www.enersearch.se). Its research affiliates and shareholders are spread over many countries: its shareholding companies include IBM (US), Sydkraft (Sweden), ABB (Sweden/Switzerland), PreussenElektra (Germany), Iberdrola (Spain), ECN (Netherlands), and Electricidade do Portugal. Essentially, EnerSearch is a knowledge creation company, knowledge that must be transferred to its shareholders and other interested parties. Its website is one of the mechanisms for this. However, it is rather hard to find information on certain topics – the current search engine supports free text search rather than content-based search. Therefore, the EnerSearch case study evaluates the *On-To-Knowledge* toolkit to enhance knowledge transfer to (1) researchers in the EnerSearch virtual organization in different disciplines and countries, and (2) specialists from shareholding companies interested in getting up-to-date information about R&D results on IT in Energy.

Some first results of *On-To-Knowledge* techniques are shown in Figure 2. It shows two semantic structure maps of the EnerSearch website, produced by the WebMaster tool of AIdministrator [6], and based on a domain ontology concerning important IT in Energy research topics. Every node represents a webpage (that can be directly opened in a browser by clicking on the node); edges denote hypertext links. Left, we see a map of subtypes (subtopics) of agent research by EnerSearch. It is easy to see how subtopics are related and find the relevant webpages. Right, we see an interactively generated sitemap showing how the topic of e-commerce intersects with other topics (dark blue nodes). A nice feature of the visualization is that spatial proximity correlates very well with semantic closeness.

Methodology. In addition to the toolset and the OIL language, *On-To-Knowledge* is developing an associated methodology for ontology-based knowledge management. Input to this are existing European research results, such as the *CommonKADS* approach to knowledge engineering and management [3], experiences from knowledge-based software engineering [12] and tool development [5-10], ontology composition [2] and information retrieval techniques [14], and feedback from industry case studies. The methodology will also cover how to develop the business case for ontology-based knowledge management.

Conclusion. World-Wide Web and company intranets have boosted the potential for electronic knowledge acquisition and sharing. Given the sheer size of these information resources, there is a strategic need to move up in the data – information – knowledge chain. As a necessary step, *On-To-Knowledge* provides innovative tools for semantic information processing and thus for much more selective, faster, and meaningful user access.

References

[1] D. Fensel: *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce*. Springer-Verlag, Berlin, D, to appear (2000).

[2] P. Borst, J.M. Akkermans, and J.L. Top: Engineering Ontologies, *International Journal of Human-Computer Studies* **46** (1997) 365-406.

[3] A.Th. Schreiber, J.M. Akkermans, A. Anjewierden, R. de Hoog, N. Shadbolt, W. Van De Velde, and B. Wielinga: *Knowledge Engineering and Management.* The MIT Press, Cambridge, MA, 2000.

[4] U. Reimer (Ed.): *Proc. 2nd Int. Conf. on Practical Aspects of Knowledge Management* (PAKM'98), Basel, Switzerland, October 1998. URL: http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-13/.

[5] D. Fensel, S. Decker, M. Erdmann, H.-P. Schnurr, R. Studer, and A. Witt: Lessons learnt from Applying AI to the Web. *Journal of Cooperative Information Systems*, to appear (2000).

[6] F. van Harmelen and J. van der Meer: WebMaster: Knowledge-based Verification of Web-pages. In *Proceedings 2nd Int. Conf. on The Practical Applications of Knowledge Management* (PAKeM99, London, UK, April 1999), pp. 147-166. The Practical Applications Company, Blackpool, UK, 1999.

[7] J. Davies, S. Stewart, and R. Weeks: Knowledge Sharing over the World Wide Web, *WebNet '98*, Florida, USA, November 1998. (also at http://www.bt.com/innovation/exhibition/knowledge_management/).

[8] J. Davies: Supporting Virtual Communities of Practice, in R. Roy (Ed.): *Industrial Knowledge Management*, Springer-Verlag, London, forthcoming (2000).

[9] B. Bremdal, F. Johansen, C. Spaggiari, R. Engels, R. Jones: Creating a Learning Organisation Through Content-Based Document Management, OECD HRP Meeting, Loen, NO. CognIT Report, Oslo, May 1999.

[10] A. Maedche, H.-P.Schnurr, S. Staab, and R. Studer: Representation Language-Neutral Modeling of Ontologies. In: Frank (Ed.), *Proceedings German Workshop Modellierung 2000*. Koblenz, D, April 2000.

[11] I. Horrocks, D. Fensel, J. Broekstra, S. Decker, M. Erdmann, C. Goble, F. van Harmelen, M. Klein, S. Staab, and R. Studer: *The Ontology Inference Layer OIL*, On-To-Knowledge EU-IST-10132 Project Deliverable No. OTK-D1, Free University Amsterdam, Division of Mathematics and Informatics, Amsterdam, NL, 2000. Available from http://www.ontoknowledge.org/oil.

[12] F. Ygge and J.M. Akkermans: Decentralized Markets versus Central Control - A Comparative Study, *Journal of Artificial Intelligence Research* **11** (1999) 301-333. (Also available from http://www.jair.org/).

[13] J. Angele, D. Fensel, and R. Studer: Developing Knowledge-Based Systems with MIKE, *Journal of Automated Software Engineering* **5** (1998) 389-418.

[14] Y. Ding, G. Chowdhury, and S. Foo: Bibliometric Information Retrieval System (BIRS): A Web search interface utilizing bibliometric research results. *Journal Am. Soc. for Information Science*, to appear (2000).