Interoperability and scalability of On-To-Knowledge tools

Frank van Harmelen — **VU Amsterdam**

Arjohn Kampman, **Jeen Broekstra** — **AIdministrator Amersfoort**

Summary. In this report we describe a number of decisions that have been taken in the On-To-Knowledge consortium in order to ensure

- that the software tools deliverd by the separate partners will be interoperable;
- that they will make uniform assumptions about their deployment environment;
- that they will scale to the levels required by the case-studies performed in the On-To-Knowledge project.

The contents of this document has been agreed upon by the partners during the consortium meeting in Karlsruhe in January 2001.

Document Id.	deliverable X1
Project	EU-IST On-To-Knowledge IST-1999-10132
Issuer	Vrije Universiteit
Date	24th April 2001
Status	Final
Distribution	Restricted

On-To-Knowledge Consortium

This document is part of a research project partially funded by the IST Programme of the Commission of the European Communities as project number IST-1999-10132. The partners in this project are: Vrije Universiteit Amsterdam VUA (coordinator, NL), University of Karlsruhe (Germany), Schweizerische Lebensversicherungs- und Rentenanstalt/Swiss Life (Switzerland), British Telecommunications plc (UK), CognIT a.s. (Norway), EnerSearch AB (Sweden), AIdministrator Nederland BV (NL).

Vrije Universiteit Amsterdam (VUA)
Division of Mathematics and Informatics W&I
De Boelelaan 1081a
1081 HV Amsterdam
The Netherlands
Tel: +31 20 4447718, Fax: +31 20 872 27 22
Contactperson: Dieter Fensel
E-mail: dieter@cs.vu.nl

Schweizerische Lebensversicherungs- und Rentenanstalt / Swiss Life
Swiss Life Information Systems Research Group
General Guisan-Quai 40
8022 Zürich
Switzerland
Tel: +41 1 284 4061, Fax: +41 1 284 6913
Contactperson: Ulrich Reimer
E-mail: Ulrich.Reimer@swisslife.ch

CognIT a.s
Busterudgt 1.
N-1754 Halden
Norway
Tel: +47 69 1770 44, Fax: +47 669 006 12
Contactperson: Bernt. A.Bremdal
E-mail: bernt@cognit.no

AIdministrator Nederland BV
Julianaplein 14B
3817 CS Amersfoort
The Netherlands
Tel: +31 33 4659987, Fax: +31 33 4659987
Contactperson: Jos van der Meer
E-mail: Jos.van.der.Meer@aidministrator.nl

University of Karlsruhe
Institute AIFB
Kaiserstr. 12
D-76128 Karlsruhe
Germany
Tel: +49 721 608392, Fax: +49 721 693717
Contactperson: R. Studer
E-mail: studer@aifb.uni-karlsruhe.de

British Telecommunications plc
BT Adastral Park
Martlesham Heath
IP5 3RE Ipswich
United Kingdom
Tel: +44 1473 605536
Fax: +44 1473 642459
Contactperson: John Davies
E-mail: John.nj.Davies@bt.com

EnerSearch AB
Carl Gustafsväg 1
SE 205 09 Malmö
Sweden
Tel: +46 40 25 58 25; Fax: +46 40 611 51 84
Contactperson: Hans Ottosson, Fredrik Ygge
E-mail: hans.ottosson,fredrik.ygge@enersearch.se

1 Interoperability

The On-To-Knowledge project will produce a number of software tools to support the use of ontologies for knowledge management tasks:

- ontology extraction from text,
- ontology editor,
- ontology storage and retrieval,
- ontology-based information navigation and querying,
- ontology-based visualisation of information.

These tools will be developed as an "integrated tool-suite". In other words: they will not be delivered as a single piece of software, but they will be separate programs that will interoperate.

In this section we describe a number of decisions that have been taken in the On-To-Knowledge consortium in order to ensure this interoperation. By "interoperation" we mean in this context that:

- the results of one program can be used as the input for another,
- that the proper communication protocols are present to enable this exchange

For example, one might use the editor to modify an automatically extracted ontology, and subseqently visualise the result.

Notice that the decisions described below will only ensure interoperability of the tools in this sense. They are *not* intended to ensure a uniform user interface, or any sharing of datastructures and/or code beyond what is required to ensure interoperability. This is the essense of developing the On-To-Knowledge tools as a "suite of independent tools" instead of as a single large application.

The decisions ensuring interoperability are:

- Maximal reliance on existing standards
- Using a single ontology representation language
- Client-server architecture

1.1 Maximal reliance on existing standards

A crucial engineering decision for all the On-To-Knowledge tools is to rely as much as possible on existing standards. The current set of standards in use by the consortium are:

- XML as a universal format for data interchange and storage. `http://www.w3c.org/XML`
- RDF for the representation of semi-structured data `http://www.w3c.org/RDF`
- RDF Schema for describing simple vocabularies of RDF data `http://www.w3c.org/RDF`
- HTTP as the basic protocol for communication between Web-servers and -clients `http://www.w3.org/Protocols/`
- JDBC^{TM} as the Java API for database access `http://java.sun.com/products/jdbc/`
- RQL for querying RDF/RDF Schema documents `http://www.ics.forth.gr/proj/isst/RDF/RQL/rql.pdf` RQL is currently a proposed standard in the IST-project Memuse.

Reliance on these open standards (with the exception of JDBC, which is ^{TM} of Sun Microsystems) is an important step towards interoperability between On-To-Knowledge tools. At least as important is the fact that it will also greatly enable the exploitation of the On-To-Knowledge tools outside the consortium.

1.2 Using a single ontology representation language

All tools in the On-To-Knowledge consortium will exploit DAML+OIL as the language for representing ontologies.

DAML+OIL has been defined by the The Joint United States / European Union ad hoc Agent Markup Language Committee (`http://www.daml.org/committee/`). This committee was created in October 2000 by Jim Hendler of DARPA and Hans-Georg Stork of the European Union Information Society Technologies Programme (IST). Members of the On-To-Knowledge consortium are actively participating in the committee.

The specification of DAML+OIL can be found at `http://www.daml.org/language`. It is closely modelled on the OIL language as developed by the On-To-Knowledge consortium (`http://www.ontoknowledge.org/oil`).

Since DAML+OIL is expected to be one of the starting points for the recently announced Semantic Web activity of the W3C (`http://www.w3.org/2001/sw/`), it is a good language of choice for the On-To-Knowledge consortium.

An important sublanguage of DAML+OIL is OIL Core. OIL Core coincides largely with RDF Schema. More precisely: OIL Core is identical to RDF Schema, with the exception of reification and containers, which are not part of OIL Core (due to their dubious semantics in RDF Schema). This means that even simple RDF Schema agents are able to process the DAML+OIL ontologies, and pick up as much of their meaning as possible with their limited capabilities.

Support for OIL Core is mandatory for On-To-Knowledge tools. Support for the extension to full DAML+OIL is optional.

1.3 Client-server architecture

Figure 1 depicts the conceptual architecture of the On-To-Knowledge tool-suite. Many of the decision above can already be seen in this figure:

- The ontologies in the ontology- repository are represented in DAML+OIL
- annotated data from documents are uploaded to the data-repository as RDF
- the user interface queries both repositories using RQL

The technical architecture which we will use to implement the above conceptual design is shown in figure2. This figure shows the following important design decision w.r.t. figure 1:

- ontology repository and data repository are merged into a single repository
- the software components are organised in a client-server architecture, where the repository is the server to which individual tools (ontology-editor, ontology-extractor, visualisation tools, etc) connect.
- the HTTP protocol `http://www.w3.org/Protocols/` will be used as the basis for all client-server connections, since it is a simple, open, standardised and well-known protocol, and is expected to provide sufficient expressiveness for our needs. In the final year of the project, we expect to add support for further communications protocols, such as SOAP, EJB, Corba and RMI.

2 Scalability

In order to assess the scalability of the On-To-Knowledge tools, the following orders of magnitude have been provided at the January 2001 meeting in Karlsruhe:

1. Ontologies are expected to be $O(10^3)$ classes. This number is relevant for the ontology-editor and for the ontology-storage. OntoEdit can comfortably deal with this number of classes. Also, this number is small enough to allow Sesame to cache the entire schema in memory, which will significantly improve the performance of the query engine.

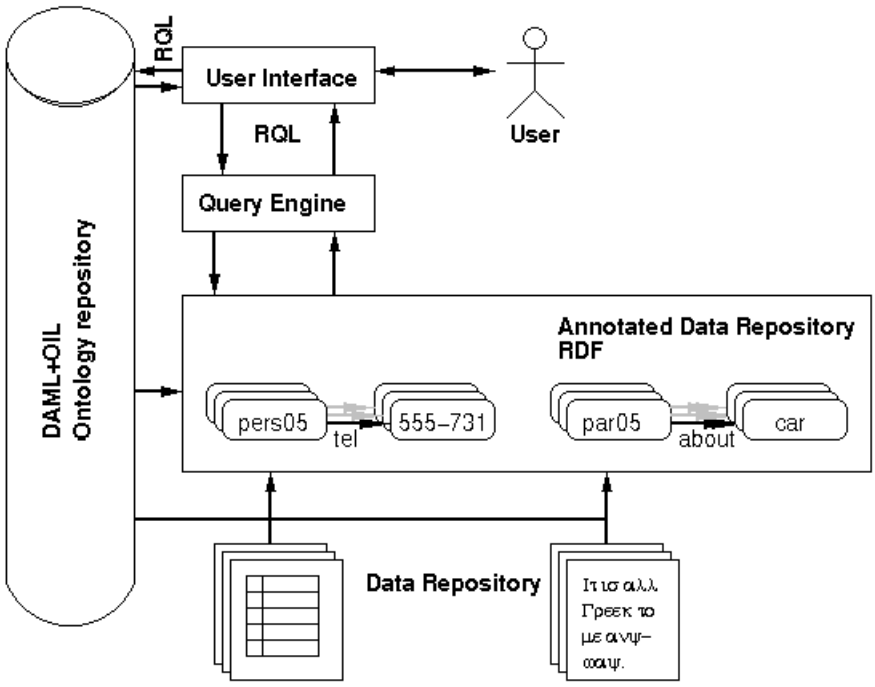

Figure 1: The conceptual architecture of the On-To-Knowledge tool suite

2. Intranets are expected to be up to $O(10^4)$ pages. In fact, the intranets from the three case-studies will more likely be of the order of $O(10^3)$ pages.
3. WebFerret can index up to $O(10^6)$ pages.
4. The ontology-extraction service from CognIT analyses $O(1)$ pages per second.
5. the same ontology-extraction service generates $O(10)$ – $O(10^2)$ RDF triples per page, and more likely to be $O(10)$ then $O(10^2)$.

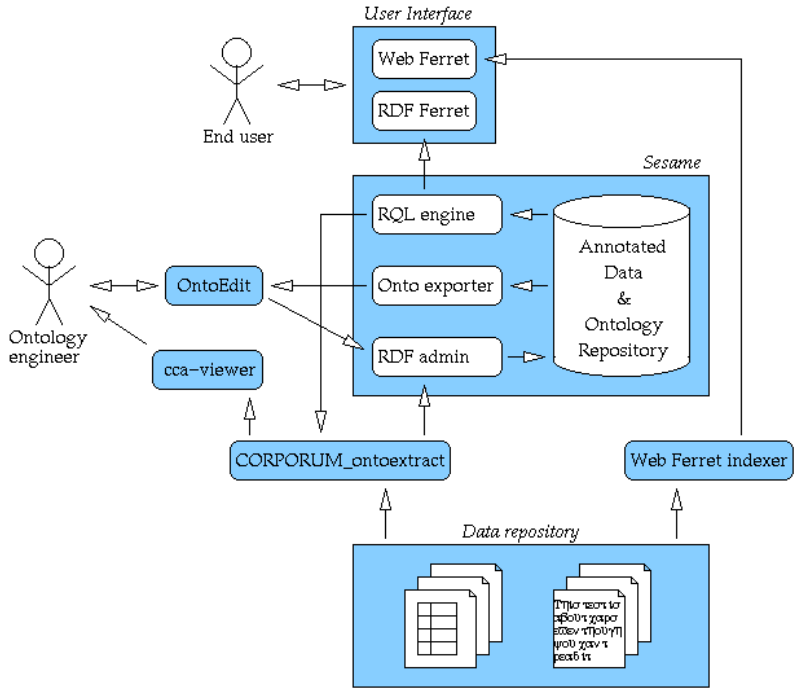

Figure 2: The technical architecture of the On-To-Knowledge tool suite

6. Combining [2] and [5] results in $O(10^5)$ triples to store in Sesame (with a maximum of $O(10^6)$). The next version of Sesame is expected to be able to cope with this. The current functional, but not scalable, version has already been tested with up to $O(10^5)$ triples.

7. Combining [2] and [4] results in $O(10^4)$ seconds for the ontology extraction service to analyse a complete intranet. This amounts to a few hours to 1 day for a full analysis. This is acceptable since such a full extraction service can be done off-line, and needs to be done at most once a day (and more likely much less often).

The above calculations lead us to believe that the tools that will be developed in the On-To-Knowledge project will be able to deal with the demands of the three case-studies as included in the project, assuming no more than standard desk-top hardware.