Enhancing Information Management in the Water Sector

An Exploration of the Semantic Web option

by Otto Krüse

Enhancing Information Management in the Water Sector

An Exploration of the Semantic Web option

Cover note

This bundle contains the writings I've made for my graduation of the masters programme Systems Engineering and Policy Analysis. This programme is offered by the faculty of Technology, Policy and Management that is part of the Delft University of Technology.

Instead of a conventional masters thesis I have written a scientific paper. This because the department where I graduated felt that a scientific paper would have more value than a full-blown masters thesis. A paper could possibly be published in a journal or in the proceedings of a convention. Also the knowledge in a paper has a larger chance to be reused since a paper is attractive to read because of its reduced size and its summarizing nature. I feel that a masters thesis is cumbersome to read and will most likely end up in some dusty cabinet at the faculty library. Perhaps stated a bit bold here, but the point is made. Also a scientific paper may be smaller in size but is more difficult to write. Thus, writing a sound scientific paper is more proof of a students capabilities.

Because the scientific paper focuses almost solely on the content and findings of my research I have also written a process report. The process report describes the way and the setting in which I have performed my research.

The research I have done was aimed at exploring the possibilities of a new technology called the Semantic Web within a real organization, Rijkswaterstaat. It was an interesting challenge to put theory to practise.

I wish to thank the following people:
The graduation committee: Jos Vrancken, Stefan Fischer, Sam Muniafu and professor Wim Vree for their help, feedback and advice.
Larissa Zegveld and Michel Kuypers, for reading early drafts of my work and providing feedback on it.
Gaston Lamaitre for useful discussions we had concerning my research.
Aart Vermetten for his help and feedback.
The rest of the KANS team for providing inspiration through discussions concerning KANS.
The section IBW from the AGI Rijkswaterstaat for their help, but also for the pleasant and friendly manner in which they took me up in their group.

Delft, January 13, 2005

Otto Krüse

Enhancing Information Management in the Water Sector: an exploration of the Semantic Web option

Otto Krüse
Delft University of Technology
Faculty of Technology, Policy and Management
otto_kruse@hotmail.com

Abstract. This paper is about applying the Semantic Web to Information Management in the water sector. Information management in this sector is complex resulting in problems concerning information exchange and information retrieval. The possibility for using Semantic Web technology in solving these problems is analyzed in this research. The technology has been applied in the fields of Enterprise Application Integration (EAI), Information Management and Knowledge Management. This case study encompasses the construction of a prototype and concludes that Semantic Web technology is not mature enough for EAI but already a valuable option for Information and Knowledge Management.

1. Introduction

Water management institutes are concerned with both the quality and quantity of water. Especially in the Netherlands water management has great importance: the Netherlands houses the mouths of 4 major European rivers and lies for 26% below sea level. Securing ‘dry feet’ for the Dutch citizens is a challenging task. Also, in a country where water has such a large presence and is so widely used people have to *live* with water. This means that water must be of sufficient quality and that safe and efficient traffic on waterways must be possible. Rijkswaterstaat is the governmental institute that is responsible for water management in the Netherlands. Part of this water management task is Information Management. Information Management within Rijkswaterstaat is a complex undertaking. Inherent to this complex nature is the existence of a number of problems. The research described in this paper aims to determine what value the upcoming Semantic Web technology can have in solving these problems and reducing complexity.

One complexity in information management is formed by the sheer amount of information that’s used. The water management tasks of Rijkswaterstaat encompass many business processes which require large amounts of information. For providing this information even larger amounts of data are generated (e.g. through measurements) and processed. A number of typical examples include: water levels of rivers and the sea, heights of dikes and dunes and concentrations of heavy metals in surface and ground water. A second complexity in information management arises from the fact that the consumed and produced information is of diverse nature. Managing these varying kinds of information is difficult. The complexities in information management result in problems for information exchange as well as for information retrieval.

A project currently carried out within Rijkswaterstaat aims to reduce the complexities and alleviate resulting problems. This project, called KANS[1], strives to implement two technological solutions: a Web Service system that will make information exchange more easy and an information catalogue that will facilitate better information retrieval. In this paper an analysis will be presented of how Semantic Web technology can be used for these solutions. The Semantic Web promises to make information exchange easier and to enable more intelligent information retrieval than what can be done with search engines like Google today. Obviously two valuable properties for the KANS project.

Through literature study, action research and the construction of a prototype this paper will show that Semantic Web technology is mature enough to be used for the information catalogue but is yet too immature to be usefully applied in the context of the Web Service system. With the creation of a prototype it will also be shown *how* Semantic Web technology can be used. This research concludes that Semantic Web technology in its current state is mature enough for use in Information and Knowledge

[1] Abbreviation for (Dutch): Koepel Architectuur Natte Sector. This translates roughly to *Encompassing Architecture Wet Sector.*

Management but yet too immature for use within Enterprise Application Integration.

This paper will continue in part 2 with a more elaborate discussion of the problems and complexities concerning Information Management at Rijkswaterstaat. Part 3 will give a short introduction into the Semantic Web. An analysis of and how the Semantic Web can be used for the Web Service architecture is made in part 4. The same analysis is done with the information catalogue in part 5. General conclusions are presented in part 6.

2. Background and setting

The fundamental goal of the KANS project is to make a structural step forward in overall information management. KANS tries to both solve current problems as well as explore new options. As said, KANS hopes to solve two problems; information exchange between information-systems is inflexible and information itself is difficult to pin point and find. This section will elaborate on the KANS project and its two solutions

Let's first take a closer look at the problems in information exchange. The current information-system landscape of Rijkswaterstaat comprises over 100 systems. Most of the systems are not monolithic in the sense that they need information from other systems to function properly. Thus there exist many interfaces between these systems. Having evolved without a structural vision or architectural approach these interfaces are largely heterogeneous and have a 1-to-1 mapping with information systems. All this results in a spaghetti-like network of systems that:

- hampers the evolution of systems because of inherent change propagation to other systems.
- makes addition of new systems more difficult because new specialized interfaces have to be build.
- has high and rising maintenance costs due to the heterogeneous nature of the interfaces.
- makes the overall information-system landscape complex, and thus difficult and costly to manage.

To overcome these problems KANS aims to implement a uniform and generic application interface, based on Web Service technology[2]. For this two objects will be created: the generic application interface itself (referred to as GAIN) and a register of services (referred to as SRN). In this new architecture every information-system uniformly interfaces with other information-systems, instead of having a unique interface for each information-system it interfaces with (*n* complexity instead of n^2). The SRN keeps track of the different information-systems (especially the services they offer) and stores information about their interfaces which can be accessed by other information-systems.

The second problem KANS perceived is the difficulty in pin pointing and finding information itself. Pin pointing here means to describe what information exists and where it can be found. Decision processes take place at system level, while the related problems mostly exist at information level. It would provide managerial benefits to make a clear separation between the information and system level. Difficulties in information retrieval arise from the fact that Rijkswaterstaat is a large organization (over 10.000 employees) and information available in one place of Rijkswaterstaat may not be visible to other parts. This hampers business processes or leads to recreation of information at additional effort and costs. The solution KANS aims to implement to alleviate these problems is an information catalogue (referred to as ICN). This catalogue explicates what information exists and what information-system provides it, thus both pin pointing information and enabling it to be found.

Research questions

The following question is to be answered in the research described in this paper:

"In which way can Semantic Web technology improve information exchange and information retrieval ('searchability') as aimed for in the KANS project?"

This question falls apart in two sub-questions:

- *What can Semantic Web technology do in the area of information-exchange for the KANS project?*
- *What can Semantic Web technology do in the area of information searching for the KANS project?*

Answering these two questions will have value in itself for the KANS project, also it allows for the induction of more general conclusions

[2] Web services provide a standard means of interoperating between different software applications, running on a variety of platforms and/or frameworks. [1] Web services are based on Internet and XML technology.

concerning the applicability of Semantic Web technology above and beyond KANS. The first sub-question will be implicitly answered in part 4 of this paper. There an analysis is made of the applicability of Semantic Web technology for the Web Service architecture which covers 'information exchange as aimed for in the KANS project'. Likewise the second sub-question will be answered in part 5 of this paper, where 'information searching' is applied in light of the information-catalogue.

Methodology

In carrying out this research a number of steps were followed according to a specific and tailor-made methodology. The first element in this methodology concerned a theoretical (literature) study of the Semantic Web. This study was necessary for determining the different potential applications that Semantic Web technology has. After gathering knowledge on the entire application spectrum the second phase in the followed methodology was aimed at determining which of these applications were relevant within the context of the KANS project. By working as part of the KANS project team the author of this paper already gained a rich impression of what the needs within KANS were. This indeed enabled the determination of the most relevant Semantic Web applications for KANS. The third phase of the methodology had the goal to analyse the feasibility of the relevant applications. The concept feasibility captures more factors that determine the value of Semantic Web technology for KANS than just the relevance of its applications alone (maturity of the technology for example).

The fourth phase of the methodology was aimed at putting the relevant and feasible application(s) to practise through the construction of a prototype. This had a number of goals. First the construction of a prototype serves as a kind of proof that Semantic Web technology can be applied for Rijkswaterstaat[3]. Secondly it serves to demonstrate this particular application. Also by putting the Semantic Web theory to practise the validity of the theory itself will be implicitly tested. The construction of a prototype also leads to new insights and idea's and stimulates creativity. The prototype has a largely conceptual nature. This is sufficient because for answering the research questions above it only needs to show that *and* how Semantic Web technology can be used. The fifth and final phase of the methodology aimed to make final conclusions about the current applicability of the Semantic Web.

3. Semantic Web in short

The Semantic Web is the encompassing name for the products of the Semantic Web research initiative conducted by the World Wide Web Consortium. The goal of this initiative is broad: to create a universal medium for the exchange of data. The main underlying instrument for this is an extensive use of meta-data that enables computers to 'understand' information and process it more automatically.

Currently the Semantic Web consists of a number of recommendations [2] that define three languages that can be used to describe meta-data and meta-data schema (known as ontologies):

- Resource Description Framework (RDF) [3]
- RDF Schema [4]
- Web Ontology Language (OWL) [5]

3.1 RDF

RDF is a language with which information can be described. More specifically, RDF is a meta-data framework. RDF was designed to be read and understood by computers as well as humans.; it is therefore called machine-processable. RDF describes information by making statements consisting of three elements: subject, predicate and object. RDF statements represent a graph. Although statements are commonly serialized in XML form the graph is the essence. Multiple statements can be made to make larger graphs and thus more elaborate descriptions of information.

An important ingredient of RDF is the Uniform Resource Identifier (URI). A URI uniquely identifies some thing or concept and is often presented in URL form. By using URI's ambiguities in the semantics of terms are resolved because terms are uniquely identified.

3.2 RDF Schema & OWL

RDF Schema is an RDF vocabulary with which other RDF vocabularies can be defined. The RDF Schema vocabulary consists of a number of terms that enable for the description of simple meta-data schemes as well as more advanced taxonomies. As such RDF Schema can do more than just define simple meta-data schema; it can classify information in a tree structure, thus providing more accurate descriptions of information.

[3] As such it could also be seen as a proof-of-concept.

OWL goes even further in accurately describing information, it is an RDF vocabulary that enables quite precise and semantically rich descriptions of information. Such a semantically rich description is called an ontology, hence the name Web *Ontology* Language. An ontology might be described as an extended and enhanced taxonomy (likewise taxonomies are sometimes called 'light-weight'ontologies). OWL is largely based upon so called 'description logic': a subset of logic that's used for Knowledge Representation. Because OWL is based upon Description Logic, which has a formal grounding in artificial intelligence, it is possible to let inference engines reason with OWL ontologies[4] [6]. This means that computers can automatically check for the consistency of the ontology and infer the class hierarchy; important issues regarding scalability and maintenance of large ontologies.

3.3 Application area's

Semantic Web technology is rather fundamental and can thus be applied in numerous ways. A summary of its most promising applications, some proposed in [7] and [8]:

- Data storage and integration: by storing data in a uniform way it can be integrated easily, providing for uniform access by applications. Also, ontologies can provide mappings between different RDF vocabularies and XML-Schema.
- Enhanced information retrieval: by explicating the semantics of terms, as well as classifying them searching can be more efficient and effective.
- Web Service descriptions: Semantically rich descriptions can augment current Web Service technology in facilitating a greater automation in their use.
- Shared vocabularies, grammar and process descriptions: the Semantic Web languages can be used to describe these things in a machine processable way (like an enhanced RosettaNet).
- Categorization and indexation: the ability to define meta-data schemes as well as taxonomies allows for categorization and indexation of information.
- Website structuring: An ontology (or taxonomy) can be an instrument to structure Websites. Super/sub-class relations as well as properties can map to hyperlinks between Webpages.
- Information disclosure: by offering information to the outside world in RDF form other parties can easily integrate it into their own applications (B2B, shopbots).

Not all these applications are relevant for the KANS project Obviously the Web Service description deserves attention in light of the Web Service architecture. Enhanced information retrieval and categorization and indexation (which are actually rather inherent to each other) are interesting for the information catalogue. The other applications lie beyond the current scope of KANS though they may very well become more relevant in the future.

3.4 Current shortcomings

The Semantic Web is still in a development phase and as such has some shortcomings. When analyzing the value of the Semantic Web for the KANS project these are important to consider:

- No rules language yet: it remains impossible to say for example that "if a and b are true then c must also be true" [7]. In many applications there is a great need for such rules.
- No standard query language: querying will be an important element of practically *any* Semantic Web application. However, no standard query language exist (like SQL for relational databases). Only a handful support RDF Schema semantics, none support OWL[5].
- No mature software implementations: there currently exist very little mature software solutions that successfully deploy or use Semantic Web technology. Software that does exist mainly resides in academic/scientific scenes and thus remains relatively obscure.
- No industry-wide support: the mainstream industry has yet to embark upon Semantic Web technology. Having industry support could be vital in sustaining critical mass; Semantic Web technology will flourish more when it is commonly used.

[4] Note that this is only possible for OWL Lite and OWL DL, subsets of OWL Full. Elaboration upon this subject lies beyond the intended scope of this paper.

[5] Although this might not be needed/applicable for a query language since it requires inference.

The Semantic Web has considerable potential to gain widespread use in the future. It already has a significant support in the academic and scientific world. Also, because the Semantic Web languages are serialized in XML, they conform to and can coexist with one of the most dominating technologies of these days resulting in advantages as compatibility and technology reuse.

4. Web service description

As mentioned in the previous section, Semantic Web technology can be applied within KANS for describing Web Services. A Web service is a kind of procedure, method or function accessible via internet protocols (e.g. http) that sends its data in XML form, also via Internet protocols. Web services can be very simple, like a service that when invoked returns the current water height of the Rhine at Lobith. By combining multiple Web services complex Web services might be build. This process of combining and invoking different Web services is often called orchestration or composition.

The most important standards concerning Web services are SOAP [9], WSDL [10] and UDDI [11]. SOAP is a standard that specifies the way messages are exchanged between Web services. WSDL is a language for describing (the interface of) Web services and UDDI is a protocol for registers of Web services (popularly referenced as the Web services yellow pages). Note that currently Rijkswaterstaat has indeed chosen to implement the KANS Service register (SRN) according to the UDDI protocol, this also implies the use of WSDL upon which UDDI is dependant.

4.1 Problem of current technology

Searching for a particular Web Service in a UDDI register is not a trivial task because:

- both WSDL and UDDI have little descriptive capabilities; information that is not well described is hard to retrieve/find [12].
- current UDDI registers provide limited searching facilities [12].

The second point forms the smallest problem since future versions of UDDI registers might incorporate better searching facilities. The first point is rather fundamental however. Neither WSDL nor UDDI are able to describe (information on) Web Services in a semantically rich way. Information retrieval in a register that has little semantic content will be difficult, no matter how advanced its searching facilities might be.

In [13] four requirements for adequate description of information over Web Services are recognized:

1. High degree of flexibility and expressiveness
2. Ability to express semi-structured data
3. Support for types and subsumption (categorization)
4. Ability to express constraints

UDDI and WSDL indeed don't answer well to these requirements. WSDL provides mostly information on the interfaces of Web services and UDDI's descriptive capabilities are limited to keywords and a simple categorization (which is cumbersome to use). Neither facilitate the representation of semantics. There are also other Web service-related initiatives, like ebXML and RosettaNet, that provide similar technology. Although these initiatives are in some aspects superior to UDDI and WSDL they also don't fulfill the requirements [12].

Because Web services are currently not described in an adequate way heavy human involvement is required. Computers cannot automatically interpret WSDL documents because underlying semantics are not represented nor well-defined. Additional information is needed, accessible via documentation which is only readable/understandable for humans. As such the use of Web services is difficult to automate. The same goes for the use of UDDI registers. UDDI in itself lacks an expressive representation language and human knowledge is required to interpret search results.

4.2 Semantic Web services

The full potential of Web services can only be realized when greater automation of their use becomes possible. More automatic use of Web services requires them to be described in a rich, accurate and machine-interpretable manner. For creating such descriptions Semantic Web technology can be used. The Semantic Web provides excellent means for describing the capabilities and properties of Web services; RDF, RDF Schema and OWL fulfill all four requirements stated above. Moreover, automation in Web Services requires intelligent agents to reason with descriptions and make inferences based upon them. OWL, as described earlier, is grounded in Description Logic and is engineered to enable inference. This makes the

Semantic Web an excellent candidate for enhancing Web Services.

Rich semantic descriptions of Web services could support a greater automation in service discovery, selection and invocation and makes (semi-)automated service composition possible. More down the stretch automation could be possible of things like verification, simulation, configuration, supply chain management, contracting and negotiation of services [15]. When the Semantic Web paradigm is applied to Web services they are called Semantic Web services or Semantic Web Enabled Web services. The figure below, taken from [15] shows where Semantic Web services (here referred to as 'Intelligent Web Services' stand with respect to current standards:

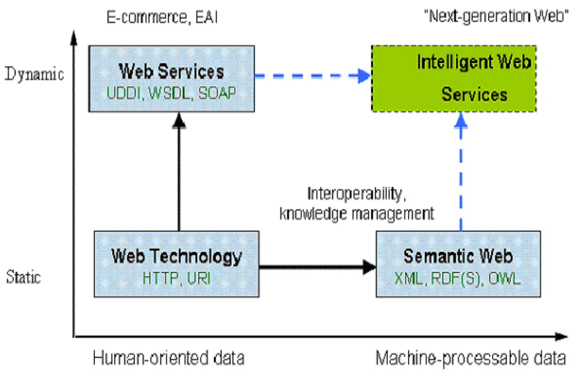

Figure 1: Semantic Web services; "Next generation Web"

For enhancing the current Web services with Semantic Web technology an OWL vocabulary called OWL-S is under development [16]. OWL-S is a service description framework and aims to provide the descriptive facilities that current standards lack, thus enabling the more automatic use of Web services. Because UDDI and WSDL are by now widely accepted standards, OWL-S is at first intended to be used besides them. For this, OWL-S can be mapped to UDDI and WSDL as is shown by [15].

OWL-S consists of three ontologies that describe three different things: service profiles, service models and service groundings. A service profile indicates what a service does for purposes of advertising and matchmaking. The service model (also called process) describes how a service works to enable invocation, composition and monitoring. The service grounding describes the data-formats and protocols used and forms the link with WSDL.

Besides OWL-S a large number of *shared* ontologies will also be needed. Consider the water height example above. Automatic discovery and invocation of a Web Service that, with an up-date frequency of ten minutes, provides water heights of the Rhine at Lobith is only possible if the following shared ontologies[6] are available and used:

- An ontology in which time and up-date frequency are represented
- A geographical ontology in which the river Rhine and city Lobith are represented, or at least the concepts 'river' and 'city'.
- An ontology in which the term 'water height' is represented.

4.3 Semantic Web Services for KANS?

Although Semantic Web enhanced Web Services have great potential, their particular use for KANS will not be further explored in this research. The technology is as yet to immature to be usefully applied:

- Semantic Web Service technology is still in development. OWL-S is still highly immature (as its developers also implicitly acknowledge [17]) and has several accompanying shortcomings. Also besides a language for describing Web services, software is needed to process these descriptions. Such software does not yet exist or it resides in an experimental phase.
- In the (automatic) use of Web Services there is a great need for the specification of rules ("*if* a *then* b"). As we saw earlier, there is no Semantic Web standard for the specification of rules yet and this thus hinders the use of Semantic Web technology for Web Services.
- The development of Semantic Web Services involves much time and effort. Because it is uncertain that the technology will live up to its promises it seems wiser to wait and see what happens in the future. Rijkswaterstaat is not an organization who needs to be on the cutting edge of technology.

Rijkswaterstaat as yet has contact with several contractors for the development of the SRN and GAIN according to UDDI, WSDL and SOAP standards. The Semantic Web approach can exist alongside these standards so it could be applied

[6] Note that these things could be represented in a single ontology but this is unlikely for reasons of scalability and reusability.

in the future without breaking the Web service system that is currently being developed by the contractors.

5. A Semantic Web based Catalogue

Semantic Web technology can play an important role for the KANS information catalogue (ICN). We have already seen that categorization and indexation as well as enhanced information retrieval are prime applications of the Semantic Web. These applications are highly relevant for the ICN as they are for any catalogue.

Although Semantic Web technology is as we saw too immature for use with Web Services it appears mature enough for use within applications like the ICN. The prototype elaborated on below indeed proves this. It is an implementation of the information catalogue based on Semantic Web technology that is currently available.

5.1 Structure of the KANS Information Catalogue

The ICN has two important tasks: it must make explicit what information exists and where it can be found. This means that the entire array of available information within the water management related part of Rijkswaterstaat has to be categorized and indexed. Indeed, an OWL (or perhaps even RDF Schema) ontology can serve well in representing this categorization and indexation and in facilitating enhanced information retrieval of its contents. However, an ontology is not enough; it must be queried and answers must be conveniently presented to users. The prototype Semantic Web ICN consists of three parts:

1 Ontology layer: An ontology will form the knowledge repository that is the essence of the ICN.
2 RDF Query layer: A query engine enables users to query the ontology so that they can retrieve information.
3 GUI layer: The ICN application enables easy access to the knowledge in the ontology through a sophisticated graphical user interface.

5. 2 Ontology layer

The first step in developing the prototype ontology was determining its contents. Besides the contents that are determined by the requirements of the KANS project also other elements might be incorporated. On a meta-level the concepts in the ontology can be divided in two groups. One group consists of all the classes, properties and instances that represent the information the ICN is intended to contain. The other group consists of classes, properties and instances that facilitate the retrieval of elements from the first group.

Group 1: Information-units, systems and meta-data
The primary ICN goal imposes the main content of the ontology: descriptions of the different pieces of information used/generated within Rijkswaterstaat (from now on called information-units), as well as references to the information-systems that provide them[7]. But a library catalogue that only stores book-names with references to shelf-numbers is very limited, additional meta-data as ‘author’ and ‘year’ are also valuable to incorporate. For analogous reasons several meta-data items about information-units and information-systems are represented in the ontology. Examples of meta-data items for systems are name, Website and project leader. Examples of meta-data items concerning information-units are name, up-date frequency and format.

Group 2: Knowledge grid
Besides housing relevant information essential to its primary goal the ICN should also enable this information to be retrieved easily. The ICN ontology must also facilitate easy searching for users with limited specific knowledge. Consider a user searching for information regarding water quality. A query for the term ‘water quality’ will not retrieve the information-unit ‘heavy metal-concentrations’ although it is obviously very relevant. A possible solution for this is to include in the ontology a meta-data field with keywords. However, this approach lacks in intelligence and scalability. A Semantic Web ontology makes it possible to represent the fact that heavy metal concentrations are related to water quality. The query facility of the information catalogue can then exploit this knowledge and retrieve the information-unit. Also it can be represented that cadmium is a kind of heavy metal making it possible that a search for water-quality will result in a reference to the information-unit ‘cadmium concentration’. The part of the ontology that contains all classes, properties and instances that facilitate this enhanced searching is called in this research the *knowledge grid.* It can overcome traditional information retrieval problems inherent to hyponym/hypernym cases.

[7] It is here assumed that information is always produced by some information-system.

Also synonym/homonym problems can be tackled in the knowledge grid; it might for example be represented that mercury and quicksilver are synonyms. An example of a miniscule part of the knowledge grid is given in the following figure:

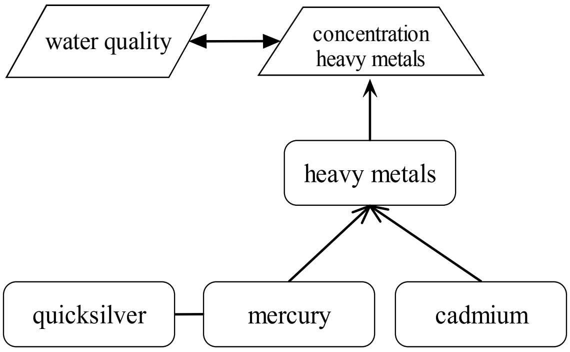

Figure 2: Example knowledge grid

It is possible to use a way of representing concepts in the ontology to make both retrieving and adding information easier. Consider the meta-data element 'business process' that is a property of information-units. The possible values of this meta-data field belong to a predefined space; within Rijkswaterstaat a set of defined business processes is maintained. Such a predefined and maintained set of terms is called a controlled vocabulary. The terms of controlled vocabulary can be represented as instances of a class, thus having a URI in stead of only a string value. This makes input easier since now a simple reference to those instances can be made. Because each term has a URI it also become possible to retrieve only those information-units that correspond to a specific term (thus to a specific URI). This could also be possible under a non-controlled vocabulary that's not represented by URI's, but it would in practice be very difficult. Note that by using URI's it is possible to up-date and change the controlled vocabularies quite easily. Up-dated terms can be given a new URI, thus not breaking references to the URI of the old term. The ICN ontology will use controlled vocabularies for its meta-data fields wherever possible and desirable. Other examples are: units (m., kg, l., etc.), enterprise clusters and enterprise area's of attention.

Determining information-units

The construction of the ontology requires the representation of information from Rijkswaterstaat about information-units and systems. Within Rijkswaterstaat information concerning information-systems is widely available. It is known what information-systems exist and various categorizations of them are already made which can be integrated in the ontology quite easily. Information concerning information-units however does not yet exist. In fact it is not even known what information-units exist at all. This explicit recognition of information-units is new to Rijkswaterstaat, the organizational culture has always been focused on information-systems and not on information itself.

Determining what information-units exist is not a trivial task. Making a categorization of information is difficult in itself due to issues of aggregation. Also, different people have different opinions on what information should be represented in the ICN ontology. This research copes with the aggregation problem in a pragmatic way; recognizing information-units should go as deep as is practically possible and desirable. Problems related to the different visions that people have are more difficult to overcome. The ICN ontology cannot be constructed by one person alone, input from others is needed. It is important that this input is consistent, so a definition to which it must conform was devised in this research:

> *An information-unit is data in processable form, provided by a certain information-system, that can be seen as separate from other data and that has an sich value for Rijkswaterstaat.*

This definition takes a first step in making clear what an information-unit is, and thus what kind of information must be represented in the ICN. The second step is taken by the acknowledgement that an information-unit is in fact a combination of an object (eg. 'water' or 'heavy metals') and information about that object (eg. 'quality' or 'concentration'). Both this vision and the definition took shape through numerous discussions with members of the KANS project team and can considered supported compromises.

Information model ICN

The following figure represents the information model of the ICN ontology. It shows on a meta-level the information and inter-relations that are represented in the ontology:

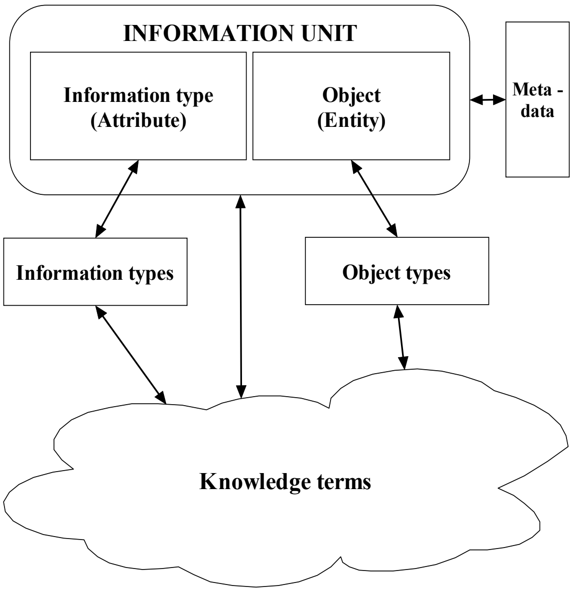

Figure 3: Information Model ICN

Information-units are indeed represented as the combination of an object and an information type (which some prefer to call entity and attribute). Objects and information types are independently represented by two vocabularies. This explicit separation is in accordance with the vision on information-units as described but also enables useful queries to be made. For instance a user can query for all information types regarding a specific object.

Knowledge terms (the collection of which was earlier called the knowledge grid) can be linked with object types, information types and information-units. All these links make the information grid a powerful tool in information retrieval; the result of a query over a knowledge term might point to object or information types, as well as directly to information-units.

The ontology in OWL

Since an OWL ontology is written in XML any basic text editor can be used to create one. This is convenient since it makes it possible for anybody to directly examine and edit an ontology using notepad or so, no special software is needed. Of course it is much more practical to use a specialized tool. Most ontology engineering tools are open-source packages which are installed easily and can be used freely. The ICN ontology in this research was constructed using Protégé 2000 [19]. With the OWL plug-in Protégé facilitates the construction of Semantic Web ontologies. The ICN ontology was created with the following top-level classes:

- `InformationUnit`: represents all information-units
- `InformationSystem`: represents all information-systems
- `ObjectType`: represents all objects (of information units)
- `InformationType`: represents all information types
- `Cluster`: represents all enterprise clusters
- `AreaOfAttention`: represents all enterprise area's of attention
- `AquoTerm`: represents all definitions known as the Aquo-standard
- `BusinessProcess`: represents all recognized business processes
- `KnowledgeTerm`: represents all knowledge terms

These classes might comprise subclasses (most notably the class `KnowledgeTerm`). Most classes are also defined as the domain of certain properties. For example, the meta-data elements of information-units are represented as properties with the class `InformationUnit` as their domain. A number of these properties are:

- `hasObjectType`: represents the object type of the information-unit. Refers to an instance of class `ObjectType`
- `hasInformationType`: represents the information type of the information-unit. Refers to an instance of class `InformationType`
- `hasRelatedKnowledgeTerm`: refers to a related knowledge term represented as an instance of class `KnowledgeTerm`.
- `relatedBusinessProcess`: refers to an instance of class `BusinessProcess`
- `providedBy`: refers to the provisioning information-system represented by an instance of `InformationSystem`).

By defining classes and subclasses multiple taxonomies can be created. Properties can connect concepts in these different taxonomies. This clearly enables a much richer knowledge representation than is possible with taxonomies alone and also enables integration of multiple taxonomies into a single system.

Fully describing all classes and properties of the prototype ontology lies beyond the scope of this paper. Only a short elaboration on the class `KnowledgeTerm` will now be made that suffices in demonstrating the use of OWL for describing information. The class KnowledgeTerm represents the knowledge grid as described earlier. This knowledge grid can be represented in an OWL ontology relatively easily since OWL already provides a number of useful terms. Hyponym - hypernym relations (e.g. ‘heavy metals’ and ‘cadmium’) could be represented by either using `rdfs:SubClassOf` or by declaring something an instance of some class (e.g. ‘cadmium’ as an instance of the class ‘heavy metals’, ‘heavy metals’ here being some subclass of `KnowledgeTerm`). Through `owl:SameAs` synonyms (e.g. quicksilver and mercury) can be represented. Other relations in the knowledge grid might be represented by newly defined properties; for example the property relatedInformationUnit which is the inverse of the property hasRelatedKnowledgeTrem mentioned above.

Protégé 2000 proved a mature and usable tool for engineering the ontology. It can also be used for maintenance and up-dating.

5.3 Repository and querying

As said, for retrieving the information represented in the ontology a query language is needed. A recent survey of different query languages [20] shows that RQL and SeRQL are relatively mature query languages. They score high on most aspects examined in the survey. RQL and SeRQL both support RDF Schema semantics which is highly desirable. Without RDF Schema support a query to retrieve an instance of class A will not retrieve an instance of A’s subclass B, although that is also an instance of A. RQL and SeRQL are both only implemented in an according RDF repository: RQL in ICS-Forth’s RDFSuite [21] and SeRQL in the Sesame [22] repository[8] (SeRQL indeed stands for Sesame RQL). For the prototype ICN in this research the Sesame repository and SeRQL were used.

Sesame is an open-source RDF database system. In this research it was operated on top of a MySQL database. This means that all the RDF statements that together form the ontology are stored in the MySQL in stead of in an XML-file only; by doing this performance advantages of current database systems are achieved. Although not used in this prototype, Sesame offers a flexible access API, that supports access through mechanisms as HTTP and SOAP.

An example of an SeRQL query is presented below. This query retrieves information-units that are related to knowledge term “water quality”. It presents the name, description and providing information-system of those information-units:

```
select distinct Name, Description,
             ProvidingSystem
from {x} <rws:relatedInformationUnit>
                               {InfoUnit}
        <rws:name> {Name};
        <rws:description> {Description};
        <rws:providedBy> {ProvidingSystem}

where x like "*water quality*"
using namespace rws =
<!http://www.rws.nl/ontology#>
```

Using SeRQL we can effectively query the ontology in many ways, a number of examples:

- Retrieve all information-units that are linked with knowledge term ‘heavy metals’
- Retrieve all information-units that are generated by information-system WADI.
- Retrieve all information-units that have object type ‘water’, information type ‘height’ and an update frequency of less then 10 minutes.
- Retrieve all information-units that belong to business process ‘ABP’.

In [23] it is shown that queries related to the graph properties of an ontology cannot be expressed in current query languages (including SeRQL) although that is desirable. It remains to be seen what kind of such queries need to be asked within the KANS ICN. If necessary extensions to SeRQL must be developed.

[8] Actually Sesame supports other query languages as well but SeRQL is intended as its prime one.

5.4 GUI layer

Querying the ontology requires knowledge about the SeRQL query language as well as about the structure of the ontology. While some Rijkswaterstaat employees will indeed possess this knowledge, the typical ICN user will not. For this reason some user-interface must be developed that enables users to query the ontology in an easy way. The GUI shown below does precisely that:

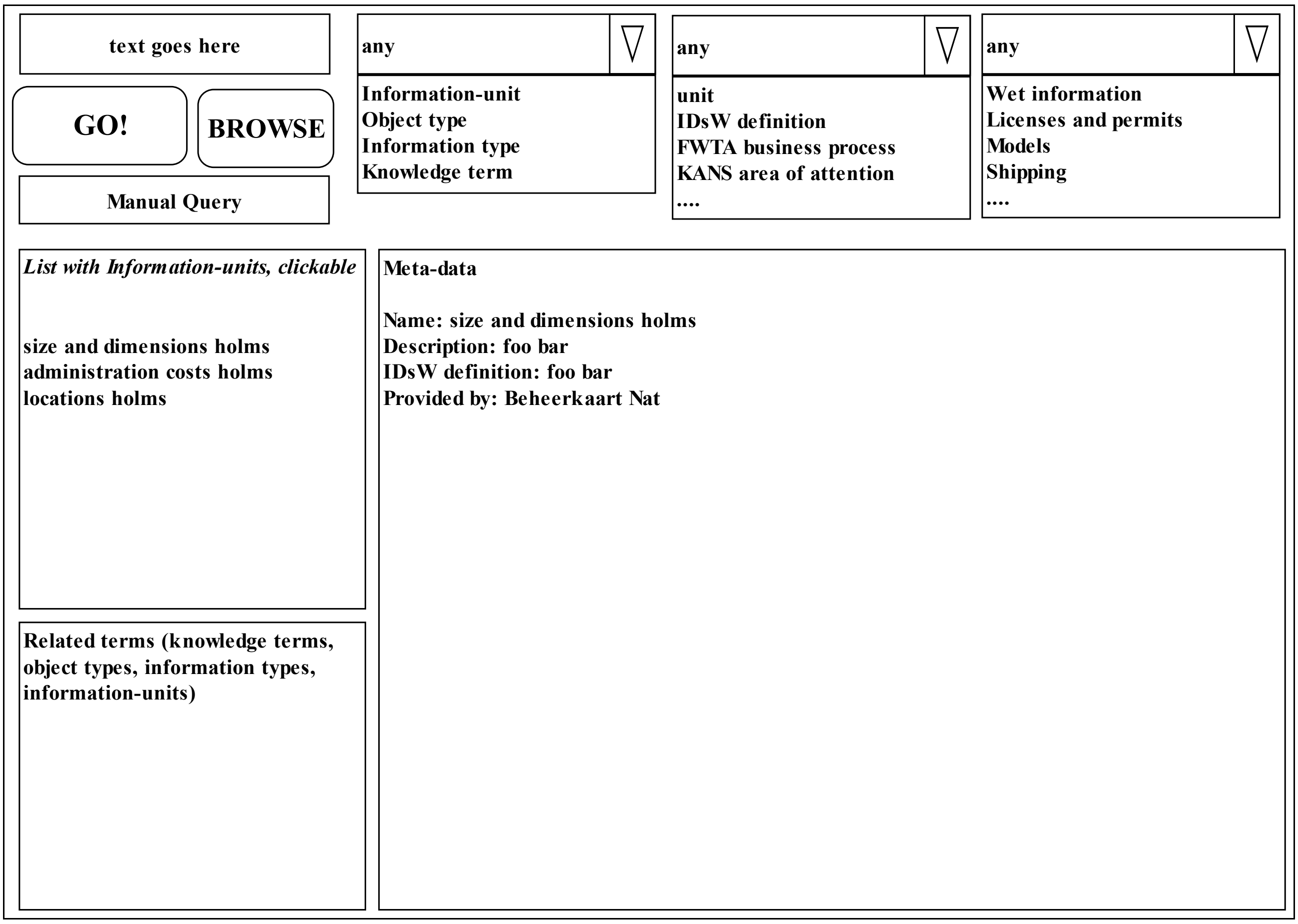

Figure 4: GUI

This GUI facilitates text based querying with support for semantics. The upper left box is a text entry field where a user can type his search term. Through the dropdown box right of it the semantics of the term might be indicated. Also constraints on the query can be conveniently expressed. The bottom three boxes represent the results plane. The list at the bottom left corner of the GUI displays related terms. It shows terms which are connected to the search term in the ontology. This allows for example a user to search for knowledge term 'water' and be shown knowledge terms 'water quality' and 'water quantity'. Clicking an item in this list will initiate a new query with the selected item as search term.

A simple keyword-based approach alone doesn't use the full capabilities of an ontology. It is highly useful to present a visualization of the ontology (as is elaborated upon by [18]). Thus a user can easily asses the knowledge grid and the relations it comprises. The prototype GUI designed in this research supports this visualization, as can be seen in the next figure.

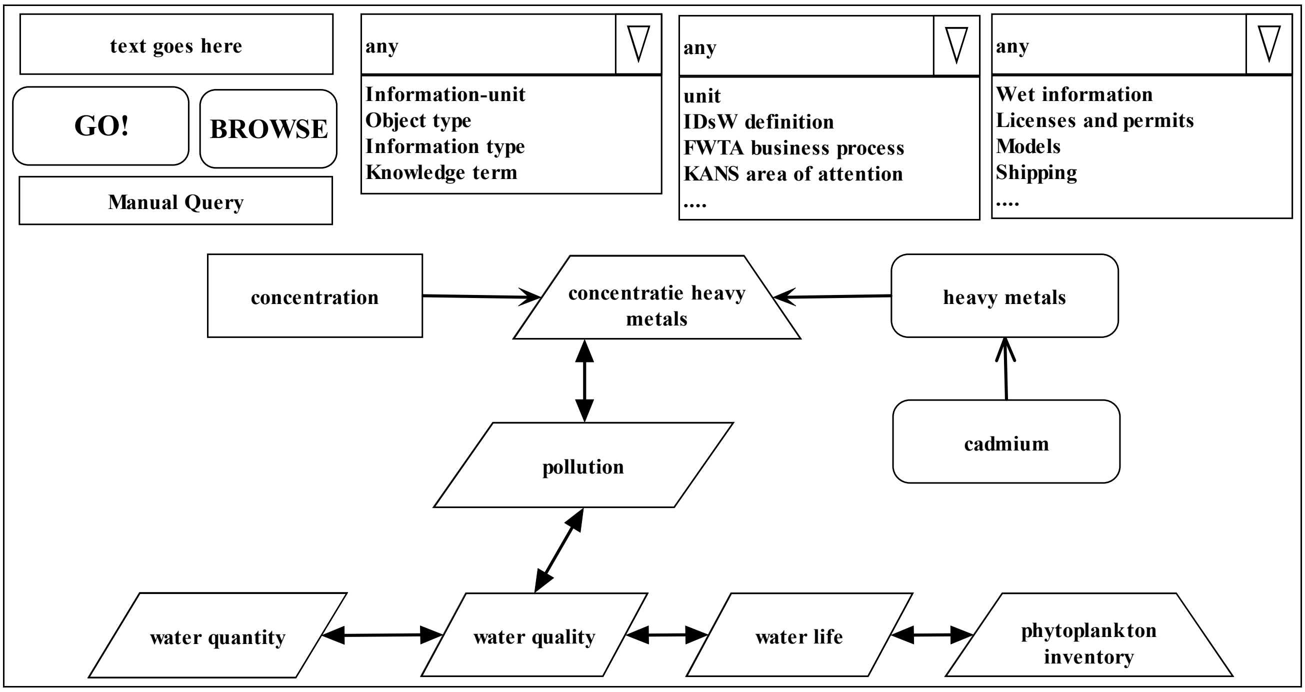

Figure 5: Browsing the ontology

Using a point-and-click method users can browse through the related concepts. Clicking a concept will centralize focus on it and show its related concepts. In the example above focus is set on the knowledge term ‘pollution’.

5.5 Validation of the Prototype

Four criteria for validating the prototype are recognized. The first one is a direct translation from the primary goal of the ICN which states that it will show what information exists and where it can be found:

1. The prototype must explicate the links between information-units and information-systems (what information exists and where it can be found).

The second criterion originates from the intent to test and demonstrate the enhanced information retrieval capabilities of the Semantic Web:

2. The prototype must facilitate searching for information-units and information-objects using an approach that enables queries to be made not only for their particular names, but also for related terms. Traditional information retrieval problems (hyponym, hypernym, synonym, homonym) must be sufficiently tackled for effective retrieval of information-units.

The third criterion is also related to enhanced information retrieval and aims to exploit the structured way in which ontologies store information:

3. Query facilities must allow the specific retrieval of information that satisfies certain criteria. Constraint based searching must be possible (eg. retrieve only information-units provided by information-system “A” and that are related with business process “B”).

The fourth criterion originates from the acknowledgement of the typical ICN user that has no knowledge of the Semantic Web. Technical complexities must be hidden to users.:

4. Easy searching must be facilitated, users with no understanding of the Semantic Web and the structure of the ontology must still be able to use the Semantic Web ICN.

Validation

For validating the prototype it will be determined whether it satisfies the criteria. The validation is here first represented in tabular form after which a more extended explanation is then given.

Criteria:	**Satisfaction**:
1: What information and where?	Yes: in ontology instances of information-units are linked with instances of information-systems.
2:Enhanced information retrieval	Yes: related terms are represented and traditional information retrieval problems are tackled.

3:Constraint based searching	Yes: in the ontology everything structurally links with everything else.
4: Non-expert searching	Yes: GUI facilitates information retrieval for users with no knowledge about the Semantic Web

The ICN prototype clearly satisfies criteria one; in the ontology instances of information-units are explicitly linked with instances of information-systems. The second criteria is also satisfied. Traditional information retrieval problems are tackled in the ontology through categorizations of hyponyms and hypernyms as well as through the representation of homonyms and synonyms. In the ontology related concepts are represented thus supporting them to be retrieved more easily. Because an ontology in cooperation with an adequate RDF query language/engine (eg. SeRQL) behaves much like a relational database constraint based searching is possible, thus satisfying the third criteria. The graphical user interface hides technological complexities of the Semantic Web and facilitates information retrieval for users unfamiliar with the technology.

The prototype indeed shows that Semantic Web technology has advantages for use within an application like the information catalogue. Semantic Web technology can be applied to facilitate advanced information retrieval in such applications. An ontology provides a natural place to put knowledge in. The more knowledge is put in the ontology the less knowledge is needed for querying it.

Maturity of the Semantic Web ICN

As described in this paper, Semantic Web technology is yet too immature to play a role in the use of Web Services. At the beginning of this section it was said that the technology did appear mature enough for use within the KANS ICN. The question whether this assumption was correct can now be answered:

Ontology construction and repository: There already exists a number of relatively mature tools for the engineering of ontologies; Protégé 2000 used in this research is a fine example. Also there exist software products that can combine an OWL ontology with the performance of relational database systems thus providing a usable repository.

Querying: Although there exists no standard query language, mature ones are available that can be used in knowledge management applications like the ICN. Both SeRQL and RQL support RDF Schema and can be used in this light. Queries related to graph properties are not supported.

GUI: There are no standard or off-the-shelve Graphical User Interfaces for visualizing, browsing and querying ontologies. Thus custom GUI's must be developed.

It can be concluded that at it current state Semantic Web technology is mature enough to be used within applications like the KANS ICN. The lack of a standardized query language is not a big problem, only custom GUI's will have to be developed.

5.6 Semantic Web in perspective

Using Semantic Web way technology in building the information catalogue has many benefits and opens up new possibilities. Much attention has been paid to these benefits and possibilities in this paper so far. A short summary:

- Structured information storage and knowledge representation
- Structured and enhanced information retrieval
- Standardized explication of semantics
- Computer inference
- Reducing knowledge need for users
- Convenient extensibility of data model (ontology)
- Easy storage and access of information in XML files
- Coexistence with other XML related technology

Of course using Semantic Web technology has downsides as well. Currently its immaturity is perhaps the most dominant one. But looking further also other issues exist for applying the Semantic Web. With respect to applications related to information exchange it can be observed that these require ontologies to be *shared*. Sharing ontologies between different businesses or organizations is not trivial. It requires organizations to make agreements. This is comparable with the negotiation effort needed for organizations before exchanging XML-files. However since RDF is more flexible then XML the negotiation process is easier. In general, the Semantic Web would increase in value when a large number of standardized ontologies were available. Such ontologies can then be used by organizations sparing them from costly and time-consuming development. Efforts are underway in creating such ontologies but their number is small. It can be expected that more of these initiatives shall emerge when the Semantic Web gains more momentum.

For effective Information and Knowledge Management ontologies have to be shared by all layers in an organization. People 'in the field' who generate data and information must adhere to the ontologies as well as people who manage information. Classification schemes (ontologies) can only be of real value when they are supported within the entire organization.

In general it can be said that use of Semantic Web technology requires substantial effort and commitment. This will come automatically when the Semantic Web gains more momentum. Relevant for attaining momentum are network effects; when the user community of the Semantic Web grows the value of the technology will also grow attracting more users in turn. It remains to be seen what the critical mass for the Semantic Web movement will be.

A practical downside of the Semantic Web is that the three languages RDF, RDFS and OWL are not purely layered on top of each other. Even more, there exist three species of OWL. The reason for this all has to do with computability problems concerning inference engines. To guarantee computability (termination of inferences in finite time) constrained versions of OWL must be used. In such constrained versions it might not be possible to represent knowledge as one might wish.

A final important observation is that the Semantic Web is *facilitative* technology. It enables knowledge and information to be stored and retrieved in an advanced way. The Semantic Web cannot be used for determining what knowledge and information exists within an organization. This is a difficult task as was experienced with the prototype construction; defining information-units proved a non-trivial undertaking.

Alternatives

Indeed, an ontology provides a very natural place to put knowledge in but as just showed the Semantic Web has downsides as well. The KANS ICN could also be implemented using other technologies. A full comparison lies beyond the scope of this paper. These remarks suffice:

- HTML pages: This would be easy to construct and easy to access and use. It doesn't provide enhanced information retrieval nor the structured information representation of a Semantic Web ontology. Note that this approach could possibly be augmented by the use of technology like topic maps [24] to improve information retrieval but this would mean much extra work for doing something an ontology does by itself.
- 'Simple' XML files: This would create a structured data format for information that would enable more structured information retrieval. RDF has a number of advantages over XML though [25]. For example the data model of an XML file always adheres to a tree-form; thus only taxonomies can be represented accurately. In general XML technology is more rigid than RDF technology and is therefore more difficult to use.
- Relational databases: Relational databases are proven technology and offer structured data storage. Semantics related advantages of the Semantic Web (such as enhanced information retrieval) cannot be achieved. Also, working with Semantic Web files is easy since they are basic text files, data stored in databases is not so easily accessible. Relational databases can have a role in Semantic Web based systems though, as the prototype developed in this research demonstrates.

6. Conclusions

This research was focused on answering the question: *"In which way can Semantic Web technology improve information exchange and information retrieval ('searchability') as aimed for in the KANS project?"* As said, this main question falls apart in two separate questions, treated below.

Question 1: "What can Semantic Web technology do in the area of information-exchange for the KANS project?"

The part of the KANS project concerned with information-exchange was about the implementation of a Web Service based system. Current Web Service standards mostly lack the ability to describe Web Services in an accurate and semantically rich way. Semantically rich descriptions of Web Services could enable a much greater automation of their use. Semantic Web technology could be used to provide for such descriptions and would enable intelligent agents to reason with these descriptions.

At this point in time Semantic Web technology is not mature enough to use in this light. Firstly, an ontology which can be used to describe Web Services is needed. OWL-S is currently being developed for this but it is not mature enough. Secondly, there are no software/inference engines available to truly exploit Semantic Web Service descriptions. Software is needed that understands the Semantic Web descriptions of Web Services and is able to process them automatically. The same goes for other applications that require intelligent agents to infer and act with respect to Web Services. Fortunately, Semantic Web technology as applied to Web Services can be applied alongside current Web Service standards as UDDI, WSDL and SOAP so when the time comes it can be gradually implemented without breaking current Web Service systems.

It can be said that Semantic Web technology is too immature to be used for Enterprise Application Integration at all. These applications require an advanced use of the Semantic Web, one where intelligent agents make inferences. As yet such an advanced use is not possible with the technology. When it becomes possible remains to be seen; first serious business use within 10 years is a popular estimation.

Recommendation: Though not a viable option now, organizations like Rijkswaterstaat should monitor the evolution of Semantic Web Services. They may have tremendous added value for Web Service systems that are currently being developed. Semantic Web services may drastically reduce the complexity of the information system landscape within organizations.

Question 2: "What can Semantic Web technology do in the area of information searching for the KANS project?"

The second part of the KANS project was concerned with information retrieval searching) embodied by the creation of an information catalogue. Semantic Web Ontologies form an excellent knowledge base for the catalogue. An ontology provides a structured and extendable place where information can be stored. Because it explicates semantics an ontology has enhanced information retrieval possibilities as compared to text-based (syntax) search mechanisms. Traditional information retrieval problems related to synonyms, homonyms, hypernyms and hyponyms can be overcome in a Semantic Web ontology.

Since Semantic Web technology for creating and working with ontologies seemed mature enough to be used, a prototype was constructed in this research. Technology for creating ontologies indeed exists, as well as for querying them. There exists yet no standard query language but this posed no problem for our prototype as the language used proved sufficiently capable. The prototype shows how Semantic Web technology can be used in making an information catalogue that scores high on 'searchabilty'. The construction of the prototype showed that there exists as yet no off-the-shelve tools for adequately visualizing ontologies. However, information retrieval from ontologies will be much more powerful when adequate visualizations of the ontology can be shown to users. Despite some immaturities and with some programming effort, a useable information catalogue based on Semantic Web technology can already be built as the prototype demonstrates.

It can be said that Semantic Web technology is mature enough for use within any Information or Knowledge Management application. The Semantic Web provides languages that enable semantically rich descriptions of information that can be exploited quite easily by such applications. When the technology matures further the Semantic Web may well become a standard for applications in the Information and Knowledge Management field.

Recommendation: For creating an information catalogue that is highly searchable the Semantic Web is a valuable option. There exist few other technologies that offers the enhanced information retrieval of the Semantic Web; those that exist are inferior in important aspects. If 'searchability' is a determining factor for the information catalogue, the Semantic Web option cannot be overlooked.

Future research

This research had an explorative nature and opens a number of options for future research. Most notably perhaps is the development of a *fully working* prototype. The prototype in this research has a conceptual nature and has paved the way for the implementation of a real working system. The construction and use of a fully working system will no doubt lead to valuable insights concerning the Semantic Web as an enhancer of information retrieval. With the further developed prototype workshops can be held with potential users to gain more practical insights.

Other applications of the Semantic Web might be further analysed within a real-life context like the KANS project. For organizations like Rijkswaterstaat the abilities of the Semantic Web to integrate data might for example be very valuable to explore. When Web Service systems are fully implemented research can be carried out to study how they can be further enhanced by Semantic Web Services. For Rijkswaterstaat, the Semantic Web could be the link between the

Web Service system and the information-catalogue the KANS project ultimately seeks.

[1] http://www.w3.org/2002/ws/Activity
[2] http://www.w3.org/2001/sw/
[3] http://www.w3.org/TR/rdf-concepts/
[4] http://www.w3.org/TR/rdf-schema/
[5] http://www.w3.org/TR/owl-features/
[6] I. Horrocks, P.F. Patel-Schneider and F. van Harmelen, "From SHIQ and RDF to OWL: The Making of a Web Ontology Language", 2003, http://www.cs.vu.nl/~frankh/postscript/JWS03.pdf
[7] F. van Harmelen and G. Antoniou, "A Semantic Web Primer", 2004, MIT Press
[8] D. Fensel, Ontologies, a Silver Bullet for Knowledge Management and Electronic Commerce, 2001, Springer-Verslag
[9] http://www.w3.org/TR/soap/
[10] http://www.w3.org/TR/wsdl
[11] http://www.uddi.org
[12] V. Terziyan and O. Kononenko, Semantic Web Enabled Web Services: State-of-Art and Industrial Challenges, 2003, http://www.cs.jyu.fi/ai/papers/ICWS-2003.pdf
[13] D. Trastour, C. Bartolini and J. Gonzalez-Castillo, "A Semantic Web Approach to Service Description for Matchmaking of Services", 2001, http://www.semanticweb.org/SWWS/program/full/paper52.pdf
[14] D.Martin, M. Paolucci, S. McIlraith et all., "Bringing Semantics to Web Services: The OWLS Approach", 2004, http://www.daml.org/services/owl-s/OWL-SWSWPC2004-CameraReady.doc
[15] D. Fensel, "Intelligent Web Services", 2002, Microsoft Powerpoint Presentation at http://informatik.uibk.ac.at/users/c70385/ftp/slides/finland.ppt
[16] http://www.daml.org/services/owl-s/
[17] The OWL Services coalition, "OWL-S: Semantic Markup for Web Services", 2004, http://www.daml.org/services/owl-s/1.0/owls.pdf
[18] C. Fluit, M. Sabou and F. van Harmelen, "Supporting User Tasks through Visualisation of Light-weight Ontologies", 2003, http://www.cs.vu.nl/~frankh/postscript/OntoHandbook03Viz.pdf
[19] http://protege.stanford.edu/
[20] P. Haase, J. Broekstra, A. Eberhart and R. Volz, "A Comparison of RDF Query Languages", 2004, http://www.aifb.unikarlsruhe.de/WBS/pha/rdf-query/rdfquery.pdf
[21] http://139.91.183.30:9090/RDF/index.html
[22] http://www.openrdf.org
[23] R. Angles, C. Gutierrez and J. Hayes, "RDF Query Languages Need Support for Graph Properties", 2004, http://www.dcc.uchile.cl/~cgutierr/ftp/graphproperties.pdf
[24] http://www.topicmaps.org
[25] http://www.semaview.com/img/rdfxml_web.gif

Process Report

Introduction

In this report a description is made of the research process of my final graduation project. It accompanies and complements a scientific report which focuses on the actual content and findings of my research. This report describes the describes the way and the setting in which I have performed my research.

1. Preliminary course of actions

Choosing a graduation project

Deciding what kind of graduation project I wanted to do was difficult in itself due to the fact that my interests are broad. Reflecting on this broad interest was my list of elective courses, it also had a broad nature. In shaping my preferences and the quest for a graduation project I talked with different people. A number of possibilities arose, the two most prominent options were a project at Rijkswaterstaat and one at the municipality of Delft. Both organizations had an interesting assignment to embark upon. Choosing between them was difficult but after careful consideration I decided to go for the project at Rijkswaterstaat. I made this choice for two reasons. The project of the municipality of Delft was in my opinion rather unclear. Besides this the Rijkswaterstaat organization appeared more professional to me, better enabling me to do a more professional research.

The first research proposal

The assignment at Rijkswaterstaat concerned a comparison of middleware technology. A project at Rijkswaterstaat called KANS aimed to implement a middleware solution in order to solve problems related to information exchange. Because it was uncertain which particular middleware solution would be best there existed a need for an exploration and comparison of options. After reading some relevant literature and KANS documentation I wrote my first research proposal and went to work at Rijkswaterstaat. The first two weeks I spent deepening my knowledge on middleware and read more KANS documentation. Also I talked with people involved with the KANS project. It became clear to me that Rijkswaterstaat had already unofficially chosen to implement Web Service middleware. This choice had considerable momentum and the need for a comparison of middleware technologies was in essence non-existent. After realizing this two options arose. I could continue the middleware research but now in light of finding a justification for the choice of Web Services. Also, I could abandon the middleware research and try to find a new research topic within the KANS context. Justifying the choice Web Service seemed not very interesting so I decided to look at alternative topics.

The second and final research proposal

Through talking with my supervisor from university who was also involved with KANS I became aware of Semantic Web technology. The Semantic Web seemed to be of potential value for KANS. There was little knowledge within Rijkswaterstaat about the Semantic Web, therefore doing an exploration of this technology would have great value. After consulting my other supervisors as well as different people involved in the KANS project I decided to indeed do a research into the Semantic Web. I wrote a new research proposal centred around the question: *"In which way can Semantic web technology improve information exchange and information retrieval as aimed for in the KANS project?"*. The new research proposal was then officially approved by the graduation committee.

Later on my research into the Semantic Web gained additional value when the 'trend-watch' section of Rijkswaterstaat contacted me. The 'trend-watchers' were concerned with spotting and exploring new technologies. They had been planning an exploration of the Semantic Web but didn't have the time and manpower to conduct such a research. When they learned that a student of the TU Delft was doing precisely that research for a Rijkswaterstaat project they were very pleased. I met up with them and we agreed that they'd use my report and findings to write their own 'trend-watch' report.

2. Research into the Semantic Web

The research I conducted consists of a number of phases. In practise these phases were not fully subsequent but got intertwined and ran parallel at times. For descriptive purposes I will treat them separately below. First I will describe some complementary work I did at Rijkswaterstaat.

Working with the KANS team

Besides doing my research I also helped with the KANS project team. Besides more general KANS matters I was mostly involved with the team concerned with building an information catalogue. I took part in numerous discussions and helped answer a number of questions, among these:

- What information should be contained in the information catalogue?
- What is an information-object[9]?
- What kind of meta-data should be represented for information-objects?
- What should the information catalogue look like?
- How should information-objects be represented in the information catalogue?
- What is the link between the information catalogue and the service register[10]?

Besides the fact that being part of the KANS team was interesting in itself, it also provided valuable insights into the project and Rijkswaterstaat. Being closely involved enabled me to gain a thorough understanding of the possibilities of using Semantic Web technology within KANS. I got to understand the problems that existed in a project like

---

[9] In my research paper information-objects are called information-units. I believe unit is a more accurate and unambiguous description than object. The original name as used in the KANS project is information-object though.

[10] The service register is a computer-readable catalogue of Web Services which is part of the Web Service architecture as implemented in KANS.

KANS and could adapt my Semantic Web research to them. Also by working in KANS I got to understand its needs, requirements and constraints much better than I would have without this high level of involvement. Another benefit of working in the KANS team was that I could discuss the Semantic Web research with my team members and receive their valuable feedback.

All in all I believe that the close involvement I had with KANS enabled me to do a more qualitative research than I could have done without it. Theoretical thinking and practical appliance are easily misaligned, the close involvement enabled me to prevent this.

Not only the KANS team provided me with feedback on my research. The members of the exam committee that were not involved with KANS were also very helpful in providing advice. Due to the independence and distance from KANS their advices and feedback were very valuable.

Literature study

At the beginning of my research I knew very little about the Semantic Web. This required me to do an extensive literature study. Understanding the Semantic Web is difficult because it encompasses many aspects. But basically Semantic Web technology can be seen as a combination of a modelling paradigm and Internet technology. At the university I had already met often with both Internet technology and modelling paradigms which thus made understanding the Semantic Web easier.

A difficulty I experienced during the literature study was that there only exists little literature on the Semantic Web. Just a handful of books exist that elaborate on the Semantic Web. Unfortunately most of them are too high-level or too application specific making it difficult to apply to the KANS project. The first book that describes the Semantic Web in detail has only recently been published[11]. Luckily there exist lots of scientific papers available on the Internet that deal with different aspects of the Semantic Web. Also the official W3C Website provides valuable material for understanding RDF, RDFS and OWL.

I've studied almost all available material of significance and have formed an elaborate picture of the possibilities of the Semantic Web for KANS. It can be of use for both the Web Service Architecture and the information catalogue. Since I determined in the literature study that the technology as yet is too immature for use with Web Services I shifted full focus to analysing its application for the information catalogue.

Besides material on the Semantic Web I also read many KANS documents in order to gain a deeper understanding of what had to be done in the project. Complementary to this I performed a causal analysis for determining success factors for the information catalogue. Also I've made an actor analysis to chart and understand organizational issues that might pose problems for application of the Semantic Web.

---

[11] "A Semantic Web Primer" by Frank van Harmelen

A prototype Semantic Web information catalogue

When I realized that the information catalogue could be represented by an ontology I constructed several draft ontologies. Through my work in the KANS team I knew broadly what kind of information they should contain and what kind of knowledge they should represent. I also examined several Rijkswaterstaat data sources to make sure that their data-model could be copied or reused in the ontology. This would make transfer of information into the ontology and its use afterwards easier. A Python script was used for extracting all definitions from the Aquo standard[12] and putting them in the ontology according to the same data-model. After constructing the draft ontologies and examining their pros and cons I constructed the final ontology for the prototype.

Many languages exist for querying the ontology. I experimented with a few promising ones and found, in accordance with other research, that SeRQL is a very capable language. Thus I choose to implement the prototype with SeRQL. SeRQL is implemented in the Sesame RDF repository so I had to install it. Installing Sesame and getting it to work is not very simple (the same goes for most current Semantic Web software): the Java SDK must first be installed as well as the Tomcat application server, some .jar files have to be copied into another folder and a MySQL database must be installed in combination with a Java MySQL API. Finally when it all run SeRQL could be used on the ontology and did indeed work fine.

Based on possible usage scenario's for the information catalogue and keeping in mind the possibilities of the ontology in combination with SeRQL I developed a GUI. Through this GUI users should be able to easily retrieve information from the ontology. The first version of the GUI combined a keyword based information retrieval approach with the advantages of an ontology. Because it is useful for information retrieval I also added a visualization option to the GUI.

All the knowledge I had acquired about The Semantic Web and KANS enabled me to determine relevant criteria for validating the prototype. After its validation I extrapolated the findings to draw conclusion about the broader applicability of the Semantic Web, thereby concluding my research.

[12] The Aquo standard is a list of all definitions used within the 'wet' Rijkswaterstaat.