Chapter 5

Associative Writing and the Semantic Web

The Semantic Web is a vision of the future of the Web, instigated by Tim Berners-Lee: "The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation" (Berners-Lee et al., 2001). Most of the content on the Web is designed for humans to read, not for computer programs to manipulate meaningfully, since there is no reliable way for computers to process the complex semantics of the text or its impenetrable layout. Computers are therefore only used as devices which process and render information — they don't have access to the knowledge contained within it. The vision of the Semantic Web is a global and intelligently linked *knowledge* base (where machines can help users to retrieve information from the Web), as opposed to the disjointed and poorly understood *document* base which we use today.

Since the aim of this work is to examine the issues and problems surrounding associative writing in the context of the Web, it is important to consider this next major stage in the evolution of the Web — an evolution which the recent work in this area (Horrocks and Hendler, 2002) shows is already gathering much momentum. The aim of this chapter, therefore, is to position this work (and other work related to Associative Writing) relative to the current state of Semantic Web research.

Early Semantic Web initiatives focus purely on using machine-understandable knowledge to help users find the facts they are looking for. In contrast, Associative Writing focuses on links as the 'glue' integrating new contributions with existing work in the Web, in an attempt to provide a deeper (human) understanding of the topic in hand. More recent Semantic Web work, however, seems to demonstrate how both machine-understandable knowledge and human-understandable hypertext can be brought together. In order to expand on this perspective, it is necessary first to briefly outline the technology underpinning the Semantic Web in more detail.

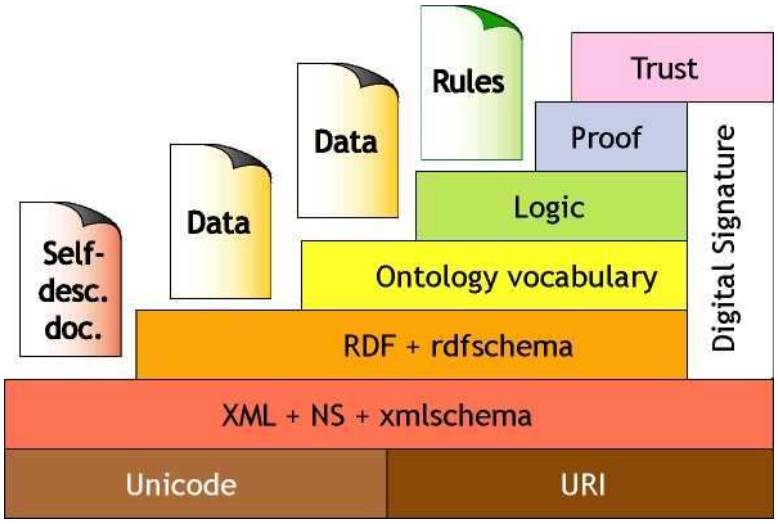


FIGURE 5.1: Semantic Web architecture (Berners-Lee, 2000).

5.1 Overview of Semantic Web Infrastructure

The infrastructure of the Semantic Web has been described as consisting of several discrete, interacting layers (Figure 5.1); this section focuses on the 3 major enabling technologies embodied in these layers:

1. Knowledge Representation (XML and RDF layers)
2. Ontologies (Ontology layer)
3. Agents (Logic, Proof, and Trust layers)

5.1.1 Knowledge Representation

The function of the Semantic Web relies on *knowledge representation*: computer-accessible structured collections of information that can be used to conduct automated reasoning (Berners-Lee et al., 2001). Two important technologies for knowledge representation on the Semantic Web are already well defined: the eXtensible Markup Language (XML), and the Resource Description Framework (RDF) (W3C, 1999b).

XML allows users to add arbitrary structure to their documents, which can be used by computers in sophisticated ways. However, the writer of the computer program must know what the creator of the document has used each structure for — XML alone does not capture information about what the structures *mean*. RDF, however, does capture meaning — users use *metadata* ('data about data') to describe Web resources and improve a machine's 'understanding' of them. The XML syntax is not redundant, since it can be (and often is) used to encode RDF.

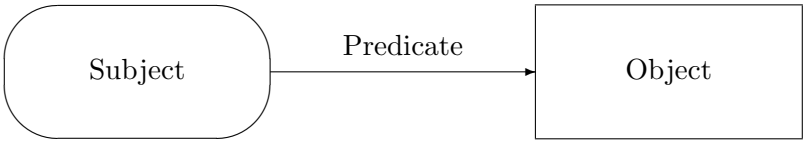


FIGURE 5.2: Simple RDF triple template.

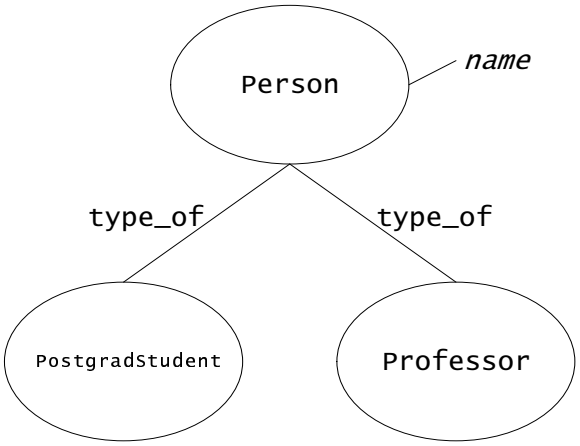
Meaning is expressed in RDF as encoded triples (Figure 5.2) which make assertions that particular resources (subjects) have properties (predicates) with certain values (objects). A resource may be an entire Web page, a part of a Web page, a collection of pages (for example, an entire Web site), another RDF triple, or even non-electronic media such as a printed book. A property is a specific aspect, characteristic, attribute, or relation used to describe a resource. As an example, the Dublin Core Metadata Element Set (DCMI, 1999) is a simple vocabulary (which can be encoded in RDF) for describing fundamental properties of Web resources such as the *creator* ("The person or organization primarily responsible for creating the intellectual content of the resource") and *date* ("A date associated with an event in the life cycle of the resource"). By agreeing on shared metadata vocabularies, machines can infer relationships between resources because their metadata declares that they describe the same concept. The triples of RDF therefore form webs of information about related things.

5.1.2 Ontologies

Knowledge representation technologies such as XML, RDF and Dublin Core allow machine-processable metadata to be added to Web resources — because the metadata declares that the resources describe the same concept, machines are able to infer relationships between them. In order to interconnect resources describing *different* concepts, however, machines need to be able to discover common meanings between different metadata definitions. The next important enabling technology for the Semantic Web is therefore collections of information which describe common meanings and relationships between resources on the Web, called *ontologies*. In philosophy, an ontology is a theory about the nature of existence, of what types of things exist; ontologies as formal models facilitate knowledge sharing and reuse, and enable reasoning over concepts and objects that appear in the real world (Uschold and Gruninger, 1996; Guarino, 1998; Noy and McGuinness, 2001).

Semantic Web researchers use the term "ontology" to describe a document which formally defines the relations among classes of objects (concepts). Typically, Semantic Web ontologies consist of a taxonomy (the set of concepts and the relations among them) and a set of axioms (simple rules governing relations). Ontologies can therefore be used to relate the information on a Web resource to associated knowledge structures.

The RDF Schema language (RDFS) (W3C, 2002) provides a means for encoding ontolo-




```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
  <rdfs:Class rdf:about="http://www.example.com/example-ns#Person">
    <rdfs:label xml:lang="en">Person</rdfs:label>
  </rdfs:Class>
  <rdfs:Class rdf:about="http://www.example.com/example-ns#PostgradStudent">
    <rdfs:subClassOf rdf:resource="http://www.example.com/example-ns#Person"/>
    <rdfs:label xml:lang="en">PostgradStudent</rdfs:label>
  </rdfs:Class>
  <rdfs:Class rdf:about="http://www.example.com/example-ns#Professor">
    <rdfs:subClassOf rdf:resource="http://www.example.com/example-ns#Person"/>
    <rdfs:label xml:lang="en">Professor</rdfs:label>
  </rdfs:Class>
  <rdf:Property rdf:about="http://www.example.com/example-ns#name">
    <rdfs:label xml:lang="en">name</rdfs:label>
    <rdfs:domain rdf:resource="http://www.example.com/example-ns#Person"/>
  </rdf:Property>
</rdf:RDF>
```

FIGURE 5.3: Encoding a simple ontology using RDF Schema.

gies by extending the RDF model to enable a collection of resources to be described according to a simple class hierarchy. Other more complex languages for encoding ontologies include DAML+OIL (van Harmelen et al., 2001) and Web Ontology Language (W3C, 2003). As a simple example of the RDFS encoding, Figure 5.1.2 shows a simple ontology defining three concepts (*Person*, *Student*, and *Professor*), and the relations among them (*Students* and *Professors* are *types of Person*, and so inherit the *name* property).

5.1.3 Agents

We have now seen how knowledge on the Web can be represented using metadata vocabularies such as RDF, and interconnected using ontologies formalised using languages such as RDF Schema. However, the real power of the Semantic Web will be realised when autonomous computer programs (or *agents*) are created that collect and reason over diverse Web content, exchanging data and working cooperatively with other agents. Agents may form a "value chain", in which sub-assemblies of information are passed from

```
<HTML>
<BODY>
<H1>Home page of Jane Smith</H1>
<P>I am a postgraduate student and my supervisor is John Doe</P>
<USE-ONTOLOGY
  ID="cs-dept-ontology"
  PREFIX="cs"
  URL="http://www.example.com/cs-dept-ontology">
 <CATEGORY NAME="cs.PostgradStudent">
 <INSTANCE KEY="http://www.example.com/pg/jsmith/">
   <RELATION NAME="cs.name">
    <ARG POS="TO" VALUE="Jane Smith">
   </RELATION>
   <RELATION NAME="cs.supervisor">
    <ARG POS="TO" VALUE="http://www.example.com/staff/jdoe/">
   </RELATION>
  </INSTANCE>
 </BODY>
</HTML>
```

FIGURE 5.4: Embedding knowledge in the content of a Web page using SHOE markup.

one agent to another, each agent "adding value" by deducing new knowledge according to its own internal reasoning. Semantic Web researchers predict the emergence of a "services architecture" (McIlraith et al., 2001), advertising the services of "trillions of small specialised reasoning services" (Fensel, 2000).

5.2 Semantic Web Initiatives

This section briefly describes some of the initiatives and research projects that have given momentum to the Semantic Web vision.

5.2.1 SHOE

Simple HTML Ontology Extensions (SHOE), an early and influential initiative, allows authors to mark up their Web pages according to a domain-specific ontology by embedding special SHOE constructs in HTML markup (Luke et al., 1996; Heflin et al., 1998). Subsequently, a crawler agent gathers marked up knowledge from the pages, and populates a central knowledge base. By visiting a SHOE *portal*, users can access this knowledge and use sophisticated queries to discover new facts about the domain.

Figure 5.4 demonstrates how knowledge can be embedded in a Web page using SHOE. The SHOE constructs first define ontology being used, and then describes the page as an instance of the *PostgradStudent* concept. The *name* and *supervisor* of this instance are then declared. The latter property describes a relation between this *PostgradStudent* and the concept instance identified by the key `http://www.example.com/staff/jdoe/`.

5.2.2 OntoBroker & $(KA)^2$

The OntoBroker project (Fensel et al., 1998) uses a similar approach to SHOE, and has been used to model the Knowledge Acquisition (KA) research community (Benjamins et al., 1998). The $(KA)^2$ initiative (Knowledge Annotation for Knowledge Acquisition) aims to support this community in building a knowledge base of its own research by populating a shared ontology. The knowledge base is constructed by authors embedding OntoBroker markup (analogous to HTML `META` tags) in their Web pages, which is gathered by an OntoBroker crawler. As with SHOE, the result is a portal where users can access and query this knowledge.

5.2.3 ScholOnto

The Scholarly Ontologies project (ScholOnto) uses a Semantic Web approach to provide tools for tracking debate and analysing ideas in the domain of scholarly literature (Buckingham-Shum et al., 2000). At the heart of the project is an ontology-based *claims server*, ClaiMaker (Li et al., 2002), which manages a shared semantic web of claims, discourse, and perspectives.

Using a Web form-based interface (Figure 5.5a), users can add or import metadata about a document, and create *concepts* (succinct summaries of a document's contribution to the literature), *claims* (how concepts relate to the literature in general, which may support or contest existing claims made by other users — there is no requirement for consensus), and *sets* (collections of concepts and claims). This "claim space" forms a semantic web of inter-linked concepts across the literature, and provides the basis for literature-wide interpretation and analysis using visualisations (Figure 5.5b), hypertext navigation, and sophisticated ontology inference rules to answer questions regarding intellectual lineage (*Where did this idea come from, and has it already been done before?*), the impact of ideas (*What reaction was there to this idea, and has anyone built on it?*), and inconsistencies (*Is there contrary evidence to this claim?*). Different perspectives (schools of thought) can also be identified using ClaiMaker — a particular perspective comprises a common set of concepts and claims on which a number of researchers have built their work; the concepts and claims that they collectively contest represent an alternative perspective.

Concepts and claims are modelled according to an ontology of rhetorical relations (Figure 5.6). In contrast to SHOE and OntoBroker which use a domain ontology to model *what* is being discussed (which, by the nature of research, may be in constant flux), ScholOnto focuses on discourse and argumentation — *how* the domain is being discussed (relatively stable).

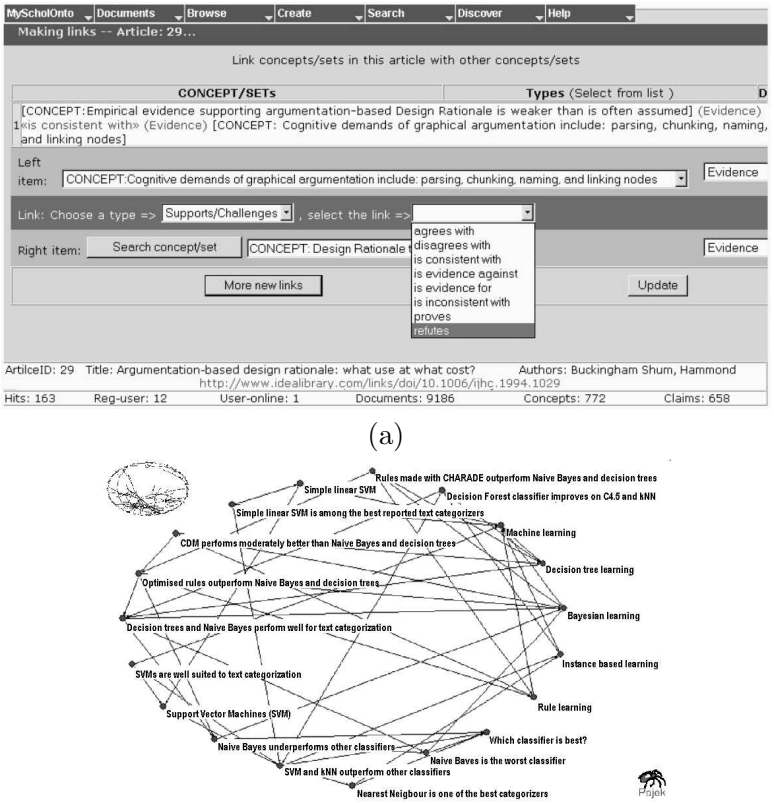


FIGURE 5.5: Using ClaiMaker to make claim about a document's contribution (a) and visualise other user's claims (b)

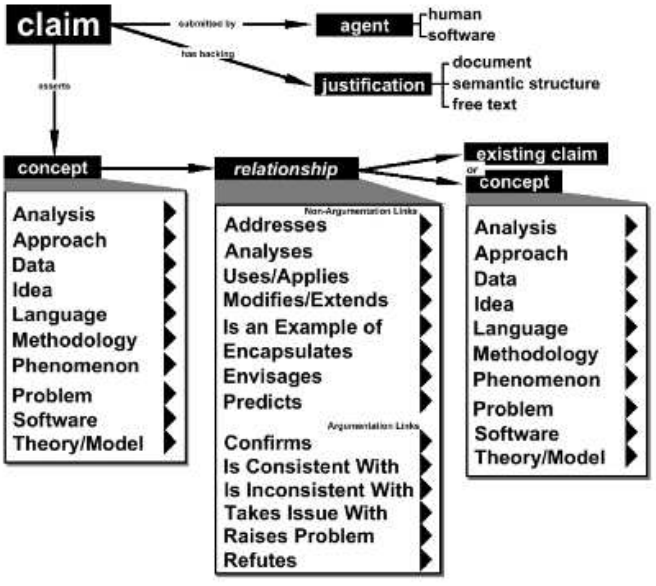


FIGURE 5.6: Ontology of rhetorical relations used by the ScholOnto project (Buckingham-Shum et al., 2000)

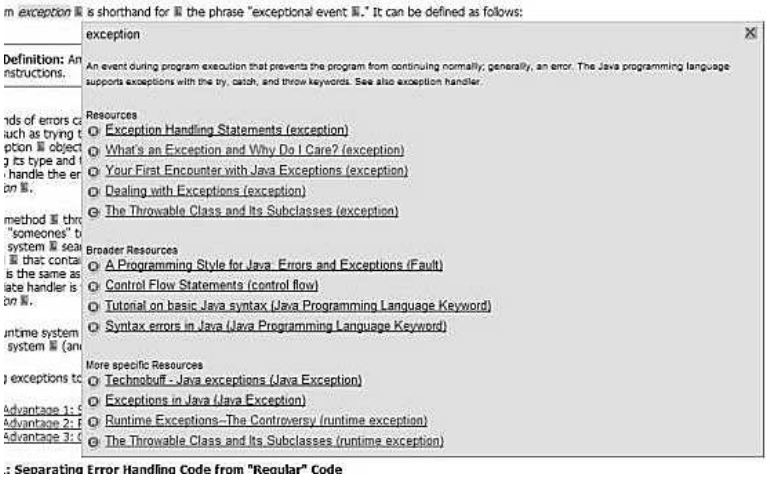


FIGURE 5.7: A COHSE concept link.

5.2.4 COHSE

Open hypermedia systems such as Microcosm (Fountain et al., 1990), which support *generic linking*, rely on keyword matching to determine which links can be applied to documents. The Conceptual Open Hypermedia Services Environment (COHSE) demonstrates how an open hypermedia link service — in this case, the Distributed Link Service (Carr et al., 1995) — can be extended with Semantic Web technologies in order to improve the quality and scope of generic links (Carr et al., 2001). Instead of simply matching keywords in documents, the extended link service is able to extract and understand the *concepts* embodied in a document.

An *Ontology Service* manages ontologies describing the linguistic terms in the domain(s) of interest (for example, a thesaurus model where concepts are connected by *broader* and *narrower* relations), and is used to extract the concepts in a document. A *Resource Service* maps the list of concepts to Web resources which discuss the concept in detail, and this information is inserted into the document as a pop-up 'concept link' (Figure 5.7). An *Annotation Service* stores additional (human-generated) metadata which may alter the behaviour of the link service (for example, if the user has defined a region of a document as an instance of a specific concept).

5.2.5 OntoPortal & ESKIMO

The OntoPortal project (Miles-Board et al., 2001; Kampa et al., 2001) demonstrated how Semantic Web technologies could be used to enhance a Web portal by providing a principled and structured approach to navigating and uncovering related information (Figure 5.9), as opposed to categorised lists of links provided by popular portals such as Yahoo!. A domain ontology was used to describe scholarly Web resources (Figure 5.8). Unlike SHOE, OntoBroker, and ScholOnto which use a community of metadata au-

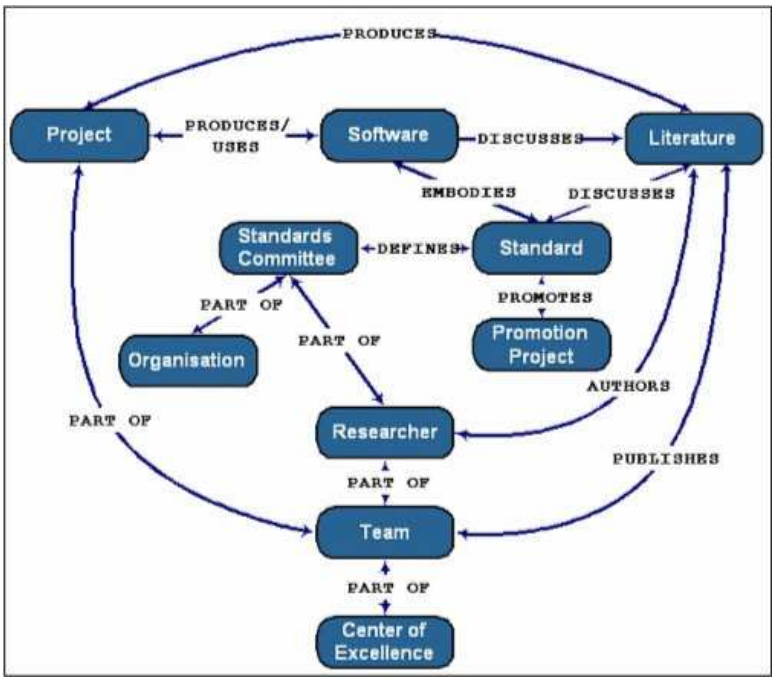


FIGURE 5.8: OntoPortal project ontology of concepts and relations surrounding scholarly Web resources.

thors, metadata about each Web resource in OntoPortal (including relations with other instances) is entered manually by a small group of expert editors who are able to translate their knowledge of the domain into the knowledge structures required by the portal.

The E-Scholar Knowledge Inference Model initiative (ESKIMO) (Kampa, 2002) extends OntoPortal to include a more complex academic community ontology and knowledge inferencing services in a similar vein to ScholOnto (for example, *what other* adaptive hypermedia papers *has* this research team *produced?*, *who are the experts in* hypertext*?*, *what are the seminal papers in* metadata research*?*).

5.2.6 WebVise+OHIF

(Grønbæk et al., 2000) demonstrate how hypermedia structures created using WebVise (links, annotations, collections, guided tours, contexts) can be encoded as metadata for the Semantic Web using the using the Open Hypermedia Interchange Format (OHIF). OHIF shares some similarities with XLink, but is able to work with non-XML data and provides a richer set of structuring mechanisms. Using WebVise, OHIF structures can be authored, imposed on Web pages, and linked to as Web resources in their own right. By extending WebVise to support Web-based Distributed Authoring and Versioning (WebDAV) (Whitehead, 1998), user can create, manipulate, and share OHIF structures together with Web pages and Microsoft Office documents stored on WebDAV servers, allowing fully distributed open hypermedia linking between Semantic Web metadata, Web pages and WebDAV aware applications.

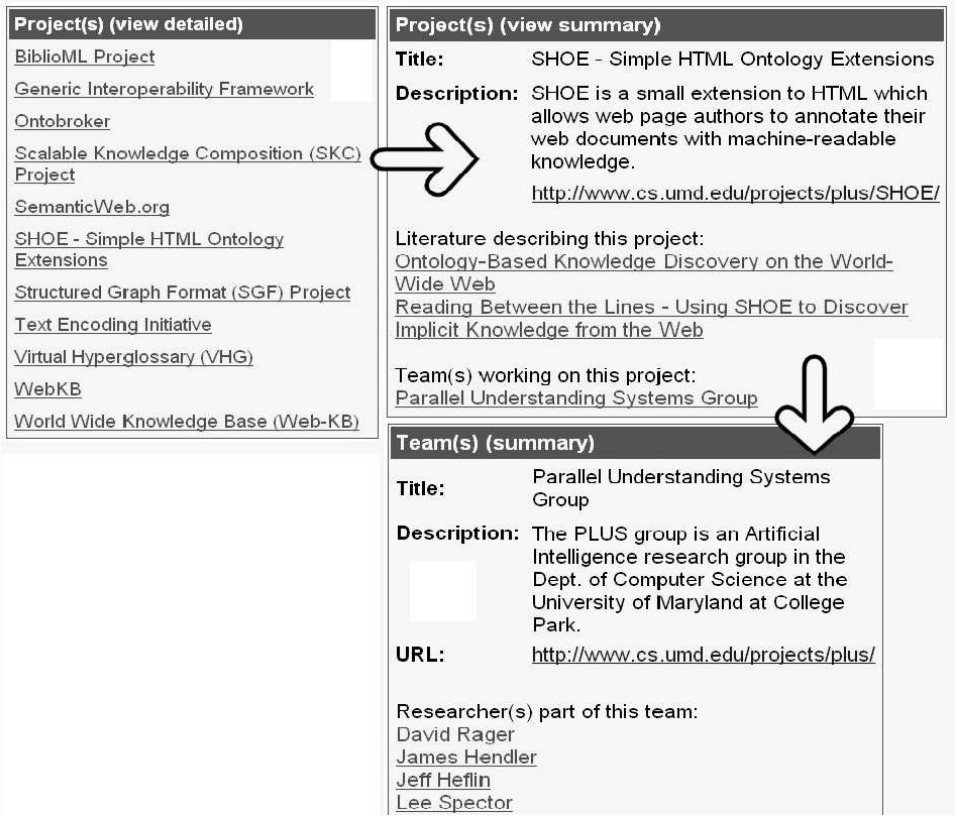


FIGURE 5.9: Structured navigation of Web resources using OntoPortal.

5.3 Discussion: Associative Writing Relative to the Semantic Web

Perhaps an initial response to the Semantic Web architecture and initiatives outlined here is that the focus is on *knowledge*. Web content is designed for humans to read, and not for machines to manipulate meaningfully, therefore the help that machines can provide to help users find information is limited. Search engines such as Google do a good job of identifying potentially useful candidates from billions of possibilities, and search engines on specialised Web sites may provide even more useful results. However, users often do not have the specific vocabulary required to use specialised search engines (consider a researcher looking for results in sites developed for other research communities), and general searches using Google return thousands of possibilities (Hendler, 2003). The Semantic Web therefore focuses on allowing *machines* to understand information on the Web in order to help humans find the facts that they are looking for. In initiatives like SHOE and OntoBroker, the role of Web documents are simply as 'carriers' for semantic information — the knowledge portal is where users go to find answers to their specific questions about the domain.

Associative Writing on the other hand, focuses on *links*. Associative Writing focuses on integrating new hypertexts with existing work (using associative links) in an attempt

to improve *human understanding*. For the writer, associative links provide a means to demonstrate the conceptual foundation being built on, and the innovation and significance of new ideas; for readers the links help provide a deeper understanding of the topic in hand. In contrast to SHOE and OntoBroker, the documents are of primary importance to the user — Associative Writing facilitates users' understanding of the documents themselves.

However, further consideration of ScholOnto, OntoPortal, ESKIMO, COHSE, and WebVise+OHIF led to a revised response: Semantic Web initiatives may focus not only on knowledge, but also on links. As well as using knowledge about documents to help users find answers to specific questions (ScholOnto and ESKIMO), these systems also use knowledge to help users understand and navigate documents more effectively — rather than simply acting as carriers for semantic information, the documents themselves are part of the Semantic Web experience. ScholOnto helps users navigate the claim space, exploring and understanding other researchers' interpretations of the literature; OntoPortal and ESKIMO demonstrate how an ontology-based Web portal can provide structured navigation of an information domain. Rather than centre user interaction around a knowledge portal, the COHSE and WebVise+OHIF initiatives demonstrate a convergence of open hypermedia and Semantic Web technology: COHSE uses an ontology-based link service to enhance the browsing experience; and WebVise+OHIF provides a platform for bridging the gap between advanced hypermedia structures and Semantic Web knowledge.

Recall that the Semantic Web is an extension of the current Web; it does not replace it — the underlying information Web remains. Perhaps what we are seeing here is a convergence between supporting *human understanding* in the information Web and supporting *machine understanding* in the Semantic Web — *machine*-supported *human* understanding in the information Web. This is a theme that is revisited in Chapter 11.

5.4 Summary

This chapter has positioned this work relative to the current state of Semantic Web research. This major stage in the evolution of the Web aims to give information on the Web well-defined meaning so that computers can manipulate it meaningfully, and so help users retrieve the exact information they are looking for.

The infrastructure of the Semantic Web has been described in terms of the three main enabling technologies: knowledge representation, ontologies, and agents. Users use metadata standards such as the Resource Description Framework (RDF) to describe Web resources and hence improve a machine's 'understanding' of them. In order for machines to be able to discover common meanings between different metadata definitions, users define ontologies: collections of information which describe common meanings and

relationships between resources on the Web. Ontologies as formal models of information on the Web enable enable autonomous agents to collect and reason over diverse Web content.

Early Semantic Web initiatives such as SHOE and OntoBroker focused on demonstrating how this knowledge carried by documents could be used to construct a 'knowledge portal' where users could go to find the answers questions about a domain. In contrast, the focus of the work reported in this thesis is on using hypertext links to assist the user's understanding of the documents themselves. More recent Semantic Web work, such as ScholOnto, OntoPortal, ESKIMO, COHSE, and WebVise+OHIF, however, has demonstrated how a convergence of machine-understandable knowledge and human-understandable hypertext can be used to help users understand and navigate documents more effectively.