WHO ARE THE EXPERTS?
E-SCHOLARS IN THE SEMANTIC WEB

By

Simon Robert Kampa

B.Sc. (Hons)

A thesis submitted for the degree of

Doctor of Philosophy

Department of Electronics and Computer Science,

University of Southampton,

United Kingdom.

October 2002

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING

ELECTRONICS AND COMPUTER SCIENCE DEPARTMENT

Doctor of Philosophy

Who are the Experts?
e-Scholars in the Semantic Web

by Simon Robert Kampa

Scholarly research is the sum of diverse activities and leads to the dissemination of a large amount of material. Traditional approaches to exploring and becoming proficient within an esoteric research field rely on slow and sometimes ineffective discourse, and depend more on a scholar's detective skill, effort, and perseverance. However, the Web has introduced the potential for improved accessibility, interconnectivity, and more efficient and effective communication; we are becoming e-Scholars.

Current efforts on the Web have focussed mainly on improving the accessibility of on-line research material and as a result, researchers have yet to realise the full implications of the new medium. Consequently, the emphasis must shift to improving and enhancing access to scholarly material; this research proposes a novel approach by presenting researchers with the facility to comprehensively, extensively, and rationally explore their research field and ask intricate questions about it and the individual facts and issues raised by it. This is realised through the integration of principles from the hypertext, Semantic Web, and digital library fields to interconnect and analyse *all* scholarly material in the academic domain. The e-Scholar Knowledge Inference Model (ESKIMO) demonstrates the approach and provides a platform for evaluation and further research.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank Leslie Carr for being a constructive, insightful, and *entertaining* supervisor who has done an excellent job in supervising his first Ph.D. student. Thanks must also go to all the people in the group who have helped me along the way and always provided assistance when required. In particular, I would like to thank Danius Michaelides and Timothy Miles-Board for spending a considerable time proofreading my thesis and supplying useful feedback, and also Guillermo Power, Gareth Hughes, and Mark Weal for sharing a bay and making my working environment a more pleasant, entertaining, and rewarding one. And lastly but by no means least, I extend an enormous thanks to my parents who have always supported and encouraged me in my work and my girlfriend Claire who was extremely supportive throughout and tolerated being completely ignored during the final months of writing this thesis. (I think it's now my turn to do the dishes for the next few months.)

Chapter 1

Introduction

1.1 Research Motivation

Scholars (researchers, scientists, or academics) are individuals involved in advanced learning within a well-defined speciality area who desire in-depth information to support their research and enable the contribution of further ideas, thoughts, theories, and observations. Their prevailing activity is becoming proficient within their chosen esoteric research field; a task that necessarily involves exploring and examining the material that the discipline has produced. This involves finding papers in a library, discussions with peers, and attending events like conferences. However, this approach is slow and sometimes ineffective and relies on scholars using detective skills (e.g. *What did this author go on to write?*, *What other papers discuss this issue?*, and *Who else works on this project?*) and considerable effort in investigating the field.

While conducting research, scholars frequently make pertinent and intricate questions about the material they encounter and their discipline in general, such as *What has this author gone on to write?*, *What are the significant papers in this field?*, *Who collaborates with this person?*, and *How has this theory affected this field?* New postgraduate students in particular, must determine the seminal papers and experts in their research field to enable them to become proficient and contribute to the field. Reviewing papers requires researchers to be aware of significant and related material and be able to position the paper in the context of the entire field. Answering these questions relies on peers, and significant effort and detective ability.

The World Wide Web, a global and distributed information repository, has the potential to improve support for scholarly research by providing improved accessibility and interconnectivity to enormous amounts of scholarly material. This has already resulted in thousands of electronic journals and digital libraries appearing, new peer review potentials being explored (Harnad, 1991), and more efficient and effective discourse (Valauskas, 1997; Sumner & Shum, 1998). Indeed, these unique opportunities prompted Dewar (1998) to note that "there are some provocative parallels" with the Scientific Revolution. As a result, we are becoming electronic scholars (e-Scholars).

Most of the initial effort has been on placing scholarly literature on-line and referencing it using technologies like the Digital Object Identifier (DOI), although less research has been directed at how the access of on-line papers can be enhanced by fully exploiting the multifaceted potentials that the medium provides. The emphasis must therefore shift from placing literature on-line, to *improving* and *enhancing* access to it.

However, as in paper-based research, the scholarly material on the Web is centred on literature, and related and peripheral material and information is not interconnected. For example, information on researchers, activities, organisations, societies, conferences, and journals are rarely explicitly linked, partly due to conflicting rights of publishers, and partly due to the lack of infrastructure, technology, and experience. Thus, researchers view papers as isolated documents and fail to further analyse, examine, and draw from its many associations to additional information. When a paper is located, it is printed and read, although, aside from following citations, the scholar rarely is able to further explore related material. However, with an explicit understanding of a paper's authors, where they work, and what projects they work on, it would be possible to effectively answer questions such as *What other projects has this organisation produced?*, *What papers has this project produced?*, and *Who are the colleagues of this researcher?*

The pertinent and intricate questions scholars ask about the material in their research field are poorly supported by the keyword-based search engines prevalent on the Web; an observation supported by an experiment discussed in Chapter 6.

Therefore, e-Scholars are locked in a similar situation as their paper-based counterparts; finding, examining, and asking questions about scholarly material require significant effort and detective ability.

This research therefore proposes a novel research environment that provides scholars with the facility to explore *all* the material in their research field extensively and intelligently, and ask intricate questions about it. The approach draws from three distinct research fields: hypertext, the Semantic Web, and digital libraries.

Hypertext is the organisation of information fragments into connected associations. Initially, this was used to connect documents, although the use of link semantics to accurately specify the intended semantics of a relation, and link abstraction to enable links to be manipulated, processed, and analysed as independent link specifications, have been proposed and are applied in this research.

Hypertext also introduces a new form of searching information by navigating the interconnected hypertext fragments. This form of information seeking is suitable for scholars who frequently do not have specific information on papers or researchers and instead browse scholarly material and home in on relevant work. However, poorly constructed hypertexts introduce problems of disorientation and cognitive overload (Conklin, 1987; Cockburn & Jones, 1996) and effective scholarly hypertexts have yet to be fully realised on the Web (Baragar, 1995; Theng, 1999).

The Semantic Web is an initiative to extend the Web with machine readable and understandable knowledge about its content. Fundamental to this initiative are ontologies; a mechanism to provide an explicit conceptualisation of a domain and enable machines to understand the information described on the Web. The Semantic Web dramatically improves the potential to support scholars. This has already been demonstrated to improve interconnectivity of documents (Carr *et al.*, 2001) and provide semantic access to annotated knowledge from Web pages (Fensel *et al.*, 1998).

In this research, an explicit model of scholarly material is used to automatically construct scholarly hypertexts and present researchers with a principled, consistent, and highly interlinked hypertext; a concept termed *ontological hypertext*. Furthermore, machine reasoning is employed to respond to the intricate questions researchers ask about their domain.

Two systems, OntoPortal and ESKIMO, have been implemented to embrace these principles and provide scholars with a novel exploration environment that allows them to comprehensively and intelligently explore their research field and ask pertinent questions about it.

ESKIMO demonstrates the potentials that the Semantic Web provides to support scholarly research and provides the platform for evaluation and further research. Facilities not available in traditional paper-based research or on the Web are demonstrated, and evaluations indicate that these encourage researchers to take a more active and involved role in their information exploration task. Furthermore, new interoperability standards for scholarly data (Hellman, 1999; Lagoze & de Sompel, 2001; Brody *et al.*, 2001b), mean it will become possible to dynamically update and maintain these advanced services and allow scholars to track the developments in their field as they happen.

1.2 Research Statement and Contribution

The objective of this research is to explore and identify methods of supporting and enhancing scholarly research on the Web. The Semantic Web has introduced a framework to enable machines to understand the content of scholarly material, and a novel approach is proposed that draws from this to present scholars with new research possibilities. This involves the intelligent interlinking of scholarly material and machine analysis to respond to typical research questions.

The contributions of this work are:

- Assessing how the Web currently measures up to a scholar's research activities.
- Proposing *ontological hypertext* as a method of interlinking complex research fields to construct a principled and consistent scholarly hypertext.
- Identifying typical research questions and providing the facility to resolve them efficiently.
- Implementing a scholarly research environment to enable researchers to explore their research field and *all* its related artifacts and objects, and ask pertinent and intricate questions about them.

1.3 Thesis Structure

Chapters 2, 3, 4, and 5 cover important background material and related work. No new contributions are presented in these chapters.

Chapter 2 introduces the hypertext concept and identifies the key issues used in this thesis. Significant historical and contemporary hypertext systems are described.

Chapter 3 introduces the Semantic Web initiative. The Semantic Web is promoting an intelligent Web where navigation, information retrieval, and machine processing will benefit from machines being able to understand the content of the documents.

Chapter 4 describes ontologies, which are a fundamental part of the Semantic Web. They provide a mechanism for defining complex domains for machine understanding.

Chapter 5 introduces scholars and their research activities. Traditional research techniques are introduced as well as the new approaches possible on the Web.

Chapters 6, 7, 8, and 9 detail the contributions made from this thesis work.

Chapter 6 discusses a user experiment that was conducted to measure the support the Web affords to scholarly research.

Chapter 7 introduces the key concepts from this research, namely the interlinking of scholarly material and how inquiry can be supported.

Chapter 8 discusses the OntoPortal system, which demonstrates how complex research fields can be interlinked and presented to the researcher in a principled and consistent manner.

Chapter 9 covers the ESKIMO system, which provides a comprehensive support environment for scholars that enables them to explore their research field and make pertinent questions about it. An evaluation of this system is also presented.

Chapter 10 concludes this thesis by reviewing the work and proposing ideas for further work.

1.4 Declaration

The work contained within this thesis has been carried out within a collaborative research group. It is all the author's work, with the exception of Chapter 8 where

Dr. Leslie Carr and Timothy Miles-Board were involved in discussions and implementation of the OntoPortal system as part of a contract between IAM and DERA, UK.

Chapter 2

Hypertext and the Web

2.1 Introduction

This chapter introduces the hypertext concept, the organisation of information fragments into connected associations, its advancements, and its use in a global information system, the World Wide Web.

Hypertext is essential to this research as it provides the facility and technology to interconnect diverse scholarly material and present researchers with a complete, coherent, and traversable view of the material they are interested in. Developments in hypertext, particularly adding semantics to links and abstracting links from the documents they appear in, have also introduced new approaches to presenting, managing, and analysing hypertext that this research draws from and evolves into the *ontological hypertext* concept described in Chapter 7.

A brief historical background to hypertext is presented, followed by an overview of the significant advances, particularly those in the context of this research. The concepts of hypertext linking and how explicit link semantics are used to improve an author's and user's view of connected material are then explored. Hypertext also enables navigation as a new paradigm for exploring documents, this is explained, and the problems it introduces outlined.

2.2 The Birth of Hypertext

Hypertext[1] is the organisation of information into connected associations. It is about non-sequential text. One could argue that the roots of hypertext are in

[1]The term hypermedia is used when referring to any type of media.

ecclesiastical readings such as the bible, where cross-referencing and annotations are common.

However, if we presume that hypertext refers to the non-sequential reading of digital documents, then the roots are found approximately 57 years ago. In 1945 the director of the Office of Scientific Research and Development, Vannevar Bush, published his seminal paper describing what he called the "Memex".

2.2.1 Memex

Memex (Bush, 1945) ("memory extender") was a revolutionary, albeit fictional, device to store all an individual's records, books, and literature and provide an access mechanism based on linking everything together. Bush proposed microfiche as the storage mechanism (resulting in stored information being off-line with reduced access speeds).

Bush was struck by the "growing mountain of research" and realised that the plethora of publications inhibited scholars making "real" use of it. It appeared to Bush that the main problem was that of selection, or information retrieval, and this he noted was down to the storage, structure, and tagging of the information.

The Memex was envisioned to provide a global storage system and enable users to make connections between fragments of information. Bush referred to these connections as trails and the practice of making these resulted in a "trail of interest" being constructed. Trails represent an individual's path as they traverse the literature and are recalled at a later stage to help re-navigate the information. Subsequent systems, such as MEMOIR (Pikrakis *et al.*, 1998), have used this notion to analyse user trails and suggest further paths worth exploring.

2.2.2 NLS/Augment

The first successful implementation of the hypertext concept was not presented until 1963, when a research team lead by Douglas Englebart (better known for introducing the mouse as a revolutionary pointing device) developed the NLS (oNLine System) to realise his vision of using machines to provide instantaneous connections between documents (in contrast to Bush's proposal of microfiche). In tribute to this work, the NLS system was chosen as the second node on the ARPANET network, the predecessor to the Internet.

NLS provided cross-referencing of research papers for sharing among geographically distributed scientists, which allowed early digital libraries to be created and their papers retrieved using hypertext linking. The objective of NLS was therefore to boost significantly an individual's, a group's, and an organisation's performance by providing fast access to interlinked material.

The system consisted of three major components: a database of text fragments, view filters (used to view particular information sets), and views (used to structure information). Documents in NLS were marked up with anchor information to form the endpoints of links. NLS later evolved and was commercialised into Augment (Engelbart *et al.*, 1973).

2.2.3 Xanadu

In 1965 another visionary, Theodor Nelson, first coined the word hypertext (as "non-sequential writing") and proposed the publishing system Xanadu (Nelson, 1980; Nelson, 1987) as a unified literary environment. The purpose of Xanadu was to establish Nelson's view of a 'docuverse': a searchable global library of all documents ever published interconnected using hypertext technology. Xanadu has so far failed to be fully implemented although it has been released into the open source community.

Xanadu represented the first concerted effort towards developing a complex and comprehensive hypertext system. Nelson proposed a mechanism by which any version of a document was stored only once and linked to whenever it was required, resulting in a significant reduction in storage requirements. Nelson referred to this as transclusion and the documents where the actual text was archived as permascrolls.

Transclusion offers many advantages aside from storage. Firstly, as text fragments are only stored once, versioning becomes manageable as newer versions replace earlier text fragments and links are provided to the previous versions. Authors transcluding these text fragments are therefore guaranteed that they are always including the latest version. Secondly, transclusion enables readers to view all documents in the docuverse that contain the transcluded fragment, or view the source text where the fragment originally occurs. Finally, it also makes the collection of royalty fees for published works easier and more accurate. Each time a fragment

of copyrighted text is transcluded into a document this can be recorded and used when royalty fees are calculated.

Nelson introduced the concept of different link types between documents in Xanadu but stated that the definition and use of these link types was the responsibility of the author.

2.3 Popularising Hypertext

This section outlines the significant work that was published in the early stages of hypertext research and was influential in the use and promotion of the hypertext concept for a variety of different scenarios and applications.

2.3.1 KMS

KMS (Knowledge Management System) (Akscyn *et al.*, 1988) was a highly developed commercial hypertext system designed to "help organisations manage their knowledge" and was used in various applications such as electronic publishing, online manuals, project management, and software engineering documents. It was based on ZOG, which provided users with a task management system and interlinked on-line manuals. Work on ZOG began in 1972 and was later tested on the aircraft carrier USS Carl Vinson. Based on its success, Knowledge Systems was established to commercialise ZOG as KMS.

KMS provided a simple paradigm to organise information. The main metaphor was that of a frame (Figure 2.1). This was the only node (document) type defined in KMS and enabled the representation of a broad range of knowledge. A frame could contain any combination of text and graphics and could be linked to any other frame. Like NLS, link information was embedded within the frames.

Linking in KMS provided the primary form of navigation. Two link types were supported: tree links and annotation links. Tree (structural) links connected lower level-frames in a hierarchy and were used to for example, connect chapters of a book. Annotation links indicated associative relationships and pointed to comments, definitions, and cross-references. Annotation links were distinguished by placing a '@' character as the first character in the link title.

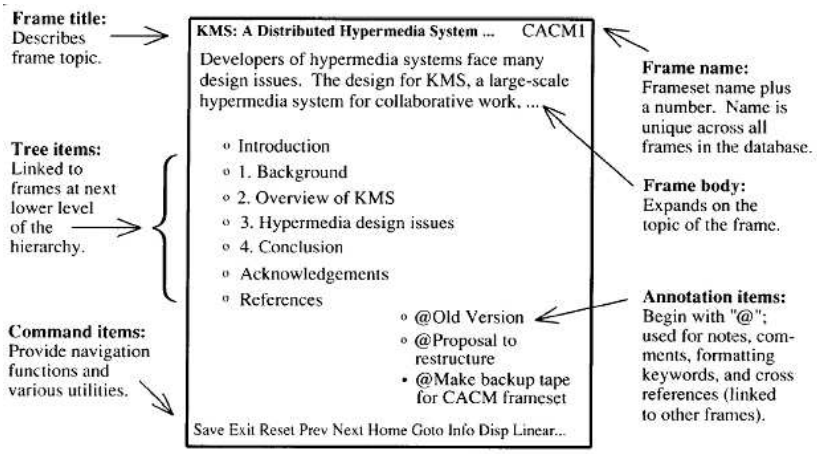


Figure 2.1: A KMS frame

2.3.2 Notecards

Notecards (Halasz *et al.*, 1987), developed at Xerox PARC, represented the first widely used hypertext system. It was based on four basic constructs designed to help researchers and designers organise and develop their ideas: notecards, links, browsers, and fileboxes. Users created a series of typed, linked nodes (notecards) to create a web of interconnected information. Notecards could be collected into fileboxes and a browser was used to view the resulting web. Using only notecards, fileboxes, and links, users navigated the information contained in the system.

The Notecards browser displayed a graphical overview of the link structures between the documents, using different line patterns to represent the various link types. This work led directly to Halasz publishing his seminal paper on identifying seven issues for hypertext systems, *Reflections on NoteCards: Seven issues for the next generation of hypermedia systems* (Halasz, 1988). The seven issues described by Halasz are: search and query, composites, virtual structures, computation in/over hypertext networks, versioning, collaborative work, tailorability. Although Halasz speculated that hypertext as a field would vanish by 1992, this did not happen and the seven issues are still relevant to modern hypertext systems.

2.3.3 Intermedia

Developed at Brown University, Intermedia (Yankelovich *et al.*, 1988) provided a classroom hypertext system to support teaching and learning. Intermedia used a dramatic approach in providing hypertext features by supporting it at the system level: a shell running on A/UX 1.1 (Apple's implementation of UNIX). This enabled

hypertext facilities to be used by any application running on the system. A suite of applications were provided including a text and graphics editor, a 3D model viewer, and a video editor. Each of the different media types viewed in these applications could be linked.

Linking data was stored separate from documents and multiple sets of these *linkbases* were allowed. This enabled users, such as students, to create their own web and link documents they did not own. Histories, tours, local maps, timelines, and other concept maps assisted users in understanding their context while navigating. Context improves a user's cognitive understanding of where they are and where they have been and thereby reduces the possibility of becoming lost.

2.3.4 HyperTIES

The HyperTIES (Interactive Encyclopedia Systems) (Shneiderman, 1987) work started in 1983 and was based on the metaphor of the electronic encyclopedia. It was designed to allow authors to easily create and publish hypertexts using a simple user interface. In fact, it was used in 1989 to create the first commercial electronic book, Hypertext Hands-On! (Shneiderman & Kearsley, 1989). Also, after the system was commercialised in 1987, it was used by Hewlett-Packard to distribute electronic documentation for its LaserJet printer in 15 languages.

In HyperTIES, each document was referred to as an article, and cross-references between these were implemented as highlighted text links or image maps. Indeed, Shneiderman invented the idea of using the text itself as the link marker (later called embedded menus or illuminated links as users were presented with a preview of a link before following it). Browsing was supported through the navigation of these links and backtracking was provided as HyperTIES recorded the paths users followed while traversing a hypertext.

Significantly, HyperTIES provided many of the ideas that would later become important in the success of the Web. Aside from providing text link markers and simple navigation, HyperTIES provided a markup language (incidentally also called HTML - HyperTIES Markup Language), image maps, and a history facility.

2.3.5 Hypercard

Hypercard (Smith & Bernhardt, 1988) was not strictly a hypertext system in itself, as it did not explicitly provide a navigation mechanism. However, it was important

as it popularised the hypertext concept. This was in no small part due to the free bundling of the system with Apple computers. Hypercard used a 'stack of cards' metaphor, each of which contained text, pictures, and other interface elements. Hypertext-like functionality was provided through scripts attached to buttons on a card which could point to other cards.

The only in-built navigation in Hypercard was the search and history tools. However, as the basic scripting used in the system was within the grasp of most developers, it was easy to use HyperCard as the framework for a hypertext system. A serious limitation, however, was that anchors could not be placed within text but were restricted to interface objects like buttons. Therefore, any changes to a card's text required the interface objects to be manually moved.

A similar, although relatively unused, environment to Hypercard, LinkWay, was created for the Microsoft DOS operating system and pages and folders were used as the metaphor.

2.4 Towards Open Systems

Early hypertext work demonstrated the feasibility and practicality of linking documents using hypertext. In addition, the basic idea of hypertext being a connection between two documents was extended with novel concepts such as link typing and abstracting the link from the content to which it applies.

Since the early work, hypertext research has progressed (Figure 2.2). It has been applied to large information systems (Berners-Lee *et al.*, 1993), to a variety of different media (Davis *et al.*, 1993), and other disciplines such as software development environments (Anderson *et al.*, 1994) and educational systems (Furuta *et al.*, 1997). Theoretical models of hypertext have been proposed in the Dexter Hypertext Reference Model (Halasz & Schwartz, 1994) and FOHM (Fundamental Open Hypertext Model) (Millard *et al.*, 2000) which aim to capture the abstractions found in hypertext systems, notably the link and anchor, with the aim of providing a framework to compare systems and develop interoperable standards.

A significant shift has also occurred in the architecture and representation of hypertext systems. Early monolithic systems, such as NLS and KMS, were termed *closed* hypertext systems as they represented rigid, encapsulated programs that made interoperability difficult as links and anchors were specific to an application

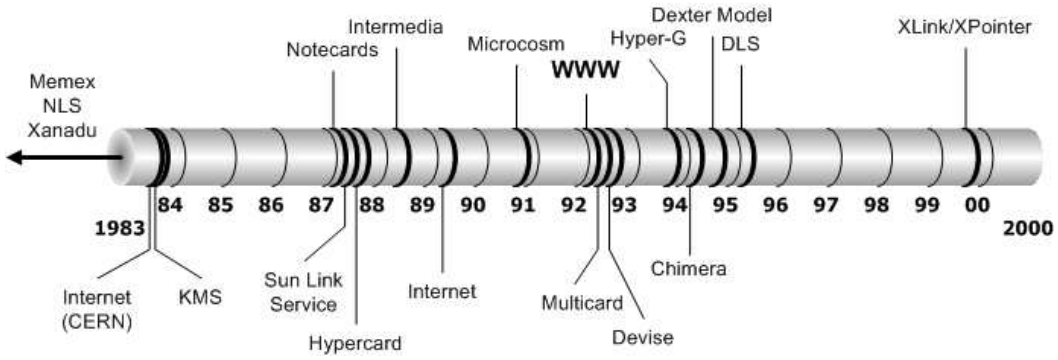


Figure 2.2: Hypertext timeline

and embedded within documents, meaning custom applications were required to view the resulting hypertext. Closed systems also rarely worked outside the scenario for which they were constructed. On the other hand, recent hypertext systems aim to provide a strict separation between the various hypertext constructs, the content, and the applications. These *open* hypertext (hypermedia) systems have the ability to integrate hypertext data created in different applications and formats and thereby improve the *interoperability* of hypertext services.

This section presents an overview of open systems that demonstrate the principles of open hypermedia in different applications and scenarios.

2.4.1 Sun Link Service

The Sun Link Service (Pearl, 1987) was the first example of the open hypermedia concept and was a protocol and toolkit to add hypertext functionality into normal applications on a Sun workstation. The service also demonstrated the link service concept, which manipulated links as first class objects, resulting in nodes and links being stored separately. Links were then combined with the nodes when the node was requested.

The link service was a novel concept as links were represented as individual objects rather than fixed embedded connections between static documents. In a link service, the link is a specification of a relationship where the source and destination of a link are specified and applied to any number of documents, providing a highly flexible linking mechanism.

The independence of the link also meant that the link service could be integrated into other applications and enabled them to benefit from hypertext without requiring an extensive programming effort. This is particularly useful when applied

to applications with no in-built hypertext functionality, such as text editors and spreadsheet programs.

2.4.2 Microcosm

Microcosm (Fountain *et al.*, 1990; Davis *et al.*, 1993; Hall *et al.*, 1993), developed at the University of Southampton, was an open hypermedia system which employed a dynamic link service and was aimed as an educational tool. The system's open architecture enabled the integration of third party applications and these were used to view linked material.

Document viewers communicated with Microcosm through a messaging system to which a chain of filters was attached (e.g. link service, link history facility). When a user opened a document in a Microcosm compliant viewer, it requested links to be added to the document. The series of filters then processed the request (e.g. by adding, modifying, or deleting links) and returned the relevant links.

However, a major problem introduced by link services is context. Linkbases are potentially very large, meaning the number of inserted links has to be somehow controlled, or the user faces filtering through out-of-context links. A coarse-grained solution would be to identify the user's context and then toggle the state of the appropriate linkbases. For example, if the user is viewing pages on sailing, then the boat and weather linkbases are turned on, while the computer terms linkbase is switched off. Microcosm required users to select linkbases manually.

Microcosm provided a range of dynamic link types:

- *Specific links* are links from a specific point in a source document to a specific destination document.
- *Local links* specify a link from any location within a specified document to a specific destination document.
- *Generic links* have a fixed destination while the source anchor can occur at any position in any document.

Generic links enable authors to construct 'declarative' links that can be applied to multiple documents. For example, a generic link can be constructed to link every occurrence of the phrase 'IAM Group' in any document to the IAM Group homepage. This unique concept added a high degree of openness to hypertext systems.

Microsoft has introduced a similar technology termed Smart Tags, which applies generic-style linking to its products, like Office and Internet Explorer. Hughes *et al.* (2002) note the political and technical criticism that this has attracted. Primarily, critics argue against altering content (in this case adding links) without the explicit permission of the document author. Smart Tags are also open to be exploited by commercial companies for advertising purposes, (e.g. a link on a keyword takes the user to an on-line shop) or they could raise contentious issues (e.g. the word 'visa' links to a credit card company's site). While generic linking on a global scale raises serious issues, a controlled linking system where (i) the default setting of generic linking is disabled, (ii) users have the ability to switch the linking on or off when they desire, and (iii) users choose which linkbases they require, is credible.

2.4.3 Multicard

Multicard (Rizk & Sauter, 1992) was a hypermedia toolkit to allow programmers to construct and manipulate hypermedia structures. As it followed an open systems approach, it did not handle the actual content of nodes and users could use any compliant authoring or programming tool. Multicard communicated with compliant editors by using a communication protocol similar to that used in Microcosm: M2000. Any editor that provided at least the minimum M2000 support could begin to take advantage of the hypermedia functionality provided by Multicard.

The core of the toolkit consisted of the representation of hypermedia objects. These included nodes, links, groups (logical collection of nodes and other groups), and anchors. Nodes, groups, and anchors could have scripts attached to them to extend their behaviour, for example, by manipulating a document's content. Multicard also provided a standard authoring environment (itself implemented using Multicard) to assist in creating hypermedia objects and scripts.

2.4.4 Chimera

Chimera (Anderson *et al.*, 1994) provided hypertext services within a Software Development Environment (SDE) to visualise and capture the multifaceted relationships between the objects in these environments. An SDE is highly heterogeneous as different editors are used to manipulate objects which include multiple versions

of prototypes, design specifications, requirements documentation, code, test information, and scripts. The relations between these objects are varied and complex, and lead to a cognitive load on the software engineer.

Chimera proposed the use of hypertext services to capture and visualize these relations, enabling software developers to locate the required objects and understand their relation to a programming issue. Chimera only handled the linking between the objects, leaving the presentation of the content and the linking to the application. This meant that the particular applications could customise the link presentation to best suit the type of content.

Significantly, Chimera did not restrict the anchors of links and instead modelled them as a collection, enabling *n-ary* links to be created. When the user selected a link with multiple ends, all the destinations could be viewed at once, for example, by starting several viewing applications. This is useful when a link from a code fragment points to a relevant specification and an accompanying example. However, n-ary linking places a sizeable demand on the screen real estate.

2.4.5 The Distributed Link Service

The Distributed Link Service (DLS) (Carr *et al.*, 1995) (later commercialised as Webcosm) abstracted linking from a document's content, and provided a hypermedia linking service to be used in *conjunction* with Web (Section 2.5) document servers to improve the overall connectivity of on-line documents. It was based on work from the Microcosm project and supported generic linking. While the Sun Link Service and Microcosm ran on local machines and added hypertext to local applications, the DLS linkbases were distributed and links were retrieved from them and added to documents to compliment the links already provided within them. An interface agent was provided that ran on the user's browser and requested links from the DLS when a document downloaded. This added a high degree of openness to the Web.

The DLS had advantages to both users and authors. As links were added based on simple textual matches (e.g. the word 'programming'), it was possible that the linkbase(s) would contain multiple destinations for this keyword. Therefore, the user would be presented with multiple possible destination documents, as opposed

to only being presented with a single one. Users were also provided with weak contextual support; the link servers supported multiple linkbases and users manually selected the appropriate linkbases to reflect their current task context. For example, if a user explored information on computer graphics, then the graphics linkbase was used enabling the user to create a reader-driven (as opposed to author driven) personalised Web space. Alternatively, the Queries In Context (QuIC) (El-Beltagy *et al.*, 2001) system analyses Web pages to automatically derive a context and then uses this to determine which links to add to the page.

With the DLS, authors were given flexibility and a reduced authoring overhead. Rather than having to edit individual documents and mark them up with linking information, links were only edited in the linkbase through the interface agent. If the author wished to modify the destination location for the word 'programming', then this only needed to be edited once in the linkbase, rather than for each time the word appeared in a document.

2.5 The World Wide Web

To date, the most popular example of a hypertext system is the World Wide Web (WWW or Web), created by Tim Berners-Lee in 1989 in response to the difficulties of exchanging resources on the Internet (Berners-Lee *et al.*, 1993).

The roots of the Internet can be traced back to 1969, when the US Defence Department wanted to control computer systems remotely and provide remote access to information. Therefore it created a network of interconnected systems called the ARPAnet[2]. During the 1970s the network grew to include several research institutes and laboratories. Important communication standards, such as the TCP/IP stack of protocols, were also developed during the ARPAnet years.

In 1989, ARPAnet was shut down and the Internet emerged out of the loosely connected networks and the newly created backbone network, NSFNET. Three years later, Berners-Lee, a software engineer at the Centre for European Particle Physics (CERN), developed a hypermedia system that has now become universally referred to as the World Wide Web, and with it, three simple, albeit significant, standards: the Uniform Resource Locator (URL) (Berners-Lee *et al.*, 1994) for

[2]Named after the funding body, the Advanced Research Projects Agency.

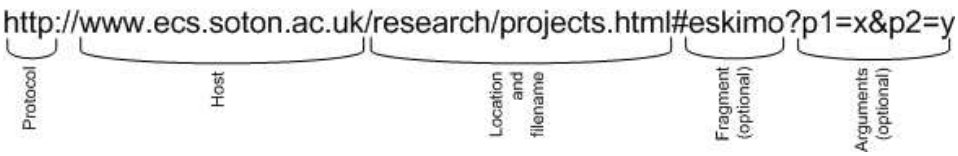


Figure 2.3: Structure of a URL

addressing, the Hypertext Transfer Protocol (HTTP) (Berners-Lee, 1991) for exchanging documents, and the Hypertext Mark-up Language (HTML) (Berners-Lee, 1992) for encoding and marking up documents. These standards were fundamental to the success of the Web as they presented a standardised global protocol to view and exchange documents.

The protocols meant that it was simpler to create Web documents, link them (retrieving a document was now simply a matter of clicking on a link), and address servers around the world. This simplicity has made the Web a more attractive proposition to inexperienced users. However, limitations are noticeable in the protocols, which reduce the overall effectiveness of the Web. These protocols and their restrictions are briefly discussed.

A URL is used to specify an address of a Web page, and its structure is illustrated in Figure 2.3. The protocol type is specified before the '://', in this example HTTP. The text between the protocol and the first '/' indicates the host which represents a server on the Web. The location and name of the file to retrieve is specified between the '/' and the end of the address or the '#' or '?' characters. The fragment identifier is used to specify the location of the link anchor. However, the fragment identifier only works in HTML files where the anchor end has been explicitly marked up inside the document. If the page author fails to maintain the anchor, then the fragment identifier part of a URL will fail to function. Arguments (e.g. CGI[3] parameters) can be specified to be passed on to the server for use in responding to the request.

HTTP is an application protocol for exchanging files on the Web. Typically, an HTTP client, such as a browser, contacts an HTTP server (as specified in the domain part of a URL) and requests a page on the user's behalf. Using the file location part of the URL, the HTTP server locates the file and returns it to the

[3]Common Gateway Interface (CGI) is a standard method used to pass a user's request to an application program.

user. The protocol is stateless, meaning each request is executed independently without any information about previous requests. This was intended to improve performance by reducing a connection's overhead. However, by not retaining any state information it is difficult to retain user information and customise the content to their preferences, for example when using dynamic linking. This restriction has been addressed in technologies such as Java, JavaScript, and HTTP Cookies, which enable state to be saved between different Web pages and sessions.

Content on the Web is authored in HTML, which, through a series of mark-up symbols, formats a file for presentation in a user's browser. HTML is a presentation oriented format meaning it contains little structural coding. An HTML file can include references to a variety of other media including graphics and sound. Linking is accomplished by explicitly inserting a URL within HTML link mark-up. This enables the user to construct simple, binary, unidirectional links that act like GOTO commands in programming languages.

However, Web addresses are sensitive to change and decay as the resources they point to are removed or altered, meaning link integrity becomes a serious issue. In addition, by embedding URLs directly in Web content, maintaining links becomes difficult as changes to an address have to be made to all files that reference it. Furthermore, only authors can modify links, removing the benefits that link services provide. This also removes the possibility of bi-directional linkage between different Web sites as authors are required to access and modify both the source and destination documents.

The Web represents an enormous collection of linked resources and is a shift from earlier hypertext systems, which generally began simply as a vision of interconnected reference material. However, while the Web is a hypertext system, many researchers are disappointed by the lack of *hypertext features* inherited from classic systems (Bieber *et al.*, 1997a; Ladd *et al.*, 1997), such as link and node semantics (Section 2.7), transclusion, openness, dynamic linking, and linkbases. The next two subsections discuss two efforts to extend the Web with greater hypertext functionality, although both have so far failed to be used on a large scale.

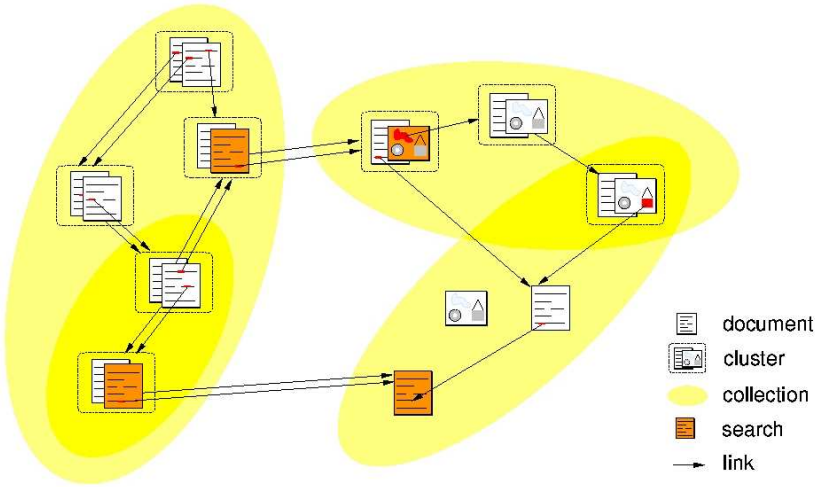


Figure 2.4: Hyper-G data model

2.5.1 Hyper-G

Hyper-G (Kappe *et al.*, 1993; Andrews *et al.*, 1995) (later released as HyperWave) is an open hypermedia system for the Internet that allows users to interact more with the underlying hypertext data. Unlike previous client and server systems on the Internet (e.g. Gopher, Web) which had no graphical navigational aids, were read-only, and supported a single structure, Hyper-G provides a large-scale, distributed, multi-user, *structured* hypermedia information system.

Hyper-G's data model (Figure 2.4) is far richer than previous Internet systems. The main concept in the Hyper-G environment is a hierarchy of collections, similar to the class system in object-oriented programming languages. Documents are grouped into collections which themselves can belong to further collections. A cluster is a special collection used to create multimedia aggregates.

Binary links have the usual source and target anchors. However, the target does not have to point to a document or fragment within it, but can also point to a collection. Furthermore, links are stored and managed separately meaning the possibility of the dangling link problem (Davis, 1999) is reduced as authors only have to manage a link database as opposed to all the individual documents the links appear in.

The Harmony browser on the UNIX platform and the Amadeus browser on the Windows platform provide access to Hyper-G functionality, although standard

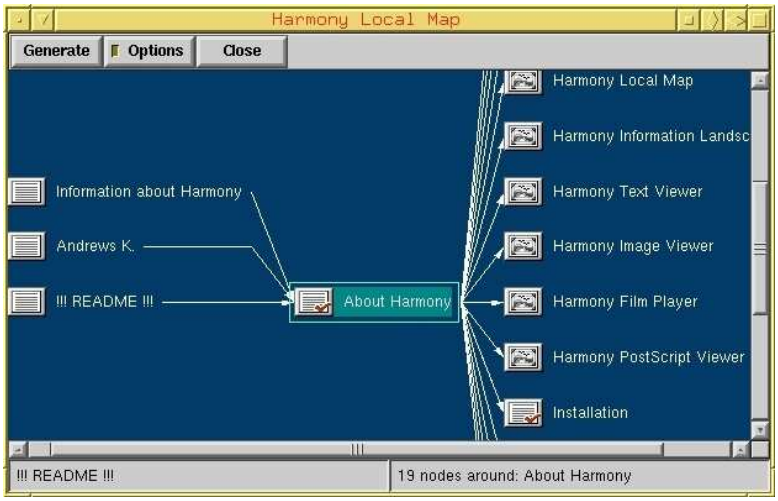


Figure 2.5: Harmony's location map orientation tool

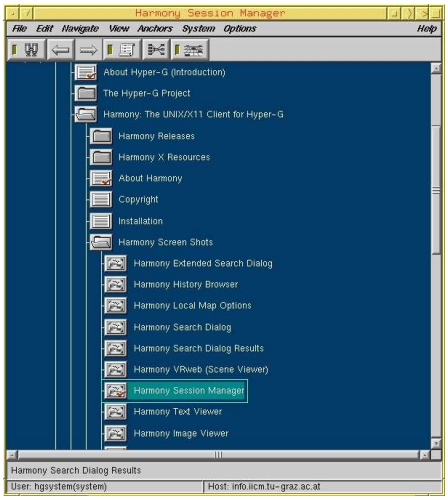


Figure 2.6: Harmony's collection browser

browsers can still access the information without the extended features that Hyper-G provides. A local map to visualise the link neighbourhood of the current document (Figure 2.5) is provided as a context tool to reduce the possibility of disorientation. The collection browser (Figure 2.6) is used to display neighbouring collections.

2.5.2 Devise Hypermedia

Devise Hypermedia (DHM) (Grønbæk & Trigg, 1994), based on the Dexter Reference Model, is a system based around a client-server model and provides hypermedia functionality through multiple database servers. It supports collaborative system development and design, which enables users to cooperatively use and edit a collection of pages. The system highlights a number of areas where the Dexter model

proves insufficient, for instance, in modelling generic links.

Rather than propose a replacement Web, Grønbæk *et al.* (Grønbæk *et al.*, 1997) propose augmenting it with DHM services and thereby adding some of the hypertext functionality that the Web lacks. (The general framework of DHM enables it to model the Web.) This would provide Web users with four significant advantages:

- Links can be created on documents that users do not own
- Simple Web and DHM-based links (e.g. bi-directional links with multiple endpoints) can be traversed.
- Documents can be grouped into collections.
- Cooperation between users on shared pages (e.g. to use and edit a shared site).

A platform independent solution is proposed by Grønbæk *et al.* that uses traditional Web browsers in combination with technologies such as Java and JavaScript[4]. This approach does not rely on custom browsers (c.f. Hyper-G enabled browsers) or on browser patching (c.f. Microcosm, DLS).

2.6 Linking on the Web

Hypertext links explicitly assert a relationship between a source and a target node, usually due to a semantic relationship that exists between them. Link and node type, strength, and other properties of a relationship are supported by some hypertext systems (Trigg, 1983; Fountain *et al.*, 1990; Robert & Lecolinet, 1998), although not by the Web.

This section comments on the linking facilities on the Web and contrasts these with the facilities in other hypertext systems.

2.6.1 Anchors

The source and target of a link refers either to a document or to a fragment within it (e.g. section of text). If a text fragment forms one of the endpoints, then an *anchor* is required to specify the exact position within the document. This leads to four types of linking:

[4]However, in the end a Netscape specific technology was used (LiveConnect) so the solution, while platform independent, required a Netscape browser.

1. Fragment to Document - the phrase 'IAM Group' to the homepage of the IAM Group
2. Fragment to Fragment - the phrase 'contact the IAM Group' to the specific part of the IAM Group homepage that lists contact details
3. Document to Fragment - the 'IAM Group' homepage to the address section of the contacts page information
4. Document to Document - the outdated 'MMRG' homepage to the 'IAM Group' homepage

Hyper-G and the Web support the first two linking types. Hyper-G also supports bi-directional linking, while only uni-directional linking is possible on the Web. The final two linking types have the source anchor set to the entire document. While this facility could be desirable in some scenarios, it is difficult to conceive how this could be applied consistently to a document interface system such as a Web browser (perhaps by adding a special link symbol at the top of a document). Also, and more significantly, a document, unless it only contains a few words, will have many potential anchors within its text, and so constructing a source anchor that represents an entire document implies that link which uses it is (i) especially important and noteworthy and (ii) indicative of the nature of the linked documents. Therefore, any misuse of this type of link would confuse the hypertext reader and detract from the actual intention of the author.

2.6.2 Link Construction

Links and anchors can be created explicitly when a document is authored, or added dynamically when a document is retrieved. DeRose (1989) proposes a simple link taxonomy to divide links into these two general categories.

- Extensional: Static link created and stored at authoring time.
- Intensional: Implicit link, not stored, but dynamically computed and added to the document later.

The majority of linking on the Web is extensional, as these links are easier to create and the intended relationship can be precisely defined. However, storing and manipulating extensional links and ensuring link integrity are serious issues, and are the primary reason for the emergence of open systems. Intensional linking, on

the other hand, avoids these problems although requires an understanding of the structure of a document as a computational process must identify the position of a link anchor and work out where to link to.

2.6.3 Declarative Linking?

Bieber *et al.* (1997b) suggest that linking on the Web is comparable to second-generation programming languages where only low level functionality is available. As we move beyond second generation languages, the emphasis should shift to what the author wants (declarative), rather than the method used to produce it (procedural). He claims that hypertext authors should adopt the same style and define *what* they mean by a link, rather than simply producing the code to create the connection. Declarative linking is an important issue frequently visited throughout this thesis, and is used in ESKIMO to more effectively organise scholarly material.

Although the linking facilities on the Web are limited, the link makes the Web unique, as without it, it is merely an information repository like FTP. The HTML and URL standards provide authors with a simple method to quickly construct links and thereby create a set of interconnected documents they can publish. This inherent simplicity has ensured the success of the basic hypertext capabilities available on the Web.

2.7 Link Semantics (and the Web)

Representing and modelling links as simple, binary, uni-directional relationships precludes their obvious semantic nature. A simple link does not embody all complex relationships that exist between resources as well as those that the author wishes to define (Golovchinsky, 1997) and as a result is inadequate as a navigational mechanism (Conklin, 1987). Instead, a link must inform users how the content of the destination node is meant to change their perception (or interpretation) of the source node (Kopak, 1999).

A link defines a meaningful connection between two related documents and the exploitation of this meaning has resulted in the creation of a rich set of tools, such as contextual navigation, advanced authoring, and information retrieval (Bieber *et al.*, 1997a), and these are discussed shortly. Landow (1997) acknowledges that when clearly labelled, link typing aids navigation as users are able to build better

cognitive maps of the information space. Typed links also address the potential problem of disorientation and cognitive confusion arising from large associatively-linked hyper-sets, since the user is able to predict the effect of traversing a link before the act of traversal has actually taken place (DeRose, 1989).

Nanard *et al.* (1993) believe that semantically typed nodes and links help authors organise information more effectively and provide context to readers, and recommend using link types as early in the design process as possible (even when the underlying hypermedia system fails to support the concept) (Nanard & Nanard, 1995). This assists authors in structuring and organising their work at a cognitive level.

Baron (1994) identifies two categories of link types in her research: organisational and content-based. Organisational links tie hypertext nodes together at the surface structure, such as index pages and navigational buttons (e.g. next, previous). Content-based links cover the specific relationships that exist between nodes in a hypertext. This link category is further divided into semantic, rhetorical, and pragmatic links. Semantic links are used to describe (in basic terms) the relationships between concepts. Rhetorical links provide ancillary information such as illustrations, summaries, and definitions, and pragmatic links define relationships with practical results, such as a warning message.

The ubiquitous HTML standard used on the Web includes mechanisms to specify link typing, with the CLASS, REL, and REV attributes of the anchor (A) and link (LINK) tags. However, a lack of standardisation on their values and meaning has resulted in them being largely ignored.

In this section, systems that explicitly incorporate and use link typing are discussed. This leads to the new linking standards being developed for the Web at the World Wide Web Consortium (W3C).

2.7.1 Textnet

One of the earliest works to address link typing in hypertext is Trigg's doctoral work (Trigg, 1983). Trigg created Textnet, a system implemented in Lisp, to support scientific activity in text creation, footnoting, and annotation. Trigg believed that eventually all scientific activity would move on-line and typed hypertext would provide the necessary infrastructure to support it.

There were four main concepts in Textnet: chunks, tocs, links, and paths. Chunks were fragments of text (e.g. part of a scientific publication), tocs corresponded to table of content entries, links connected chunk and toc nodes, and paths defined a series of connections defined by links (i.e. ordered list of nodes).

Trigg realised the value of "make[ing] explicit the relationship between two nodes" and introduced link typing through a taxonomy of 75 basic link types. These are broadly divided into normal links and commentary links. Normal links connect nodes of scientific work using types such as solution, summarization, and continuation. Commentary links connect statements with the node they refer to (e.g. solves, contradict, mislead, rambling).

Subsequent systems, such as Notecards (Halasz *et al.*, 1987) and VIKI (Marshall *et al.*, 1994), have further explored Trigg's link typing concept to improve authoring and navigational aspects of hypertext systems.

2.7.2 gIBIS

gIBIS (graphical Issue Based Information System) (Conklin & Begeman, 1988; Conklin & Begeman, 1989) was an argumentation system that used node and link typing. It graphically presented these and used the links to represent Rittel's IBIS model relationships (Rittle, 1972) developed in the early seventies. IBIS was based around the concept of an issue having statements (positions) made about it. Each position may attract arguments that either support or refute it. In gIBIS, issues, positions, and arguments form the three node types. There were nine types of links that could be specified between them: responds to, supports, objects to, questions, replaces, generalises, specialises, suggested by, and other.

gIBIS was used in organisational settings to improve meetings by capturing the essence of a meeting in real time (Conklin *et al.*, 2001). Early work, however, showed users unwilling to use gIBIS due to the inherent cognitive overhead. Therefore the system evolved into a broader approach of *collective sense making* called Compendium (Conklin *et al.*, 2001) that reduced cognitive overhead and increased flexibility by providing rapid benefits (e.g. in the quality and productivity of the meeting) and minimal learning and behavioural changes by the team to start using

it. Compendium has been used in over 60 settings during the last 10 years including lunar work simulations at NASA and strategic planning for a consortium of universities.

2.7.3 MacWeb

MacWeb (Nanard & Nanard, 1995) was a hypertext development environment that managed a network of nodes and links and allowed authors to type links and nodes giving them a high degree of semantic specification. Links in MacWeb were explicitly anchored to pieces of text and were binary and directional, although it was possible to traverse the link in the reverse direction. The authors saw the typing mechanism as a tool to be used at creation time when an author's intentions were clear, rather than after the hypertext had been created.

Types could be freely defined and were expressed using an object-oriented type structure, allowing the author to express complex relationships between types. In effect, each type behaved as a class meaning inheritance relationships could be defined between types. Methods represented in a scripting language could also be attached to types (and nodes) to specify their intended semantics and behaviour.

MacWeb was demonstrated using an electronic car repair manual to represent the knowledge in a task oriented manner. Using the typing mechanism, different information nodes were presented to a novice user as opposed to a professional mechanic. This enabled the hypertext to be adapted to different users and tasks.

2.7.4 Aquanet

Aquanet (Marshall *et al.*, 1991) was a visual mapping tool that used hypertext to help people explore the structure of knowledge and thereby assist them in interpreting information and organising their ideas. It drew from the work in Notecards (information structuring) and gIBIS (creating knowledge structures) and proposed a system with a strong distinction between nodes (basic objects) and links (relations).

All objects in Aquanet (nodes and links) were typed, structured, frame-like entities (i.e. classes). These were arranged in an object-oriented style meaning that objects could inherit properties. Each object had several slots (properties) which contained other objects (e.g. a link) or plain data. As link objects were added as slot values, it was possible to make changes to a link object and have that change affect all objects it appeared in. This enabled structural changes to be made on a larger

scale and highlighted Aquanet's focus on the importance of the entire knowledge structure.

Unlike previous systems, Aquanet introduced the concept of a schema to hypertext. Every Aquanet session was controlled by a schema, which defined the permissible object and relation types. This was used, for example, to create the Toulmin argumentation structure (Toulmin, 1958) that defines a basic argumentation model. The schema for the Toulmin structure declared one type of relation with five properties, one for each of the argumentation types. However, the schema language in Aquanet was somewhat restrictive as structural constraints could only be applied at a local level. Global restrictions, such as ensuring that only one instance of a type can be created, were also not possible.

2.7.5 XLink and XPointer

A serious problem with linking on the Web is the lack of functionality provided by the standard HTML linking mechanisms. Proprietary systems, such as Hyper-G, have proposed ways of adding functionality to the Web; however, these have not been adopted on a large scale. Changes must to be made to the underlying Web infrastructure to enable features such as n-ary linking and link abstraction.

The World Wide Web Consortium (W3C) has released two linking standards which were heavily influenced by earlier linking work in HyTime (Newcomb *et al.*, 1991): XLink and XPointer. XLink manages actual linking between resources while XPointers handle the precise anchor locations. XLink and XPointer operate on documents based on the eXtensible Mark-up Language (XML), a language that provides a document with a machine processible structure.

XLink (W3C, 2000c) provides a framework for specifying both simple and complex links (e.g. bi-directional and n-ary links) and attaching semantic information to these link specifications. XLink uses a powerful addressing language in XPointer, to accurately, reliably, and flexibly point to precise parts of a structured document. Hypertext functionality in XLink is divided along three axes: link location, link behaviour, and link complexity. Link location defines whether a link appears within a document or is recorded separately in a linkbase, link behaviour refers to the action associated with a link (e.g. when a link is selected), and link complexity defines how widely ranging in scope and application the link is.

The simplest hypertext style link is represented in XLink as:

```
<director xmlns:xlink="http://www.w3.org/XML/XLink/0.9"
          xlink:type="simple"
          xlink:href="fincher.xml">David Fincher</director>
```

The 'xmlns:xlink' and 'xlink:type' attributes can be removed if they are declared in an element declaration beforehand. This simple link can be enhanced to include semantic information, for example:

```
<director xmlns:xlink="http://www.w3.org/XML/XLink/0.9"
          xlink:type="simple"
          xlink:href="fincher.xml"
          xlink:title="A link to the director"
          xlink:role="http://www.host.org/roles/director"
          xlink:show="new"
          xlink:actuate="onRequest">David Fincher</director>
```

The title attribute presents a human-readable description of the link. The *role* attribute points to a URI describing the property (the *arcrole* property can also be specified if *role* alone does not suffice). The *show* and *actuate* properties define the behaviour of the link. In this instance the show value indicates that a new window is to be opened for displaying the content of the link destination. Other possible values include replace, embed, and other. The *actuate* property defines how the link is actuated. This conventionally happens when a user selects (clicks on) a link. This behaviour is equivalent to the *onRequest* value used in the example. However, the behaviour could have been defined to activate on loading the document (e.g. Xanadu's transclusion).

It is also possible to define links to be stored in linkbases. If we have the local resource:

```
<manager xlink:label="george">
   <first_name>George</first_name>
   <surname>Burley</surname>
</manager>
```

A definition of a remote resource:

```
<team xlink:label="itfc" xlink:href="itfc.xml"/>
```

The resources are bound using the *arc* type:

```
<manages xlink:type="arc"
         xlink:from="george"
         xlink:to="itfc"
         xlink:arcrole="http://pip/roles/manages"/>
```

The endpoints of the link are defined using the *from* and *to* properties. The *arcrole* property points to a URI describing the relation.

Unfortunately, XLink has yet to be implemented in any of the major commercial browsers[5], although a number of processors are available, such as X2X (Empolis, 2001) and XLiP (Fujitsu, 2001). However, a general lack of software has resulted in XLink not being fully utilised yet. Experiences in using XLink in an open hypermedia environment have demonstrated its suitability for capturing the link structures in an open hypermedia system, as well as the reverse, using open hypermedia systems to create XLink structures (Halsey & Anderson, 2000).

While XLink provides the mechanism for linking documents, it does not specify how to link to a particular part of a document. For this, XPointer (W3C, 2001b) has been defined which extends the XPath (W3C, 1999b) standard commonly used to address parts of XML documents.

An XPointer represents the fragment identifier (the part after the '#') of a URL. A simple XPointer is illustrated below.

```
#xpointer(/section[@name='foo'])
```

This XPointer addresses an element named 'section' that has an attribute 'name' with a value of 'foo'. A more complex example is:

```
#xpointer(string-range(//*, 'text'))
```

This expression matches any occurrence of the word 'text'. It is also possible to specify the precise part of an XML tree.

```
#xpointer(//AAA/BBB[2])
```

The second element named 'BBB' with a parent named 'AAA' is returned. While useful in some situations, this type of expression can only be used on documents that are unlikely to change. Any change in the structure of the document is unlikely to be reflected in the expression.

XPointers are combined with XLinks by appending the XPointer to the XLink. In the example below, the XPointer is used to define the destination anchor as simply the first 'section' (if there is one) of the document.

```
<team xlink:label="itfc" xlink:href="itfc.xml#xpointer(//section)"/>
```

[5]Partial implementations have been added to the Mozilla and Amaya browsers.

In earlier work, a significant problem with dynamic linking was the positioning of anchors. Systems such as the DLS used a combination of the phrase at the anchor position and a numerical offset. While this worked, it restricted the type and flexibility of anchors. For example, it was impossible to specify an anchor that matched the first heading in a document, or the first time the phrase 'IAM Group' appeared after the phrase 'Southampton University'. XPointer provides this level of functionality and enables both very general, as well as very specific, generic links to be created.

XLink and XPointer go a significant way to providing the hypertext functionality on the Web that the hypertext research community has been demanding for years. Semantics can be attached to links to describe their intended purpose which provides several key advantages: users can be notified of the intention of a link, links can be processed to adapt them to users and their tasks, and authors can organise their information more effectively. XLinks also do not need to be embedded within documents, but can be stored in separate linkbases to improve interoperability and integration. This enables users to create customised Web spaces through dynamic linking and reduces authoring overhead as modifications to links only need to be made in the linkbases and not individual documents. However, semantics describing how these facilities should be used have not been published.

A major obstacle for large-scale adoption, however, is that XLink and XPointer only function on structured XML-based documents, and not HTML. An XML-compatible version of HTML has been proposed in XHTML, and the adoption of this is crucial for the success of these new linking standards.

2.8 Navigating the Web

The networks of links created in hypertexts have created a new paradigm for browsing documents: navigation (or browsing). Retrieving documents becomes a matter of following links and is similar to Bush's "trail of interest". This is particularly useful when a user is unsure of the exact material required and uses navigation to explore the information space and home in on the relevant information.

2.8.1 Navigation vs. Retrieval

It is important to draw a distinction between navigation and retrieval as significant differences exist. Navigation is the act of traversing a hypertext either with an eventual goal in mind or simply to familiarise oneself with a topic area (e.g. finding general information on knowledge management). It is an exploration tool. The massive network of hypertext links on the Web provides extensive navigation opportunities.

Retrieval is the process of extracting precisely defined information from a system. Users who know the exact details of the resource they are looking for (e.g. a paper title) use this approach. In these situations, unless the hypertext has been very well engineered, information retrieval will lead to faster results (Waterworth & Chignell, 1991). A popular example of information retrieval on the Web is the service provided by search engines such as Google.

The major distinction between navigation and retrieval therefore lies in the nature of a user's information-seeking objective: unspecific vs. specific.

2.8.2 Navigation Issues

Navigating the ubiquitous information space on the Web and locating resources requires effort and perseverance (Nielsen, 1990; Cockburn & Jones, 1996; Andrews & Dieberger, 1996), something that has often been referred to as *lost in hyperspace* (Conklin, 1987). It is difficult to pinpoint the exact causes of this problem, although the following two points are frequently made:

- Hypertext systems cause disorientation and cognitive overload by overwhelming users with linking options and forcing them to make many decisions (Conklin, 1987; Young, 1990; Zellweger, 1991).
- The enormous amount of information on the Web leads to information overload as users find it difficult to cope with the over-abundance of information (Nelson, 1994; Cockburn & Jones, 1996; Lang, 1996).

Conklin (1987) has defined disorientation as "the tendency to loose one's sense of location and direction in a nonlinear document" and cognitive overload as "the additional effort and concentration necessary to maintain several tasks or trails at one time".

The link typing research discussed earlier is directed at the first point. By informing users of the purpose and intention of a link (through a link label for instance), users are able to decide whether a link is worth following. Systems can also analyse the link typing and only propose links when they are suitable for the user and task. However, although link typing improves a user's understanding of the hypertext, the underlying problem remains. An abundance of linking options still requires a significant overhead in analysing the links potentially worth traversing, and then making a decision. Indeed, link typing could cause users to become interested in more link destinations, which then leads to the problem described in the second point of information overload.

It is difficult to estimate the real size of the Web as many parts are not accessible to automated indexing agents, either because they are dynamically generated or because entry credentials are required. A survey carried out by Bright Planet (Planet, 2000) estimates that in reality that Web contains 500 times more information than previously imagined. This *deep* Web contains about 550 billion unique documents, compared to the estimates for the *surface* Web of 1 billion. Even if we look at dedicated information sources, such as digital libraries, a plethora of knowledge is available. For example, the ACM digital library[6] contains over 69,000 full-text articles while the physics archive, arXiv[7], contains almost 200,000 documents.

However, some fellow hypertext researchers dispute these problems. Bernstein (1991) states that "while the so-called 'navigation problem' has come to dominate hypertext research, evidence for its existence and nature is distressingly thin" and Landow (1990) comments on his experiences with Intermedia and proposes that "navigation and orientation are not in fact a major problem." While this may be the case for carefully constructed smaller hypertexts, such as the hypertext narratives that both Bernstein and Landow construct, this is not neccessarily true for the large-scale hypertext linking evident on the Web.

Research has been ongoing into methods of reducing disorientation and information overload and these are presented.

[6] `http://www.acm.org/dl`
[7] `http://www.arXiv.org`

Metaphors

Using metaphors as navigational aids has proven popular as they provide users with a recognisable and (therefore theoretically) intuitive interface for navigating the Web. This approach was used in systems such as OpenBook (Ichimura & Matsushita, 1993) and VOIR (Visualisation of Information Retrieval) (Golovchinsky & Chignell, 1996), which used a book and newspaper metaphor respectively. Highly related or sequentially arranged material is particularly suitable for this technique, due to the many metaphors that are available for these information structures. However, most approaches require resources to be marked-up with some form of additional information to enable the metaphor systems to parse and present the documents in a customised way.

Overview Maps

Overview maps reduce cognitive overload by presenting neighbouring Web sites in addition to the currently viewed one, thereby aiming to improve contextual awareness. A proximity measure is usually used to present only the relevant links, for example, only including the direct links. Robert *et al.* (1998) propose a system whereby an overview of all neighbouring sites is displayed including a zooming facility to increase or decrease the number of visible nodes. This concept has been improved by representing landmark nodes on the visualisation: nodes which have a high connectivity and access frequency (Mukherjea & Hara, 1997). This approach makes it easier for users to focus on only the relevant resources.

Cybermap (Gloor, 1991) uses automatic indexing and clustering techniques (Salton, 1989) to partition related nodes into *hyperdrawers* and then use fish-eye view filtering (Furnas, 1986) to produce a visualisation. In author co-citation analysis (Chen & Carr, 1999b), citation links in published literature are used as the proximity measure, mapping authors who frequently cite each other together in the same visualisation space, making fields of research identifiable.

Spatial hypertext is a particular type of overview map where spatial and visual cues are provided for browsing the structures evident in hypertext. For example, VIKI (Marshall *et al.*, 1994) uses visual symbols to construct hypertext structures and facilitate the exploration and understanding of the context surrounding any node.

Critics note the additional overhead and complexity of overview maps as processes have to analyse content to produce an effective map (Bernstein, 1990). Maps also become quickly cluttered as the number of nodes and links increase.

Adaptive Hypermedia

Adaptive hypermedia systems tailor hypermedia facilities to suit particular features of a user. The main objective is to increase the navigation efficiency of a user, either by reducing the time required to locate relevant information, or by increasing the amount of information the user can consume.

For example, Bailey *et al.* (2001) propose a system where link augmentation is used to insert relevant links directly into documents. Unlike systems such as Microcosm however, where simple link insertion algorithms can lead to incorrectly placed links due to the system's failure to understand a user's browsing context, a user profile is created by analysing the navigation trails of previous sessions. However, systems that transparently collect user information (in this case, the documents viewed) raise issues of confidentiality.

Adaptive hypermedia has been an effective tool for presenting educational material, and indeed it has been mainly used as a didactic tool (Calvi & Bra, 1997; Bra & Calvi, 1998; da Silva *et al.*, 1998). Students are directed towards appropriate educational material through the accurate placement of links. Walden's Paths (Furuta *et al.*, 1997) enable teachers to use path authoring tools and a path server to create fixed paths for students to navigate through educational resources. Teachers carefully select the most appropriate path that will give students a successful learning experience.

Collaborative-based Linking

Collaborative systems such as MEMOIR (Managing Enterprise Multimedia Using an Open Framework for Information Re-use) (Pikrakis *et al.*, 1998), use the collective experience of a group of registered users to assist individuals. If a user is searching for specific information, it is feasible that another member of the community has already located a relevant document. To be successful, this method requires a large user-base with a common interest and goal.

Alexa (Alexa, 1997) is a collaborative Web service that, like MEMOIR, learns from its users' surfing habits and uses this information to suggest related sites, in

addition to usage statistics, site ratings, and site information. In addition, trails are analysed for patterns which are used to suggest popular sites.

2.9 Summary

This chapter has presented an overview of hypertext research. Early visionaries, in particular Vannevar Bush and Douglas Englebart, are credited with the hypertext concept. Theodor Nelson later proposed and defined the term as "non-sequential writing". Early hypertext systems, such as NLS and KMS, were primarily constructed to explore the hypertext concept and consequently were monolithic in design and were therefore harder to maintain, extend, and integrate (e.g. share linking data). These systems are termed closed hypertext systems as they represent rigid encapsulated tools that make interoperability difficult (e.g. links are embedded within mark-up in a document) and rarely work outside the scenarios for which they were constructed.

Notecards, Intermedia, and Hypercard were popular hypertext systems that introduced the concept to a larger audience. Notecards and Intermedia were especially influential as they introduced implementations of early concepts such as openness (e.g. abstracting the link) and link typing. More recently, hypertext research has continued to demonstrate the benefits of openness (Microcosm, Multicard, Devise Hypermedia) and adding semantics to links, anchors, and nodes (gIBIS, Aquanet, MacWeb). The former systems are referred to as open hypermedia systems as they have the ability to integrate hypertext data created in different applications and formats and thereby improve *interoperability* between hypertext services. Open systems typically abstract hypertext services into three layers: application, link, and storage.

Hypertext enables a unique method of navigating material: traversing the links created between documents. This is mainly useful for users wishing to 'browse' for information (e.g. to familiarise oneself with a subject), although it has introduced problems of user disorientation and cognitive overload.

The hypertext concept is used in a global distributed information system, the World Wide Web, although many important hypertext advances, such as openness and link typing, are not supported. Nevertheless, the Web provides access to an enormous amount of interlinked information, which has made it a popular medium.

Three main research issues are drawn from this chapter:

- Semantics: Adding typing (or semantic) information to links, anchors, and nodes enables more insightful navigation, improves contextual awareness, and assists authors in organising and creating a purposeful structure to a hypertext. Link semantics are later used to connect scholarly material using the *ontological hypertext* principle.
- Abstraction: Separating the different layers of a hypertext service improves interoperability, customisability, integration, and enables the layers to be processed and analysed (e.g. link structures can be analysed to help propose a better link set to the user). This concept of viewing links as independent semantic specifications has also influenced the *ontological hypertext* principle.
- Navigation: By following the hypertext links between documents, a powerful method of viewing and exploring interrelated information is possible that allows users to quickly gather and relate information. This form of information exploration is especially suitable to scholars who frequently explore their research field without details of specific papers or projects, and therefore home in on relevant and related research.

These issues highlight that the real power of hypertext lies in its dual nature: simultaneously a mechanism for mapping and visualising an information space, as well as a tool for representing its underlying semantics.

Hypertext plays a central role in this research; ESKIMO uses it to interconnect intricately related scholarly material on the Web based on the semantics of their relationships, and present researchers with a coherent and intuitive exploration environment.

The next chapter introduces the cutting edge of the Web, the Semantic Web, and further explores the advantages of adding semantics to documents and the relationships between them.

Chapter 3

The Cutting Edge of the Web: The Semantic Web

3.1 Introduction

Hypertext enables documents to be organised in a non-sequential manner and presents users with a novel way of exploring information spaces. This technology is being used on a global scale to create the Web. However, hypertext navigation introduces problems of information and cognitive overload (Conklin, 1987; Cockburn & Jones, 1996), especially on the Web where hypertext advances, such as using explicit link semantics, have not been adopted.

The Web is predominantly intended for human consumption and understanding, and machines are unable to mimic this. Therefore, users rely on primitive indexing and/or considerable human effort to navigate the Web and retrieve information from it. Rutherford D. Roger has previously succinctly pointed out that "we are drowning in information and starving for knowledge".

This chapter discusses Tim Berners-Lee's vision of the Semantic Web (Berners-Lee, 1998; Berners-Lee *et al.*, 2001), also known as the Programmable or Knowledge Web, which advances the Web and its basic hypertext foundation, to provide facilities such as intelligent navigation and information retrieval, automated use of distributed information sources, and knowledge services.

The main principles of the Semantic Web are the representation and application of Web-based knowledge. These are used in ESKIMO to provide a comprehensive research tool that (i) produces a unique approach to interlinking scholarly material

by creating a consistent, well-linked, and *principled* hypertext and (ii) provides a scholarly inquiry service to enable researchers to ask pertinent questions about their research field.

The Semantic Web is not a new Web, but an extension that solves the real problem on today's Web: the communication of information and knowledge. The W3C, highly active in this initiative, sums up the Semantic Web activity as:

> Facilities to put machine-understandable data on the Web are becoming a high priority for many communities. The Web can reach its full potential only if it becomes a place where data can be shared and processed by automated tools as well as by people. For the Web to scale, tomorrow's programs must be able to share and process data even when these programs have been designed totally independently. The Semantic Web is a vision: the idea of having data on the web defined and linked in a way that it can be used by machines not just for display purposes, but for automation, integration and reuse of data across various applications (Miller *et al.*, 2001).

The consequence of this realisation will have a profound effect on the Web and the way we access it. While doubt has been voiced over its feasibility and practical benefits as social and technical challenges have to be met (Uschold, 2001; Haustein & Pleumann, 2002), many researchers and corporations are taking the proposal seriously. For instance, the Defence Advanced Research Projects Agency (DARPA) has spent 80 million dollars on researching issues and technologies for the Semantic Web (Fensel, 2000).

The first section of this chapter discusses metadata, a fundamental building block of the Semantic Web, and the significant metadata standards, which have evolved. Then the architecture and enabling technology of the Semantic Web are introduced.

3.2 Before the Semantic Web: Metadata

Metadata is *structured* data about data. Usually it is descriptive information about some resource for improving a machine's understanding of it. Metadata is vital for

the fabric of the Semantic Web as it enables machines to understand the content of Web resources.

While the term metadata is relatively new, the concept has been around for some time with the canonical example being library catalogues. Librarians have used metadata to classify and index their literature, most notably using the MAchine-Readable Cataloging (MARC) standard (Mar, 2000). As the name suggests, the standard proposes a machine-readable format for the representation and communication of bibliographic data.

As this section will demonstrate, research in using metadata on the Web has been ongoing for many years and has provided the necessary grounding for Semantic Web research. Earlier work on hypertext link semantics (Trigg, 1983; Marshall *et al.*, 1991; Nanard & Nanard, 1995) could also be considered metadata as semantics were *explicitly* added to links, although the Web did not adopt this feature. However, often the semantics were embedded in the hypertext system itself and not exposed to external processes, meaning the metadata was not machine-readable.

Metadata is usually associated with Web resources using one of the following approaches:

- Distributed: metadata is embedded along with the information it describes
- Centralised: metadata is stored separate from the information it describes in a centralised repository

Early systems adopted a distributed approach and metadata was stored alongside the content it described. Indeed, the basic metadata constructs provided by HTML are inserted directly into documents. The advantage of this approach is that each resource is a self-contained unit that can be easily accessed by processes. In addition, in an environment where a large collection of unrelated metadata is authored by many users, it is impractical to have a single metadata repository as the successful integration of the diverse users and their metadata is likely to fail. Instead, it is more practical for each user to add the necessary metadata to their data as required, resulting in a community driven approach. However, using a distributed approach increases the maintenance overhead as the metadata has to be retrieved from all the documents/sources it appears in and any changes to it must be made to all affected files.

The second approach results in a scalable, reusable, and more maintainable system. Furthermore, it is easier for search engines and other indexers to locate the metadata as they simply refer to the metadata repository rather than having to locate and parse a large collection of documents. A centralised approach also usually guarantees that all metadata is represented using the same standard and syntax so translation problems are unlikely. However, metadata processes have to be aware of the location of the repositories or face disregarding large quantities of metadata. Frequently accessed repositories can also lead to communication bottlenecks as they struggle to complete all requests.

A large number of different metadata proposals and standards are available, which reflects on the activity of this research area and the various and diverse applications that require metadata. In this section, common metadata standards are reviewed and a discussion of the new standards aimed at the Semantic Web is presented.

3.2.1 Attribute-based Metadata

Early metadata standards provided a method of assigning textual values to attributes (or properties) of a document. For instance, a document may have the properties 'creator' and 'date', and respective values of 'Tom Mills' and '12/05/1995'. This form of data is simple to author and parse by processes, and is also suitable for basic indexing by search engines.

Global Information Locator Service (GILS)

The Global Information Locator Service (GILS) (GILS, 1997) is a metadata proposal first presented in 1994 as an open standard to define the most commonly understood concepts people use to find information (mainly literature), such as *title* and *author*. Its objective is to assist organisations in providing a way for their users to use standard methods to find information within an organisation's knowledge base. However, the standard does not specify the representation or semantics of these properties. This is the responsibility of the particular implementor, who simply guarantees to support the GILS elements.

Dublin Core Initiative

In March 1995, invited professionals from disparate computer science fields met in Dublin, Ohio[1], to discuss improvements in the description, access, and discovery of resources on the Web. This resulted in the proposal of a simple set of elements (or properties) suitable for both experienced and naive users, to describe on-line resources. These elements became known as the Dublin Core Metadata Element Set (DCMES) (DCMI, 1999). The initiative is similar to the GILS standard but is aimed directly for use on the Web.

The DCMES provides a vocabulary for describing core properties of Web resources, such as 'creator' and 'date'. As with GILS, the initiative only specifies the properties, rather than the syntax required to represent them. For example, a representation has been proposed using the Resource Description Framework (RDF), a framework for specifying metadata, and an example is illustrated below.

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:dc="http://purl.org/dc/elements/1.1/">

  <rdf:Description rdf:about="http://media.example.com/audio/guide.ra">
    <dc:creator>Rose Bush</dc:creator>
    <dc:title>A Guide to Growing Roses</dc:title>
    <dc:description>
      Describes process for planting and nurturing different kinds of
      rose bushes.
    </dc:description>
    <dc:date>2001-01-20</dc:date>
  </rdf:Description>
</rdf:RDF>
```

The above code fragment defines metadata about the Web page located at `http://media.example.com/audio/guide.ra`. The Dublin Core elements are identifiable by the *dc* namespace tag before the property names. In this fragment, four Dublin Core elements have been used: 'creator', 'title', 'description', and 'date'.

The complete DCMES contains 15 elements. While this sounds sparse, the objective of the initiative is to create a simple and interoperable element set that can be understood and used by as large a group of users as possible. In contrast, complex standards such as the Machine-Readable Cataloging (MARC) standard are difficult to adopt for ubiquitous use on the Web. It is precisely the simplicity

[1]The meeting was held in Dublin; the location of the headquarter for the Online Computer Library Center (OCLC) organisation. The OCLC is a nonprofit organisation exploring technologies related to library catalogues.

of Dublin Core that enables it to be easily represented in most metadata formats and is the reason for its prominence.

IAFA templates

The IAFA templates (Deutsch *et al.*, 1995) are designed to index FTP archives, inheriting much from the Linux Software Map (LSM) (Kopmanis & Wirzenius, 1994) proposal. IAFA propose 14 templates, such as user, organization, service, document and software. Each template contains attributes used to describe that template type. For example, the event template is used in the following example to describe a file for a conference call. The property names of this template are indicated as words at the beginning of a line followed by a ':'.

```
Template-Type: EVENT
Description: Call for papers for the ParCo'95

Topics: Applications and Algorithms; Systems Software.

Deadlines: Abstracts: 31st January 1995; Notification: 15th April
1995; Posters: 30th June 1995.

Author-Email: a.n.author@host.site.country
Author-Name: A. N. Author
Title: Fifth International Conference on Parallel Computing
X-End-Date: 1995-09-22
X-Start-Date: 1995-09-19
Last-Revision-Date-v0: Wed, Jan 11 11:24:39 1995 GMT
```

Summary Object Interchange Format (SOIF)

The Summary Object Interchange Format (SOIF) (Wessels, 1996) is used by the Harvest (Bowman *et al.*, 1995) system which is an integrated set of tools to gather, organize, and search for information across the Internet. SOIF is based on work from the IAFA templates and the bibliography utility, BibTeX, however, unlike these standards SOIF is designed to support streams and binary content. This means that it can be used to describe video, images, compressed files, and postscript documents as well as text documents like program code, HTML, and raw data.

Harvest *Gatherers* construct summaries for objects and record these in the SOIF format which *Brokers* collect and index. Simple and structured queries can then be issued to the broker to search for information. The metadata format is in the form of attribute-value pairs. The following example illustrates the use of SOIF to declare the title and author of a Web page.

```
@DOCUMENT { http://www.best.com/~jocelyn/resdogs/index.html
title{20}: Rescuing English Springer Spaniels
author{29}: Jocelyn Becker
}
```

Meta tags in the Hypertext Markup Language

The Hypertext Markup Language (HTML) (Berners-Lee, 1992) is the de facto language used on the Web to mark up billions of documents, therefore representing a paradigmatic standard for metadata publication. Unfortunately, HTML has been designed almost entirely as a presentation format and therefore contains little in the form of semantic markup. However, two attributes ('description' and 'keyword') of the 'meta' tags can be included in the HTML header to describe, in basic terms, the content of the document. For example:

```
<HEAD>
  <TITLE>GPRS</TITLE>
  <META name="description" content="This page discusses GPRS">
  <META name="keywords" content="gprs,mobile,2.5g">
</HEAD>
```

These properties enable authors to include basic metadata, although this is not an ideal solution due to the lack of structure and detail in the content fields. There is also no defined method or standard describing what the content should contain. Nevertheless, several search engines (e.g. AltaVista) use these tags to improve document indexing.

The metadata facilities available in HTML were improved with the introduction of the Platform for Internet Content Selection (PICS) (Miller, 1996). Originally designed to add access control to sensitive documents (e.g. pornography), it has also been used for code signing and privacy.

The syntax for PICS is compatible with HTML:

```
<META http-equiv="PICS-Label"
      content='(PICS-1.1 "http://www.rsac.org/ratingsv01.html"
      comment "RSACi NorthAmerica Server"
      for "http://www.foobar.org"
      on "200.06.16T10:30-0500"
      ratings (v 3 n 4 s 3 l 2))'>
```

The example PICS code is inserted into the head section of an HTML document where PICS aware processes parse them and use them to determine if the content of the corresponding resource is suitable for viewing. Each PICS rating is assigned an

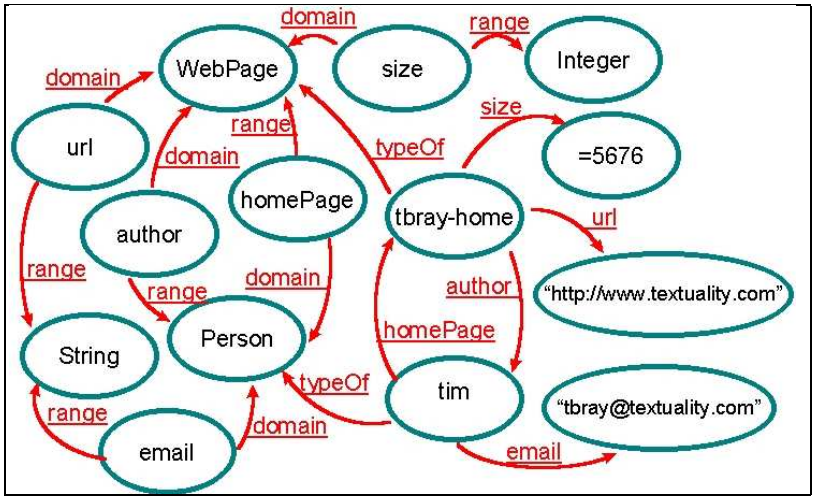


Figure 3.1: MCF as a directed linked graph

integer value from 0 to 4. In this example, the page `http://www.foobar.org` has been assigned a violence (v) rating of 3, a nudity (n) rating of 4, a sex (s) rating of 3, and a language (l) rating of 2 (so probably a page minors should avoid!).

3.2.2 Object-based Metadata

Attribute-based metadata restricts authors to making simple statements about Web pages, such as a 'has creator Richard'. Object-based metadata provide more flexibility by introducing the ability to make further statements about the objects in documents. For example, in addition to stating that a document has an author 'Tim Berners-Lee', further statements can be made indicating the author's e-mail address and homepage.

Meta Content Format (MCF)

The Meta Content Format (MCF) (W3C, 1997) is an XML-based metadata proposal. In essence, the model represents a directed labelled graph as illustrated in Figure 3.1, where nodes contain either objects or properties.

Objects can represent anything, although usually point at Web resources. Significantly, MCF metadata is stored *external* to the documents it describes. For example, the following code describes two Web pages. Each page is of type 'TextualityPage'. In this case, 'TextualityPage' is used to represent the general description of a Web page on the Textuality Web site. It inherits the author and copyright properties which it sets to 'TextualityInc' and 'TextualityServices, Inc. All rights reserved.' respectively.

```
<WebPage id="w0001">
  <url>http://www.textuality.com/</url>
  <typeOf>TextualityPage</typeOf>
</WebPage>

<WebPage id="w0002">
  <url>http://www.textuality.com/Lark/</url>
  <typeOf>TextualityPage</typeOf>
</WebPage>

<Category id="TextualityPage">
  <superType unit="WebPage"/>
  <inherits propertytype="AuthorOrg"  unit="TextualityInc"/>
  <inherits propertytype="CopyrightNotice">
   TextualityServices, Inc. All rights reserved.
  </inherits>
</Category>
```

A metadata process parses this description to quickly gain a complete appreciation of documents on the Textuality Web site, rather than locating and downloading the individual pages which make up the site. The latter method also depends on the accuracy and completeness of the hypertext linking between the site's documents, as this is used to locate all the pages.

XML as a metadata language

Several standards have recently emerged from the World Wide Web Consortium (W3C) directed at the issues of metadata and semantic interoperability. Arguably, the single most influential standard to emerge is the Extensible Markup Language (XML) (W3C, 2000a), a cut-down Web-ready version of the Standard Generalized Markup Language (SGML) (International Organization for Standardization, 1986). XML is used to describe the structure and content of a document making it possible for machines to parse it.

The syntax used by XML is similar to HTML. Content is surrounded by element tags and attributes and entity references are supported. However, unlike HTML, XML enforces *well-formedness*, a set of rules to guarantee the consistent syntactical representation of data (e.g. all tags are closed, attributes are quoted, nesting is valid). For example, to mark up a short document describing a video, the following representation is possible.

```
<?xml version="1.0"?>
<!DOCTYPE video SYSTEM "http://www.imdb.com/video.dtd">
```

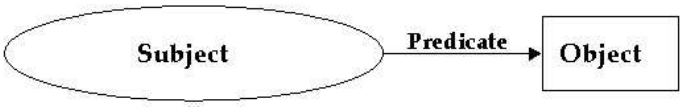


Figure 3.2: The RDF data model

```
<video>
  <title>Don't let the Bedbugs bite</title>
  <producer>Mark Weal</producer>
  <year>2001</year>
</video>
```

The *DOCTYPE* line declares the type of the document and points to a Document Type Definition (DTD) where the valid elements and entities of this XML file are declared (i.e. its grammar). The inclusion of a DTD is optional, although the benefit of defining and including a DTD is to promote interoperability. Other users then use the DTD to produce semantically equivalent documents that the same processes are able to understand. DTDs also enable the production of international standards.

The elements in the example, video, title, producer, and year, are used to define the properties of a video. The above representation is not archetypical for expressing video details in XML. In fact, there are an infinite number of ways this could have been represented.

For reasons that will be clarified later, XML is not ideal for metadata purposes. However, XML represents a general, extensible, and open standard that provides an effective framework to define further metadata standards.

Resource Description Framework (RDF)

The Resource Description Framework (RDF) (W3C, 1999a) is a further standard proposed by the W3C and is based on XML. RDF is aimed explicitly at handling metadata and is viewed by many as the ideal foundation for the Semantic Web. The basic RDF data model consists of a triple: subject, predicate, and object (Figure 3.2).

These can also be considered as a resource, property, and literal respectively. A property about a resource (i.e. a statement or assertion) is represented conceptually

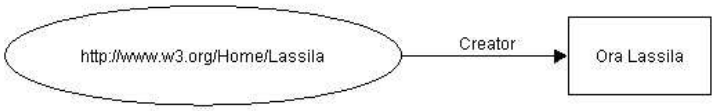


Figure 3.3: An example of the RDF data model

using a directed labeled graph as illustrated in Figure 3.3. The corresponding RDF is presented below.

```
<rdf:RDF>
  <rdf:Description about="http://www.w3.org/Home/Lassila">
    <s:Creator>Ora Lassila</s:Creator>
  </rdf:Description>
</rdf:RDF>
```

The resource identified by `http://www.w3.org/Home/Lassila` has a 'creator' predicate with a value 'Ora Lasilla'. Alternatively, the 'creator' property could point at a resource.

Fellow researchers have frequently commented on the necessity of RDF given the prominence of and similarity to XML. This is an important issue to address as it highlights two fundamental aspects in representing metadata. Firstly, while XML is a suitable interchange format, it is unsuitable for metadata purposes due to the way data is modelled in XML. XML does not set restrictions on the structure of the representation of its data. For example, the video example can be represented in multiple ways:

(1)
```
<video>
  <name>Don't let the Bedbugs bite</name>
  <producer>Mark Weal</producer>
  <date year="2001"/>
</video>
```

(2)
```
<video name="Don't let the Bedbugs bite" year="2001">
  <producer>Mark Weal</producer>
</video>
```

(3)
```
<videos>
  <video name="Don't let the Bedbugs bite"
         year="2001"
         producer="Mark Weal">
</videos>
```

The various ways of representing this data greatly increase the parsing and (necessary) transformation overheads. It also makes querying complex and inefficient,

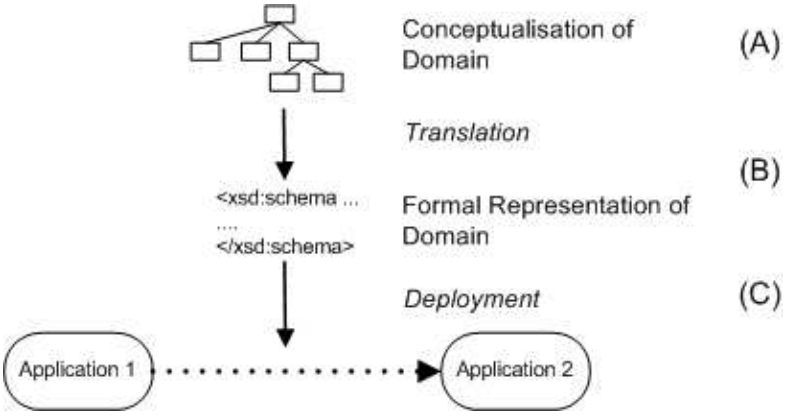


Figure 3.4: XML Deployment

as queries must function over the different possible structures. Metadata is typically accumulated in large quantities and therefore a simple and efficient format is required to enable rapid parsing.

Secondly, XML is aimed at defining the structure of documents rather than imposing any common semantic interpretation. This is illustrated in Figure 3.4 and expanded on in (Decker *et al.*, 2001). When defining complex metadata, a conceptualisation of a domain is defined (A). This is then translated into a schema, which is used to structure the XML to represent the metadata (B). However, when the XML is deployed, all semantic meaning that was originally declared in the schema is lost (C). It is impossible to determine the original relationships and hierarchies that were defined in the schema. In RDF this is not the case as its object-attribute structure is represented naturally; metadata is mapped directly into RDF's data model and so its semantics are not lost.

However, opponents of RDF also criticise the multiple ways in which RDF document trees can be represented (Haustein, 2001), causing problems with transformation languages such as XSLT. There is some foundation to this claim as there are two main syntax styles for RDF: an abbreviated form and a basic form. These syntactical variations do add additional overhead to parsers and style sheets that must parse the two structural variations. In addition, RDF also evades the primary objective of XML, which is a compromise between a human and machine-readable representation, as RDF is an order of magnitude harder to comprehend. XML and RDF also raise concerns over their modelling capabilities, a pivotal requirement for representing knowledge on the Semantic Web, and this is an issue addressed in the

next chapter.

3.2.3 Metadata on the Web

There are limitations in both attribute and object-based metadata. Attribute-based metadata provides an approach to attaching properties to documents but further statements cannot be made about the properties. Object-based metadata enables objects to be attached to documents and further statements can be made about them, meaning that a more accurate representation of a document's data is possible. However, object-based metadata is limited, as for instance, hierarchical relationships are not naturally supported and objects cannot be constrained (e.g. stating that a paper must have exactly one title). This means that it is impossible to describe accurately a document in terms of the domain it applies to and results in object-based metadata lacking the rigour required to support knowledge services.

The Web contains very little metadata (Pam, 1995) and the metadata that is available is mainly basic and attribute-based. A few popular tools use metadata, such as the AltaVista search engine which uses the metadata tags in HTML for indexing documents, while many other metadata systems (especially those employing more advanced forms of metadata such as concept-based metadata) are designed for restricted applications or user base (Deutsch *et al.*, 1995; Wessels, 1996). As a consequence, tools such as search engines are inaccurate as users are swamped with erroneous results that they must filter through (Eastman, 1999). Surveys have further demonstrated that users are not satisfied with results from search engines (Kobayashi & Takeda, 2000; Kwok *et al.*, 2001), citing the high number of incorrect results as the main problem.

Comprehensive metadata that accurately describes a document and the concepts in it, enables search engines to (i) reduce the number of erroneous search results by being able to understand the context of documents, (ii) respond to more intricate or involved queries (e.g. 'Which documents discuss pollution levels from cars manufactured in the Far East?'), and (iii) suggest related sites. Furthermore, processes with an understanding of the environment in which they operate can automate negotiation, create intelligent hypertexts, improve information retrieval,

accurately classify documents, and provide knowledge services to analyse information and uncover facts. For this level of functionality, the Semantic Web has been proposed and is the focus of the next section.

3.3 The Semantic Web

The Semantic Web initiative proposes a knowledge Web. Berners-Lee predicts:

> Properly designed, the Semantic Web can assist the evolution of human knowledge as a whole (Berners-Lee *et al.*, 2001)

In Berners-Lee's "Semantic Web Roadmap" (Berners-Lee, 1998), he envisions a Web where all content is machine understandable. Currently, the vast majority of the information available on the Web has been designed for human consumption and understanding. Machines process, analyse, and index these pages, however they cannot appreciate and comprehend their content and therefore cannot engage in any meaningful discourse about them. This is because (i) natural language processing facilities are not, and will not in the near future, be adequate for accurately discerning a document's content and (ii) Web documents lack structure and computer consumable *knowledge* for machines to readily parse.

The Semantic Web movement is comparable to the *'knowledge is power'* craze that started over 20 years ago. At the time, two communities quickly arose: the knowledge acquisition and knowledge representation communities. Knowledge acquisition proved costly and the systems developed for knowledge representation were mainly small and brittle and provided moderate solutions to minor problems. While the tasks and aims are similar to the Semantic Web, it did not have the advantage of an entire Web and its *inexpensive* workforce of millions to perform the knowledge acquisition. Social issues are influential in this task as users strive to make their information available on a global scale. As in the earlier movement, the acquisition and representation communities are at the centre of research for the Semantic Web (Fensel & Musen, 2001).

The metadata envisaged for the Semantic Web is more comprehensive and intricate than the simple attribute-value pairs offered by standards such as GILS, Dublin Core, and SOIF, and the basic object model proposed in MCF and RDF. To enable the advanced services for the Semantic Web, metadata must be expressive enough

to represent and model the domain it applies to. For example, adding metadata to a document stating that it has an author 'Brett Reynolds' and covers the topic 'pollution', assists index and search engines in classification (and indeed this was the objective of early metadata work), but this form of metadata does not enable high level cognitive questions to be posed such as *How does the combustion of fossil fuels affect the environment?*, *Has Reynolds published any papers on alternative energy technologies that produce less pollution?*, and *Who are the colleagues of Brett Reynolds that also research the effects of pollution?* This is because processes do not understand what an author, document, and topic are, how they are related, and how they relate to concepts in other documents on the Web.

For this functionality to be realised, metadata has to be able to state and represent the concepts: person, pollution, energy, energy technology, environment, and fossil fuel and specify the relations that exists between them (e.g. energy technology can produce pollution, pollution affects the environment). It is also impossible to constrain properties in conventional metadata, for example, by stating that an energy technology uses a maximum of one type of renewable resource.

Armed with an accurate model of a domain and the constraints that help define it, search engines can improve the indexing of resources and respond to user requests accurately, trading agents can intelligently negotiate over products, and information can be accurately personalised and modified to best meet a user's task. The basis of the Semantic Web then is the ability to represent and accurately model real-life domains and enable machines to gain a complete understanding of the environment in which they operate.

This section discusses Berners-Lee's vision in more detail and outlines the proposed architecture.

3.3.1 Scenario

A scenario demonstrating the benefits of a Semantic Web was published by Berners Lee *et al.* in May 2001 in Scientific American (Berners-Lee *et al.*, 2001).

> The entertainment system was belting out the Beatles' "We Can Work It Out" when the phone rang. When Pete answered, his phone turned the sound down by sending a message to all the other local devices that had a volume control. His sister, Lucy, was on the line from the doctor's office:

> "Mom needs to see a specialist and then has to have a series of physical therapy sessions. Biweekly or something. I'm going to have my agent set up the appointments."
>
> At the doctor's office, Lucy instructed her Semantic Web agent through her handheld Web browser. The agent promptly retrieved information about Mom's prescribed treatment from the doctor's agent, looked up several lists of providers, and checked for the ones in-plan for Mom's insurance within a 20-mile radius of her home and with a rating of excellent or very good on trusted rating services. It then began trying to find a match between available appointment times (supplied by the agents of individual providers through their Web sites) and Pete's and Lucy's busy schedules. (The emphasized keywords indicate terms whose semantics, or meaning, were defined for the agent through the Semantic Web.)
>
> In a few minutes the agent presented them with a plan. Pete didn't like it...He set his own agent to redo the search with stricter preferences about location and time. Lucy's agent...automatically assisted by supplying access certificates and shortcuts to the data it had already sorted through.
>
> Almost instantly the new plan was presented.

This scenario describes a knowledge driven world where any information is instantly available and machines are able to understand and process knowledge with minimal human interaction. The scenario presents the eventual goal that the Semantic Web may deliver although this is unlikely to occur in the short to medium term future, and even then, it relies on the widespread adoption and use of the relevant technologies. However, projects and technologies that provide some of this functionality have been developed and will be explored in the next chapter.

3.3.2 Architectural Model

Figure 3.5 illustrates the seven architectural layers that Berners-Lee envisages for the Semantic Web. The foundation layer of this architecture provides the basic addressing protocol (URI[2]) and document encoding method (Unicode [3]). The other

[2]Uniform Resource Identifier (URI) provides a more general addressing scheme than the URL standard.

[3]Unicode is a system for the interchange, processing, and display of texts from diverse languages.

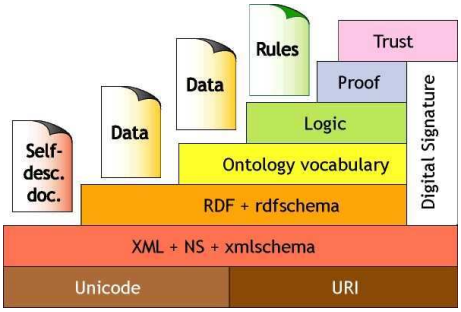


Figure 3.5: Architectural model of the Semantic Web (Berners-Lee, 2000)

layers build on this to provide the metadata and knowledge aspects of the Semantic Web. The importance of the underlying XML technology is immediately apparent.

The Schema Layers

The schema layers supply the structural and basic modelling capabilities. They are general and provide the foundation onto which further layers are added. The schema layer defines objects and relationships and constrains these so an accurate representation of a domain is made. For example, when describing the concept of a son and a father, it is important to add the constraint that a son can only have one father. It is also useful to represent hierarchies, for example, a *Snake* is a member of the *Reptile* family of animals and therefore inherits its properties.

XML Schema (XMLS) (W3C, 2001c) has been proposed as the underlying schema language. DTDs, while popular, lack the expressiveness to properly specify and *constrain* the elements in a document and its syntax greatly differs from that of XML, meaning that DTD files cannot be parsed and processed by the same processes that read the XML files.

However, XMLS does not contain any provision to model classes as it is only used to define a grammar. The RDF Schema (RDFS) (W3C, 2000b) specification extends XMLS and complements RDF to provide a data typing model and basic object oriented facilities. It also adds basic modelling capabilities as the next example illustrates.

```
<rdf:RDF xml:lang="en"
         xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">

<rdf:Description ID="Person">
  <rdf:type resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
  <rdfs:subClassOf
```

```
   rdf:resource="http://www.w3.org/2000/01/rdf-schema#Resource"/>
</rdf:Description>

<rdf:Description ID="Student">
  <rdf:type resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
  <rdfs:subClassOf rdf:resource="#Person"/>
</rdf:Description>

<rdf:Description ID="Course">
  <rdf:type resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
  <rdfs:subClassOf
   rdf:resource="http://www.w3.org/2000/01/rdf-schema#Resource"/>
</rdf:Description>

<rdf:Description ID="attends">
  <rdf:type
    resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
  <rdfs:domain rdf:resource="#Student"/>
  <rdfs:range rdf:resource="#Course"/>
</rdf:Description>
```

Significantly, concepts are defined as classes or subclasses of other concepts. Relationships between classes are possible. In the example, the *attends* property declares a relationship between a *Student* and a *Course*. Properties are attached to classes by defining a property element and declaring its domain and range.

The schema is then used to construct RDF statements as the following example demonstrates. This defines a relationship between a particular student and a course.

```
<rdf:RDF xml:lang="en"
         xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:st="http://www.soton.ac.uk/myschema/student#">

<st:Student rdf:about="http://www.soton.ac.uk/home/srk">
  <st:attends rdf:resource="http://www.soton.ac.uk/course/cm142"/>
</st:Student>

</rdf:RDF>
```

The RDF data model can be used to reduce this expression into an assertion (or triple); ideal for efficient processing and analysis by subsequent processes and is later demonstrated in ESKIMO.

```
triple('attends', 'http://www.soton.ac.uk/home/srk',
                  'http://www.soton.ac.uk/course/cm142').
```

The result of these schema layers is a widely interoperable and self-describing document that forms the basis of a Semantic Web resource.

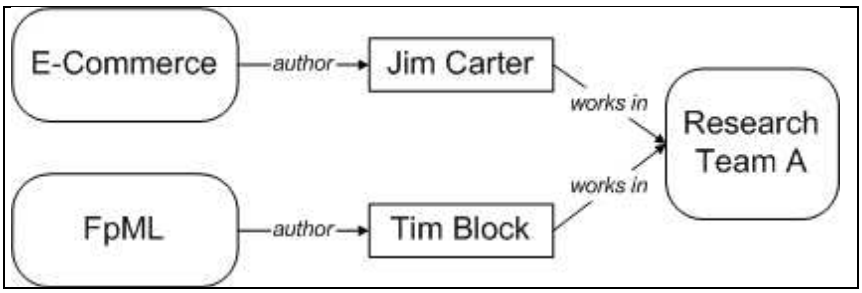


Figure 3.6: Simple inference example

Ontology Layer

The ontology layer adds further meta-information to define concepts and relationships accurately using constructs not available in the schema layers. An ontology is a conceptualisation of a domain and has been extensively used to define complex domains for the purposes of machine understanding and processing. Technologies designed for this layer focus on representing ontological constructs in a machine-readable format. Chapter 4 explores this layer in more depth as it plays a central role in this research and is later used to model artifacts in the academic domain.

Logic Layer

The logic layer is tightly integrated with the ontology layer and adds inference ability (the derivation of new data from existing information) to the Semantic Web. This is achieved by using the declarative language used in the lower layers to define concepts and relationships in a resource, and converting it into a Turing-complete (i.e. computable) logic language.

For instance, Figure 3.6 illustrates an example of how five objects are related. The following four statements can be drawn from this:

- E-Commerce has *author* Jim Carter
- Jim Carter *works in* Research Team A
- FpML has *author* Tim Block
- Tim Block *works in* Research Team A

From these statements it is possible to postulate that *e-Commerce* and *FpML* are books with similar content because they were written by authors who work in the same research team. Working in the same research team implies that the two researchers work in similar fields, and it is likely that the authors have published a book on a topic in this field.

SiLRI (Simple Logic-based RDF Interpreter) (Decker *et al.*, 1998) is an RDF logic interpreter which uses the basic constraints available in RDF Schema to provide this inference ability. SiLRI is used in ESKIMO to make such deductions on scholarly knowledge and uncover further implicit relations.

It is impractical and infeasible to have one all-encompassing schema or ontology to define every known concept and relationship in the world, or even one global schema to define just a specific domain, and therefore there will be many schemas describing similar domains. This mandates the requirement of converting between the different versions and types of schemas. For instance, documents written in an earlier schema version must still be readable. Thus the newer schemas contain logic statements to formally describe how the schema has evolved. For example, if version 1 of a schema has the concept of *time stamp*, while version 2 has modified this to be labelled *time*, the following statement provides a conversion expression:

$$\forall\ x,\ y,\ \textit{if } x \textit{ is the 'time stamp' of } y,\ \textit{then } x \textit{ is also the 'time' of } y$$

Proof Layer

The proof and trust layers, and the digital signature proposal, have so far received limited attention and uncertainty surrounds their precise implementation. The proof layer enables processes to confirm whether a statement is true by using a series of inferences. For example, it may be necessary to prove that A is a type of B. The way that this might be determined is by inspecting two documents from a trusted site, one which states that "A is a type of C", and another which states that "B is equivalent to C". From this we could conclude that "A is of type B".

Trust Layer

At the top of the architecture is the trust layer, which provides processes with the ability to guarantee resources and the statements that they contain. In the Semantic Web millions of statements and assertions will be made and some of these are likely to contradict each other. The trust layer provides a mechanism to establish the validity of statements to form a "Web of Trust".

Digital Signature

The digital signature spans four layers and makes it possible to use public key cryptography to secure a document. The security potential becomes especially effective

when a *logic of trust* is introduced. If keys are represented as first class objects, then reasoning engines are not tied to only the signature verification systems. Instead, documents are parsed into trees of assertions about who has signed what. This results in a system capable of expressing and reasoning about relationships across a whole range of trust systems.

3.4 Technologies for the Semantic Web

In addition to the underlying standards and architecture, the realisation of the Semantic Web requires significant effort on the technology front to process and analyse the semantic resources and to provide services that exploit them. These are outlined in this section.

3.4.1 Processing Documents

Tools are required to parse and process semantic documents. The resources are based on RDF and XML, meaning XML based tools can be used for parsing Semantic Web resources. (This is a major advantage that RDF has over other metadata standards.) Two predominant access methodologies are typically used to parse and represent XML data: the Simple API for XML (SAX) and the Document Object Model (DOM).

SAX (Megginson, 2000) is a proposal introduced by the XML-DEV mailing list. Unlike DOM, it is event driven meaning supporting methods are fired when a particular event is encountered (a node in the XML tree). This means that SAX is fast and very memory efficient as the whole document is not represented in memory, resulting in the ability for very large documents to be processed. Unfortunately, SAX is read only and provides no random access to the document.

DOM is an officially ratified standard released by the W3C. Unlike SAX, the entire XML document is represented in memory and is accessed using the document's tree structure. Unfortunately, DOM has a higher memory overhead meaning that large documents are inefficient to parse.

An alternative to using XML parsers is to employ a dedicated parser for RDF data, such as SiRPAC (Simple RDF Parser and Compiler) (W3C, 2001a). This parser converts RDF statements directly into the triple format, ready for injection into an inference process.

3.4.2 Querying Documents

Query facilities are used to extract and manipulate information from semantic documents. Effective query mechanisms are essential as the amount of metadata to parse can be large.

The XML Path Language (XPath) (W3C, 1999b), as its name suggests, uses path expressions to access specific parts of an XML document and also forms the basis for the XPointer standard. An XML document is represented as a tree similar to the one created using the DOM. An XPath expression starts from a context node (e.g. the root) from which the document tree is searched, using principles of child, ancestor, and sibling nodes. An example XPath is presented below:

```
/doc/chapter[3]/section[@type='m']
```

The XPath statement points to the element 'section' with an attribute 'type' and value 'm' in the third chapter of the document. XPath statements are flexible and support basic math functionality.

```
//BBB[position() mod 2 = 0 ]
```

Unlike XPath, XML-QL (Deutsch *et al.*, 1998) does not make use of path expressions. Instead, its syntax is similar to Simple Query Language (SQL) statements used in databases, as it uses patterns to match fragments from XML documents. XML-QL has two significant advantages over a path expression language as it can:

- construct a completely new XML fragment to return
- combine query data from multiple different sources

A simple XML-QL example is illustrated below:

```
WHERE
  <PERSON>
    <NAME>Ted Nelson</NAME>
  </PERSON> CONTENT_AS $p IN document.xml
CONSTRUCT  <PERSON> $p </PERSON>
```

Quilt (Robie *et al.*, 2000) is based on XML-QL but also incorporates path expressions and so results in an effective query language that provides the benefits of both languages.

3.4.3 Infrastructure: Agents and Web Services

Agents are encapsulated computer systems that are capable of autonomous behaviour and processing. They interact with one another to solve common problems; a process which will necessarily involve a degree of negotiation, cooperation, and coordination. Uschold (2001) suggests that agent technology might be the *killer application* for the Semantic Web. There will be "trillions of small specialised reasoning services" (Fensel, 2000).

Agents collect and analyse Web content, exchange results, and work together. For example, the Medical Literature Search Agent (MELISA) (Abasolo & Gomez, 2000) uses the medical literature in the Medline document database to provide an information retrieval agent. The agent is aware of the terms in the medical domain and understands well formulated queries to accurately locate information.

Web Services have become a prominent feature in the Web community and promise to deliver a new level of interoperability between applications. Similar to mobile agents[4], Web Services are a self-describing, self-contained, module that are able to provide some application logic. However, the main difference between Web Services and agents is XML. Unlike agents, Web Services are directly aimed at providing services on the Web by using XML. They are used for a variety of services such as e-procurement, weather reporting, and logistics.

Web Services use the Simple Object Access Protocol (SOAP), an XML and XMLS based messaging protocol that supports remote procedure calls[5]. SOAP is a simple and lightweight mechanism for exchanging structured and typed information (as well as remote procedure calls and responses). A SOAP message consists of an envelope, a set of encoding rules, and the body. The following example illustrates SOAP being applied to exchange share price information.

```
<SOAP-ENV:Envelope
 xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
 SOAP-ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"/>
  <SOAP-ENV:Body>
   <m:GetLastTradePriceResponse xmlns:m="http://www.ns.com/ns">
     <Price>34.5</Price>
   </m:GetLastTradePriceResponse>
```

[4]An agent is mobile if it is an unimportant factor where it resides. Thus a mobile agent can be located on a desktop machine, mainframe, or a mobile device.

[5]Remote Procedure Call (RPC) is a platform independent protocol that programs use to request services from other programs located elsewhere on the network.

```
  </SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

SOAP is based on HTTP meaning integration with the Web is possible. The SOAP syntax is also simpler than RDF and contains less nesting, which makes parsing tools more efficient. Although some researchers promote SOAP as a suitable standard for the Semantic Web (Haustein, 2001), it lacks concrete support for the knowledge representation and logic facilities that are an important constituent of the initiative. Nevertheless, it is a core communication language for Web Services.

Web Services require two main technologies:

- A registry service to register and advertise the service offered by a Web Service.
- An interface to the service that describes its functionality.

UDDI (Universal Description, Discovery, and Integration) is the proposed protocol for the first technology requirement. It is a platform-independent, open framework for describing and publishing services using the SOAP standard as the communication mechanism. It enables a business to describe itself and the services it offers, and discover other businesses that offer useful services, and thereby integrate the businesses. Many large companies such as Microsoft, IBM, Sun, SAP, and Hewlett-Packard support UDDI.

The second technology is provided by WSDL (Web Services Description Language). This is an XML-based format for describing a Web Service's functionality and the methods to access it. After a service has been located using UDDI, the service is contacted directly and communicates information about its interface using WSDL.

Eventually, the Web could consist of many of these agents and Web Services collecting, analysing, and most importantly, cooperating, to help provide automated services. These types of services are an ideal platform for the Semantic Web (McIlraith *et al.*, 2001) and demonstrate the possibility of intelligent and automated interaction of distributed processes.

3.5 Summary

The Semantic Web promises to advance the Web to an intelligent form that both humans and machines can read, understand, and take advantage of. Machines will use the knowledge that the Semantic Web provides to assist users in ways such as searching, hypertext navigation, e-commerce, finance, and automated negotiations.

The architecture of the Semantic Web contains seven distinct layers. These cover basic structural issues and encoding schemes, to facilities that enable the representation of complex data structures. Logic is used in the Semantic Web to enable inference expressions to translate between different schema versions, uncover new information, and be used for verification purposes.

The initiative revolves around the representation and interchange of knowledge using metadata: a method for representing information for machine consumption. Early work on metadata has resulted in various standards being proposed, however these are too basic to represent the intricate knowledge structures required by knowledge-intensive services. Metadata capable of representing the concepts and relationships that make up a domain, as well as the necessary constraints to restrict them, is necessary.

The principles of the Semantic Web are used in ESKIMO to represent, manage, and analyse on-line scholarly information. This material contains many inter-related concepts that form a complex and intricate network of knowledge that the Web currently fails to exploit and present in a principled and controlled manner. ESKIMO also uses knowledge services to analyse scholarly metadata to uncover patterns and further facts to assist scholars in their research activities.

The representation of knowledge by ontologies is fundamental to the Semantic Web and is discussed in the following chapter; the use of such knowledge by scholars (scientists and researchers) is then outlined in Chapter 5.

Chapter 4

Knowledge in the Semantic Web

4.1 Introduction

The Semantic Web introduced in the previous chapter, described a knowledge Web that provides computer processes with machine-readable knowledge to realise intelligent Web services. Metadata is the key technology for representing this knowledge, although current attribute and concept-based metadata proposals lack the necessary expressiveness to describe a domain accurately and rigorously. For this task ontologies have been proposed.

The term ontology was independently coined in 1613 by two German philosophers, Göckel and Lorhard. It was first recorded in English in 1721 by the Oxford English Dictionary as 'an account of being in the abstract'. Today, ontology is the study of *things that exist* that began as a branch of philosophy but has migrated to the field of knowledge management (Guarino, 1998) and is considered an integral part of the Semantic Web initiative. Ontologies provide a conceptualisation of a domain and facilitate knowledge sharing and reuse, and enable reasoning facilities to be used to uncover information and translate between different information structures.

This chapter first introduces the possible approaches to structuring data. Then the process of creating and using an ontology and its role in the Semantic Web are described. The chapter concludes by exploring several systems that use ontologies and knowledge to provide novel and useful services.

4.2 Structuring Data and Ontologies

Ackoff (1967) categorises the content of the human mind as:

- data: raw symbols that have no significance beyond their existence (e.g. '4', 'Smith', 'hypertext')
- information: data that as some useful value (e.g. 'the author is Peter')
- knowledge: the application and conception of data and information to provide a greater understanding of some area (e.g. 'an author, such as Peter, is responsible for maintaining and publishing a document')
- understanding: an analysis of knowledge to enable new knowledge to be synthesized by building on currently held information, knowledge, and understanding (e.g. the author of a document will know how to publish it)
- wisdom: evaluating understanding to produce a new understanding for which there has previously been none (e.g. predicting how the publishing process will change with the introduction of a revolutionary new medium)

Ackoff believes that wisdom is something that machines cannot possess, as it is uniquely human. It requires extrapolative understanding and solutions to problems where there are no easy (and often correct) answers. However, machines are capable of representing and using data, information, knowledge, and understanding. Modelling techniques, such as ontologies, are used to represent and provide knowledge while logic facilities are used to add a degree of understanding to it.

There are several approaches to structuring data to create knowledge for machines to read and understand. These are listed in an order of increasing semantic detail.

1. list - an enumeration of words (e.g. shopping list)
2. vocabulary - an enumeration of words within the same domain (e.g. technical computer terms)
3. thesaurus - an enumeration of words within a domain containing a degree of structure (e.g. medical terms dealing with surgery)
4. taxonomy - terms are organised in a structured hierarchical fashion (e.g. a directory of car parts)
5. ontology - a taxonomic structure with constraint mechanisms and further relationships between terms (e.g. a model of the scholarly publication process)

Ontologies are therefore an advanced method of representing knowledge that provide a common understanding of a subject area, enabling knowledge sharing and reuse, and improving communication between people (e.g. designers, users) and software entities (e.g. agents). Although potentially difficult and time consuming to construct (Farquhar *et al.*, 1996), ontologies have proven highly successful in many disciplines, such as bioinformatics (Stevens *et al.*, 2000) and multi-agent systems (Falasconi *et al.*, 1996). They are also proving suitable for e-commerce (Jennings *et al.*, 2000) where agents negotiate and discuss transactions and therefore require an understanding of a domain.

Ontologies also provide better semantics than the unconstrained vocabularies used in metadata standards such as Dublin Core and GILS, where properties (e.g. title, creator, subject, date) are used without specifying how they relate to each other or the rest of a domain. For example, it is not possible to specify that a subject has a title and that titles must be at least 3 characters long.

The main modelling primitives in an ontology are concepts, instances, relationships, and constraints. A concept represents any *thing* in the domain (e.g. laptop, person, date, Jupiter). Individual instantiations of concepts are referred to as instances (e.g. the laptop on my desk, the person called 'Simon Kampa'). Concepts are related to each other by specifying a relationship between them. For example, a 'Person' *is born on* a 'Date'. Concepts and relationships are constrained using constructs such as cardinality and existential/universal quantifiers (e.g. 'any concept A can only relate to a maximum of 6 concept Bs').

Figure 4.1 displays a conceptualisation of a simple ontology that captures a contact directory and describes a person with attributes title, first name and last name. A person relates to an e-mail, group, and location in the indicated ways.

As an ontology provides an explicit representation of a domain, it enables machines to reason over the structure of a domain. For example, if Person P1 and Person P2 both work in group G1, then we could assume (reason) that they are colleagues. This could then be used by a system when collecting information about community structures or in assessing collaboration.

Three types of ontologies are proposed by Uschold (1996):

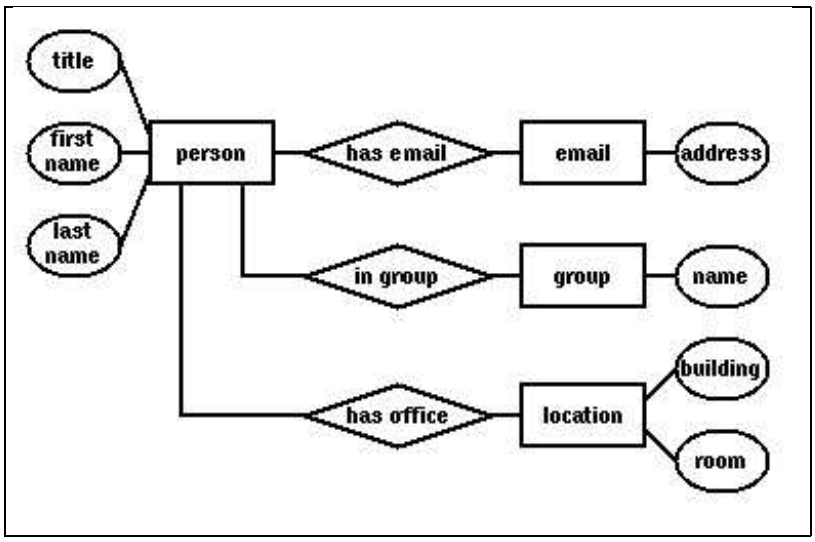


Figure 4.1: Simple ontology example

- Domain Ontology - the application domain is considered over the eventual task
- Task (or Problem) Ontology - the application domain is considered in respect to the task
- Meta-Ontology - the mechanism to define an ontology (e.g. knowledge representation language)

The structures of ontologies vary greatly, but can usually be classified along three axes (Uschold, 1996):

- Formality
- Purpose
- Subject Matter

An ontology can be highly informal and conveyed in natural language. Such an ontology is often used in glossaries as in the Medical Subject Heading (MeSH) (Schulman, 2000) project. On the other hand, a rigorously formal ontology is used when system integration or machine processing is vital (Guarino, 1998; Cui *et al.*, 2001).

The purposes of ontologies also differ. An ontology may be used to reach a consensus on a particular subject or be used by software agents to perform tasks such as negotiation and inference. Therefore, Noy *et al.* (2001) propose using several competency questions to determine the purpose of an ontology and focus developers on the nature of the task.

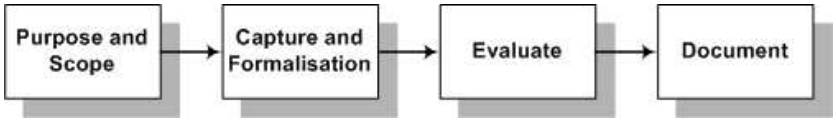


Figure 4.2: Stages in ontology construction

Finally, the subject matter influences the construction process. For instance, the ontology may be highly complex detailing thousands of medical terms. In this case the construction process is complex and can take several years to complete. On the other hand, a small general ontology may be sufficient to represent a video collection.

A detailed discussion on ontologies is presented in (Noy & McGuinness, 2001), (Uschold & Gruninger, 1996), and (Guarino, 1998).

4.3 Ontology Construction

Ideally, an ontology is constructed in a collaborative environment of domain experts, end-users, and computer specialists. It is vital to have as wide a range of experts as possible, to ensure all aspects, issues, and perspectives of a domain have been discussed.

Unfortunately, a standard methodology for the construction process has so far failed to materialise, although guidelines have been presented (Uschold & Gruninger, 1996; Gomez-Perez, 1996). Figure 4.2 summarises the overall steps usually involved in the creation of ontologies, which are not dissimilar to software engineering approaches. First the purpose and scope of an ontology is defined, before the domain under investigation is captured and formalised. Then the ontology is evaluated to determine if it meets the requirements of the task. Finally, the ontology process and modelling decisions are documented.

Similarly, Figure 4.3 illustrates the ontology lifecycle proposed by Maedche *et al.* (2001). This starts with an initial discussion where the task and domain are identified and any other issues addressed. The scope and purpose of the ontology should also be outlined as this focuses the reasoning behind the ontology and highlights any major misunderstandings or misinterpretations. Finally, a first attempt at modelling the domain (i.e. identifying the concepts and relationships) is commenced.

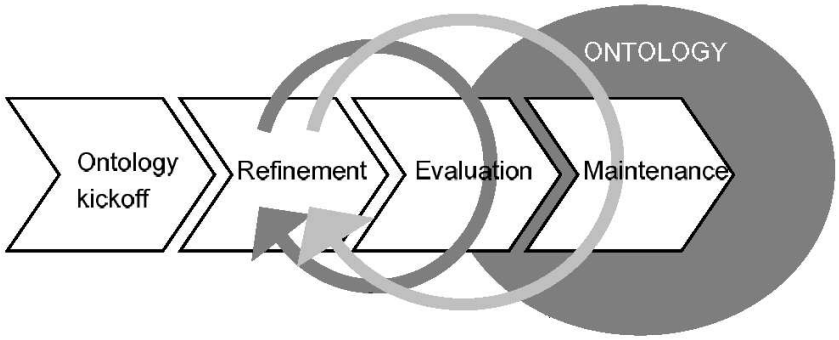


Figure 4.3: An ontology construction process

After the first attempt, the model is iteratively refined and evaluated until a satisfactory model is produced. Gomez-Perez (1995) has proposed methods for ontology evaluation, for example, by inspecting the definitions in the ontology for consistency, syntax, structure, and completeness. This iterative stage is the most time consuming, especially for complex or multifaceted domains. During the life of an ontology, any modifications require further refinement and evaluation.

Although not included in Figure 4.3, documentation is an important part of any knowledge project. The documentation should not only include a detailed discussion on the makeup of the ontology and the construction process, but also outline any assumptions that were made.

4.3.1 Conceptualisation and Capture

The majority of the effort dispensed in constructing an ontology is used conceptualising and capturing the domain. This takes place at the beginning when the ontology is first proposed, and during the iterative refinement stage. Before commencing with this however, it is vital to consider reusing entire ontologies or fragments of them as reuse significantly reduces the authoring overhead (Lee & Malone, 1990; Gruber, 1993). Even if an appropriate ontology is not located, a similar one may still provide a useful insight and interpretation into the domain. For example, a hypertext domain ontology is presented in Chapter 9 that is based on the ACM subject classification index. Ontology reuse does not work in situations where the eventual purposes of the ontologies diverge greatly as the defined concepts and relationships are unlikely to be suitable for both tasks.

Major concepts and their relationships are initially identified through debate and brainstorming sessions. This stage can become quite unstructured and inefficient as knowledge engineers argue over their differing perspectives. Van der Vet *et al.* (1998) propose a bottom-up approach to concept identification, opposing the more commonplace top-down approach. Alternatively, Uschold and Gruninger (1996) propose a middle-out approach where basic concepts within a domain are recognised first, and are then generalised and specialised. For example, the concept article can be generalised (broadened) to literature and specialised to technical article. In addition, as ontologies usually model real-life domains, object-oriented design methods are applicable. In this case, inheritance and encapsulation improve the understanding of a concept, which in turn benefits usability and readability.

At the conclusion of this stage, a graphical representation of the ontology is delivered, together with some form of textual representation to explain the conceptualisation in a more formal and rigorous format (although still only intended for human understanding). For example, Figure 4.4 graphically illustrates part of a newspaper ontology created using the FRODO RDFSViz visualisation tool (developed as part of the FRODO Project (van Elst & Abecker, 2001)). Skuce (1996) has presented work on an intermediate textual ontology representation and an example document detailing the scholarly community ontology used in ESKIMO is presented in Appendix D.

4.3.2 Formalisation

Once a consensus has been reached on the conceptualisation of the domain, a formalisation of the ontology is necessary when machines require it to process the ontological structure and use it for knowledge applications.

A knowledge modelling language, such as the Operational Conceptual Modelling Language (OCML) (Motta, 1998) or the Knowledge Interchange Format (KIF) (Genesereth & Fikes, 1992), is used to formally define the conceptualisation. Unfortunately, many researchers lack the expertise and experience to adopt this approach and therefore several systems provide users with tools to construct and browse ontologies (Farquhar *et al.*, 1996; Mahalingam & Huhns, 1997; Motta *et al.*, 1999), leaving the translation into a knowledge modelling language to the

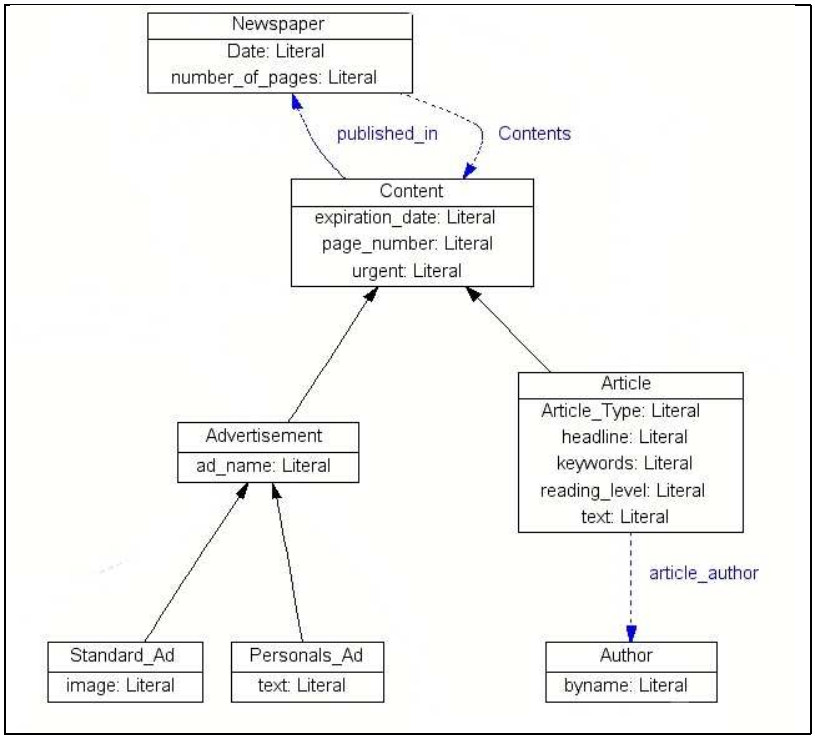


Figure 4.4: Part of a newspaper ontology

program. The formal representation of ontologies for use on the Web (i.e. the Semantic Web) is discussed in Section 4.5.

4.4 Ontology Tools

Ontology construction is a complex and intricate task that requires tools to help construct, visualise, and verify ontologies, to assist the knowledge engineer in constructing an accurate model. Indeed, such tools are used in the design, evaluation, and representation of the scholarly ontology constructed in Chapter 9. Furthermore, as an ontology represents an agreed understanding between various parties, tools to support this collaborative, and potentially distributed, environment are available.

4.4.1 Editors

Protégé 2000

Protégé 2000 is a widely used ontology editor that allows users to construct ontologies, produce knowledge acquisition forms, and populate ontologies. Ontologies are stored using the native Protégé format or other knowledge formats such as RDFS. The editor provides an extensible plug-in architecture that can be used to

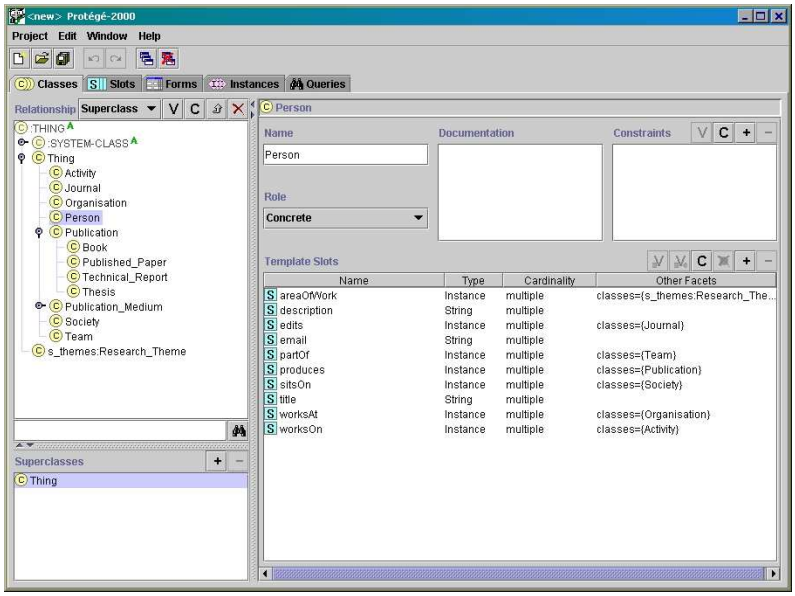


Figure 4.5: Protégé 2000 interface

tailor Protégé to a specific task or domain (e.g. additional reasoners or visualisers, support for new knowledge formats).

Figure 4.5 displays the Protégé 2000 interface. The *Classes* tab presents the definition of the ontology and its concepts, relationships, properties, and constraints. The other tabs are used to populate the ontology, create knowledge acquisition forms, and manipulate the knowledge base using logical queries.

Protégé 2000 is particularly suitable for exploring the many evolving ontology languages, as it allows the user to think about the domain at the conceptual level without worrying about the syntax and semantics of the particular knowledge format that will be ultimately deployed on the Web (Noy *et al.*, 2001). Protégé 2000 has been used during this work for ontology construction, evaluation, and experimental analysis and this is reviewed in Chapter 9.

OILEd

OILEd is a lightweight ontology editor for creating ontologies using the ontology language OIL (see 4.5.5). The editor has been described as the 'notepad' of ontology editors as there is no support for versioning, inference, large-scale ontology construction, or integration. However, it does support the Fast Classification of Terminologies (FaCT) (Bechhofer *et al.*, 1999a) reasoner that is used to verify the consistency of an ontology. For example, if the concept *Cow* is introduced as a type of mammal with the property and constraint *only eats vegetation*, and the concept

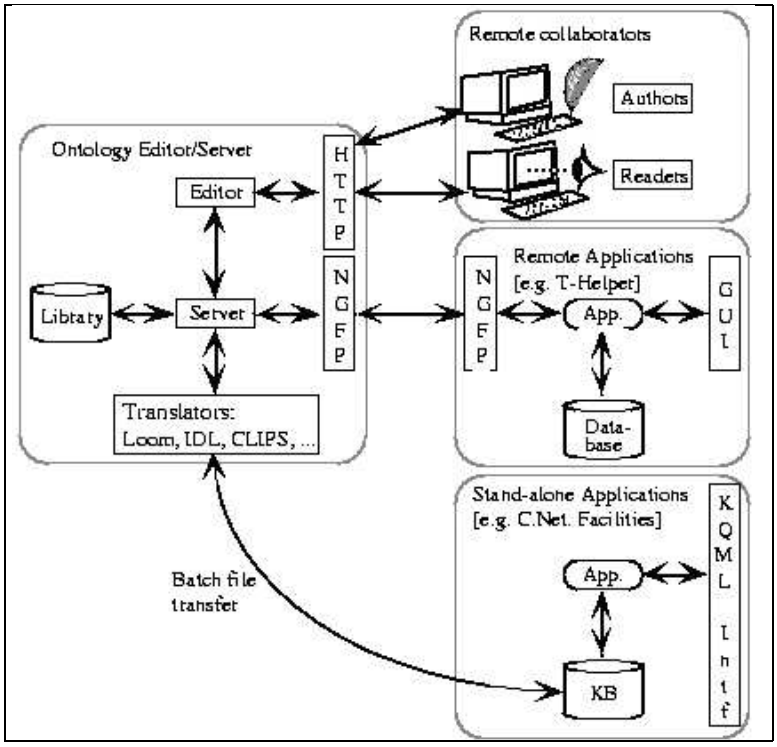


Figure 4.6: Ontolingua Architecture

Mad Cow is defined as a type of *Cow* with the property *eats Meat*, the reasoner will raise an inconsistency to the knowledge engineer as every concept of type *Cow* only eats vegetation.

4.4.2 Collaborative Environments

Ontolingua

Ontolingua (Farquhar *et al.*, 1996) provides a client/server environment enabling users connected to the Web to create, browse, extend, and customise ontologies in a collaborative setting (Figure 4.6).

Ontologies are recorded by Ontolingua using the Knowledge Interchange Format (KIF) (Genesereth, 1998), but translation routines are available to output the ontology to a variety of languages capable of representing knowledge (e.g. Prolog (Warren, 1977), LOOM (MacGregor, 1990), Epikit (Genesereth, 1990)). Ontolingua is especially useful in large collaborative construction environments where authors modify and share ontologies with other users and domain experts in the community.

Tadzebao and WebOnto

Tadzebao (Domingue, 1998), Chinese for 'Big Character Poster', is a collaborative ontology environment that supports both synchronous and asynchronous discussion

on ontologies. Tadzebao enables knowledge engineers to refer to ontologies directly in their messages through its integrated ontology support in a text and drawing editor. All dialogue is focused around a notepad, which contains a mixture of text, images, hand drawn sketches, and ontology fragments represented in OCML.

WebOnto (Domingue, 1998) was created to complement Tadzebao by supporting collaborative browsing, creation, and editing of ontologies. A graphical environment is provided for users to define concepts and relationships, and then select and drag these into position to produce a visual representation of the domain. The defined ontology can be saved in an OCML representation and used by Tadzebao, or exported to the Ontolingua format.

4.5 Ontologies in the Semantic Web

Ontologies are the mechanism added to the schema layers to provide high level facilities for modelling domains. However, for ontologies to be realised in the Semantic Web (and in applications like ESKIMO), a Web-based ontology definition language is required.

This section discusses several approaches and standards for an ontology representation language for the Semantic Web.

4.5.1 SHOE

SHOE (Simple HTML Ontology Extensions) (Luke *et al.*, 1996; Heflin *et al.*, 1998) was one of the earliest and most influential works on ontological metadata, and ontologies were used by authors to describe a document's content.

SHOE also enabled simple inference rules to be defined. For example, if a student is a member of a research group, and that research group is a member of a particular department, then it can be inferred that the student is also a member of the department. This simple, yet useful machine reasoning ability reduces the number of explicit metadata statements an author has to make.

SHOE elements are embedded within the HTML of a document meaning that Web documents contain both presentational information and SHOE knowledge. An example Web page describing a student/advisor relationship is illustrated below.

```
<HTML>
  <HEAD>
    <META HTTP-EQUIV="SHOE" CONTENT="VERSION=1.0">
```

```
  </HEAD>
  <BODY>
    <H1>Home page of Jane Smith</H1>
    <P>I am a graduate student and my advisor is John Smith.</P>
    <CATEGORY NAME="cs.gradStudent">
    <INSTANCE KEY="http://www.cs.umd.edu/users/jsmith/">
      <USE-ONTOLOGY ID="cs-dept-ontology" VERSION="1.0"
                    PREFIX="cs"
                    URL="http://www.cs.umd.edu/ont/cs.html">
      <RELATION NAME="cs.name">
        <ARG POS="TO" VALUE="Jane Smith">
      </RELATION>
      <RELATION NAME="cs.advisor">
        <ARG POS="TO" VALUE="http://www.cs.umd.edu/users/jdoe/">
      </RELATION>
    </INSTANCE>
  </BODY>
</HTML>
```

The document describes a graduate student called 'Jane Smith'. The location of the ontology used to define concepts in the document is located at '`http://www.cs.umd.edu/ont/cs.html`' and is assigned the namespace 'cs'. The page is classified (referred to as a category in SHOE) as being about a graduate student. The concept instance described by the Web page is also assigned a unique ID (e.g. the URL of the page) which is used in other documents to refer to it. Then the instance is assigned a name property of 'Jane Smith' and is related to an advisor instance that is uniquely defined by '`http://www.cs.umd.edu/users/jdoe/`'.

Significantly, ontological metadata is added to the content of the document it relates to. The SHOE crawler then parses the document to extract this knowledge. This has the advantage that any user can add SHOE tags to their documents, resulting in metadata being distributed across the Web (i.e. there is no heavy burden on the SHOE system). However, a crawler is required to retrieve the knowledge and any document that has not been parsed is disregarded by SHOE (i.e. a similar scenario to current search engines). A search engine is provided to enable users to semantically search the collected knowledge.

4.5.2 Ontobroker Project

Like SHOE, the Ontobroker Project (Fensel *et al.*, 1998) uses ontologies to annotate Web documents and provides an ontology-based query facility. An extended HTML

syntax is proposed for users to mark up documents with ontological constructs, which the Ontobroker crawler then processes.

Ontobroker provides a framework to create a knowledge environment and has been used to create (KA)2 (Benjamins *et al.*, 1998); an application which defines seven ontologies to model the knowledge acquisition (KA) community.

Ontobroker introduces the ONTO attribute of the anchor tag in HTML to add Ontobroker constructs. The following example uses Ontobroker metadata to annotate a researcher's homepage.

```
<HTML>
<HEAD>
  <A ONTO="page:Researcher"></A>
<HEAD>
<BODY>
<H1>
  <A HREF="pictures/id-rich.gif>
    <IMG SRC="pictures/richard.gif">
  </A>
  <A ONTO="page[photo=ref]"
     HREF="http://www.iiia.csic.es/~richard/pictures/richard.gif"></A>
  <A ONTO="page[firstName=body]">Richard</A>
  <A ONTO="page[lastName=body]">Benjamin</A>
</H1>
</BODY>
</HTML>
```

In the head of the page, the document is declared as being of type 'Researcher'. Within the body of the page, the properties of this researcher are defined (e.g. first name is 'Richard' and last name is 'Benjamin').

The Ontobroker crawler parses annotated pages and represents the knowledge internally as F-Logic, a language for reasoning about objects. By combining the concepts and relationships specified in the ontology with the facts collected from the Web pages, Ontobroker enables users to pose a variety of complex queries, such as: *What are the titles of all the projects where Richard Benjamin is a member?*

The Ontobroker (and SHOE) approaches promote a community authoring effort (Staab *et al.*, 2000). The process of annotating Web resources is a joint effort between all the members of an esoteric community. This scenario is particularly suitable for a group of users with overlapping interests who wish to create a community-oriented research service.

4.5.3 RDFS

The previous chapter introduced RDFS as a schema language to compliment RDF. RDFS is used to create a class hierarchy with simple constraints (range, domain) that enables the construction of basic ontologies. Unfortunately, while RDFS provides modelling support, it is limited in its constraint mechanism (e.g. property constraints are not possible) and complex class expressions (e.g. the statement 'a herbivore *is a* type of animal and is *not* a carnivore' cannot be expressed). Therefore, several proposals extend RDFS to provide additional constructs and these are discussed next.

4.5.4 DAML

The DARPA [1] Agent Markup Language (DAML) (DARPA, 2000) is a US government sponsored endeavour which has a large and diverse following from academia, government, and industry. It aims to develop languages, tools, and methodologies for the Semantic Web. Two languages are proposed:

- DAML-ONT: An ontology language which extends RDFS and its object-oriented type system to enable agent and service inter-operation. The language draws from object-oriented modelling, frame systems, and conceptual schemas.
- DAML-L: The logic language to provide constraint mechanisms and the ability to define inference rules.

An example class definition that defines the concept *Female* in DAML-ONT is listed below.

```
<Class ID="Female">
  <subClassOf resource="#Animal"/>
  <disjointFrom resource="#Male"/>
</Class>
```

The class *Female* is defined as being a subclass of *Animal* and is disjoint from the class *Male*.

DARPA is actively encouraging the use of DAML for a wide spectrum of applications and therefore a wide range of tools have been released by the community, including a crawler, browser, inference engine, ontology analyser, and an annotation tool.

[1] Defence Advanced Research Projects Agency (DARPA)

4.5.5 OIL

The Ontology Inference Layer (OIL) (Fensel *et al.*, 2000; Bechhofer *et al.*, 2000) is a response from the academic world for an ontology representation language. It extends RDFS and draws from work in three distinct disciplines.

- Frame-based systems

  The central modelling primitives in frame-based systems are classes (which can be defined as subclasses of other classes) and properties of classes (or slots).

- Description logics

  Description logics (DL) describe knowledge in a similar approach to frame-based systems. The important feature of DL is that the meaning of any expression can be described in a mathematical way and processed by machines.

- Web standards

  OIL requires a Web-based format for representing the ontology. For compatibility and openness, OIL builds on XML and RDFS. XML has the advantage of being a successful interchange format while RDFS provides the basic modelling primitives (e.g. *instance-of* and *subclass-of*).

A brief example of an OIL ontology description that defines the concepts *animal*, *plant*, and *tree* is illustrated below. The OIL specific extensions to RDFS are identifiable by the OIL namespace.

```
<rdfs:Class rdf:ID="animal"/>
<rdfs:Class rdf:ID="plant">
  <rdfs:subClassOf>
    <oil:NOT>
    <oil:hasOperand rdf:resource="#animal"/>
    </oil:NOT>
  </rdfs:subClassOf>
</rdfs:Class>
<rdfs:Class rdf:ID="tree">
  <rdfs:subClassOf rdf:resource="#plant"/>
</rdfs:Class>
```

4.5.6 DAML+OIL

A more recent proposal has emerged from a collaborative effort from the DAML and OIL groups: the DAML+OIL (van Harmelen *et al.*, 2001) proposal. Although

inheriting much from the OIL proposal, DAML+OIL has moved away from the frame-based approach to a more description logic approach. General assertions are used to constrain classes, rather than specifying them within their respective class as in frame-based languages. While computationally this makes little difference, it results in conceptually losing some of the intention and understanding of the concepts they constrain, leading to an adverse effect on authoring and exchange of ontologies (Bechhofer *et al.*, 2001).

DAML, OIL, and DAML+OIL provide rigorous and formal semantics to represent knowledge. Much of this stems from the artificial intelligence community, which has extensive experience in machine processing of knowledge. However, defining the ultimate ontology language for the Semantic Web is proving to be a difficult task with new proposals being frequently published. In addition to OIL, DAML, and DAML+OIL, the W3C has recently released a working draft for the Ontology Web Language (OWL) (Heflin *et al.*, 2002) which builds on the experiences of OIL+DAML and tightly interoperates with standards from the W3C.

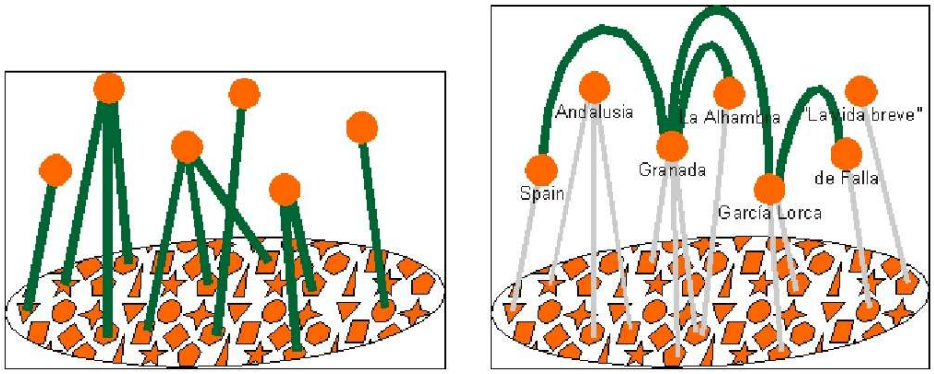
4.6 Alternatives to Ontologies

Ontologies have evolved out of the artificial intelligence and knowledge management communities. However, alternatives from different communities are available that have less of a mathematical and logical influence, and these are presented in this section.

4.6.1 Database Schemas

Schemas have been popular in database systems for describing the structure and semantics of data destined for a relational database. Information is viewed as a series of rows in tables, but constraints (aside from property types) are not possible. In addition, new information sources, such as the inter-related document centric information common on the Web, do not fit well into this rigid scheme.

In Chapter 8, the OntoPortal system is described that represents an ontology using a relational database. The additional overhead required to recreate the ontological concepts and relationships in the database is significant. However, OntoPortal does not permit hierarchies as representing these using only tables is difficult and



(a) Topic occurrences (b) Topic associations

Figure 4.7: Topic occurrences and associations

adds further processing overhead. ESKIMO is discussed in Chapter 9 and supports hierarchical relationships using an indirect approach.

4.6.2 Topic Maps

Topic Maps (Rath & Pepper, 2000) is the subject of an international standard (ISO/IEC 13250) and is viewed as an alternative to RDFS and a building block of the Semantic Web. It is based on SGML and HyTime and was originally designed to merge indexes, but has since been demonstrated to be useful for general navigation on the Web (such as in the creation of tables of content, glossaries, thesauri, and cross-references). Similar to ontologies, topic maps are capable of representing concepts and their relationships. There are three major constructs:

- Topic
- Topic occurrence
- Topic association

A topic, like a concept in an ontology, can represent anything (e.g. Manchester United). A topic occurrence is an addressable instance of a topic (e.g. home page of Manchester United). Relationships between topics are called associations (e.g. Manchester United *is in* Manchester). A collection of topics, occurrences, and associations is called a topic map. Figure 4.7 illustrates the notion of occurrences and associations. The various topic occurrences are depicted as different shapes. Each topic (Figure 4.7a) is represented as a node with connections to its related occurrence(s). Topic associations are illustrated as arcs between the topic nodes (Figure 4.7b).

The abstraction of topics and their occurrences enables topics maps to be reused in different scenarios as associations apply only to topics and not the individual occurrences. Facets (a restriction such as cardinality or value-type) are used to add constraints to topic maps and are also used to assign them to different applications.

To enable topic maps to fit into the Semantic Web activity an XML-based form has been proposed. XML Topic Maps (XTM) (Pepper & Moore, 2001) provides a universal grammar for interchanging topic maps. However, there is significant overlap between RDFS and XTM:

- Both are able to provide the foundation for the Semantic Web.
- Both provide simple, but elegant, models.
- Both models can be used to build a semantic network.

However, RDFS is extensible and has been extended with richer and more rigorous semantics in the form of OIL, DAML, and DAML+OIL. RDFS also has the advantage of being endorsed by the standards consortium, W3C.

The overlap between RDFS and Topic Maps has caused the research community to investigate ways of integrating the standards. For example, Lacher *et al.* (2001) propose a mapping between the two data models while Freese (2000) proposes selecting the resource feature of RDFS and the topic concept in topic maps to define a new standard.

4.7 Critique

The primary advantage of ontologies is facilitating improved communication between users and machines through a common understanding and formalisation of a domain. This enables users to conform and agree to a standard language and machines to integrate and discourse over a domain. Three diverse applications areas are proposed in the ontology triangle by Fensel (2000) (Figure 4.8). This wide range of potential uses has promoted the adoption of ontologies and the publication of ontologies for reuse (e.g. the Virtual Business effort at `http://www.ontology.org/`). The introduction of an ontology into knowledge systems enables the system to provide new levels of functionality as it better interacts with users, intelligently integrates information from a variety of sources, and effectively responds to more intricate queries that require an understanding of the underlying domain.

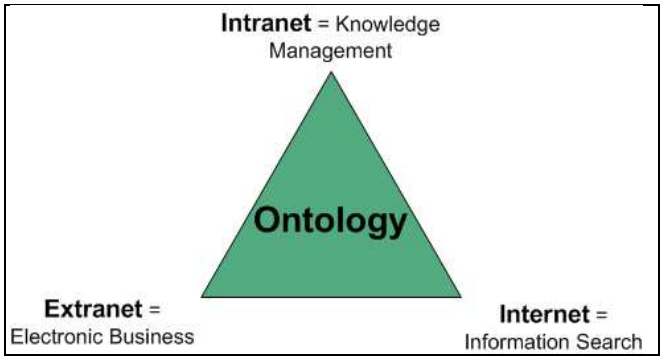


Figure 4.8: The Ontology Triangle

The disadvantage in the use of ontologies lies mainly in the construction and maintenance overhead. Construction, especially of larger ontologies, is a time consuming and complex task. Many concepts and relationships must to be identified and formally defined and as there is no *correct* solution for the resulting structure, a 'best effort' approach must be adopted. A solution is to adopt a minimal ontological commitment, used by Motta *et al.* (1999) in the definition of an ontology to represent the claims between scholarly papers. In this approach, just enough of the domain is made explicit to be usefully expressive. Incremental formalisation (Shum *et al.*, 1999) has also been used where an ontology is constructed in gradual and distinct phases.

4.8 Researching the Semantic Web

The Semantic Web initiative has prompted research into projects and tools to explore its potential and these are discussed in this section.

4.8.1 Ontobroker Project

The Ontobroker project enables users to annotate documents with ontological metadata. The Ontobroker crawler then captures this knowledge and exposes it to users in four ways:

- Hypertext link: A dynamically evaluated link into the knowledge base. For example, a link on *Projects* results in a query for all known projects.
- HTML form: A form to specify which instances to return. For example, locating publications that are about Ontolingua and have an author named Farquhar.
- Hyperbolic view: Spatially visualise the ontology as a hierarchy of concepts.

- Expert mode: Issue F-Logic statements directly.

These interfaces enable users to search the knowledge collected by Ontobroker and use it to understand the content of documents. The Ontobroker project has been succeeded by the Karlsruhe Ontology (KAON) project (Handschuh *et al.*, 2001).

4.8.2 COHSE

The Conceptual Open Hypermedia Services Environment (COHSE) (Carr *et al.*, 2001) improves the quality and scope of linking on Web documents, by combining an ontology service with a dynamic link service to add links to documents based on the concepts that appear in it.

Dynamically providing links to resources using keyword-based matching has been popular for several years (Fountain *et al.*, 1990; Carr *et al.*, 1995). However, consistent keyword descriptions are difficult to create and maintain leading to a poor conceptual model of the concepts and hence resulting in poor linking. To overcome this some specialist communities have created a taxonomy or thesauri of the linguistic terms in their domain. However, without the rigorous interpretation and reasoning ability provided by an ontological representation, this reduces the browsing and querying effectiveness. The alternative, to associatively link similar documents by hand, is error prone and inconsistent (Ellis *et al.*, 1996).

COHSE draws on two technologies to overcome this and demonstrate an effective application of the Semantic Web.

- Ontological reasoning service: Represents a conceptual model of document terms and their relationships using DAML+OIL as the representation language.
- Open hypermedia link service: Links are represented as first class objects and stored, analysed, and manipulated separately from the documents in which they appear (e.g. DLS (Carr *et al.*, 1995)). In COHSE, this is used to offer a range of *link providing* facilities.

COHSE integrates the ontology and open hypermedia link service to form a conceptual open hypermedia system.

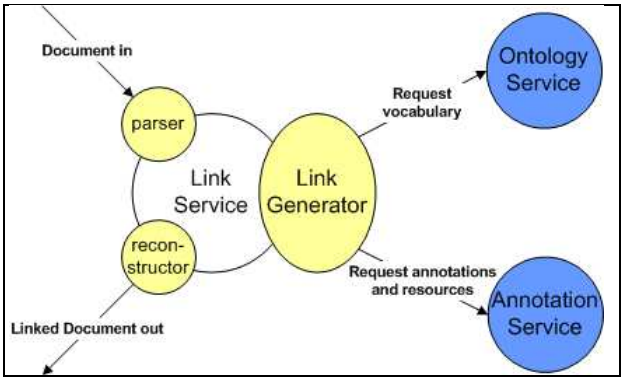


Figure 4.9: COHSE architecture

Figure 4.9 illustrates the COHSE architecture. It uses an augmented Web browser or a proxy server, controlled by a link service that uses two independent knowledge services, to manipulate the document and add ontologically-controlled hypertext. Two knowledge services are used: an ontology and annotation service. The ontology service loads DAML+OIL encoded ontologies and is then able to answer queries about them, such as retrieving all the concepts from an ontology and determining sub- and super-concepts of a particular concept. As concept names frequently do not translate directly into useful textual representations (e.g. a concept named 'PG_Researcher'), textual representations (terms) of each concept can be added to the ontology service in the form of a lexicon. The ontology service then returns these terms and provides a lookup service. For example, if the lexicon includes the term 'postgraduate researcher', the ontology service can be used to lookup this term and determine which concept it describes, in this case, the concept named 'PG_Researcher'.

The annotation service is a simple annotation engine, similar in principle to Annotea (Koivunen & Swick, 2001), except that it annotates regions of a document with a concept, rather than a simple piece of text. An XPointer is used to identify a region in the document; a fragment of RDF, which corresponds to a DAML+OIL statement, identifies the concept. The annotation service serves two purposes. Firstly, it is used by knowledge workers (or authors) to specify that a region of a document is 'about' a particular concept which enables the link service to match on the region and provide a link. Secondly, an annotation also represents a resource in its own right, meaning the annotation service also acts as a simple

librarian that is used to lookup Web pages which are examples of a particular concept (i.e. pages which can be used to illustrate a concept). Therefore, annotations are used to both identify link source anchors and link destinations.

To improve the dynamic linking on documents, COHSE combines these knowledge services with the link service. Traditional link services provide destination anchors based only on the keywords found in the source document, however, COHSE links on concepts in the ontology found in the document. The structure made explicit in an ontology enables COHSE to also broaden and narrow the destinations of a link. For example, when the concept 'Java' is matched, COHSE proposes resources about Java, as well as then consulting the ontology to discover that 'Java' is a type of 'Programming Language'. COHSE then proposes more general resources about programming languages as well. In addition, COHSE can request for the sub-concepts of 'Java', and then provide links on concepts such as 'Java Programming Concepts' and 'Java Memory Management'.

As COHSE is aware of the concepts in a document, different linking behaviours are also defined to affect the number and type of links added. For example, if a Web page is about the concept 'Java data types' COHSE would normally link its sub-concepts, such as 'int', 'float', or 'class' that appear on the page. However, the author of the page may not wish these sub-concepts to be linked, as they are likely to be explained on that page and therefore adding additional links would distract the user. The link suppression behaviour therefore prevents any sub-concepts that are within the super-concept's range (in this case the entire page) to be disregarded.

COHSE is an example of how ontologies and knowledge services are used to improve the hypertext evident in Web pages. It has been successfully used to improve the linking on the 'Java Tutorial' site and hypertext metrics were used to demonstrate evidence of improved linking (Carr *et al.*, 2002).

4.8.3 OntoPortal

OntoPortal is a collaborative project developed during this research to explore a new method of browsing hypertexts that is based on ontological structures. It uses ontologies to express 'real-life' relationships between Web resources. These resources are cast as instances of concepts in an ontology and are linked to other instances based on the relationships in an ontological model. OntoPortal is most suitable for

creating sites that provide information on complex and interrelated subjects (e.g. research portals) as it creates a principled, consistent, and intuitively linked site. OntoPortal is tightly integrated with this research and chapter 8 discusses it in more detail.

4.8.4 HealthCyberMap

HealthCyberMap (Boulos *et al.*, 2001) maps the complex field of health information using an ontological approach. A health model of various medical resources is constructed using Protégé 2000 and integrated with GIS (Geographic Information Systems) technologies to enable interactive hypermedia visualisations. The Dublin Core metadata set is used to define a resource's properties. Although this element set is limited by its simplicity, it is sufficient for HealthCyberMap as only the basic properties of resources are required.

Several interactive hypermedia visualisations are provided to demonstrate the access possibilities of the knowledge, such as browsing the health resources based on the location of their providers or on the body parts they apply to. It is thought that by combining this access framework with health information providers, a useful service for both patients and doctors can be established.

4.8.5 AKT

The Advanced Knowledge Technologies (AKT) project (Shadbolt, 2001a; Shadbolt, 2001b) is a large scale, six year, interdisciplinary project exploring all aspects of the management of knowledge. In particular, AKT is investigating the challenges of the Knowledge Management (KM) process that it defines as: acquisition, maintenance, modelling, retrieval, extraction, and publishing.

Several demonstrations of AKT technologies aimed at the Semantic Web have been implemented. For example, the ONTOlogy-based Community Of Practice Identifier (ONTOCOPI) (O'Hara *et al.*, 2002) demonstrates the communities of practice concept which defines groups of self-organising people with similar interests. It uses an ontology and populates it with community knowledge gathered from a variety of sources including Web documents. It then measures the strength of associations to discover instances that are closely related. By analysing peoples' interests, like-minded people are identified and this information could be used to locate experts in a discipline or to find people to collaborate with.

Natural Language Processing (NLP) has been used to analyse texts and automate the ontology population process. This has been used in a Web-based news server that is used to share stories between members of a research group (Vargas-vera *et al.*, 2001). As stories are submitted, they are automatically classified and the concepts and relationships within the article are extracted and used to populate the ontology with minimal interaction from the user.

The APECKS (Adaptive Presentation Environment for Collaborative Knowledge Structuring) (Tennison & Shadbolt, 1998) system enables collaborative ontology construction, and unlike general systems such as Ontolingua and Tadzebao, is aimed directly at domain experts. Rather than enforce consistency or correctness, APECKS enables multiple conceptualisations of a domain to coexist making it possible for domain experts to produce their own conceptualisation of a domain and then converse/debate over the various versions.

Knowledge technologies play a fundamental role in the Semantic Web and much of the outcome of the AKT project will be directly applicable to it.

4.8.6 OntoKnowledge

The On-To-Knowledge (OTK) (Fensel, 2000) project is similar to the AKT project in its focus on KM. It is also responsible for publishing the OIL standard and continues to develop it with facilities for the integration of concept instances (Instance OIL) and reasoning (Heavy OIL) in the ontology definition.

The OTK project is targeting three main areas of KM, highlighting the importance of ontologies in each one.

1. Acquisition: Integrating text analysis methods with ontologies.
2. Access: Improving keyword based methods of searching large collections of documents as these searches usually deliver large numbers of *potential* documents or return a document where it is the user's responsibility to then locate the search phrase in the document.
3. Maintenance: Tools to provide a systematic approach to knowledge maintenance.

OTK is also identifying approaches to improving the knowledge flow in organisations and thereby maintaining a business's competitive advantage. The tools they are creating are ideal for *non-specialist* knowledge workers who can access an

organisation's knowledge repository in an efficient way, and indeed, such tools are a vital element for the wider adoption of the Semantic Web.

4.8.7 Commercial Efforts

Network Inference (Network Inference, 2002) has released Cerebra that is a product aimed directly at the Semantic Web. It provides a machine reasoning mechanism which uses ontologies to enable machines to better interact with information. Similarly, Cycorp (Cycorp, 2002) has produced a multi-contextual knowledge base and inference engine which has been under development since long before the Semantic Web was proposed, the Cyc Knowledge Server. Its basis is a large-scale 'common sense' knowledge base that improves indexing and classification of documents, and provides knowledge services to enable other programs to access information about the concepts in documents. The common sense reasoning engine consists of a knowledge base, inference engine, the CycL knowledge representation language, a natural language parser, and semantic bus where semantic information (e.g. queries) are communicated.

Ontoprise (Ontoprise, 2002) focuses on the different roles of ontologies in the Semantic Web. A suite of products has been released to enable the creation and integration of ontologies into a wide range of applications. In fact, Ontoprise has released a commercial version of the Ontobroker system described earlier, as well as additional tools to author ontologies and annotate resources.

4.9 Summary

This chapter discussed the pivotal roles of ontologies and knowledge in the Semantic Web. Ontologies provide a conceptualisation of a domain and provide the facility to describe content semantically in a rigorous and expressive way that was not possible with earlier metadata models such as Dublin Core and GILS. However, this comes at the cost of a significant construction overhead that involves domain experts, end users, and computer specialists. Although a standard methodology for ontology construction has failed to emerge, guidelines suggest a highly iterative approach of conceptualisation and evaluation.

The use of ontologies in the Semantic Web requires a machine processible and Web-based representation language. DAML, OIL, and DAML+OIL extend the basic modelling primitives in RDFS to provide a highly expressive modelling language that enables the accurate conceptualisation and representation of domains.

The Semantic Web enables machines to analyse and reason over the content on the Web. This has been demonstrated in systems such as Ontobroker that provides extensive access to knowledge gathered from Web documents, COHSE where ontologies are used to improve the connectivity of Web documents, and AKT that is exploring methods of improving all aspects of the management of knowledge.

Chapters 7, 8, and 9 discuss the role ontologies and metadata play in constructing a principled and intuitive hypertext of scholarly material and in supporting scholarly inquiry. The next chapter introduces scholarly activities and practices and describes how the Web provides significant further opportunity to assist research.

Chapter 5

From Scholars to e-Scholars

5.1 Introduction

ESKIMO demonstrates how hypertext and the Semantic Web are integrated to assist scholars during their research activities, in particular, by providing scholars with more knowledge about the individual artifacts in their research community and how they are related. It enhances traditional scholarly activity to provide an environment in which scholars methodically and systematically explore their research field and make pertinent questions about it.

A scholar (researcher, scientist, or academic) is defined as an individual involved in advanced learning within a well-defined speciality area who desires in-depth information to support their research and enable the contribution of further ideas, thoughts, theories, and observations.

This chapter introduces the predominant activities of traditional scholars to explore and understand their research habits. This leads to a discussion on new electronic services that are emerging to support e-Scholars (electronic scholars) on the Web, and how scholarly data, such as bibliographies, are analysed to uncover patterns and useful facts. Finally, several new research projects are discussed that support scholars in their work.

5.2 The Traditional Scholar

At the heart of scholarly activity is the consumption and production of knowledge within a scholar's esoteric field. Scholars consume work published by others to appreciate new ideas and become knowledgeable in their particular field of study. They

then publish their own theories, experiments, observations, solutions, predictions, and refutations in journals and conferences. Following publication, debate ensues where peers refute, support, or modify the ideas by publishing further papers. As Bishop (1998) notes, "one begins by identifying and reading a source document and ends with the production of a document representing one's own work."

It could be argued that the first publication appeared in around 2400 BC on a Sumerian clay tablet (Kramer, 1963). However, the first serious advance in publishing came in 1452, when Johannes Gutenberg, a goldsmith and businessman from the mining town of Mainz in southern Germany, invented the printing press and enabled large-scale printing. His most significant work was the printing of a run of 300 two-volume bibles in 1456. However, the printing press was an aggregation of earlier technological advances, primarily the movable type and the proliferation of paper (as opposed to animal skin).

The printing press had an immeasurable effect on research as scholars could accurately publish their work in large quantity (Eisenstein, 1979). Before mass publication, a scholar's library consisted of a few hand-written manuscripts; with the advent of the printing press it was suddenly possible to obtain much larger amounts of knowledge and thereby improve the quality and quantity of research as communication between scholars improved. However, printing was expensive and therefore control of scholarly publishing moved to printers and publishers, a situation still evident today (Harnad, 1995b).

Furthermore, the printing press enabled the publication of non-verbal objects (e.g. diagrams, maps, images) (Cane *et al.*, 2001). A detailed account of the progress in the publication process is available in Eisenstein's book (1979). Although Eisenstein convincingly argues for the importance of technology (most notably the printing press) in promoting the scientific revolution, other historians disagree and point to the rise of universities and the changing, non-religious, attitude on books (Febvre & Martin, 1984) and the transformation of beliefs in cosmology, astronomy, and physics (Hatch, 1998).

The importance of publishing papers cannot be understated as they enable scholars to present their thoughts and claims to a large community of fellow researchers. Papers are mobile and *permanent* meaning they are consumed by scholars for many

years in widespread locations. For example, Vannevar Bush's seminal paper in 1945, 'As We May Think', which introduced the Memex concept, has been, and continues to be, widely cited. Brown *et al.* (1996) present the notion of a publication having a *social life*: a "paper transport carrying pre-formed 'ideas' or 'information' through space and time". Publications are unique in their way of binding communities together. Scholars, scattered around the globe and having never met, use publications to form a "robust social world" with a "strong sense of shared identity" (Brown & Duguid, 1996).

Collaboration also plays an essential role in research and is defined as the cooperation of two or more people in an intellectual endeavour. It takes many forms: writing a paper, working on a project, sharing expensive scientific equipment, or simply exchanging ideas. Collaboration is vital in research as it promotes the production and increases the quality of knowledge. Identifying the scholars that collaborate allows researchers to form communities of like-minded people, which then represent useful access points for knowledge in that field.

However, scholarly activity does not only revolve around publication and peer interaction. Scholars regularly meet at conferences and workshops to exchange ideas and establish collaborations. Conferences are funded and organised by societies and organisations, whose committee members are usually themselves researchers. Scholars are members of research teams (which are based at universities or organisations) that provide the environment to learn, conduct research, and interact with peers. They also work on projects which are based at research groups, funded by societies and organisations, and may be related to or controlled by other projects.

These relationships between scholarly objects (or artifacts) weave an intricate network of associations and are used by researchers to obtain a complete understanding of the material, issues, and events in their field. Literature contains many of these associations (e.g. structural, semantic, rhetorical, logic) as implicit or explicit references that researchers recognize and use. For example, researchers are identified through author lists and references, projects are described in the paper's content, research teams and organisations are mentioned in the affiliation section, and conference or journal information is outlined in the copyright declaration. An experiment by Dillon *et al.* (1989) demonstrated the extent to which scholars use

this peripheral information.

5.2.1 Research

The activity of scholarly research involves a process of systematic investigation and collection of information relevant to a particular research field. It is the sum of disparate activities and therefore becoming proficient in a field is a lengthy, on-going, and intricate task that requires a scholar's full attention. To manage this, scholars read papers to learn about a field and to reveal the perspectives of their peers on particular research issues. This process requires detective work (*What did the author of this paper go on to write? What other papers describe this project? Did this work influence any standards or software?*) because the ideas and concepts described in the literature are disconnected from the records and reports of the research activities which produced it.

Studies have been conducted to observe how scholars read and use papers. In a study at the University of Illinois, a group of researchers participated in an experiment to investigate how they used journals for research (Bishop, 1998). It was found that initially scholars skim an article to inspect if the material presented matches their own perspective. Usually this is accompanied by reading the abstract and/or introduction, or perhaps by going over the bibliography or diagrams in the article. While digesting a suitable article, readers attempt to formulate the ideas presented in the paper in their own language and understanding. They also noted that scholars predominantly locate further reading through the citations in a paper; in fact one researcher commented that 'sometimes an article is bad but the references are great'. However, the study noted that participants remarked that just following citations provided too narrow a view. One participant also used institutional affiliation information to broaden the search and find 'hot spots' of research.

Dillon *et al.* (1989) conducted a similar experiment and noted that a researcher's information gathering, while problem-driven, was not conducted in a systematic way. They noticed the participants making little use of ancillary material or tools such as citation indexes. The experimenters recorded that frequently the participants digested the author's address in order to gain an impression of their nationality and possible background. Following this, participants read the abstract,

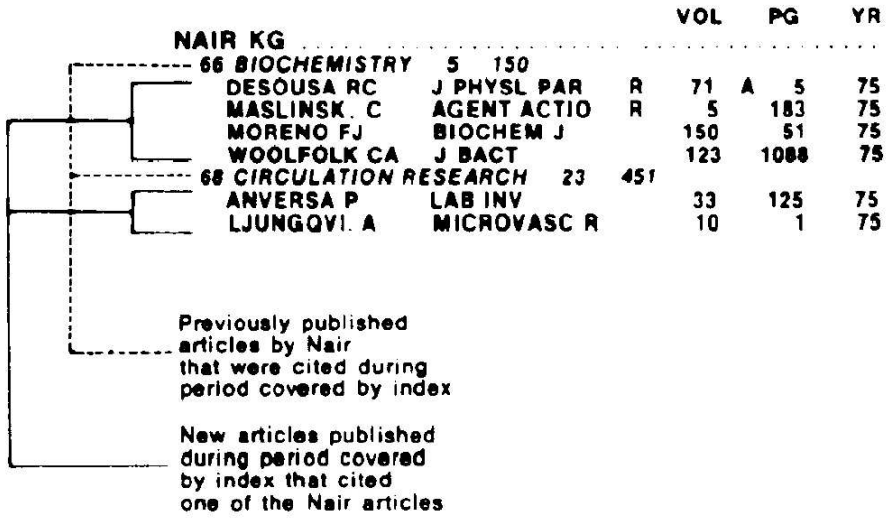


Figure 5.1: Example entry in the SCI citation index (Garfield, 1979a)

although most only skimmed it or read part of it, and then skimmed the rest of the article to ensure it was of interest, before reading it in depth.

5.2.2 The Citation

Citations are the most salient link between scholarly literature and are a prominent factor in providing the facility for scholarly debate. They have been "the way researchers have been interconnecting their writings all along" (Harnad & Carr, 2000). "Documents are not independent. Like biological organisms, every document is always related to some other" (Brown & Duguid, 1996). Indeed, the research impact of a scientific community is often assessed by the number of citations it attracts (Garfield, 1994).

Traversing citations is the primary method used by scholars to locate further literature. They enable scholars to uncover related ideas and produce a comprehensive literature survey. Citation indexes catalogue citations that a publication makes and link papers with cited works. Initially these indexes were used for locating literature and for providing a unique navigation experience. For example, the index enables a prospective search of the literature, in contrast to the conventional retrospective search, meaning scholars can establish how a paper has influenced a community and what subsequent papers and ideas it has contributed to.

One of the most common indexes in the field of science is the Science Citation Index (SCI) (Garfield, 1983) which contains references from 3,500 journals. An example entry from the SCI is illustrated in Figure 5.1. The table connects a

Citation Types	IBIS
Conceptual (Theory) / Operational (Method)	refers to / refers to
Organic (Essential) / Perfunctory (Non-essential)	responds to / generalises, other
Evolutionary (Development of idea) / Juxtapositional (Contrasting idea)	specialises / replaces
Confirmative (Support findings) / Negative (Oppose findings)	supports / objects to, questions

Table 5.1: Citation vs. Argumentation

paper published during a particular year with papers it has cited. It is organised alphabetically by the cited author, with a list of those papers that have cited it in references underneath. The SCI is frequently used by libraries to improve access to scientific information. Indeed, the National Research Library Alliance (NRLA) in America is using the SCI to create an advanced digital library for the Naval Research Laboratory (NRL) covering maritime research (Stackpole & Atkinson, 1998). However, a limitation of citation indexes is that any references to journals outside the index are excluded from its coverage.

However, the citation is not a hugely reliable indicator; as with hypertext linking, a citation bears no indication on the quality of the linked material. For example, politics play a substantial role in what papers are cited and how (e.g. colleagues will often cite each other's works while adversaries will not). In fact, a flawed work is often highly cited as peers refute the work: Is this a seminal paper?

Citation linking is also very field dependent. For example, within the biochemistry discipline the average number of citations made by a paper lies at 30, while in mathematics this is usually less than 10 (Garfield, 1979b). Also, isolated or specialised fields are likely to receive less citation attention than more general fields.

Nevertheless, a classification for citation types has been proposed by Murugesan *et al.* (1978) and parallels can be made with the IBIS argumentation model used in gIBIS (Conklin & Begeman, 1988) (Table 5.1).

The citation link is fundamental to the academic world; however, it should be treated with the utmost caution and not used as the sole mechanism for understanding and traversing scholarly material.

5.2.3 Peer Review

Peer review is the process used by publication mediums (e.g. journals) to select the papers to publish. It plays an integral part in scholarly activity and ensures a degree of quality and impartiality for published works. Peters (1995) succinctly sums up the objective of peer review:

> [It is] aimed at making a publication reflective of the peer community, not the [journal] editor's individual preferences and scope of knowledge

In general, the review process involves five stages:

1. A journal editor receives a paper.
2. The editor will usually use some basic elimination rules to weed out obvious inappropriate submissions before selecting a few reviewers (referees) from a predetermined review board to review and critique the papers.
3. Referees review the papers.
4. Referees make recommendations about the suitability of the paper.
5. Authors are informed if their paper is accepted or rejected.

The process is usually anonymous. The authors' names may be removed from the paper before the referees receive them. Similarly, the referees' names (themselves scholars) are omitted from the feedback the authors receive. However, the review process has received criticism with regard to the partiality of referees, the selection process, and the effects of the 'old-boy' network (The Economist, 1997). King (1987) proposes applying bibliometric indicators, the study of bibliographic data, to complement the peer review process and help assess scientific performance.

5.3 The Next Scientific Revolution?

The majority of universities and institutes are now connected to the Internet resulting in scholars increasingly looking to the Web to help them conduct research (Anderson, 1999). Already immense advantages in the access and dissemination of scholarly literature are noticeable with many journals and libraries providing digital equivalents of their papers.

As with many historical advances, when the printing press was invented its impact was at first difficult to gauge. The benefits were not actually realised for

more than a 100 years (Dewar, 1998) and with the advantage of hindsight, Eisenstein (1979) presents an in-depth discussion on the impact the printing press had. Indeed, Eisenstein is convinced the success of the Scientific Revolution (approximately lasting from 1550 to 1700) was partly due to the emergence of the printing press. Similar to the salient changes the Web has made, the printing press "changed the conditions under which information was collected, stored, retrieved, criticized, discovered, and promoted" (Eisenstein, 1979).

The Web delivers immediate advantages over traditional methods. "The web is changing the way that researchers locate and access scientific publications" (Lawrence & Giles, 1999). Dewar (1998) notes the similarities of the printing press and the Web; "there are some provocative parallels ... each defining technology represents an important breakthrough in the ability of humans to communicate with each other; each enables important changes in how we preserve, update and disseminate knowledge; how we retrieve knowledge; the ownership of knowledge; and how we acquire knowledge."

5.3.1 Accessibility

The Web offers scholars an increased quantity and breadth of available literature and thereby assists them in the fundamental problem: "physically to get hold of all the journal articles they need when they need them" (Hitchcock *et al.*, 1997a). The simplicity of publication and ease of access means that the corpus of literature can quickly find its way onto the Web. Indeed, this is also advantageous as it has been demonstrated that on-line articles are more highly cited than papers published only in paper form (Lawrence, 2001).

As a result of this, the Web has become a popular publishing medium for scholars in many fields (Harnad, 1995a; Harnad *et al.*, 1999). We are becoming e-Scholars, either through the actions of primary and secondary publishers placing their archives online and adapting to e-commerce opportunities, or else because of the actions of researchers themselves in using the Web to extend free access to their own work (Hitchcock *et al.*, 1997a).

5.3.2 Peer Interaction

The iterative process of debate and scholarly discourse is time consuming due to the lengthy delay between journal publications and conference proceedings, thereby

reducing a scholar's interest and momentum on a subject. The time frame between submitting a paper and receiving notification of acceptance takes from several months to years.

The Web makes this process more immediate and effective by speeding up the communication channels (Peters, 1996; Valauskas, 1997) and supplementing it with mechanisms, such as hypertext, to enable scholars to browse interconnected literature (Hitchcock *et al.*, 1997a). Electronic publication can also take advantage of many-to-many feedback, as on-line ad-hoc discussions between geographically distant scholars are possible.

Harnad (1991) proposes a revised peer review process, peer commentary or "scholarly skywriting", that takes full advantage of the Web as an interactive medium to publish unfinished works (i.e. 'working' papers or pre-prints) for peers to review and comment on before they disappear into the slow peer review process. This provides a sort of real-time, interactive, and open (i.e. visible to all) mechanism for discussing the issues in papers. However, Harnad does not propose replacing the quality control tool of peer review, but merely envisages scholarly skywriting as a supplementary tool. This unique collaborative environment is only feasible in an open and highly accessible medium such as the Web.

However, there are several serious implications with the scholarly skywriting principle. Firstly, as the material is unpublished and experts in the field have not reviewed it, there is potential for erroneous data and observations to be made in the document that can be particularly dangerous in fields like medicine. Secondly, scholars interested in the research described in a pre-print may be tempted to cite the paper. Even if the reference mentions the paper as being a pre-print, the paper's content could change dramatically (e.g. corrections of a major error or omission) and thereby possibly invalidate the context in which the citation was made. Thirdly, as a pre-print paper is unpublished the possibility of intellectual theft exists.

5.3.3 Interconnectivity

The Web is increasingly providing inter-connectivity, allowing cited research to be inter-linked so that e-Scholars are able to easily navigate through the research literature. The resulting hyper-web enables scholars to become quickly familiar with a research field in terms of its literature, activities, and authors, in order to make

subsequent pertinent and appropriate contributions. Indeed, "what would be the ideal online resource for scholars and scientists: all papers in all fields systematically interconnected, effortlessly accessible and rationally navigable, from any researcher's desk, worldwide for free." (Harnad & Carr, 2000),

Hypertext provides the facility to interlink scholarly material effectively. Link labelling and semantics, as demonstrated in hypertext system such as Aquanet (Marshall *et al.*, 1991) and MacWeb (Nanard & Nanard, 1993), provide the mechanism to relate scholarly material, such as literature, and enable scholars to understand how material is related and improve their ability to navigate around it. This is particularly suitable for citations, and the parallels between citation motives and the IBIS argumentation structure used in gIBIS (Conklin & Begeman, 1988) was explored earlier.

Optimistically, Cameron (1997) envisions a universal citation index containing every scholarly work ever written (c.f. Memex (Bush, 1945), Docuverse (Nelson, 1980)). Cameron considers the Web a possible mechanism for this and predicts how such an index would "serve as an important catalyst for reform in scholarly communication" by enabling scholars to uncover material and to better understand its relevance and impact. However, at the moment the vast majority of published papers are unlinked, although it is anticipated within the publishing industry that links on citations within scholarly papers will be one of the primary new services driving integration between scholarly sources (Needleman, 1999).

Linking to ancillary scholarly information is not unique to hypertext. Indeed, during the printing press era, encyclopedias and dictionaries provided pointers to further information in the form of footnotes, annotations, and cross-references. The Web improves this facility by providing *immediate* access. However, even this concept of large amounts of scholarly material instantly available is not particularly novel. Vannevar Bush's Memex concept as well as H. G. Wells' World Brain (Wells, 1938), are attributed with this.

5.4 Publication on the Web

The most common facilities for digital access to scholarly literature have been digital libraries, electronic journals, and e-prints. Digital libraries and e-journals are more

advanced in their support for scholarly activities than e-prints, and usually offer more than just a document download facility.

5.4.1 E-print Archives

E-print archives are highly automated and efficient repositories providing access to *free* scholarly papers. Although e-prints are less common and provide fewer services than digital libraries and e-journals, they contain papers that have usually been self-archived by authors or institutes with the purpose of making them easily available to the research community, and thereby removing the financial barrier evident in most e-journals and digital libraries. Usually, e-print archives only offer content management services, although facilities such as citation linking (e.g. OpCit (Harnad & Carr, 2000)) and citation ranking (e.g. CiteBase (Brody *et al.*, 2001a)) are emerging. Significantly, these services are being constructed in an open nature, independent of the underlying representation in the archives.

The growth of e-print archives, due to their low-cost entry and open nature, offers huge potential for integration of cross-disciplinary archives. The Open Archives Initiative (OAI) (Lagoze & de Sompel, 2001) is developing and promoting such low-barrier entry interoperability standards aimed at facilitating the dissemination of scholarly data between archives. Its goal is to enable every content service provider to be aware of the content provided by every other service provider, thereby improving access and availability of content. OAI enables publishers to expose their scholarly material using the Open Archives Metadata Harvesting Protocol, which is based on XML and HTTP. Through the integration of archives, services such as cross-archive navigation, searching, and citation linking are possible, by using a consistent interface to access the underlying unified global literature.

5.4.2 Electronic Journals

Work exploring how journals could be presented and accessed in electronic form started as early as 1977, when Senders researched the possibilities of implementing an electronic journal (Senders, 1977). Early projects such as BLEND (Birmingham Loughborough Electronic Network Development) (Shackel, 1982) also explored the feasibility of digital journal publication. However, the main transition from paper-based journals to commercial electronic journals (e-journals) started in the early

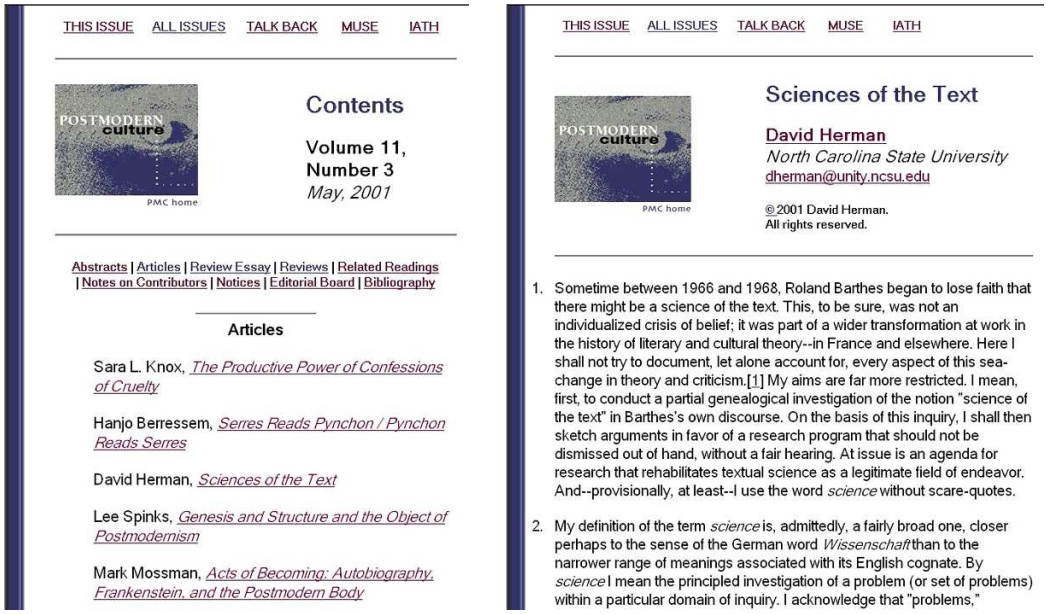
1990s with projects such as the Postmodern Culture[1], Harnad's Psycoloquy[2], and The Public-Access Computer Systems Review[3]. Aside from making publications available to a large community of users, e-journals also provide accelerated peer review and more effective discourse facilities (Valauskas, 1997). While traditional methods of publication incur a timely response lag due to the infrequency of publication and the lengthy selection process, on-line journals avoid these restrictions. Scholars are thus able to discuss and respond to papers more immediately while their ideas, thoughts, and motivations are fresh.

E-journals are becoming more advanced with many offering features other than publication. Discourse features are appearing that enable scholars to initiate newsgroup style debates about issues in literature (Sumner & Shum, 1998). Moreover, several e-journals augment papers with links to discussions, other articles, notification and alerting services, and also to auxiliary information such as dictionary definitions (e.g. Elsevier[4], IOPP[5]).

Wills believes that e-journals represent "the grandest revolution in the capture and dissemination of emerging academic and professional knowledge and information since [William] Caxton[6] developed his printing press" (Wills, 1995), although others take a more sober view. Valauskas (1997) argues that the differences between electronic journals and their paper relatives are not as radical as believed by some. He acknowledges the difference in process but remarks that the core constituent of peer review and verification remain. Valauskas believes an equilibrium between the digital and paper publication world will develop to offer scholars a rich variety of media to select from.

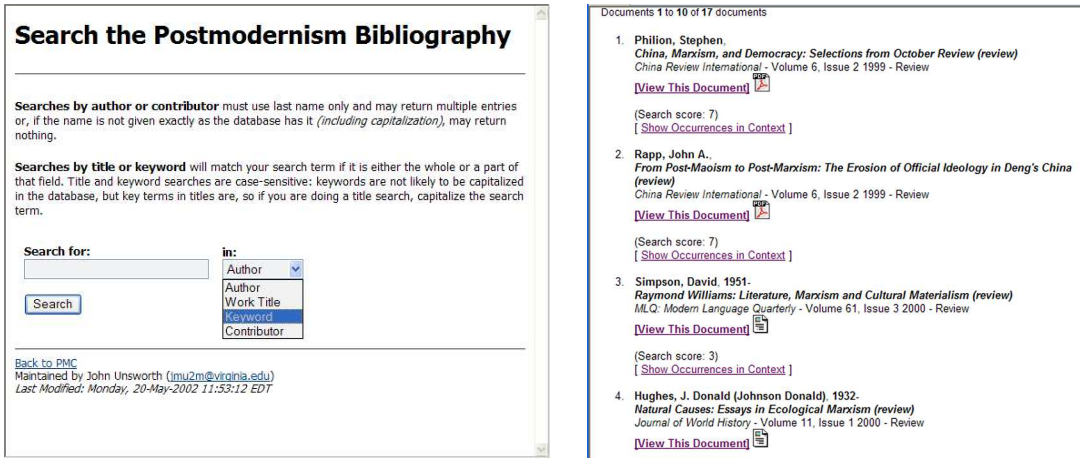
The Post Modern Culture electronic journal[7], the Internet's oldest peer-reviewed e-journal, contains peer-reviewed content in the humanities. Within articles or reviews, a few citation links are available as well as links from the authors' names

[1] `http://www.iath.virginia.edu/pmc/contents.all.html`
[2] `http://www.cogsci.soton.ac.uk/psycoloquy/`
[3] `http://info.lib.uh.edu/pacsrev.html`
[4] `http://www.elsevier.com/`
[5] `http://www.iop.org`
[6] William Caxton (1422-1491) was the first English printer. Caxton printed nearly 100 publications including the Canterbury Tales.
[7] `http://www.iath.virginia.edu/pmc/contents.all.html`



(a) Table of contents (b) A publication

Figure 5.2: Post Modern electronic journal



(a) Search options (b) Search results

Figure 5.3: Searching the Post Modern e-journal

to general information about them. It is also possible to view the entire collection in a bibliographic fashion, although again there is a lack of interconnectivity.

Figure 5.2 displays the contents page and example publication from the Post Modern Culture e-journal. The similarity to traditional journals is immediately evident as its style closely resembles that of a paper-based journal. It is possible to search the bibliographic entries by title, author, keyword, and contributor (Figure 5.3a), to which a list of potential matches is returned (Figure 5.3b).

McKnight (1997) acknowledges the benefits that e-journals deliver; they provide continual access to papers, can include other media types (e.g. video clips),

have search facilities to quickly locate material, and use hypertext to link articles. However, he notes the difficulties of reading papers from screen, the uncertainty of papers being readable/compatible with future software, and that poorly indexed and linked e-journals make it difficult for users to navigate the site.

An experiment conducted by Woodward *et al.* (1997) at Loughborough University involving staff and postgraduate students, confirmed that users preferred to read paper versions of articles. Therefore, they propose that human factors will determine the success of e-journals, and not simply the fact that a large quantity of papers are accessible on-line.

5.4.3 Digital Libraries

Digital libraries are like their traditional counterparts except they store, access, and disseminate *digital* documents. They contain literature from journals, conferences, magazines, and books. The software systems used to run digital libraries are more complex and diverse than e-journal or e-print software, and "can be among the most complex and advanced forms of information systems as they often involve collaboration support, digital document preservation, distributed database management, hypertext, information filtering, information retrieval, instructional modules, intellectual property rights, multimedia information, question answering and reference services, resource discovery and selective dissemination of information." (Fox & Marchionini, 1998). Clearly, digital library construction is an expensive and resource-intensive task (McCray & Gallagher, 2001).

Although digital libraries are used to publish academic papers, their applications are extremely diverse. The Perseus Project[8] (named after a Greek hero who explored the limits of the world) is a popular digital library, which contains resources for the study of the ancient world and beyond. As with the Post Modern Culture e-journal, the library has a familiar feel with tables of content and information retrieval (e.g. a search engine) being the dominant methods for locating literature. Both textual and visual resources are available, coupled to search facilities and secondary resources (encyclopedias, dictionaries, grammar guides). Although many of the texts are heavily linked, this is mainly to help translate texts (e.g. the Greek phrase 'luchnou' is linked to the translation 'a portable light, a lamp'). Hypertext

[8] `http://www.perseus.tufts.edu/`

(a) Introductory page (b) Papers within a particular theme

Figure 5.4: ACM Digital Library

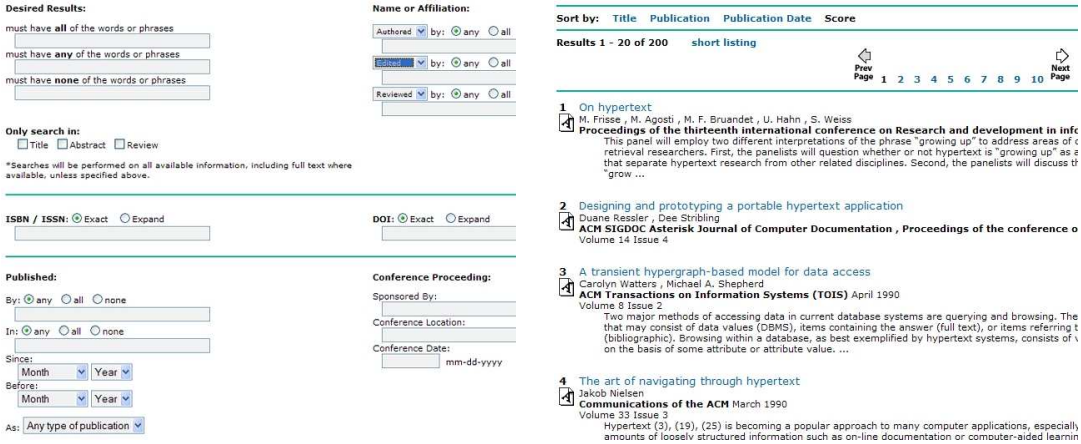
links to further material are generally lacking, meaning users rely on the search engine.

Figure 5.4 displays two screen shots of the ACM Digital Library (ACM Portal). The library contains papers from all ACM sponsored journals and conferences spanning many diverse technical topics such as communication security, document engineering, and operating system design. Figure 5.4a displays the introductory page for the digital library which enables users to select the part of the library to browse. Once users have selected which journal or conference proceeding they are interested in, they retrieve a listing of the available publications (Figure 5.4b). Users then select the appropriate link to retrieve the full-text of the article.

Figure 5.5a illustrates the (literature oriented) search mechanism provided by the ACM. The result is a list of links to full-text articles within the library (Figure 5.5b).

A study conducted by Theng (1999) concluded that digital libraries, including prominent ones such as the ACM Digital Library, caused users to become disorientated, with the main difficulties cited as the inability to (i) easily return to previously visited information and (ii) retrieve information that users believed existed. This experience is a similar predicament to the 'lost in hypertext' syndrome that afflicts many hypertext systems.

Like e-journals and e-prints, digital libraries provide a location for scholars to rapidly access a large quantity of similarly themed literature, although improved interlinking is necessary as Fox *et al.* point out: "the challenge is to find ways



(a) Search form (b) Results of a search

Figure 5.5: Searching the ACM Digital Library

to link the diverse content and perspectives provided by individual digital libraries around the world" (Fox & Marchionini, 1998). Searching by issuing a query is the most common method of finding information in a digital library although new approaches such as distributed library searching (French *et al.*, 1998), index schemes for structure-based querying (Lee *et al.*, 1996), and phrase-based searching have been proposed (Nevill-Manning *et al.*, 1997). Due to the lack of interconnection, hypertext navigation is only of limited use as an exploration tool.

5.5 Scholarly Analysis and the Web

The enormous amount of scholarly data, such as in bibliographies, enables further analyses to uncover trends, patterns, and new information. Traditionally, analysing scholarly information was a difficult task as it involved the capture and merging of data from various sources and quickly became unmanageable as the amount of information grew. However, computers improve the situation as they are able to rapidly access and process data. In addition, Web standards such as XML, OAI (Lagoze & de Sompel, 2001), and the Academic Metadata Format (AMF) (Brody *et al.*, 2001b) are improving the interoperability of scholarly sources and are making large quantities of data available in a structured and machine processible format.

5.5.1 Citation Networks

The citations that are established between scholarly works form a network of knowledge that connect a body of literature and are an obvious candidate for further analysis and visualisation. For example, Figure 5.6 illustrates a citation network

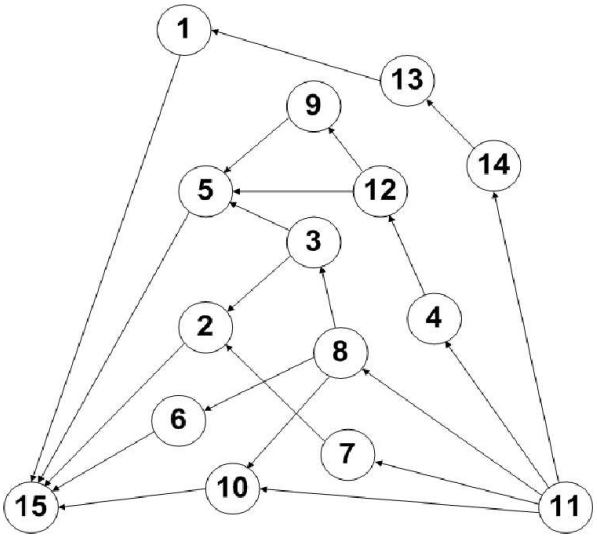


Figure 5.6: Example citation network

where publications are positioned in chronological order from left to right. The node labelled *15* represents the starting publication in this diagram. It is cited by the nodes labelled *1*, *2*, *5*, *6*, and *10*. The final publication in this network is the node labelled *11*, which cites five other papers. The network provides a clear overview of the papers in a field and how they have affected it.

Garfield (1979a) proposes that this visual representation makes it easier to grasp the historical development of a field, as well as identifying key publications and research areas. An analysis of these maps can lead to the discovery of patterns, trends, and other implicit knowledge, such as citation patterns to identify new research fields. However, as experienced with hypertext overview maps (Gloor, 1991; Robert & Lecolinet, 1998), complex interlinked structures are difficult to visualise and understand efficiently and effectively, and quickly become cluttered as the number of nodes increases.

Providing links on the Web between a paper and its citations is an active research area (Harter, 1996; Cameron, 1997; Hitchcock *et al.*, 1997a; Hitchcock *et al.*, 1997b) as it promises to provide scholars with significant benefits of accessibility and interconnectivity not available in paper-based approaches. Early work highlighting the use of citations was conducted by Eugene Garfield (1955), and has since also been demonstrated on the Web (Harnad & Carr, 2000). Carr *et al.* (1999a) explain how citations are used to uncover principal researchers based on the frequency with which their papers are cited. The Open Citation Project (Harnad & Carr, 2000)

have added citation links to a massive physics archive to link every paper to every other paper it cites within the archive, enabling physicists to navigate the literature in a unique way.

5.5.2 Bibliometrics

Bibliometrics is the study of bibliographic data using mathematical and statistical methods. Due to the difficulty in obtaining this data in large quantity and in electronic form, bibliometric analysis has received less attention than the potential advantages it offers would suggest. "Most practicing scientists seem completely oblivious to the large literature of citation and bibliometric studies" (Garfield, 1996). The attention directed at bibliometrics is divided.

Advocates stress the importance of analysing bibliographic data to uncover otherwise unknown facts and trends (Garfield, 1955; Garfield, 1996), while others believe that the citation link is too simple a semantic link to draw any accurate conclusions from and that bibliometric methods oversimplify this complex issue (Edge, 1977). Nevertheless, with an increased number of on-line bibliographies, improved citation capturing techniques, and the capabilities of computer processing, bibliometrics is becoming more practicable.

Bibliometric analysis ranges from simple analysis (e.g. a publication count for a research team) to intricate and complex analysis (e.g. to study science over time, evidence of cooperation in research, geographic distributions). Jacobs (2001) has conducted a study where he demonstrates a link between productivity and funding using bibliometrics. Holmes *et al.* (2001) explain how bibliometrics was used to predict the outcome of a Research Assessment Exercise (RAE).

There are two main areas of bibliometrics. Firstly, that characterised by the application of statistical and mathematical models to bibliographic data. Lotka (1926), Bradford (1934), and Zipf (1981) have been influential in this field (each producing a bibliometric law named after them). For example, Lotka's law measures scientific activity and states that the number of authors making n contributions is k/n^2 where k is a constant.

Secondly, there are empirical methods based on the study of relationships in bibliographies. Examples include co-citation analysis, impact factors, and collaboration measures. Garfield (1979b), Small (1973), and Price (1965) are key researchers

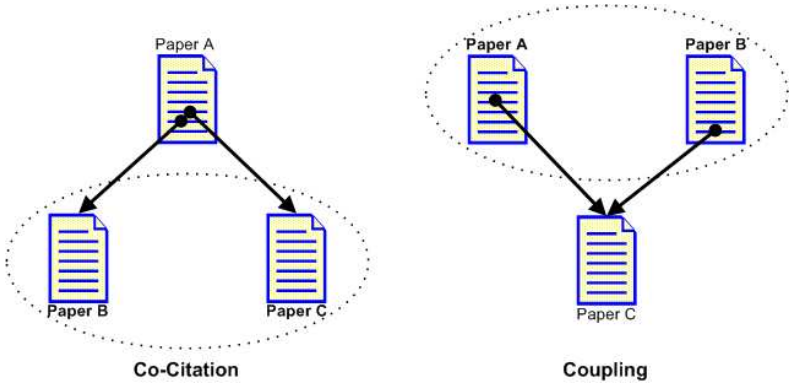


Figure 5.7: Co-citation vs. Coupling

in this area. As opposed to the statistical methods, these rules are flexible as they represent principles rather than mathematical formulae.

Co-Citation Analysis

Two papers are co-cited when a third paper cites them both. In Figure 5.7, paper A cites both papers B and C and therefore papers B and C are said to be co-cited. Co-citation is the reverse of Kessler's notion of bibliographic coupling (Kessler, 1963) which measure papers based on the similarity of their bibliographies. This means that contrary to co-citation, bibliographic coupling takes a retrospective view and is therefore less useful (e.g. it fails to uncover new trends that appear).

Co-citation analysis relates bibliographic data based on co-citation strengths (i.e. the number of times two papers are cited together). These values are used as proximity measures in visualisations where papers that are frequently co-cited are plotted near each other. The resulting graph enables research fronts to be identified, the theory being that a research front will usually emerge around a few seminal (or core) papers that are heavily co-cited (Small, 1973).

In Figure 5.8, a co-citation network of computer graphics papers from 1982-1999 is illustrated (Chen & Paul, 2001). A node represents each document, and a line between two nodes indicates that these are co-cited documents. The network makes it possible to identify highly co-cited clusters. In addition, citation impact bars above each node indicate the number of times the document has been cited, meaning research 'hotspots' can be discerned.

Co-citation analysis has received criticism (Edge, 1977; MacRoberts & MacRoberts, 1989) due to its over-simplification of the citation link, technical problems

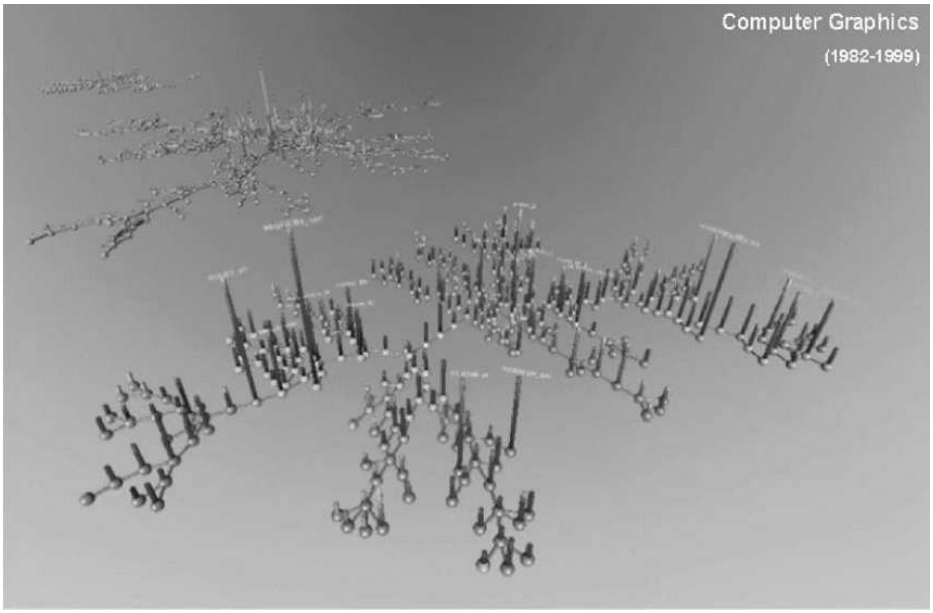


Figure 5.8: Co-citation network for the Computer Graphics discipline

(e.g. inaccurate citations), and the focus on citations when other factors (e.g. social/political motivation behind the citation) should be taken into consideration. Indeed, these factors are applicable to all bibliometric measures. However, as Garfield notes co-citation provides a perspective on scholarly material when used cautiously and wisely. "Citation analysis is a messenger. It doesn't 'prescribe' anything; it merely 'describes' " (Garfield, 1993).

Garfield has successfully demonstrated how he uncovered important historical links between research fronts using co-citation analysis that scholars had previously overlooked (Garfield *et al.*, 1964). Furthermore, ISI has used co-citation analysis in the Science Citation Index annually since 1973. This is used to identify and provide details of thousands of research fronts, emerging areas of research, and those authors and institutions that are active in these areas (Small, 1973). These could be used by funding bodies to help decide which areas of research are emerging and then direct funding towards the relevant projects.

Impact Factor

Another popular bibliometric tool is the impact factor, which applies to journals and is a measure of the frequency with which the average journal article has been cited over a year (or period). ISI publishes yearly Journal Citation Reports (JCR) drawing from over 8,400 journals worldwide and provides information on impact

factors (as well as journal sizes and the 'hottest' journals). Institutions and libraries use JCRs to evaluate and rank the significance (and quality) of journals and can use this information to, for example, make informed decisions on which journals to stock.

To calculate the impact factor for a particular journal in the year 2000, the following formula is used:

```
A = cites in 2000 to articles published in 1999
B = number of articles published in 1999
Journal Impact for 2000 = A/B
```

The impact factor tends to be less biased than other measures as it does not benefit larger journals over small ones, frequently issued journals over less frequently issued ones, or older journals over newer ones. However, as with all citation-based measures, the impact factor should only be used as a guide.

Collaboration Measures

Smith (1958) proposes using co-authorship as a measure of collaboration and this has been further advocated (Clarke, 1964; Beaver & Rosen, 1978). Altering patterns of funding, increased interdisciplinarity, escalating demands, and the growth of research have all contributed to the growing number of multi-authored papers (Katz & Martin, 1997). In addition, the increasing ease of communication between researchers across the globe has made it easier for scholars to co-publish. However, these factors have also contributed to the increasing difficulty in gauging collaboration based entirely on co-authorship.

There is little doubt that closely collaborating researchers will frequently co-author. However, it is whether the reverse holds true which has caused controversy. Often researchers simply pool their results or are only marginally involved but still receive recognition. On the other hand, two researchers might work together on a project but then publish independently.

Therefore, co-authorship is at best a partial indicator of collaboration and it is important to treat collaboration with the same caution as for all bibliometric studies.

5.6 Supporting e-Scholars

Improving support for scholarly research on the Web has been the focus of several disciplines, such as library studies, hypertext, and knowledge management. This section presents significant research in this area.

5.6.1 D³E and JIME

The Digital Document Discourse Environment (D³E) (Sumner & Shum, 1998) is a tool for document-centric discussion which applies integrated discourse facilities and supports the publication of Web-based documents.

The tool has been used in several settings including an interactive journal, the Journal of Interactive Media in Education (JIME) (Shum & Sumner, 2001). Figure 5.9 illustrates a document being viewed using JIME. The right pane is where the discourse facilities are available (between readers, authors, reviewers, and editors). Each document has a comment icon (1) that users use to comment on a paper. To aid navigation of a document, a content list (2) is presented. HTML and PDF[9] versions of a document are available (3). Citations are linked to their corresponding references (4) and a reverse link is inserted for each citation (5). An editorial note is added (6) to draw the reader's attention to important issues about the paper. The D³E environment also provides section-specific commentary (7) as well as an editorial commentary (8).

JIME makes extensive use of hypertext to present users with an interconnected approach to reading and interacting with papers, and enabling scholars to discover how fellow researchers view and have commented on the issues in them. The approach is only limited by the screen real-estate requirement, which is significant.

5.6.2 ScholOnto

The ScholOnto (Shum *et al.*, 1999) project highlights the advantage of exploring (non-citation) relationships between literature to build networks of knowledge. The project uses scholarly claims within literature, instead of direct facts about the literature or its community, to assert relationships between papers. Authors make claims about their work and usually back these up through citations to other literature. ScholOnto captures these claims and categorises them into relationships such

[9]Portable Document Format (PDF)

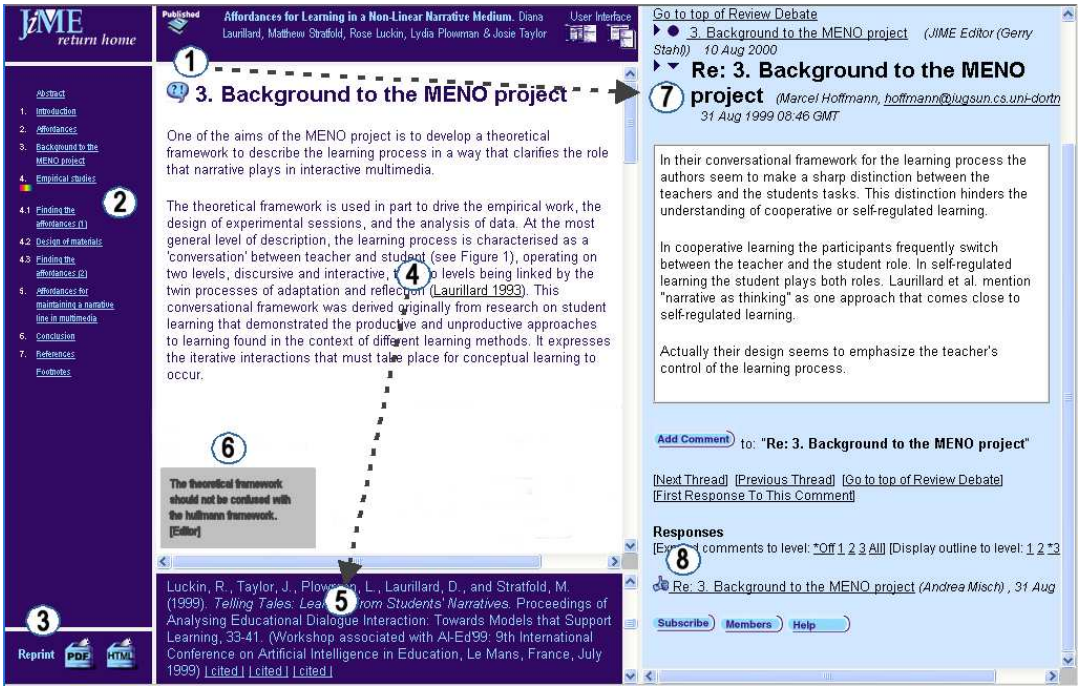


Figure 5.9: JIME interface

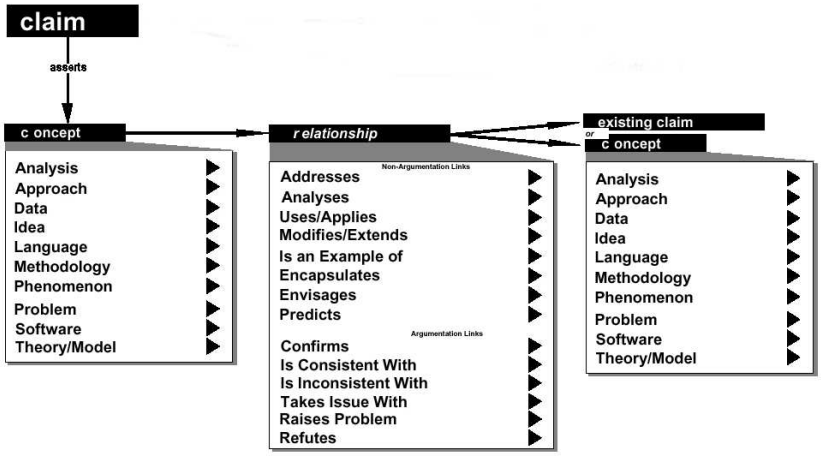


Figure 5.10: Ontology used to represent claims

as addresses, analyses, modifies, predicts, and uses, and so an intricate network of claims is established.

Figure 5.10 illustrates the claims ontology used by ScholOnto to capture the relationships between papers. Scholars make claims about a concept that appears in a paper and specify its relation to another concept in a different paper. The types of relationships are grouped into two categories: non-argumentative and argumentative. These broad claim types are able to capture most relationships evident in scholarly works.

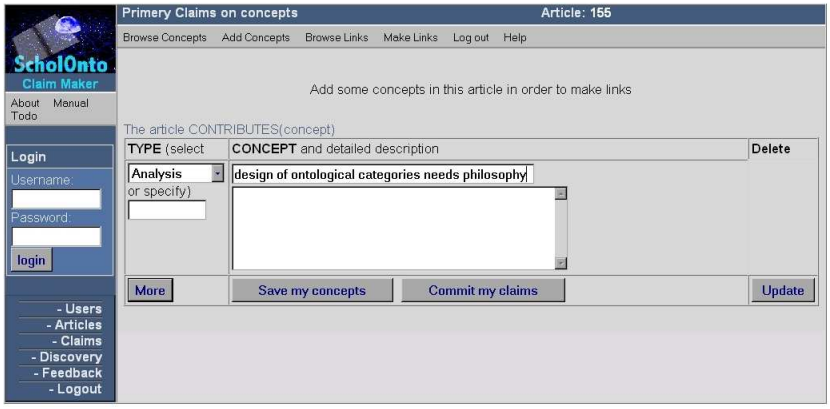
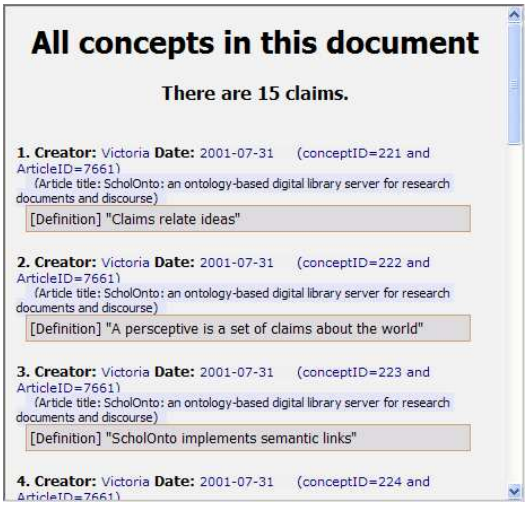
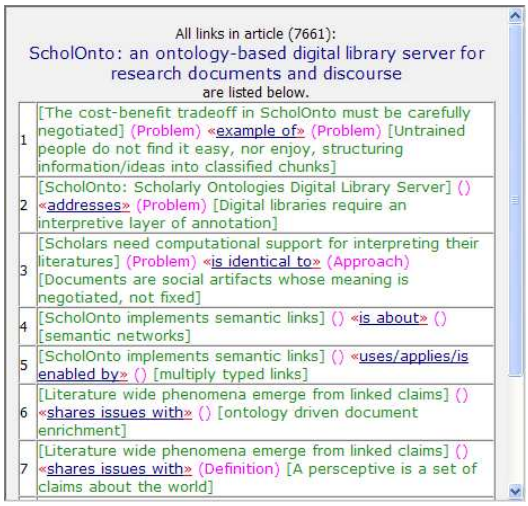


Figure 5.11: ScholOnto Claim Maker



(a) Concepts in a paper



(b) Claims in a paper

Figure 5.12: Viewing concepts and claims in ScholOnto

The claim maker tool is used to author claims between concepts in papers. To make a claim (or link) between the concepts, the relevant concepts are added to ScholOnto (Figure 5.11) or selected from a list if they already exist (Figure 5.12a). A claim is then created by supplying a claim type between the two selected concepts. All claims within a document can then be viewed to enable users to identify relationships with further papers (Figure 5.12b).

Analysing ScholOnto claims enables sociological queries to be answered, such as "What motivated the Dexter Hypertext Reference Model" and "What impact did Einstein's theory of relativity have?" To answer the former query, all claims in the ScholOnto knowledge base where the 'Dexter Reference Model' concept (a concept of type 'Theory/Model') has a claim relationship with another 'Theory/Model' concept are retrieved. A simple solution is then to return those claims with a

relationship type of 'Modifies/Extends', 'Uses/Applies', 'Takes Issues With', 'Raises Problem', and 'Refutes'.

In theory, ScholOnto provides an elegant solution to answering otherwise complex questions, although the authoring overhead in identifying claims in source documents is likely to be large. The success of queries is difficult to gauge, as they will not have concrete or obvious answers. In addition, it is difficult to judge the quality of the claims added to the system, as these depend on the personal experiences and thoughts of the claim author and therefore may be inaccurate.

5.6.3 ResearchIndex

The ResearchIndex (formerly CiteSeer) (Lawrence *et al.*, 1999a; Lawrence *et al.*, 1999b) is a large database of citations retrieved using Autonomous Citation Indexing (ACI). The ACI process used in the ResearchIndex is as follows:

1. Existing search engines (e.g. Alta Vista) are used to locate scientific papers. To improve coverage multiple search engines are used.
2. Documents are converted into text.
3. Citations are automatically extracted and the context in which they are used recorded.

By searching the Web, the ResearchIndex collects a large number of papers with their citations. As it also notes the context of a citation, the index is useful as a prospective searching tool. The automation of this process means the ResearchIndex removes the significant authoring overhead usually required when analysing scientific literature. However, due to the unregulated and non-standard methods used to record citations, this occurs at the expense of occasional incorrect entries.

Figures 5.13 and 5.14 illustrate the information that ResearchIndex displays for an article in its database. In addition to title, author, full-text link, and abstract information, extensive bibliographic information is available. There are also links to papers that have been identified as *similar*; both at the sentence and co-citation level.

Figure 5.15 displays the context mechanism which is used to view the context in which a citation is made. By monitoring the database of citations, the ResearchIndex also provides tools to alert users of new citations to specified documents.

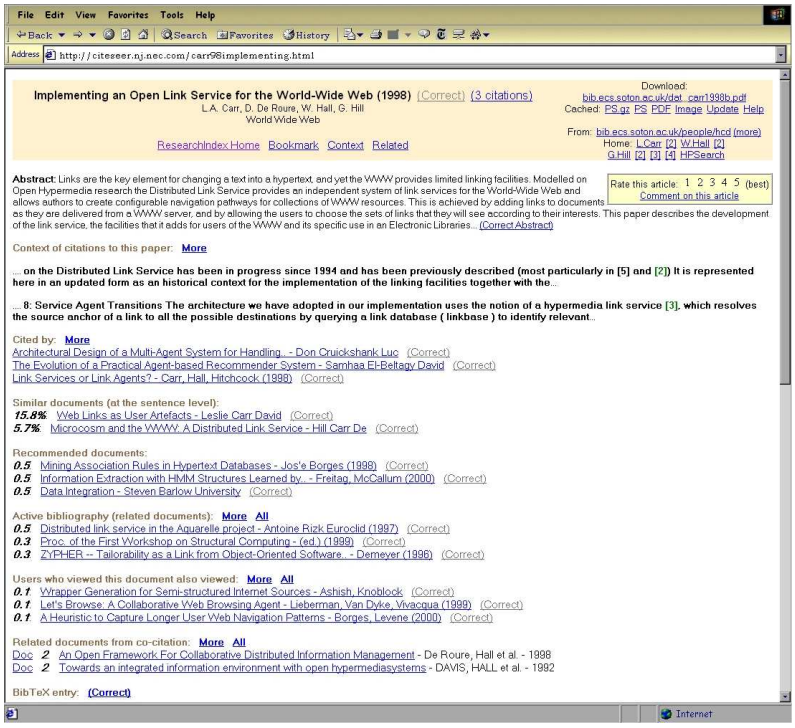


Figure 5.13: ResearchIndex - Document information

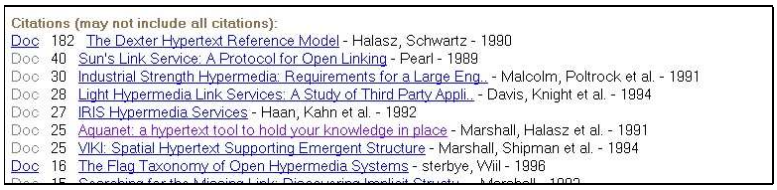


Figure 5.14: ResearchIndex - Citation information

The ResearchIndex is encouraging as it demonstrates the feasibility of automatically capturing scholarly data from the Web and offering a useful service to researchers. The automated process ensures that the index is always up-to-date and has a wide coverage. In return, users tolerate the occasional errors that are evident in the index.

5.6.4 Open Citation Project

The Open Citation (OpCit) Project (Harnad & Carr, 2000) is a three year funded project that is applying citation linking to the massive arXiv archive. The archive contains almost 200,000 papers, is growing at more than 25,000 papers a year, and has a daily user base of 35,000. It is the de facto archive for many physicists.

The project builds on earlier work from the Open Journal Project (Hitchcock *et al.*, 1998) where publishing journals in a network environment using hypertext

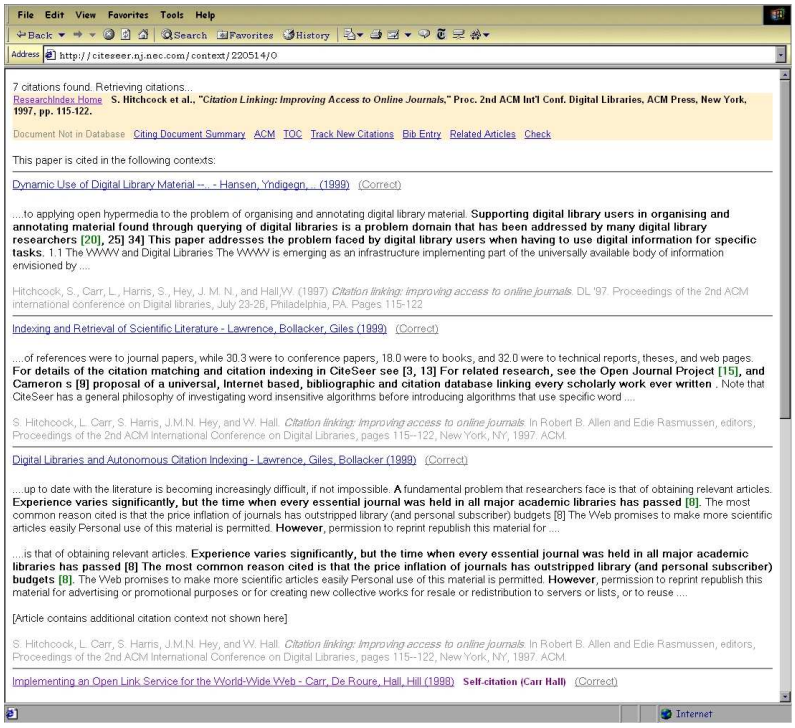


Figure 5.15: ResearchIndex - Citation information

facilities was explored. OpCit links together all the cited works in the archive and provides an effective method for scholars to browse the literature. The effort is complicated by the fact that the papers are stored in various formats, such as PDF and postscript. Automated tools to parse these documents and extract and analyse their citations were constructed. For example, the proportion of citations that point to current papers in the archive and the most frequently cited papers can be calculated. Web access patterns to the archive are also logged, enabling researchers to view how papers are cited and accessed over time. For example, in the first few weeks of publication, the access rate is high, and then quickly tails off (Harnad & Carr, 2000).

5.6.5 SLinkS

Like the OAI, the Scholarly Link Specification framework (SLinkS) (Hellman, 1999) improves interoperability between publication repositories. It defines an XML syntax for linking references between publishers. This linking has previously required publishers to agree on a linking protocol, not only due to commercial obligations, but because many reference links are created dynamically and are likely to change.

However, publishers are responsible for implementing the actual SLinkS process, which is used to convert the reference specifications into hypertext links.

SLinkS does not only specify a reference specification, but also enables metadata to be attached to links using RDF. For example, the type of publication, its title and ISBN, and what type of content the link points to (e.g. full-text article, abstract, table of contents) can be specified. This not only enables users to determine the content of the destination of a link, similar to link labeling in hypertext systems such as Textnet (Trigg, 1983) and Aquanet (Marshall *et al.*, 1991), but also allows publishers to filter references based on user profiles or access rights.

SLinkS provides the foundation for scholarly linking between competing publication mediums. The financial barriers usually imposed are avoided by using SLinkS to only link to an article's abstract. Alternatively, SLinkS provides a digital signing facility which can be used for verification, enabling scholars who have subscribed to several publishers to seamlessly browse between the repositories.

Further scholarly linking is envisaged, that extends the specification to include data on researchers, affiliations, projects, and topics. This allows researchers to request all papers published by a particular author or all papers on the topic 'network security'.

5.6.6 Foxtrot

The Foxtrot recommender (Middleton *et al.*, 2002) system suggests research papers to scholars based on their research interests. The system monitors the literature that scholars view and adds them to a dynamic shared database. The documents are classified using concepts from a topics ontology which is then used to construct a personalised profile for each scholar. The advantage of using an ontology to represent profiles, is that the relationships and structures in an ontology can be used to infer other properties of a profile.

Profiles in Foxtrot are computed daily and used to group scholars with similar research interests. Therefore, recommendations are papers on a scholar's research topics that have also been read by fellow researchers with similar interests. The profile analysis could also be used to identify potential collaborations within a community.

5.7 Summary

This chapter has introduced the research practices of traditional scholars, in particular how they conduct research, disseminate information, and interact with fellow researchers. The Web represents a revolutionary new medium that is increasingly being used as a platform for publication, integration, and the sharing of scholarly knowledge; e-Scholars benefit through the facilities of hypertext to improve inter-connectivity between scholarly information, faster communication channels to improve peer interaction, and rapid accessibility for authoring and retrieving documents on the Web.

Indeed, the support for research on the Web has advanced significantly since scholars first used e-mails to discuss and share research information.

- More electronic publications are being placed on the Web, either by publication mediums such as digital libraries and e-journals, or by authors themselves to extend free access to their research.
- Publication mediums are offering more extensive search capabilities and are making increased use of hypertext.
- Projects have already demonstrated how scholarly activities, such as discourse and research, and the publication and integration of distributed scholarly material, can be supported and implemented.

However, current scholarly benefits on the Web fall somewhat short of the potentials discussed in this chapter, partly because of the apparently conflicting rights and responsibilities of authors and publishers of the research literature, and partly due to the lack of required infrastructure, technology, and experience. While the accessibility and editorial processes of scholarly publications on the Web have improved, the full advantages scholarly hypertext could offer to the Web have yet to be fully applied or demonstrated. Without adding rich links to a wide spectrum of scholarly information, and not just the literature, researchers are unable to access, and are unaware of, the potential knowledge that is relevant to their research, and instead find themselves locked in a similar situation as in the traditional paper-based world; having to employ detective skills in locating scholarly material. In addition, the analysis of scholarly texts, such as bibliometrics, is being largely used

in limited applications and their advantages on the Web have yet to be fully exploited. These are issues addressed in ESKIMO, which demonstrates and promotes the use of scholarly metadata to provide a highly interlinked and principled research environment.

The next chapter discusses an experiment to explore the effectiveness of the current Web as a research tool, in particular, assessing its ability to answer the analytical questions scholars make about their research field.

Chapter 6

Study: Research on the Web

6.1 Introduction

This research explores new approaches to supporting e-scholars in using the Web as a research tool; a task that demands an understanding of how the Web *currently* measures up to a scholar's requirement in order to determine how to improve it. Chapter 5 discussed the numerous possibilities and advantages of supporting scholarly research on the Web, in particular the improved accessibility and interconnectivity of scholarly material. Unfortunately, many of these facilities, such as coherent interconnected access to scholarly literature and its supporting material, have yet to be fully realised.

This chapter presents an experiment that was conducted at the beginning of this research to identify potential problems in using the Web as a research environment, and to study the approaches scholars used when interacting with it. Therefore, participants were observed as they used the Web to answer nine typical research questions.

The chapter introduces the motivation, background, and structure of the experiment, and then describes the outcome and results.

6.2 Motivation

The experiment collected information on scholarly research habits on the Web to determine the type, quantity, and quality of information used. For example, do scholars mainly interact with a search engine either on the Web or within a digital library to obtain papers (i.e. information retrieval), or do they browse/traverse the

different parts and artifacts in their research community to locate papers and at the same time gain a deeper understanding of their domain (i.e. hypertext navigation)?

From personal experience (e.g. writing this thesis) and that of peers, questions such as these are common:

- Who are the experts for hypermedia reference models?
- What else has been written on this topic?
- Is there a journal edition that discusses the benefits of using ontologies in system integration?
- What are the seminal papers in knowledge management?
- Which research teams focus their research on digital library technologies?

Therefore, the motivation for this experiment was to understand the current scholarly support afforded by the Web, whether scholars are aware of current Web research tools (e.g. digital libraries, citation indexes, portals), and if they actually use them appropriately.

6.3 Background Studies

The study conducted by Theng (1999) concluded that digital libraries caused users to become disoriented, citing reasons similar to that of poorly constructed hypertexts. Therefore, although the organisation and access of on-line documents benefit from hypertext and the Web, real improvements are only realised for carefully constructed sites that have also considered human factors in the design process (McKnight, 1997).

A further study conducted by McKnight *et al.* (1992), compared the speed and accuracy with which participants read hypertext documents compared to equivalent paper versions. In the first part of the experiment, participants were asked to use the provided texts to answer ten questions and their time and accuracy were recorded. The result of this experiment demonstrated that when using paper the questions were answered significantly faster, although the accuracy ratings were similar. Participants also noted problems with navigating and using question searches in the hypertext version.

In the second experiment, subjects were asked to read a substantial document and answer an essay style question about it. Domain experts then graded the essays,

unaware of which medium was used in producing it. No significant differences between the two media were noted, although participants complained of uncertainty in locating all the necessary information when using the electronic version.

Baragar (1995) investigated how an article originally intended for paper-print, could be converted to hypertext to deliver a paper that was easier to read. She notes that this process requires a new way of thinking comparable to being a co-author; a deep understanding of the content and issues raised in the paper is required. A substantial authoring effort is needed to deconstruct a text into individual sections, and then reconstruct them using hypertext. Reconstructing the text is more than simply 'matching words', it requires a domain expert to identify new contexts and create conceptual structures. In this project, Baragar created 150 hypertext links between the different fragments of the text at a cost of 30 hours authoring effort.

Constructing scholarly hypertexts is a complex task that requires a deep understanding of the content being linked. However, due to the high authoring overhead many scholarly hypertexts are not linked to their full potential, and thus scholars fail to benefit from the full advantage hypertext offers.

6.4 Survey

An informal survey was conducted to identify some key research questions scholars make. Four professors at the IAM (Intelligence, Agents, Multimedia) Group at the University of Southampton were asked about the questions they would make in two typical research activities.

The professors were asked to comment on the following two questions:

- What are the main question(s) that you would want your new Ph.D. student to find the answers to, so that they became proficient within their new field?
- When reviewing a paper for a journal or conference, what are the sorts of question(s) you ask yourself?

The questions were aimed at capturing the essence of two common research activities: becoming familiar with a new field and refereeing papers. The answers provided for the first question are tabulated in Table 6.1. The dominant questions refer to determining the significant work and issues in a research area. One professor

Professor	Comments
Pr1	What are the canonical papers and Web sites and how can these be improved?
Pr2	Who are my counterparts? What are the respective contributions of the various disciplines to the shared problem?
Pr3	What is the leading edge? What are the research issues? What problem are you trying to solve?
Pr4	Identify the core disciplines on which the thesis will be based and be familiar with the literature of those disciplines. Do not assume that the only work worth referencing has happened over the last five years. Can you envisage what would be the novel contribution to knowledge of your thesis in three years time?

Table 6.1: Answers to survey question 1

Professor	Comments
Pr1	What is novel and significant?
Pr2	Is there sufficient theoretical elaboration? Is the problem raised/solved sufficiently important and original? Does it take into account the relevant current empirical and theoretical literature on the subject?
Pr3	Is it in scope? Does it present a new result and does it position itself with respect to other work in the field? Has it been presented elsewhere?
Pr4	Is it informed about the current work in the area? Are they clear about what has been achieved in their work and why this is an advance? Are they aware the strengths and limitations of the work? Do they have a credible plan beyond the work presented?

Table 6.2: Answers to survey question 2

noted that some of the significant work was *not* identifiable through a high citation impact alone.

Table 6.2 outlines the answers to question 2. In this case, the dominant task was to establish the research contributions evident in the paper. Most professors also noted the need to position the work with respect to the rest of the research field and to determine whether it noted other relevant empirical and theoretical work.

These answers were used to further understand the types of questions scholars make while participating in research activities. They are referred to at various points in this thesis as they provide a useful insight from experienced researchers as

to the type and level of support scholars would demand from a support environment. The experiment described in this chapter evaluated the extend to which the Web currently supports these types of questions.

6.5 Hypothesis

The hypothesis is that the Web lacks support for scholars in three predominant areas:

- Coherent facilities do not exist to enable scholars to effectively and efficiently locate and navigate scholarly information.
- Solving common analytical type queries is either difficult and time consuming or impossible.
- Scholars do not make effective use of current Web research tools.

These predictions are based on personal experience, discussions with fellow researchers, and the outcome of background studies discussed above.

6.6 The Experiment

The experiment required subjects to answer research queries using the Web. No restriction was set on how the Web could be used. While participants used the Web, all their actions, answers, and comments were logged. The log file for each participant was then analysed to determine the type of material gathered and the duration and number of unique resources used per question. The instructions handed to the participants are included in Appendix A.

6.6.1 Experiment Questions

The research questions used in the experiment were based on discussions with other researchers about the type of questions they make while conducting research. For example, reading seminal papers and becoming familiar with the prominent researchers in a field is important, especially for new students who must quickly become proficient within their field. Also, researchers are required to compare and contrast work by relating projects and the various work conducted at research teams, and by identifying collaborating researchers and their overlapping research.

The experiment was divided into two sections: task 1 and task 2. The questions are tabulated in Tables 6.3 and 6.4 respectively. Task 1 was aimed at discovering

Number	Question
1	Locate a noteworthy paper on the Ontobroker project
2	Was this paper ever presented at a conference, and if so, which one?
3	Are there any other related papers at this conference?
4	What projects are related/similar to Ontobroker?
5	Who are the researchers that are part of this project, and where do they work?
6	What other project has the institute produced?

Table 6.3: Questions for Task 1

Number	Question
1	Eugene Garfield is probably the most prominent researcher within the field of citation analysis. Find one of his seminal papers and explain why you believe this paper to be seminal?
2	Which institute participates in significant ontology research?
3	Broadly speaking, how has the perspective of hypertext changed over the last decade?

Table 6.4: Questions for Task 2

the ease at which specific resources could be located. These are typical questions scholars make as part of a broader task, such as investigating a series of related projects or perusing on-line conference proceedings. The questions were aimed at being completed through direct means without requiring a deeper understanding of a particular field. For example, question 1 required the participant to locate the Ontobroker project homepage and select a paper that had been published in a prominent conference or journal.

Task 2 posed analytical type questions aimed at forcing the user to draw from various scholarly materials rather than just acquiring an answer from a single source. These questions were intended not to be solvable simply by using a search engine, but rather a more encompassing approach was demanded. For example, to respond to question 1 several approaches were possible.

- Using a citation tool, such as the ResearchIndex, to list papers that cite papers by Garfield. A paper that is frequently cited is then an indicator for being seminal.
- Viewing several papers about citation analysis not authored by Garfield, and then determining if they frequently note Garfield's work and in which context (e.g. do they refute it, or agree with it). After several iterations, the significant papers become tangible.

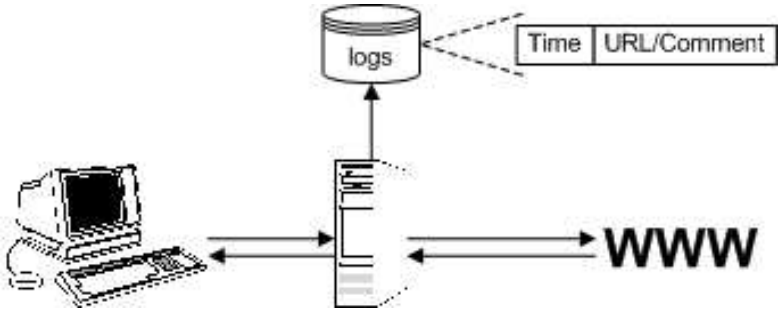


Figure 6.1: Experiment technical setup

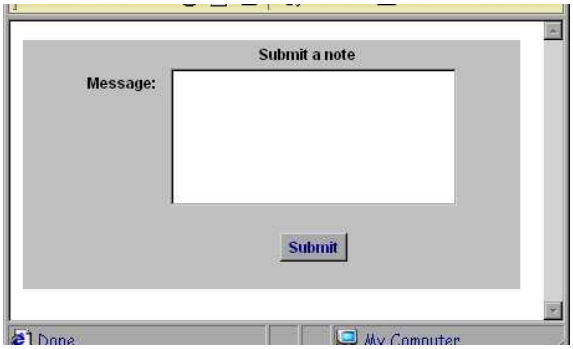


Figure 6.2: Client tool for experiment

- Visiting Garfield's homepage and identifying any papers that Garfield himself has positioned as significant. There may be a section in his bibliography entitled 'My Main Contributions'.
- Looking at several related projects, research teams, conferences, and journals that are active in citation analysis and determining if any make frequent reference to the same Garfield paper.

6.6.2 Hardware and Software

Participants were asked to use a Web browser to access any Web resource and to respond to each question in the order listed. Their computers were configured so that Web communication passed through a proxy[1] which recorded their actions (Figure 6.1). Subjects were also provided with a client tool (Figure 6.2), connected to the proxy server, which participants used to record their answers to the questions along with any additional comments. These were logged together with the web site URLs recorded through the proxy.

[1]A proxy is a server that acts as an intermediary between a workstation and the Internet.

6.6.3 User Group

Eight participants with varying levels of experience on the topics covered by the experiment were selected to take part. Participants were selected in order to obtain a diverse range of candidates with different levels of research experience. Therefore, three participants were second year and two third year postgraduate students reading for a Ph.D. The remaining three subjects were research assistants/fellows. All participants were from the IAM Group at the University of Southampton and were confident Web users.

6.6.4 Assumptions

The following assumptions and rules were made before the experiment.

- A question's start time is timed when the answer to the previous question is submitted.
- A task is deemed complete (and timed) at the point a Web document is visited immediately preceding the submission of a note with the answer.
- Web speed is constant throughout the tasks and between users' computers.
- The proxy does not inconsistently slow down the downloading of resources.
- Pages containing frames are treated as single pages. If a participant navigates within the frames, these count as new pages.
- Participants are encouraged to spend not more than 5 minutes on any one question.
- Browser specific navigation facilities (e.g. the back button) are not recorded.

6.7 Results

After the experiment, the log files of the eight participants were analysed and a qualitative and quantitative analysis was conducted. Due to the small sample set, statistical analysis was not conducted.

6.7.1 Qualitative Analysis

The data from the log files contain qualitative information in the form of user comments and the types of resources and tools they used in answering the questions. The analysis has been broken down according to the two tasks.

User	*Main* tools/sites used
P1	Google search engine, AIFB and KMI institutional sites
P2	Google search engine, AIFB institutional site, Webnet conference site
P3	Google search engine, AIFB institutional site
P4	Google search engine, AIFB institutional site, ResearchIndex
P5	Google search engine, AIFB institutional site, semanticweb.org portal
P6	Google search engine, AIFB and FSR institutional sites
P7	Google search engine, AIFB institutional site
P8	altavista and askjeeves search engines, AIFB institutional site, Kluwer online journals

Table 6.5: Primary methods used for solving Task 1

Task 1

Table 6.5 lists the primary tools and sites used by the participants in Task 1. Overwhelmingly, the task involved the use of a search engine (mainly Google) and the Ontobroker project home page (at the AIFB institute) as many questions revolved around this project. Most participants used this approach for questions 1, 4, 5, and 6. For question 4, the 'Semantic Web Community Portal' at `www.semanticweb.org` was used by five participants, although not to any great extent. In many cases question 2 could be answered with the information gained from question 1, as the conference title was indicated in the reference used in question 1. Question 3 involved participants making use of numerous tools such as digital libraries, conference sites, and the ResearchIndex, although again the use of search engines was evident.

Participants expressed uncertainty for questions 1 and 3 which could be due to the questions involving some subjectiveness (e.g. determining what a was noteworthy paper):

- 'wondering which I would call noteworthy'
- 'the SHOE project might be similar, but I am not 100%'

Participant 5 noted a preference for downloading the full-text of articles for question 4, and then analysing these in detail (in particular noting the citations) to arrive at a conclusion.

The Web sites participants visited were similar, suggesting that their search methods were comparable. Overwhelmingly, the tool of choice was Google meaning the coverage of results was largely the same. However, participant 4 used a citation

index for questions 1, 3, 4, and 5, and participant 1 used a digital library to resolve question 3.

Task 2

Overall, task 2 presented the participants with many problems. Uncertainty and difficulty was expressed for all questions, with some participants indicating that they would not normally use the Web for such a task, instead seeking to analyse papers and their references off-line.

- 'much more difficult'
- 'I can't decide which would be a seminal paper'
- 'I would not usually use the web for this'
- 'I can't answer this question in 5 minutes. I wouldn't question the web for this information'
- 'That would take a long time to answer'
- 'The second and third [were] much more difficult without a lot of reading and searching'
- 'task 2 is harder for me'
- 'my answers may not be correct'
- 'Doesn't Institute for Applied Informatics and Formal Description Methods do Ontology research?'

Table 6.6 lists the main tools and sites used by participants in completing this task. Although most participants again used the same methods, there was more diversity in their toolset than for task 1.

Question 1 was successfully attempted by most participants with the majority employing the same search method, namely Google, and visiting Garfield's on-line library[2]. Question 2 forced most participants to visit various research labs in an attempt to determine those with significant ontology research. These Web sites were located using Google. Only one participant managed to successfully complete question 3 by using a search engine, an online computer magazine, various institutional websites, and personal homepages of hypertext researchers. Three participants failed to attempt the question while two subjects tried briefly before conceding.

[2] `http://www.garfield.library.upenn.edu/`

User	*Main* tools/sites used
P1	Google search engine, The Scientist journal
P2	Google search engine, Garfield's personal pages and online library
P3	Google search engine, Garfield's online library
P4	Google search engine, IP-CNR institutional site
P5	Google search engine, Garfield's online library, ZDNet online computer magazine
P6	Google search engine, Garfield's online library, ISI site, AIFB institutional site
P7	Google search engine, Garfield's online library, AIFB institutional site, semanticweb.org portal
P8	altavista search engine, Garfield's online library

Table 6.6: Primary methods used for solving Task 2

User	Combined			Task 1						Task 2		
	T1+T2	T1	T2	Q1	Q2	Q3	Q4	Q5	Q6	Q1	Q2	Q3
P1	3:37	2:45	0:52	00:37	A	01:18	00:13	00:07	00:30	00:52	P	P
P2	17:19	14:46	2:33	00:28	06:03	X	02:27	05:48	A	02:33	P	X
P3	18:19	14:44	3:35	00:51	A	01:58	02:01	09:33	00:21	02:36	00:59	X
P4	16:47	1:26	15:21	05:21	A	05:45	05:18	08:33	00:29	07:31	03:59	03:51
P5	38:49	12:51	24:58	00:31	03:47	A	02:19	03:03	03:11	10:40	06:03	08:15
P6	20:09	13:09	7:00	00:07	A	03:25	06:43	02:44	00:10	05:40	00:40	00:40
P7	26:27	11:26	15:01	02:48	03:37	02:33	01:53	00:35	A	05:49	03:03	06:09
P8	15:37	3:29	12:08	01:58	A	X	00:10	00:38	00:43	09:08	03:00	X
Mean	26:13	12:56	13:17	1:35	1:40	2:29	2:38	3:52	0:40	5:36	2:57	4:43

Table 6.7: Duration of each question

6.7.2 Quantitative Analysis

Three tables of results were collated from the log files of the eight participants: durations per question, unique resources visited per question, and quality of results per question. Due to their considerable length, the log files have not been included in this thesis, although they are available at `http://www.ecs.soton.ac.uk/~srk/phd/experiment/log.html`.

Table 6.7 tabulates the duration of each question for all participants. An entry of 'X' indicates that the participant failed to attempt the question, 'P' indicates only prior knowledge was used (i.e. the participant did not require the Web), and an 'A' indicates that the participant was able to answer the question using information they had already acquired from answering a previous question. Entries marked with an 'X' or 'P' were *not* included when calculating the means.

Examination of Table 6.7 reveals that questions 1, 2, and 6 of task 1 demanded a comparatively short time to complete, whereas question 5 and questions 1 and 3 of task 2 required the most time. It is also evident that for each question participants required noticeably different durations to answer them (e.g. question 5 has a

User	Combined			Task 1						Task 2		
	T1+T2	T1	T2	Q1	Q2	Q3	Q4	Q5	Q6	Q1	Q2	Q3
P1	24	20	4	7	A	6	2	1	4	4	P	P
P2	33	27	6	7	9	X	1	10	A	6	P	X
P3	79	64	15	8	A	7	13	28	8	7	3	5
P4	77	49	28	9	A	13	11	14	2	12	9	7
P5	72	31	41	6	5	A	7	9	4	5	15	21
P6	67	45	22	6	A	7	16	13	3	14	2	6
P7	45	23	22	7	4	7	2	3	A	8	7	7
P8	51	30	21	16	A	X	4	5	5	15	6	X
Mean	62	37	25	8	2	7	7	10	3	9	7	9

Table 6.8: Table of unique resources visited by each participant

duration range of 7 seconds to 9:33 minutes). This is significant as the participants on the whole used an identical research method: a search engine.

Task 1 required on average 12:56 minutes to complete, while task 2, containing 50% fewer questions, required 13:17 minutes. From these durations, it is apparent that resolving the more analytical style questions on the Web took significantly longer. Overall, the tasks took participants a mean time of 26:13 minutes to complete. While this number is not overly large, there were only 9 questions (and 6 of these were basic questions from Task 1) to complete. A participant also remarked about the long duration, 'Task one was fairly straight forward, although it took me over 15 mins to find all the information'. Therefore, weaknesses are evident in using the Web to rapidly answer *these* types of questions.

Table 6.8 tabulates the number of unique resources visited by the participants for each task. Question 5 of task 1 and questions 1 and 3 of task 2 required the most unique resources. Question 2 from task 1 required the user to visit on average just two resources.

A mean of 62 unique resources was required to resolve the 9 questions. Task 1 required on average 37 resources, while task 2 required 25. Proportionally, task 2 required more resources, which is also indicative of the longer durations noticed for this task in Table 6.7.

Several Web pages supplied the answer to more than one question, therefore reducing the mean duration time and number of pages visited for some questions (as indicated with an 'A'). For example, the page located for question 1 (e.g. a bibliography listing for the Ontobroker project), also goes some way to providing an answer to question 2 (e.g. the conference is listed in the bibliographic entry). Indeed, this is what participants 1, 3, 4, 6, and 8 noticed. However, question 2 is

User	Combined			Task 1						Task 2		
	T1+T2	T1	T2	Q1	Q2	Q3	Q4	Q5	Q6	Q1	Q2	Q3
P1	1.7	1.8	1.3	2	2	2	1	2	2	1	2	1
P2	1.6	1.7	1.3	2	2	0	2	2	2	2	2	0
P3	1.6	1.7	1.3	2	2	0	2	2	2	2	2	0
P4	1.7	2.0	1.0	2	2	2	2	2	2	2	1	0
P5	1.7	1.7	1.7	2	2	1	1	2	2	2	1	2
P6	1.7	1.8	1.3	2	2	1	2	2	2	2	1	1
P7	1.7	1.8	1.3	2	2	1	2	2	2	2	2	0
P8	1.0	1.5	0.0	2	2	1	1	2	1	1	0	0
Mean	1.6	1.8	1.2	2.0	2.0	1.0	1.6	2.0	1.9	1.8	1.4	0.5

Table 6.9: Table of quality ratings for each question

unusual as the three participants who did not use previously acquired knowledge, did necessitate a notable 6:03, 3:47, and 3:37 minutes respectively to answer the question. A similar scenario is evident for question 5 and 6 and explains why questions 2 and 6 required the fewest numbers of resources to elicit an answer.

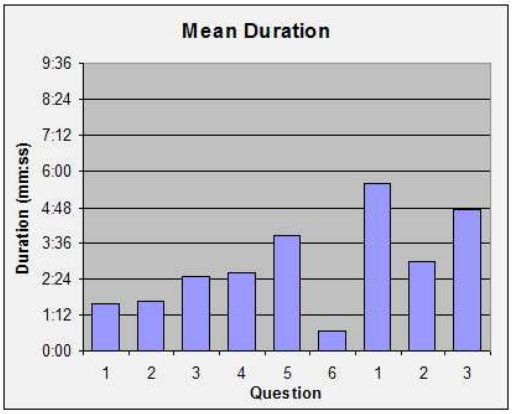
Upon further investigation, it was noticed that the results listed so far did not take into account the quality of the participants' responses. This is significant as some participants provided incorrect answers or even failed to attempt a question. Therefore, a simple grading scheme was devised with which to weigh the response to each question. It was vital that the grading scale was simple to minimize subjectivity. The scale used was:

- Did not provide any form of answer: 0
- Provided an incorrect answer: 1
- Provided an 'essentially' correct answer: 2

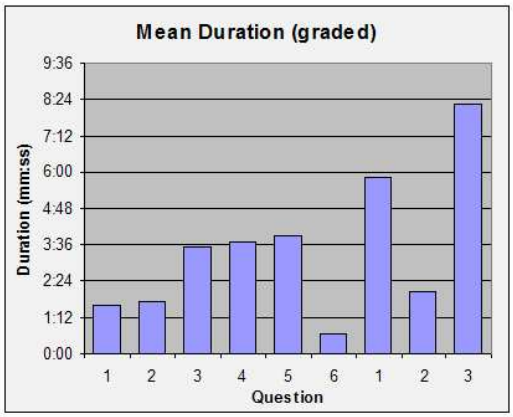
To receive a grading of 2, the participant only had to provide a partially correct response, as this experiment is not expressly concerned with the quality of responses, but rather the approach the participants used in answering the questions.

Table 6.9 tabulates the quality ratings assigned to each task. Questions 1, 2, 5, and 6 of task 1 and question 1 of task 2 had consistently high quality ratings. Question 3 of task 1 and questions 2 and 3 of tasks 2 had the lowest gradings. By excluding the results for the participants that used prior knowledge (i.e. the Web was not used to solve the questions), then the mean grading for question 2 of task 2 is reduced to 1.2, and for question 3 of task 2 to 0.4.

The advantage of grading the questions is highlighted by Figures 6.3 and 6.4 where ungraded and graded charts of the mean duration and mean number of

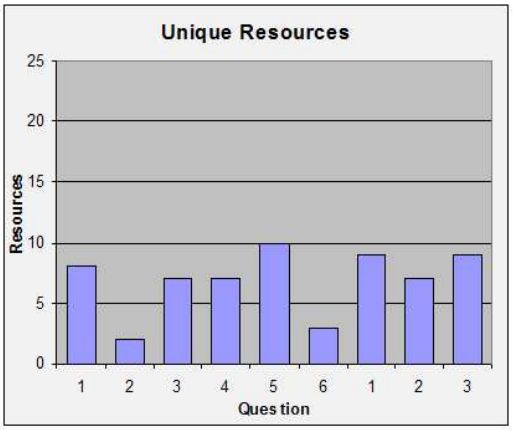


(a) Mean durations

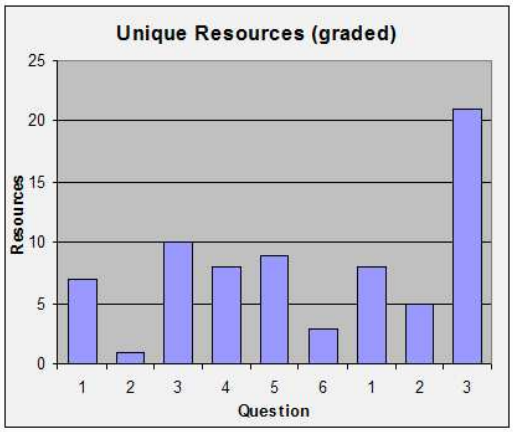


(b) Mean durations for a good response

Figure 6.3: Comparison of question durations with and without quality ratings



(a) Mean unique resources



(b) Mean unique resources for a good response

Figure 6.4: Comparison of number of resources with and without quality ratings

resources are presented. While the ungraded charts are evidence of total effort, they do not relate this effort to success.

In Figure 6.3a, question 5 of task 1 and questions 1 and 3 of task 2 have above average mean duration times. However, in Figure 6.3b where only responses with a weighting of 2 are included, only questions 1 and 3 of task 2 now stand out. For question 3, this can be explained by the fact that there was only *one* 'good' answer, which took the participant a considerable time to complete.

The difference between Figures 6.4a and 6.4b is even more marked. While the graph in Figure 6.4a failed to uncover any general patterns, Figure 6.4b indicates that question 3 of task 2 required a significantly larger number of resources. In addition, the number of resources viewed for questions 1 and 5 in task 1 have been reduced.

The quantitative analysis has indicated that difficulties in using the Web for research do exist. In Task 1, it took on average 2:30 minutes and approximately 6 resources to answer a question graded *good*, rising to 5:21 minutes and 11 resources in Task 2. For task 1, 79% of the answers were graded as *good*. However, only 46% of the answers in Task 2 were graded as *good*. If these percentages are then viewed in the context of the high duration and resource results, the difficulties are compounded.

6.8 Comment on Results

Several conclusions can be posited based on the outcome of the experiment.

1. Difficulties are apparent in using the Web as a research tool, either due to the research methods used by the participants or inadequacies in the Web. One participant appeared not to even be aware of citation indexes and commented 'I have to figure out a way to search for a paper that is cited a lot'.
2. Locating explicit resources (e.g. a paper on OntoBroker) as well as the resources implicitly related to it (e.g. the conference where it was presented) were usually possible, although they were time consuming and required the perusal of numerous resources. However, as was demonstrated with question 3 of task 1 (*Are there related projects presented at this conference?*) some less specific questions proved difficult.
3. Current scholarly tools on the Web do not appear to be extensively used, with participants instead opting for search engines.
4. Analytical questions that involved participants drawing information from multiple sources were time consuming and difficult.

The conclusions imply that using the Web as a general research tool (e.g. downloading papers, viewing project information, inspecting on-line proceedings) has its merits. This is confirmed by observing the many researchers that use the Web on a daily basis to assist them in their research. However, for tasks that are more intricate the Web appears to falter and scholars are forced to seek other means of reaching a conclusion.

For example, real difficulties were exposed in task 2 where the questions were more analytical and required a greater understanding of the resources involved.

Participants were expected to retrieve and understand numerous resources. Although answering question 1 was possible for most participants, the completion time, a mean of nearly 6 minutes, confirmed the considerable amount of effort required. This number might have been even higher if the participants had not been encouraged to restrict their effort on each question to five minutes.

Question 2 did not prove overly difficult for half of the participants, although this could be because the earlier questions required them to visit the homepages of several institutions that conducted ontology research.

Question 3 proved problematic for all but one of the participants, with many failing even to attempt a response quoting that they were unsure on how to answer it using the Web alone. As most users relied heavily on keyword-based search engines, it is unsurprising that difficulties were encountered for these types of questions where simple keywords do not suffice.

A further observation was that the search methods employed by participants were similar. Overwhelmingly they employed the same search engine and as a result, the responses to the questions were alike. This restricted the knowledge gathered, as anything not indexed by the search engine (in this case predominately Google) was ignored by the participants.

Related to this, it was disappointing to notice the practical exclusion of digital libraries, e-Journals, and other scholarly aids in a participant's repertoire of tools. This implies either that scholars are unaware of the tools available to them or that these tools do not provide the full level of support expected by them.

The experiment has limitations, which should be taken into consideration when interpreting the results. These are:

- No senior academics were involved. Senior researchers could have greater experience/awareness in using advanced Web-based research tools.
- All participants were within the computer science discipline, and a wider discipline range may have yielded different results. Researchers outside computer science are likely to have different research methods and some may be less acquainted with using the Web.
- A larger user group might have more accurately and confidently identified problem areas.

6.9 Summary

The experiment proved to be an important catalyst in this research as it suggested that e-Scholars were not completely supported in using the Web for research tasks. Resolving direct questions about research issues was largely possible although time consuming. However, ineffectiveness was discovered when participants used the Web to respond to more analytical questions about research issues, evident through the increased duration times and resource counts, and the lower quality gradings; the Web does not adequately support a scholar's higher-level cognitive processes.

Participants primarily used an information retrieval approach in answering the questions. There was limited evidence of hypertext navigation being employed as participants continuously reverted to a search engine, although this was probably due to the resources being spread across competing libraries and institutional Web sites.

The experiment also indicated that current research tools, such as citation indexes and digital libraries, are not used prominently by scholars. However, this may be unique to the participants selected for this experiment or be due to their lack of training in using such facilities. If more participants had employed these tools, the results of this experiment might have improved.

However, the Web *does* do a reasonable task in assisting researchers for certain tasks, a point made by several participants.

- 'Google is easily the best for paper searches'
- 'One of the advantages of using Google is that a search brings up the relevant homepage'
- '[What] I find [with] using Google is that a search for a particular word or name often brings up the relevant homepage'

Therefore, it seems that some of the restrictions of the Web are tolerated in return for the immense amount of information available. Search engines are also getting more effective at indexing and returning *relevant* target documents (Eastman, 1999).

The following three chapters present new methods, based in part on the results of this experiment, to assist scholars in using the Web for research. Chapter 7 introduces the principles of this approach, while their implementation in two different systems, OntoPortal and ESKIMO, are discussed in Chapters 8 and 9 respectively.

Chapter 7

Supporting Research in the Semantic Web

7.1 Introduction

In Chapters 5 and 6 scholarly activity and the support for this on the Web was explored and the difficulties highlighted; there are limitations in efficiently searching for scholarly material and resolving analytical questions.

The initial effort in providing scholarly support on the Web has been digitising and placing the literature on-line. E-Journals, e-prints, and digital libraries have provided the main platforms on which to accomplish this. While these repositories provide instant access to a wealth of information, comprehensive support to provide context to position these with respect to the rest of the literature and the scholarly domain is lacking. Inside each publication medium (as well as on the Web in general) the ability to locate scholarly information and make informed questions about it requires an efficient search engine and a scholar's intuition and detective skills.

Theodor Nelson stated that scholarly documents on the Web closely resemble their traditional paper-based counterparts, rather than taking more advantage of the potentials provided by the medium such as linking and transclusion (Nelson, 1999). A similar situation is evident with the scholarly support currently evident on the Web, as tools have yet to embrace fully the potentials of hypertext and knowledge technologies and provide scholars with extensive facilities not achievable in the paper-based world. The focus must shift from how we can publish scholarly literature on-line, to how we can *enhance* and *improve* access to it.

This chapter draws from the principles of hypertext and the Semantic Web, in particular the use of ontologies and machine-readable knowledge, to *introduce* a new approach to supporting scholarly research on the Web. This involves the formal representation of scholarly knowledge to provide a principled method of viewing academic material and analysing it to enable scholars to pose intricate questions about their research field.

7.2 Representing the Scholarly Community

As was discussed in Chapter 5, scholarly research is the sum of many activities and resources, such as: journals, conferences, projects, collaborations, peer review, papers, research teams, universities, seminars, researchers, committees, organisations, theses, ad-hoc discussions, debates. These artifacts form the body of knowledge that constitute the *scholarly community*, and examining and analysing this wealth of information in both traditional and digital mediums is an involved and complicated task as the facts and objects are disconnected from each other and the research that helped produce them.

For example, Table 7.1 compares the possible approaches to resolving four typical research queries using traditional methods and the Web. In the former instance, resolving the queries involves following references and discussions with peers. While peer interaction is essential for research, this slow communication channel is not always effective in providing the desired results in the shortest time. By using the Web, as was evident from the experiment detailed in Chapter 6, scholars resort to the use of search engines for the majority of tasks, and digital libraries and other scholarly services less frequently. Both approaches are not ideal, as they require a substantial effort and detective ability to determine the answers to even typical research questions.

The construction stages of a basic scholarly community model are illustrated in Figure 7.1. In this model, scholars read *literature* (A) and make references to further literature (B). The *Author* and its relation with literature is illustrated in (C). In (D) the conceptualisation is extended further by adding the *Project* and its relations. (E) and (F) add *Team*, *Journal*, and *Conference*. A complete scholarly ontology is presented in Section 9.2.

Query	Traditional Approach	Web Approach
What else has this author published?	Following references. Asking peers. Library search.	Visiting the authors homepage. Using a Web search engine. Visiting a digital library.
What are the projects based at this research group?	Following references. Contacting the research group.	Visiting the homepage of the research group. Using a Web search engine.
What information is there on this unpublished project?	Contacting the research group (if known). Asking peers.	Visiting the project homepage. Using a Web search engine.
Are there any papers on hypertext models at this conference?	Inspecting the titles of papers in the conference proceedings. Asking peers.	Searching the on-line proceedings or digital library.

Table 7.1: Contrasting approaches to resolving scholarly queries

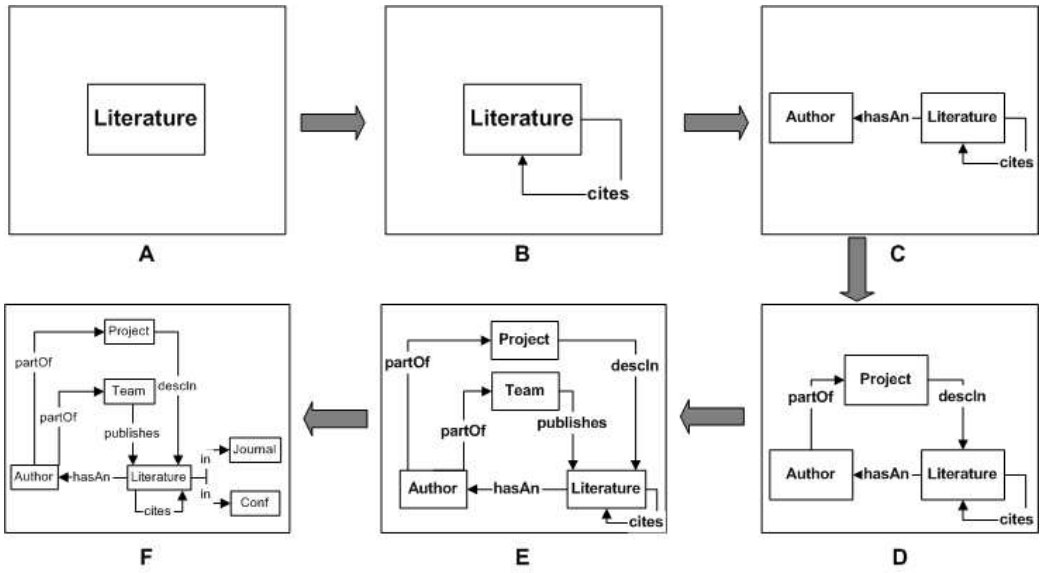


Figure 7.1: Constructing the scholarly community

The next two sections detail how a knowledge system can use the model of the scholarly community and metadata about scholarly resources to present two knowledge services: intelligent navigation and scholarly inquiry.

7.3 Ontological Hypertext

While navigation is sometimes claimed to be the most natural form of studying hypertext (Whalley, 1990) and enables scholars to explore an entire area and *home*

in on relevant material, providing scholars with an abundance of interrelated knowledge on the Web in an effective way is difficult as extensive linking and poor placement of hyperlinks can lead to information and cognitive overload. This predicament has been observed in hypertext systems (Conklin, 1987; Cockburn & Jones, 1996) and in scholarly hypertexts (Baragar, 1995; Theng, 1999), and is particularly applicable to the primitive hypertext features available on the Web (Bieber *et al.*, 1997a). After traversing several links, users do not know where they are, how they got there, how to get back, and what the current context is; a problem often referred to as "lost in hyperspace". "Just as large software programs with many patches can turn into 'spaghetti' code, so a hypertext can turn into a morass of meaningless, obscure connections and references" (Fiderio, 1988). Therefore, there is little point in providing extensive scholarly knowledge on the Web, if problems of poorly designed hypertext mean the e-Scholar becomes easily disoriented or has to discover associations through other means.

An information retrieval approach could be adopted (e.g. a scholarly search engine) with the ontological representation enabling structured and accurate querying. For example, if the search query is 'knowledge acquisition', an ontology modelling this domain could be consulted to discover that the query term is a subtype of 'knowledge management' and this term could then be used to expand the search. However, Eastman (1999) analysed popular search engines, such as Alta Vista, Lycos, and Excite, and although she concluded that they functioned well, she noted that users preferred simple and short queries over complex ones provided by advanced search engines as they required less user effort.

Furthermore, querying is not always a natural task, especially if users are unsure what to query for, or worse, are unaware of what can be asked (Bechhofer & Horrocks, 2000). The cognitive thought process is also disrupted every time the user has to suspend their thought task and issue a query. Therefore, Bechhofer *et al.* (1999b) propose an approach where the query entry interface is ontology driven. An ontology is used to help users refine their query based on the hierarchical structures evident in the ontology, for example by specialising it.

Therefore, the approach adopted in this research is to extend and use the principles and benefits of hypertext link semantics and abstraction, and promote the

scholarly ontology to the hypertext layer to act as an advanced conceptual template. This results in a richly interlinked scholarly knowledge environment that enables researchers to extensively explore their research field and directly answer questions about it. The approach is underpinned with a model of the scholarly community, which provides the framework to organise and examine the material.

In this approach, scholarly resources are identified as instances of ontological concepts and relationships between them (as modelled by the ontology) are established. This enables resources within a domain to be presented and associated in a principled and consistent manner that is based on their real-life representation in the ontology. Relationships between concepts not specified in the ontology are not possible, which enforces a consistent view on the underlying knowledge. The resulting combination of hypertext and ontologies has been termed *ontological hypertext* (Kampa *et al.*, 2001b; Kampa *et al.*, 2001a) and in effect uses "hypertext [to] mimic the brain's ability to store and retrieve information by referential links for quick and intuitive access" (Fiderio, 1988).

Ontological hypertext interconnects resources to provide scholars with a comprehensive view on their domain organised in a way to maximise contextual awareness and reduce disorientation; researchers are always aware of what to expect when selecting a link and how to get back. Vannevar Bush (1945) also recognised the limitations in just using indexes to organise information and instead proposed a means of associating ideas and building webs of trails. This notion has been expanded by Beeman *et al.* (1987) who compares hypertext to an adventure game which enables users to "wander through a world of facts and ideas" and by Baird and Percival (1989) who state, perhaps naively, that "information can be put into the hypertext in the same structure as it is in the author's head".

By constraining scholarly hypertext through an ontology, a principled, consistent, and intuitive hypertext is constructed that goes some way to reducing problems of disorientation and information overload. Naturally, the conceptualisation of the ontology is paramount: it has to make sense to users by accurately modelling the domain it describes. For example, there is no point in having a relationship between an author and a conference as there is no obvious association between these concepts.

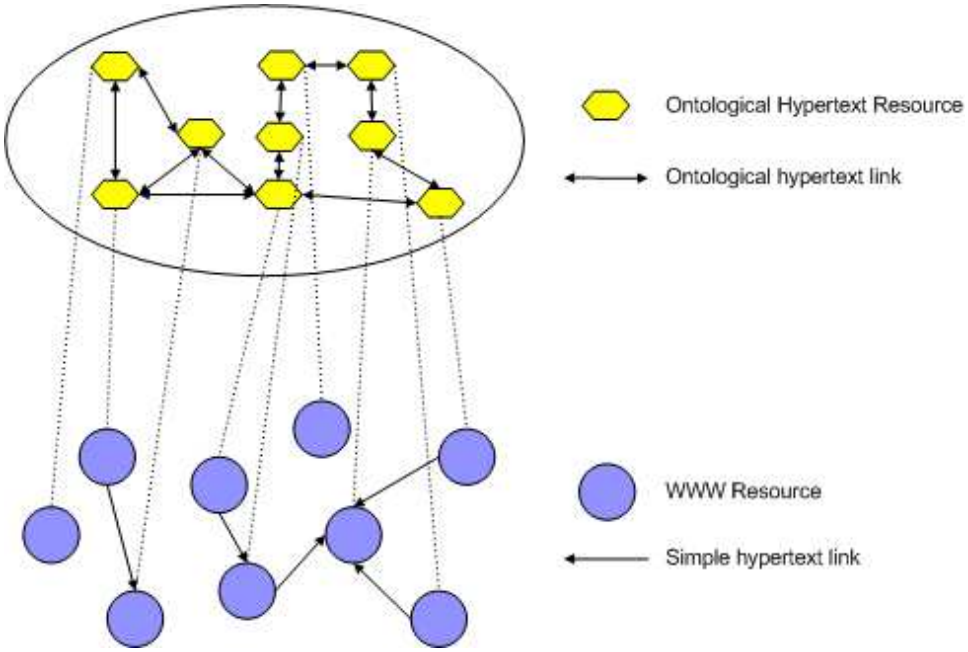


Figure 7.2: Ontological hypertext as a meta-layer over underlying Web resources

Ontological hypertext also allows the nodes and links between them to be abstracted from the content to which they apply, meaning it is unnecessary to modify original resources (c.f. Ontobroker (Fensel *et al.*, 1998)) or dynamically add hypertext links to actual documents (c.f. Microcosm (Davis *et al.*, 1993), DLS (Carr *et al.*, 1995)). Instead, metadata about instances is collected and an ontologically linked meta-layer is projected over the underlying resources (Figure 7.2). (Although instances usually refer to an underlying resource, they can be created without this association.) The approach therefore does not require access rights to resources to modify them with linking data, nor does it require a computational process to determine where in the underlying resource to dynamically insert a link.

The task of collecting this ontological metadata and identifying the concepts and relationships can be a manual (c.f. concept annotation in COHSE (Carr *et al.*, 2002)) or automated (c.f. metadata scraping in AKT (Alani *et al.*, 2002)) process and these different approaches are used in Chapters 8 and 9 where they are applied to a research portal application and a scholarly support system respectively. In both cases a declarative style of linking is possible, as promoted by Bieber *et al.* (1997b), where *what* is to be linked is specified as opposed to *how* it should be linked.

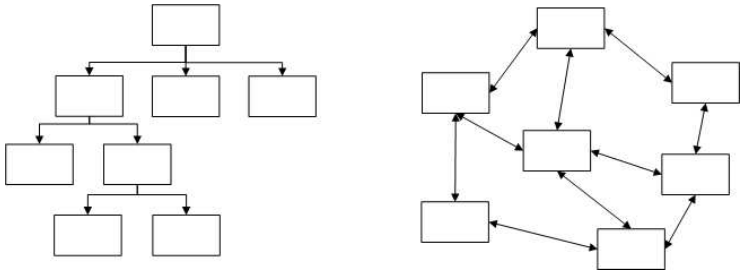


Figure 7.3: Hierarchical hypertext vs. Ontological hypertext

7.3.1 Intelligent Navigation

A system supporting ontological hypertext presents intelligently interlinked resources to researchers within the context of the surrounding ontological concepts. For instance, using the ontology defined in Figure 7.1, a literature resource is accompanied by (intensional) links to all related authors, research teams, projects, conferences, and journals. Any of these may have multiple destinations, as a literature, for instance, may have several authors. Although links are indicated as directional in the ontology conceptualisation, the ontological hypertext system may choose to present the relationships in either direction (e.g. 'literature has authors' and 'author writes literature'). Figure 7.3 illustrates the difference between documents organised using a conventional hierarchical approach where most links are based on local structural relationships and an ontological hypertext where links are determined based on the real-life model exposed in the ontology.

For example, a scholar may be viewing information on the paper 'The Dexter Hypertext Reference Model' and the ontological hypertext system provides a link to information on its author, 'Frank Halasz'. This page then provides links to the papers he has published, the projects he works on (e.g. 'Aquanet') and the research team he is a member of (e.g. 'Xerox Palo Alto Research Center'). By following the link to Aquanet, it is possible to quickly determine the other members of the project (e.g. 'Catherine C. Marshall' and 'Russell A. Rogers') and the papers that have been published about it (e.g. 'Aquanet: A Hypertext Tool to Hold Your Knowledge in Place')

This ability to navigate intelligently between resources enables scholars to quickly grasp the context of a particular scholarly object and understand how it relates to

the rest of their research field. It also encourages scholars to be serendipitous and explore paths and associations that they might otherwise have not considered. Rather than simply locating and reading a document, researchers ask questions about how and why the paper affects the research field.

Several studies point to difficulties in using hypertext networks to support learning. Gordon and Lewis (1992) compared the use of different hypertext structures (linear, hierarchy, network) for two types of tasks: factual questions and problem solving. For both tasks, the linear arrangement performed well. A hierarchical representation performed well for the problem-solving task, while the network arrangement did significantly worse in both cases. McDonald and Stevenson (1998) conducted a similar experiment where they compared the performace of participants using different hypertext structures for a learning task. The structures were: hierarchical, nonlinear (network of referential links), and mixed (hierarchical and referential). As with the earlier experiment, performance suffered with the nonlinear structure. However, the mixed arrangement performed best. Both experiments point to problems of just using nonlinear hypertext structures for learning material, and suggest the use of hierarchies to reduce the potential for disorientation. Ontological hypertext supports hierarchies, as well as the construction of consistent and intuitive hypertext structures. Furthermore, ontological hypertext (in particular the research environment produced in ESKIMO) is aimed at providing an *exploration* environment and not a learning tool.

7.3.2 Query-by-linking

Ontological hypertext enables particular questions to be resolved through linking by exploiting the relationships evident in the ontology. This *query-by-linking* principle is best demonstrated using three hypothetical queries a scholar may pose.

1. What else has this author published?
2. What other papers are published at the same conference as this paper?
3. What other papers discuss this project?

Resolving these queries in a digital library would require an effective search engine. However, with a properly constructed ontological hypertext, this is simply a matter of navigating the links between instances of the ontological concepts. In the ontology illustrated in Figure 7.1, there is an explicit relationship between the

literature and author concepts. Therefore, for every literature, there is a hypertext link to each of its authors. Following the link to an author then provides information about that author, including information on related material, such as links to each of the papers the author has published. This solves (1). Similarly, as there is an explicit link between a literature concept and a conference and journal, it is possible to determine where a literature was published and the other papers published there (2). (3) is similar to the first query. Users follow the link from the literature to the project it describes, and are then presented with all the literature that discusses the project.

7.4 Scholarly Inquiry

An ontology provides an explicit understanding of a domain and facilitates analyses to uncover implicit information and make plausible suggestions based on patterns evident in the knowledge. The Semantic Web has further introduced the foundations on which ontological metadata can be used to provide these knowledge services.

In the previous two chapters the analytical tasks facing scholars in disseminating their research was made apparent. They ask intricate queries while conducting research, many of which are either impossible to answer using the Web, or require an extensive investigation of scholarly material.

This section discusses two methods for using the ontological scholarly knowledge to support research: reasoning and augmented bibliometrics.

7.4.1 Reasoning over Scholarly Material

Drawing on the knowledge made explicit through an ontology allows various levels of inference to be provided as is later demonstrated in ESKIMO (Kampa *et al.*, 2001a). An inference is defined as passing from one true proposition, statement, or judgment to another. For example, if A is type of B, and B is a type of C, then it is possible to *infer* that A is also a type of C. Inference is used in various techniques and three methods suitable for scholarly data are presented.

Reflexivity

The reflexivity of relations can be used to determine facts in the reverse direction to how they are stated in the ontology. For example, in the relation *hasAuthor* between the concepts *Researcher* and *Publication*, it is possible to reverse the relation. If

'*Europe (A History)*' is authored by '*Norman Davis*', it is possible to formulate that '*Norman Davis*' is the author of '*Europe (A History)*'.

Although on first inspection this may seem of little apparent use, if we extrapolate this example it is possible to obtain *all* the publications by *Norman Davis*, and not just the European history book, and thereby answer the question: 'What are all the books authored by Norman Davis?'

Further examples of reflexivity are listed below.

- All the publications in the 'Communications of the ACM'.
- All the researchers working on an activity.
- All the publication mediums where the University of Southampton is represented.

Deduction

It is also possible to use deduction, a form of logic whereby a conclusion is reached by logical consequence, to uncover facts that although already 'embedded' in the knowledge, are implicit. For example, if the papers a researcher has authored and the organisation that researcher is a member of are known facts, then it is possible to deduce all the papers produced by that organisation (i.e. by all of its members). Not only does this allow facts to be uncovered that the user might otherwise have been unaware of, but importantly, it also reduces the authoring overhead. In this case, manually specifying the relationships between publications and organisations is unnecessary as it can be automatically deduced. Indeed, this approach is used in Chapter 9 to reduce the authoring overhead when populating the ESKIMO system.

Deduction can be used to solve other typical scholarly inquests, such as:

- The organisations represented in a journal or at a conference.
- The activities or projects based at a research team.
- The journals and conferences where projects are published.

Abduction

Abduction is the logic of exploratory data analysis; stated differently, that of critical thinking. It is used to suggest a hypothesis based on patterns in data. The key factor is that abduction results in plausible answers, rather than facts deduced from

logical consequences. There may be several convincing patterns in data, but only the more probable ones are *abducted*.

Abduction is a powerful facility that is used with the scholarly knowledge, heuristics, and background knowledge of scholarly activities to identify observations, trends, and patterns about a research field. For example, a novice to a particular field initially requires the seminal papers, details of the experts, and information about major projects. As following citations between papers alone cannot always provide an accurate account of this level of knowledge, scholars are forced to locate the answers in a non-coherent and inconsistent way.

However, abduction could be used to discover the *likely* experts in 'spatial hypertext' using the following rules:

1. Collect instances of all hypertext researchers
2. Retain those who have published at least x papers on spatial hypertext (x is a threshold value)
3. Retain those whose spatial hypertext papers have been cited at least x times (x is a threshold value)
4. Retain those who work on spatial hypertext projects
5. Rank those who edit journals on spatial hypertext higher
6. Rank those who work on committees involved in setting up hypertext conferences higher

The heuristics appear reasonable and are likely to indicate experts. However, better heuristics for such queries could be researched through a thorough investigation of scholarly practices, but this is beyond the scope of this thesis.

Abduction is extensively used in ESKIMO to enable scholars to ask pertinent questions about the material in their research field. Other questions that could be resolved using this method are:

- What are the seminal papers in hypertext?
- Which projects in agent systems have had the most impact in the hypertext community?
- Which activities or projects collaborate?
- Who are the experts in knowledge management that have also published a seminal paper about hypertext link semantics?

7.4.2 Augmented Bibliometrics

Bibliometrics is a useful tool for scholarly research. However, while its advocates profess that it uncovers numerous useful facts and patterns, opponents point to the unclear quality issues of citations. Furthermore, citations only connect literature, and fail to identify relationships between other objects such as authors, projects, institutions, and research groups. As a participant in the survey by Bishop (1998) noted, "just following references could give too narrow an outlook on the field." A researcher's objective is to become proficient within a chosen field and more than the literature must to be explored in order to meet this task.

Ontological knowledge from the scholarly community can be drawn on to further inspect the relationships specified through citations and reduce the reliance on it as the sole mechanism for understanding scholarly material. For example, seminal papers are sometimes not cited at a level expected or only begin to be heavily cited after a considerable time, and therefore may not appear as significant in a citation analysis of the literature. However, by analysing a paper's authors (Are they experts?), the research team(s) where the author(s) work (Are they prominent?), and the project the paper describes (Is it significant?), further noteworthy papers can be identified and more accurate and rational statements made about them.

In addition, collaboration measures can be improved over the results obtained by just using co-authorship as the determining factor, by investigating the projects scholars participate in and which research teams they are members of. For instance, two scholars that work on several projects together, are based at the same research group, and are members of the same conference committee, are likely to have formed some degree of collaboration.

Collaboration measures can also be applied to other scholarly artifacts. For example, determining the projects that collaborate is likely to group projects by the research issues they tackle and thereby provide useful sources of information. Identifying collaborating teams and organisations would also supply further insight into academic data.

Furthermore, the rules previously used for citation data can be applied to other scholarly artifacts. For example, co-citation analysis does not need to be restricted to literature, but can be applied to discover frequently co-cited projects and research

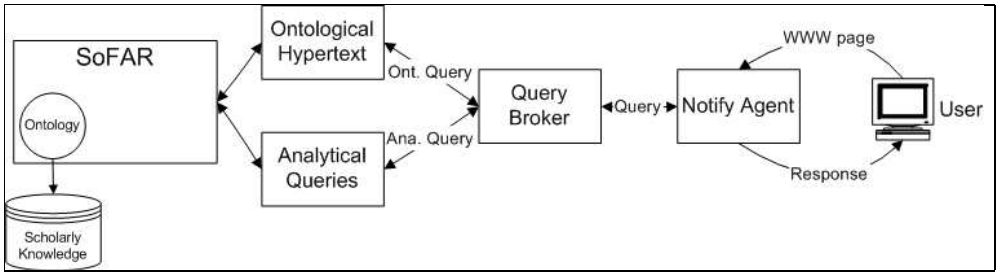


Figure 7.4: WSS architecture

teams. Indeed, investigations into author co-citation have demonstrated their feasibility and usefulness in identifying research specialities (Chen & Carr, 1999b) and as an interface tool for digital libraries (White *et al.*, 2000).

Impact factors are used to judge the significance of a journal, and are calculated by dividing the number of citations a journal's publications have received over a particular period, by the number of papers that were published during that period. This formula can also be applied to organisations, conferences, and projects, to help determine how active and significant they are.

7.5 Web Scholar System

The Web Scholar System (WSS) (Kampa & Carr, 2000) was an early prototype to explore and demonstrate the feasibility of an ontological hypertext approach. It used the Southampton Framework for Agent Research (SoFAR) (Moreau, 2000) and its support for basic ontologies, to provide the framework with which to rapidly create the prototype. An ontology similar to the one illustrated in Figure 7.1 was constructed and data from ACM Hypertext Conference 1997 was used to manually populate the system (i.e. add concept and relationship instances).

The architecture of WSS is illustrated in Figure 7.4. Users interacted with WSS by configuring their browser to use WSS as a proxy. This enabled the Web browser to communicate with the *Notify Agent*. This agent inspected the URLs of Web pages as users downloaded them to determine if it recognised any of the pages. For a recognised resource, such as a hypertext paper in the ACM Digital Library, the agent added a simple interface to the top of the page (Figure 7.5) which allowed users to find out more about the resource.

By selecting the 'More Information' link, WSS returned the knowledge for the current resource linked using ontological hypertext (Figure 7.6a). For example, after

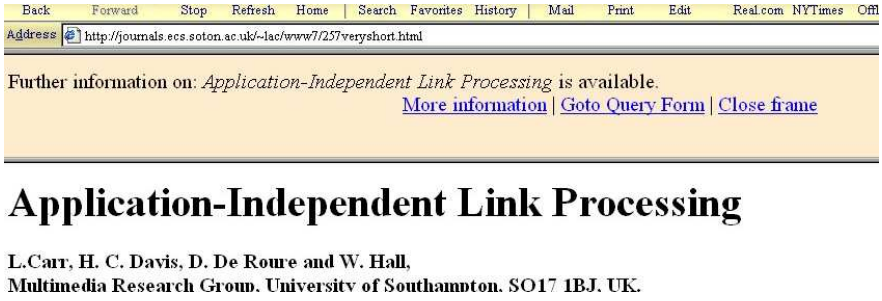
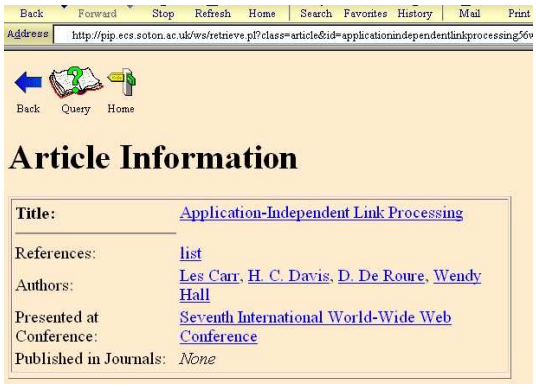
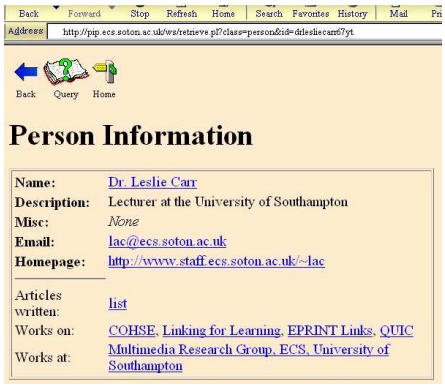


Figure 7.5: Interface into WSS



(a) Knowledge about an article



(b) Knowledge about a person

Figure 7.6: Ontological hypertext in WSS

following the first author listed in Figure 7.6a, 'Les Carr', the user was presented with information about this author together with the articles, projects, and research teams the researcher was associated with (Figure 7.6b). This is an example of the query-by-linking facility and is in contrast to the conventional query mechanisms available in digital libraries.

The 'Go to Query Form' link illustrated in Figure 7.6a provided a direct interface to the available scholarly material (Figure 7.7). Users retrieved knowledge about a known instance by providing a property or relation of it. For example, to retrieve the page on the researcher 'Les Carr', the class 'Person', and property 'Name' are selected, and the text 'Les Carr' is typed into the input field.

Finally, analysis of the scholarly knowledge was attempted. Four analytical queries, such as *Find seminal papers?*, were implemented and tested. However, the referential integrity problems in the metadata (e.g. multiple representations for the same researcher, such as 'Wendy Hall', 'W. Hall', 'Prof. Hall') made it difficult to achieve useful results.

Nevertheless, the WSS prototype demonstrated the advantages of representing

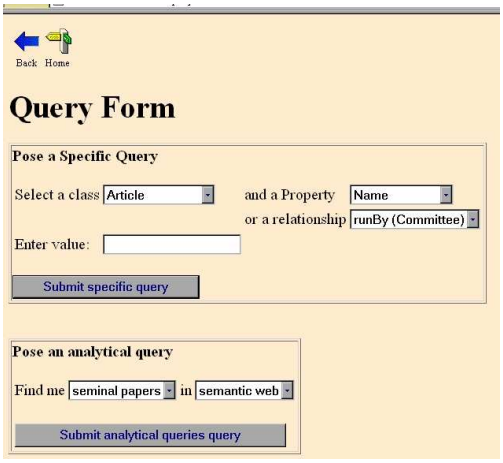


Figure 7.7: WSS Query window

the scholarly community, and in particular the use of ontological hypertext. Consequently, WSS prompted further research into this approach.

7.6 Related Work

This section discusses related research and the work that has influenced the principles discussed in this chapter, in particular the notion of using an underlying model of the domain to provide a richer information exploration environment.

7.6.1 Hypertext Semantics

Ontological hypertext is closely related to the link labelling in systems such as Notecards (Halasz *et al.*, 1987), gIBIS (Conklin & Begeman, 1989), and VIKI (Marshall *et al.*, 1994), and the explicit use of link semantics in systems such as Textnet (Trigg & Weiser, 1986), Aquanet (Marshall *et al.*, 1991), and MacWeb (Nanard & Nanard, 1993). However, these systems (i) only apply typing to links, (ii) have a pre-defined set of link types, and/or (iii) lack the rigour required for machine processing as they have been mainly designed for human use and visualisation and not for further processing and exploitation by machines. Therefore, they do not explicitly control and constrain the hypertext, or allow large hypertexts to be automatically constructed.

Furthermore, these systems do not usually consider the linking for an entire domain, as with ontological hypertext, but rather as localised typing between resources. These are usually simplistic (e.g. *is a*, *solves*, *addresses*, *contains info*)

or refer to structural relationships (e.g. *continued in*, *contains more info*, *indexed*) rather than a domain's *real-life* relations that are evident between resources. There is also no clear distinction between concepts and instances. Rather than create instantiations of concepts (e.g. the instance 'VW Golf' for the concept 'Vehicle'), new concepts are introduced for every instance, resulting in the underlying model being fine-grained and only representing a domain according to its individual instances and not its concepts. This does not result in an accurate model of the domain that machines can easily analyse.

In their defence, these systems were designed for structured authoring and annotation tasks, and to provide more effective navigation through link labelling; not for producing a rigorously controlled hypertext for organising very large collections of research material. Indeed, VIKI's and Aquanet's typing mechanism is flexible for precisely this reason. They do not require authors to construct explicit structures, but rather leave the structures implicit and then recognise them using structure finding algorithms when required.

7.6.2 Thoth-II

The ontological hypertext principle is an evolution of the semantic linking in Thoth-II (Collier, 1987). Thoth-II provided the facility to create a rich network of semantic relations between documents. Significantly, the domain being linked was modelled using a directed graph: a graph consisting of vertices and edges (links and nodes). In Thoth-II, nodes represented 'real world objects' and the relations between them were based on real relations, as opposed to structural aspects. Collier believed that hypertext should 'represent some part of the designer's conception of the topic'.

Users interacted with Thoth-II by using a browser that presented an interactive version of the directed labelled graph (nodes and links are labelled). Similar to ontological hypertext, browsing was simply a matter of traversing the network of inter-related nodes, although in Thoth-II the nodes and relationships were graphically visualised. In addition, similar to the semantic meta-layer proposed in this chapter, fragments of text or actual documents were associated with each Thoth-II node. When users selected a node, another browser was used to display the text associated with it.

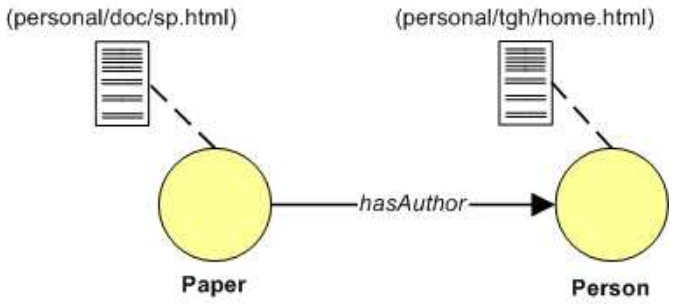


Figure 7.8: A Thoth-II network

The structure of a Thoth-II hypertext was defined in a single file using a Lisp-like notation. An example is presented below which defines the nodes *Paper* and *Person* and the *hasAuthor* link. A text file is also associated with each node.

```
(Paper hasAuthor Person)
(Paper description 'A paper from the ACM Hypertext Proceedings')
(Paper text-link 'personal/doc/sp.html')
(Person text-link 'personal/tgh/home.html')
```

Users browsed the resulting hypertext using the Thoth-II Spiders browser for displaying directed graphs. In this example the user would be viewing a graph based on the simple network illustrated in Figure 7.8.

However, there are several significant differences between Thoth-II and ontological hypertext, which make Thoth-II unsuitable for both larger applications and the Semantic Web. Thoth-II has a weak definition (model) of the underlying domain as nodes, relations, and the text associated with them (i.e. their instances) are defined and blended together. The lack of distinction between concepts and their instances make it impossible to create specific relations between instances, unless new nodes are introduced. However, this then changes the intentions of the original semantics.

In the example above, every text representing a paper and a person would be linked to the *Paper* and *Person* concepts respectively. However, this results in all papers being connected to all persons, which is not the desired effect. New paper and person concepts would be required for every additional instance, resulting in the original model being extended and thereby loosing the author's intended meaning. This limitation also produces a fine-grained model that perhaps caused Collier to notice that in Thoth-II 'it is hard to do the representation work correctly' and that 'it is easy to get lost'.

Thoth-II is also not suitable for implementation on the Semantic Web; the directed graph model, while sufficient for basic applications, is limited in its ability to model complex domains with multiple concept instances. Support for hierarchies, constraints, and properties is also lacking. Indeed, it appears that Thoth-II has been designed purely with navigation in mind, rather than enabling machines to further process and use the underlying model and its knowledge.

7.6.3 Ontobroker

The Ontobroker (Fensel *et al.*, 1998) system provides a framework to annotate Web documents with ontological metadata and a query service to access the knowledge. Although Ontobroker provides a hypertext interface to access the knowledge, this only acts as a dynamic query service into a knowledge base. For example, a hierarchical listing (i.e. index) is used to find the required instance. Ontobroker then lists the other instances it is related to, but there is no facility to then discover *further* information about them. To achieve this, the user has to return to the index listing and locate the particular instance. This does not provide a fluid or intuitive navigation experience.

The Ontobroker system has been used in the SEAL (SEmantic PortAL) (Maedche *et al.*, 2001) project where an ontology to model the domain of the research topics and administrative tasks at the AIFB Institute has been created. This ontology has been used to annotate Web pages at the institute to enable researchers to accurately access the resources they require (e.g. project, researcher, paper information) and semantic ranking and similarity measures are proposed to help position these results.

7.6.4 ConceptLab

Simpson's (2001) ConceptLab is a spatial hypertext system that uses link structures for authoring and exploration tasks. It is a research tool for organising a scholar's personally collected material, as opposed to ontological hypertext which is used to organise the material in a research field.

Figure 7.9 illustrates a screen shot of ConceptLab. Concepts are used to represent some notion that the author has defined and are depicted using different shapes. The concept type labels are in the centre of each large shape. The smaller shapes within theses denote further instances which are somehow related.

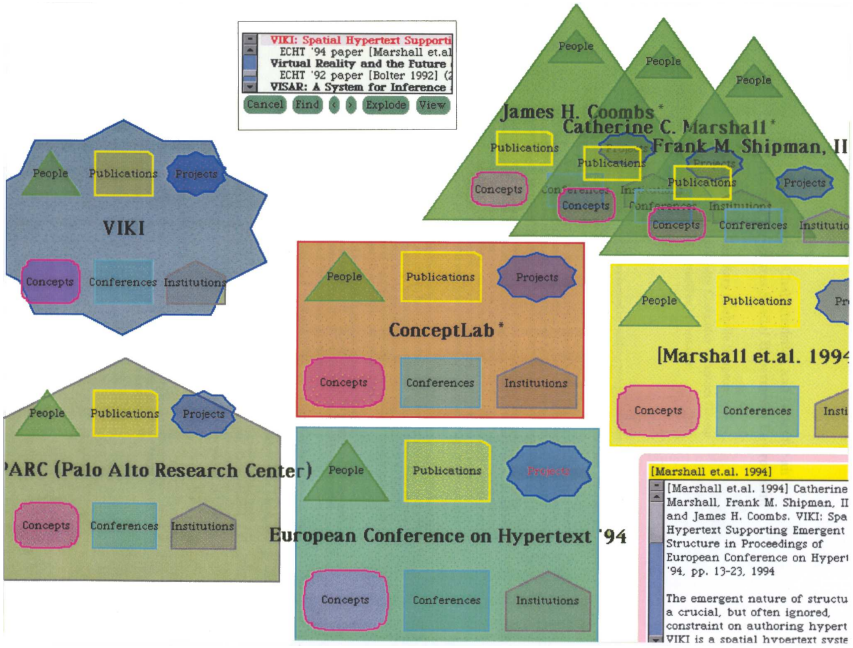


Figure 7.9: ConceptLab screen shot

ConceptLab is a useful tool for scholars to organise and arrange the information they are collecting for a specific task, although it obviously provides no assistance in locating this information in the first place. For example, when preparing a journal paper a scholar will want to discuss several projects and cite other papers from journals and conferences. ConceptLab allows this collected information to be organised for quick access and to provide a useful overview. However, for large collections this approach may prove cumbersome as only a simple hierarchical navigation style is supported. Furthermore, as few contextual cues are provided, navigating concepts in ConceptLab is likely to cause disorientation when large data sets are involved.

7.6.5 Web of Knowledge

The ISI Web of Knowledge (ISI, 2002) provides a highly customisable research tool to explore and discover research fields. It provides a general integration platform for ISI products, and in particular integrates two successful tools: Web of Science and Current Contents Connect. The Web of Science provides a Web interface to citation data from over 8,000 journals and 2,000 books. The Current Contents Connect is a current awareness resource that provides comprehensive access to bibliographic information, full-text articles, editorials, commentaries, and Web resources such preprints and funding sources. The search engine in the Web of Knowledge examines

multiple disciplines and filters are provided to restrict results by date, relevancy weighting, and source database.

A citation-based indicator tool and the citations in the journal citation reports are analysed to rank institutions, researchers, nations, and journals, although other scholarly services are not provided. Reference linking is added to allow scholars to traverse the literature, the benefits of which were explored in the OpCit project (Harnad & Carr, 2000). However, the reference linking in the Web of Knowledge is provided *outside* the documents they appear in. If users wish to follow a reference when they encounter it in the full-text, they must first locate the appropriate reference list in the Web of Knowledge, and then follow the supplied link to the paper.

However, the Web of Knowledge only provides one view on the academic data. It is almost entirely based on citations and ISI define the rules used to rank scholarly objects. Its failure to include and use other scholarly data, and instead relying on the unsure quality of citations, means its results must be observed with prudence.

7.6.6 PROPIE

The Proposed Information Environment (PROPIE) (Liew *et al.*, 2001b) enhances the interaction with electronic documents and delivers 'value-adding' services. By using a collection of workspaces and a direct manipulation mode of interface, researchers are able to quickly browse and analyse the electronic documents in their collection.

PROPIE provides four workspaces that scholars use to interact with their underlying information (e.g. electronic documents) and enhance their understanding of it. Each workspace behaves independently and serves a different function. The *InfoSphere Organiser* is used for researchers to organise their collection of documents; for instance, all those from a particular e-journal they subscribe to. Various methods of searching and visualising the collection are provided, such as map views, hierarchical views, and subject listings. When a particular item is located, it can be dragged into the *Object Viewer* workspace. This workspace enables the content of a document to be viewed and browsed. A selection can be made (e.g. a span of text, a diagram) and dragged into the *Object Explorer* workspace. This allows researchs to further analyse information objects. For example, the set of words in

a phrase can be used as the search text to search over a collection of documents. Finally, the *Structure Viewer/Organiser* provides an overview displays of an objects structure. If a document is selected, a series of thumbnail images of each page of the document are displayed. If all documents from a particular author are selected, then a timeline with each document added as an icon is illustrated.

Liew et al. (2001a) argue that it is the addition of scholarly services, such as those provided in PROPIE, that will eventually persuade scholars to abandon paper-based documents and use enhanced electronic papers instead. Indeed, empirical evaluation of PROPIE indicates that users are aware of these benefits and use them when exploring documents. The scholarly services introduced in this chapter also aim to demonstrate the advantages of moving to electronic material.

7.6.7 ScholOnto

ScholOnto (Shum *et al.*, 1999) models and captures the claims scholars make in papers. The resulting knowledge structure is analysed to provide a social perspective on the knowledge. As claims represent opinions and perspectives, rather than facts, different inferences are possible that capture the inter-social climate of a research field more accurately than factual assertions. For example:

- Has anyone challenged this publication?
- How has the perspective of digital libraries changed over the last few years?
- Do any papers build on this theory, but contradict each other's predications?
- Is there any software that tackles this problem?

While ScholOnto captures knowledge *within* scholarly papers, this research discusses how facts about and between the various scholarly objects can be used to provide a comprehensive view on a research field. However, by combining the factual knowledge of the scholarly community with the claims in the ScholOnto system a powerful application could be created that provides an insight into the facts about research *and* the issues and research opinions surrounding it. For example, the following questions could be posed:

- How have the perspectives of the experts in open hypermedia changed over the last ten years?

- What are the key claims made about link semantics in the seminal papers from the ACM Hypertext conference series?
- Is there a correlation between frequently co-cited authors and the claims they make?
- Do the researchers that collaborate with Tim Berners-Lee share his views on the Semantic Web?

7.7 Summary

This chapter has introduced a new approach to improving the Web for scholarly research by drawing on the Semantic Web and providing a principled method of navigating scholarly material and allowing scholars to ask pertinent research questions.

Nanard *et al.* (1991) posit that "user disorientation in hypertext is not due to the concept of hypertext itself, but rather generally results from the lack of a conceptual model for hypertext application." At the centre of this approach therefore, is the modelling of the scholarly community. Improved navigation is then afforded through ontological hypertext, which promotes an ontological representation to the hypertext layer to act as a conceptual template. As links are abstracted from the documents they connect, intensional hypertext links are calculated and used to interlink the scholarly resources and enable researchers to navigate their field comprehensively.

The advantage of this approach is twofold. Firstly, it forces authors of ontological hypertext to consider and appreciate the artifacts and relations in their research field. This provides a template and information gathering structure to create a knowledge repository accurately and comprehensively with all the necessary information and associations.

Secondly, researchers are presented with a highly organised, principled, consistent, and intuitive site that enables them to explore information based on real-life relationships rather than local or structural relations. These provide additional context to help reduce disorientation and cognitive overload.

In addition, the knowledge representation permits reasoning to answer scholarly questions. This allows scholars to inquire about their research field and discover

the salient facts and events. Traditional bibliometric rules can also be used and extended to assist scholars further in obtaining a wider impression of their field.

The WSS system was an early prototype to explore the ontological hypertext principle using real data from the ACM Hypertext conference series. The feasibility of analysing this formal body of knowledge was also evaluated, although data integrity difficulties prevented these from functioning correctly. However, this limitation is addressed in OntoPortal by providing manual authoring facilities to capture the ontological metadata and in ESKIMO by applying a semi-automatic approach to the data and knowledge acquisition tasks.

However, there is an increased initial authoring effort in producing ontological hypertext. By its very nature, ontology construction is a time consuming and laborious task. A standardised methodology has so far failed to emerge although a key lesson is that ontology construction is a highly iterative task and necessitates as much feedback and evaluation as possible. It is also likely that with the gradual adoption of the Semantic Web, more ontologies will be available for reuse.

The next chapter discusses the OntoPortal system, which has successfully demonstrated the ontological hypertext principle and applied it to real-world scenarios, including a commercial effort to produce a highly interlinked Web site detailing the latest research in metadata. The subsequent chapter then extends the work in OntoPortal to produce a scholarly support environment that supports ontological hypertext and scholarly inquiry.

Chapter 8

Scholarly Hypertext in the Semantic Web: OntoPortal

8.1 Introduction

OntoPortal evolved from the initial experimentation with WSS in providing scholars with a comprehensive semantic background to their research field. It is designed to create well-linked hypertext structures for navigating research material and is based on the ontological hypertext principle discussed in the previous chapter. It represents a demonstration of the principle on a large scale and in a real world scenario where its suitability for supporting scholarly activity could be further determined.

Ontological hypertext is aimed at improving the interlinking of complex and inter-related research material to provide a more effective navigation experience for scholars. It is constructed by analysing and exploiting the structures evident in ontological metadata; in OntoPortal this is *manually* authored.

As its name suggests, OntoPortal is a framework on which to build portal Web sites. A portal, similar to a Web directory, is a term used for a site that offers links to a significant amount of information within a particular domain. Portals are excellent starting points for finding information on a particular subject. Typically, a portal contains links to external resources rather than providing real content. Examples of popular portals are DMOZ, Yahoo, Excite, and The Computer Portal.com.

Chapter 2 discussed how poorly constructed hypertexts exhibit problems of user disorientation and information overload; a problem also evident in scholarly hypertexts (Baragar, 1995; Theng, 1999). This is confounded in research portals by

the fact that they provide many exit routes to external resources. Researchers may unknowingly leave the realms of the portal and thereby lose the support structures provided by it, such as linking, index pages, and tables of content. Therefore, OntoPortal uses the meta-layer approach introduced in the previous chapter to enable scholars to explore information *about* resources rather than being forced to visit the underlying resource directly.

OntoPortal demonstrates the representation and publication of ontological metadata — knowledge — in the Semantic Web and its objectives are to:

- Reduce disorientation and cognitive overload by promoting a formal model of the underlying domain to the hypertext level to provide contextual awareness.
- Reduce the effects of information overload by only presenting links to directly relevant information.

An overview of the OntoPortal system is presented in this chapter, followed by a description of four applications that were created using it.

Some of the work presented in this chapter is not solely that of the author. The reader's attention is therefore directed to the declaration in Section 1.4.

8.2 Overview

OntoPortal originated through a contract between the IAM Group and the Defence Evaluation and Research Agency (DERA), UK. DERA approached the IAM group with an initial proposal for creating a richly interlinked research portal containing research material on metadata. The project proposal stated:

> The aim of this project is to capture and summarise the current state of metadata research and present the results in a framework of Web pages that is coherently interlinked and incrementally updateable by DERA researchers.

The requirement was not to author content, but rather to provide a compilation of information and links to relevant resources on the Web. Nor were DERA researchers only interested in literature, but required information on standards, centres of excellence, experts, software, and projects. The requirement of tight interlinking and the potential size of the site, presented a formidable task. Certainly, manual creation of the resources and their interlinking was considered infeasible due

to budget and timescale constraints, and due to the increased risk of errors that a manually created hypertext presents (Furner *et al.*, 1999). Therefore, an automated system based on a systematic and methodical approach was required.

The experimental ontological hypertext system, WSS (Section 7.5), had already been successfully implemented and it was therefore decided to realise an ontological hypertext system on a larger scale and use it as a basis to create the portal.

The first DERA proposal was primarily a feasibility study and a prototype was delivered to demonstrate the concept and its potential to DERA researchers. Based on its success, a second contract was issued to create a fully functioning system and to employ expert domain authors to use OntoPortal to create a comprehensive research portal.

8.2.1 Features

The design goal of an ontological hypertext system is to implement a system to improve the navigation facilities available to researchers through links that reflect real-world relationships rather than structural/hierarchical relationships that are apparent in many current Web portals. The latter is evident in portals such as DMOZ and Yahoo, where resources are classified under topics (subject headings) and the topics are arranged in a hierarchical fashion. The topic 'Legal Ethics' in Yahoo is positioned under 'Law', which is under 'Government'.

In ontological hypertext, resources are 'classified' as *instances* of concepts in an ontology (either manually by an author or automatically through some categorising or theming process). Concepts represent real-life objects and not just arbitrary placeholders for collections of similar resources as in Hyper-G (Kappe *et al.*, 1993). The salient relationships are then identified between these instances and specified according to the structure evident in the ontology. Only concepts and relationships explicit in the ontology can be used. Relationships between ontological concepts also provide a natural link taxonomy from which the interface can determine presentation techniques for displaying different link types; the advantages of this link labelling were demonstrated in Notecards (Halasz *et al.*, 1987), gIBIS (Conklin & Begeman, 1989), and VIKI (Marshall *et al.*, 1994).

While ontological hypertext is used to control the interlinking between information, a graphical representation of the same ontology is also promoted to the

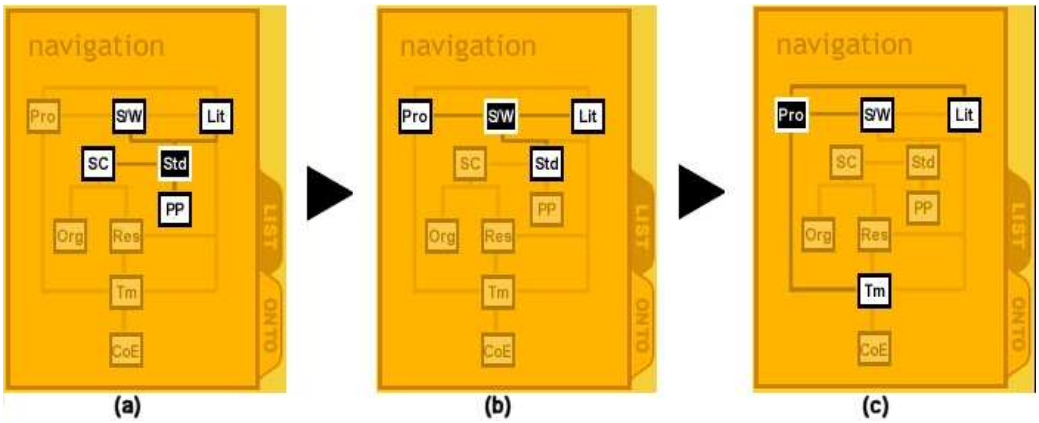


Figure 8.1: Context in OntoPortal

forefront of the OntoPortal interface as an exploration context tool. Users are then always aware of their position in relation to the rest of the portal through the highlighting of the current and related nodes in this interface. For example, the ontology illustrated in Figure 8.18 is presented to the OntoPortal user as a simplified version and adjusted as the user moves through the portal (Figure 8.1). As the user moves through the material available in the portal, the context diagram changes to reflect the new context. In Figure 8.1a a darkened node indicates that the user is viewing information on a standard. From the image it is possible to determine that a standard is related to software (e.g. software can implement a standard), literature (e.g. a paper can discuss a standard), standards committees (e.g. the committee that defines a standard), and promotion projects (e.g. a project that promotes the use of a standard). When the user moves from the standards page to a page about software that it is related to, the context is reflected in the image (Figure 8.1b). On moving from the standards page to a related project, the image again reflects the new context (Figure 8.1c).

OntoPortal also provides a threaded discussion and editorial commentary (in the form of opinions and analyses) facility to enable a discussion to ensue on any individual instance. An opinion is used to add a personal judgment, while an analysis allows for a more detailed explanation. These mechanisms allow authors to impart their editorial knowledge and experience on a resource.

As a large subject area, such as metadata research, has several sub-fields that users may wish to explore independently, OntoPortal introduces the *theme* idea to provide this vertical partitioning of the knowledge into speciality areas (or topics).

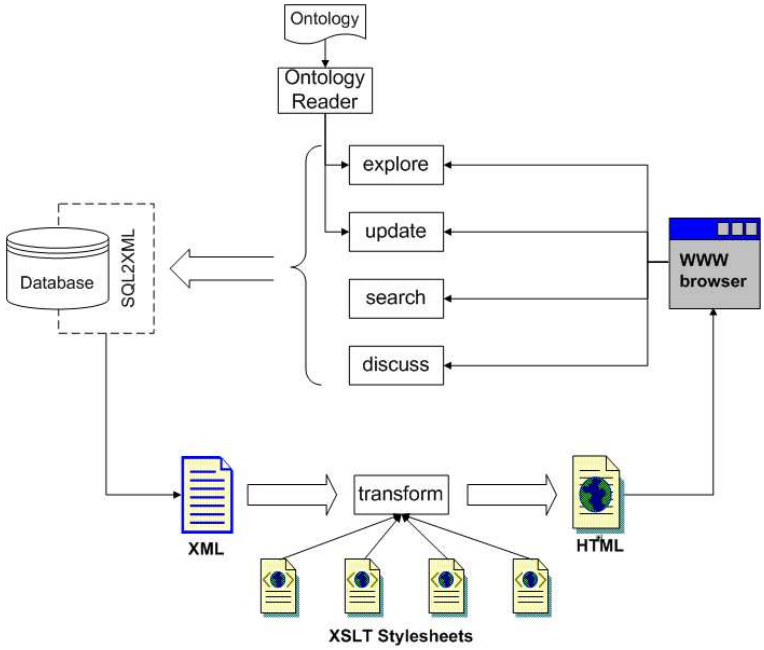


Figure 8.2: OntoPortal Architecture

When using OntoPortal, users first select the theme they are interested in, and then explore the inter-related instances within that theme.

Authoring facilities in OntoPortal enable domain experts to identify and specify concepts and relationships (i.e. manually author ontological metadata). Alternatively, in ESKIMO a semi-automatic population process is employed.

8.2.2 Architecture

This section examines the overall architecture of the OntoPortal system which has been designed to provide a *generic* framework for creating ontological hypertext portals (Figure 8.2). Users interact with the OntoPortal system through four transparent interfaces (Explore, Update, Search, and Discuss) using a Web browser.

The *Explore* interface allows the user to browse the ontologically linked resources in the OntoPortal knowledge base. When users supply authoring credentials through this interface, this causes the displayed resources to be decorated with an additional link pointing into the *Update* interface which enables resources to be edited.

Editorial credentials also results in the *Explore* interface providing links from each resource into the *Discuss* interface, allowing the user to browse and participate in any threaded discussion. The *Explore* interface also contains an entry point into

the *Search* interface which is used to query the resources stored in OntoPortal for specified terms.

The *Explore* and *Update* interfaces use the *Ontology Reader* module to access the underlying database as an ontology based knowledge repository. The module reads an ontology definition file that defines the ontology for a particular application, and uses this when communicating with the database. This ensures that the database always accurately reflects the ontology.

Each interface then translates user requests into appropriate SQL[1] statements and executes these against the database using the *SQL2XML* module. This module retrieves data from, and inserts data into the database and returns the result as an XML document. For example, an XML fragment constructed for a person instance is listed below. The instance has a single relationship to a 'tutorial' instance.

```
<?xml version="1.0"?>
  <ontoportal>
    <class type="person">
      <id>18</id>
      <title>Aaron Weiss</title>
      <email>aron@em.com</email>
      <url>http://wdvl.internet.com/WDVL/Authors/#Aaron</url>
      <short_description/>
      <relationships>
        <relationship related_class_type="tutorial">
        <id>63</id>
        <link>
          http://host/op/explore.cgi?class_type=tutorial&class=63
        </link>
        <title>
          XML via the Document Object Model: A Preliminary Course
        </title>
      </relationship>
  </class>
</ontoportal>
```

Before being returned to the user, the XML document is transformed into an HTML document through a series of eXtensible Stylesheet Language Transformations (XSLT) (W3C, 1999c) style sheets (a standard method to describe how to transform the structure of an XML document). For example, the following HTML is returned after the XML listed above has been transformed.

```
<H3>Aaron Weiss</H3>
e-Mail: <A HREF="mailto:aron@em.com">aron@em.com</A>
```

[1]The Structured Query Language (SQL) is a standard language for interacting with databases.

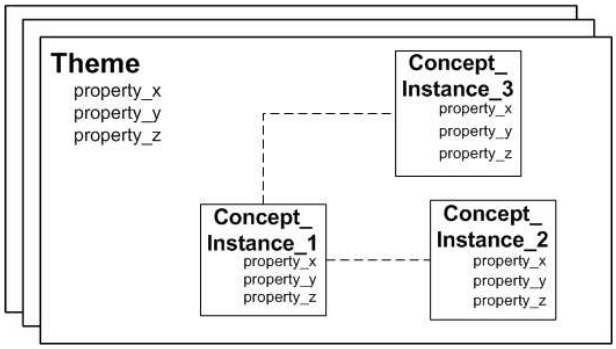


Figure 8.3: Organisation of OntoPortal information

```
<P>
  <A HREF="http://wdvl.internet.com/WDVL/Authors/#Aaron">
    Link to this resource
  </A>
</P>
<BR/>
<I><B>Tutorials</B> run by this expert</I>
<UL>
  <LI>
    <A HREF="http://host/op/explore.cgi?class_type=tutorial&class=63">
      XML via the Document Object Model: A Preliminary Course
    </A>
  </LI>
</UL>
```

OntoPortal's design results in an open framework that can be tailored to a particular application simply by defining a knowledge structure (ontology) and presentation rules (style sheets). Indeed, this flexibility has enabled the construction of four different OntoPortal applications. Full architectural details of OntoPortal are presented in (Carr *et al.*, 2001).

8.3 OntoPortal in Practice

8.3.1 Installation

An OntoPortal application is described in terms of the basic model depicted in Figure 8.3. The application is divided into a number of themes (if it is sufficiently large to warrant this) where each theme contains a number of concept instances which can be related to other instances as defined by the ontology. The ontology also defines the general properties (e.g. title, description, URL) associated with themes and concepts.

An example of a basic OntoPortal ontology, which uses an XML notation, is listed below. The concepts (classes) 'Person' and 'Paper' are introduced and a relationship specified between them. Properties and their types (e.g. text, integer) of the concepts and themes are also specified. An XML notation was chosen to represent the ontology because (i) OntoPortal ontologies were basic and could be represented without additional ontological constructs, (ii) at the time OntoPortal was implemented, the development and support of Web ontology languages was immature, and (iii) OntoPortal ontologies were not going to be used with other evaluation, visualisation, or analytical tools and therefore did not require compatibility.

```
<ontology name="authorship">
  <themes>
    <properties>
      <property name="title" type="TEXT"/>
      <property name="description" type="TEXT"/>
    </properties>
  </themes>
  <classes>
    <class type="person">
      <properties>
        <property name="name" type="TEXT"/>
      </properties>
      <relations>
        <related_class type="paper"/>
      </relations>
    </class>
    <class type="paper">
      <properties>
        <property name="title" type="TEXT"/>
      </properties>
      <relations>
        <related_class type="person"/>
      </relations>
    </class>
  </classes>
 </ontology>
```

The steps required to build an application using the OntoPortal framework are described below and illustrated in Figure 8.4:

1. Describe the ontology in XML based on the OntoPortal format.

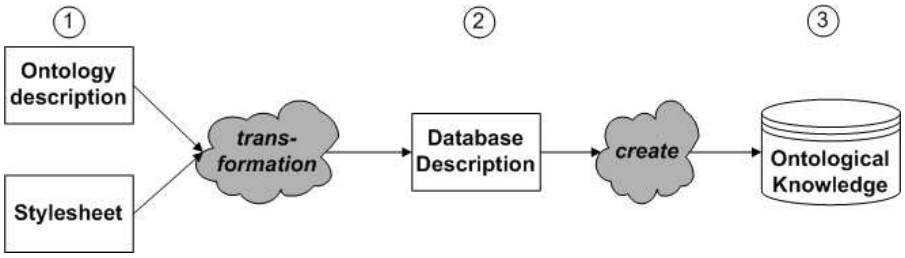


Figure 8.4: OntoPortal database construction

2. Apply an XSLT stylesheet to the ontology definition to transform the definition into an intermediate XML 'database description'. This 'database description' describes the database tables that need to be built in order to capture the knowledge structure made explicit by the ontology.
3. Translate the 'database description' into series of SQL statements which construct the tables in the database. Each concept in the ontology requires three tables in the database: one to record the general properties for each instance of that concept, one to record editorial information on each instance, and a third table to connect the instances to their respective themes. A relationships table is also added to specify the relationships between instances. However, individual relations are only ever added if they are also defined in the ontology (a condition enforced by ontological hypertext).

OntoPortal only supports basic ontological modelling; techniques such as subsumption (i.e. inheritance), constraints, and self-referential relationships are not supported. However, the model is still termed an ontology (as opposed to a taxonomy, schema, or vocabulary) as arbitrary relationships between concepts can be established to add a greater semantic value to the model.

8.3.2 Navigating and Exploring

Users begin using an OntoPortal application by selecting the theme they wish to explore (Figure 8.5). An introductory screen is presented for the selected theme (Figure 8.6) explaining its objective along with links to general overview material about the theme (e.g. FAQs, newsgroups, key documents). This document is independent of the ontological representation and is only part of the vertical partitioning available in OntoPortal. Ontologically linked resources then appear inside a theme.

Figure 8.5: OntoPortal themes

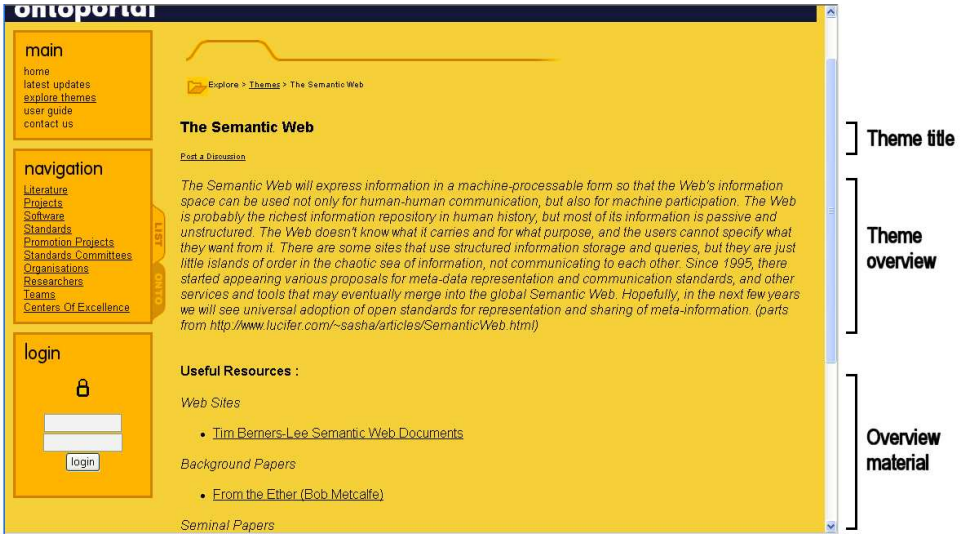


Figure 8.6: Theme introductory screen

Within each theme, users are able to browse the resources relevant to it. In Figure 8.7 the user is viewing all the literature instances in the 'Semantic Web' theme. Selecting one of the literature instances presents information on that instance and its relationships (Figure 8.8). Different labels are used to introduce each group of relationship links based on their type. These relationship types provide a natural link taxonomy from which the interface determines appropriate presentation techniques.

The literature discusses the RDF standard and as the OntoPortal author has specified a relationship between this instance and the RDF instance, OntoPortal

Figure 8.7: Trail 1/4 - Literature instances

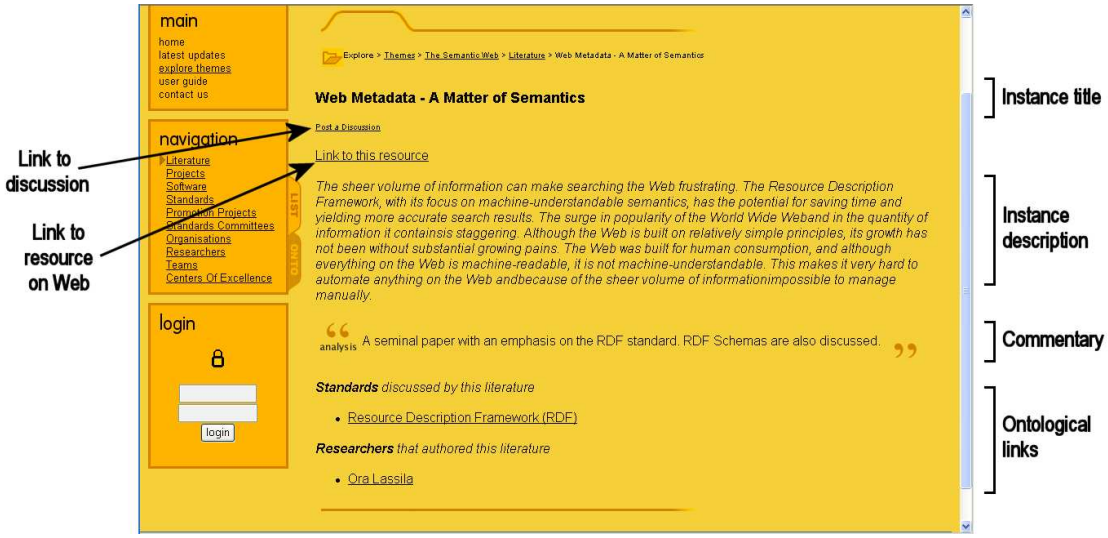


Figure 8.8: Trail 2/4 - Literature instance

Figure 8.9: Trail 3/4 - Standard instance

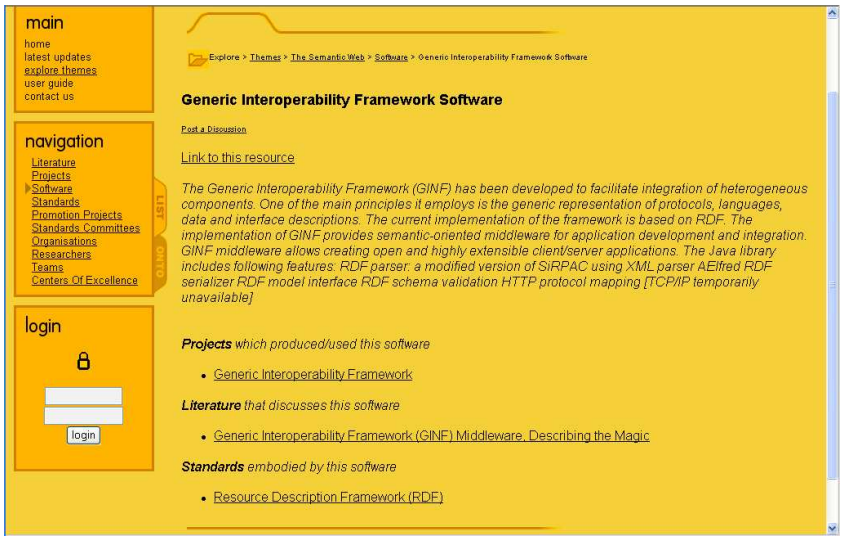


Figure 8.10: Trail 4/4 - Literature instance

adds a link to it. By traversing this link information about the RDF resource is presented (Figure 8.9). This provides a brief explanation of the standard, a link to its actual external resource (e.g. the specification), as well as the links to related instances (including the original literature instance). By continuing the exploration of the available knowledge, software that uses RDF is retrieved (Figure 8.10).

From this scenario, it is evident how a user browses 'around' an OntoPortal resource and discovers related artifacts. Starting with information on a paper, the RDF standard it discusses is viewed, followed by software which manipulates RDF. There would have been no way of discovering this software using the original paper alone, as there is no reference to the software in the paper. If the researcher

Figure 8.11: Searching in OntoPortal

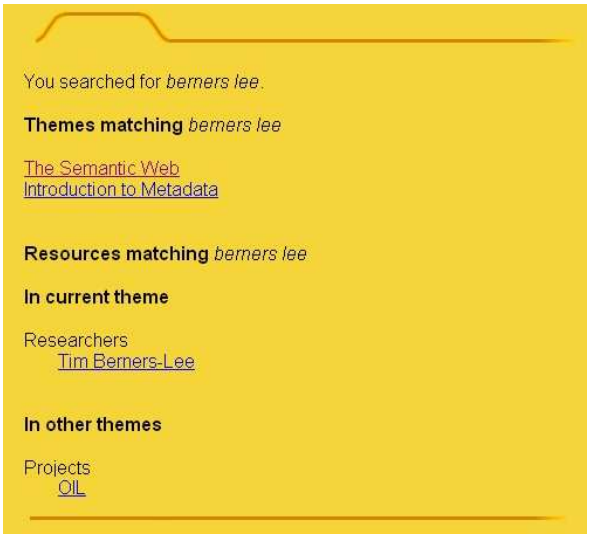


Figure 8.12: Search results in OntoPortal

reading the paper had asked *What software implements the standard discussed in this paper?* then query-by-linking could have been used to answer it efficiently. This is distinctly different to the hierarchical relationships evident in conventional portal designs, which often fail to provide links between related material.

Figure 8.11 illustrates the search mechanism that is added to the top of each page. The result of the query 'berners lee' is illustrated in Figure 8.12. Once the desired resource has been located, ontological hypertext can then be used to realise how it relates to other resources.

Registered users of OntoPortal have the ability to impart their knowledge by adding a discussion on any instance. When logged in, a link entitled *Post a Discussion* or *View discussions*, is displayed along with each instance which users can use to enter the discussion interface (Figure 8.13) and add to the discussion (Figure 8.14).

Figure 8.13: A discussion thread in OntoPortal

Figure 8.14: Adding a discussion in OntoPortal

8.3.3 Authoring

The inherent complexity of an ontological hypertext system requires an effective authoring component to aid the construction process by enforcing the ontological principles and (in effect) manually capturing the ontological metadata. When a domain expert acts as a hypertext writer to construct a highly interlinked Web site, a significant effort in the authoring and maintenance of hypertext links is incurred (Ellis *et al.*, 1996; Mendes *et al.*, 2001). The authoring tools therefore aim to reduce the technical aspects of creating links by forcing the author to consider OntoPortal resources as concepts and the links between them as relationships. The process of authoring then becomes that of identifying and creating concept instances and specifying relationships between them, rather than authoring documents and manually constructing the links and their respective anchors. Nanard *et al.* (1995) believe this approach better assists authors in structuring and organising their work at a cognitive level.

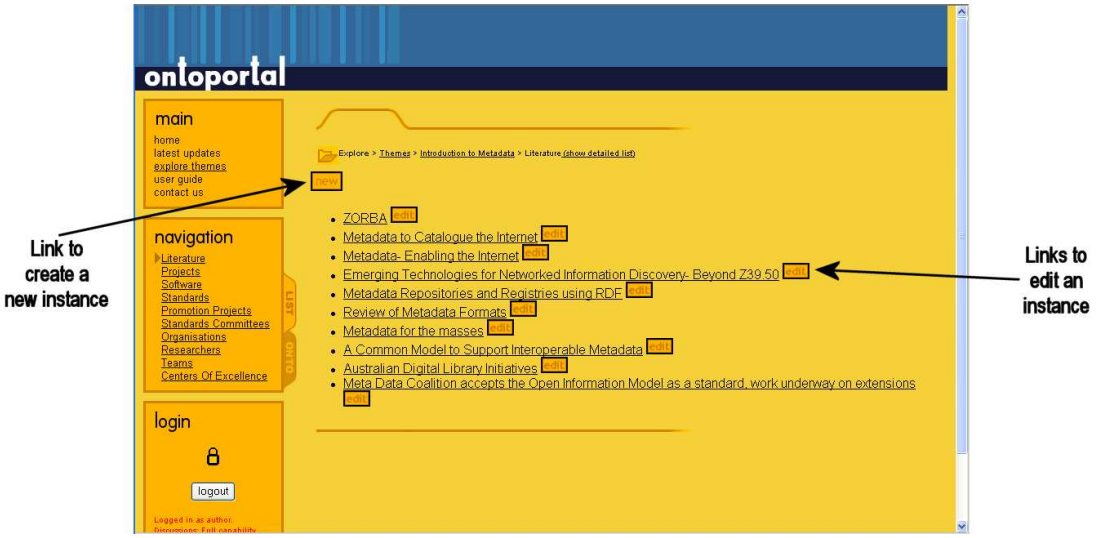


Figure 8.15: OntoPortal author mode

When logged on as an author, the usual navigation pages are annotated with buttons, which initiate creation and editing operations. If no instances exist, only a button to create a new instance is provided. Figure 8.15 illustrates the authoring mode for a list of literature instances.

The same authoring form is used to create as well as modify instances; the only difference being that edit forms will have many of the property fields already completed. Figure 8.16 displays an edit form for a literature instance. Authors also use the form to indicate how the instance is related to other instances. Only those that can be related, as defined by the ontology, are presented. In this example, the OntoPortal system recognises that a literature concept is related to teams, software, projects, standards, and researchers and therefore displays the available instances that the user can select to indicate a relationship. Once the data are specified, the form is returned to OntoPortal and captured as ontological metadata.

It is likely that by segmenting the knowledge into separate themes, some overlap exists (e.g. researchers may be active in several areas). To prevent the user creating duplicates and then dealing with the resulting maintenance problems, an existing instance can be 'imported' from another theme (Figure 8.17). Any changes made to its properties are reflected across all themes, while its relationships and editorials are theme dependent and remain unique. However, dealing with duplicates is the responsibility of the author. OntoPortal does not provide consistency checking or

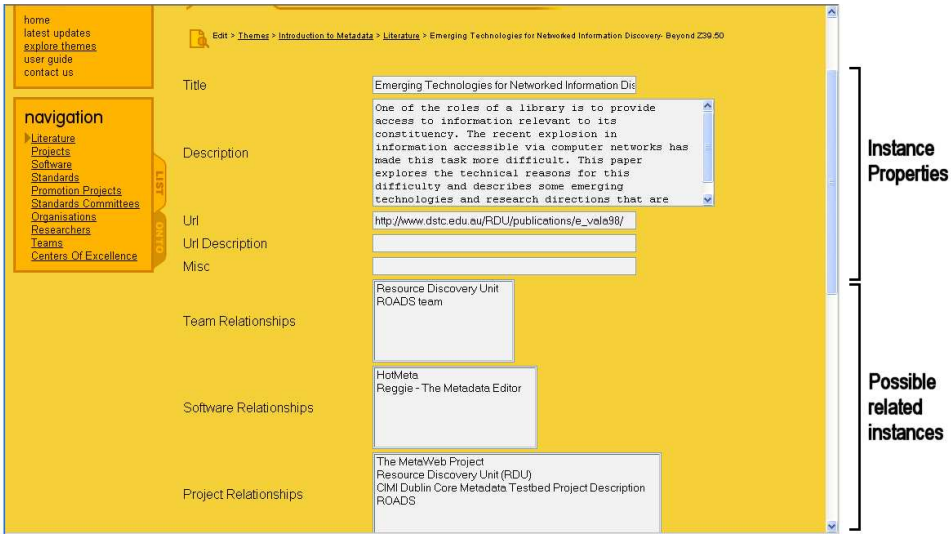


Figure 8.16: OntoPortal literature editing form

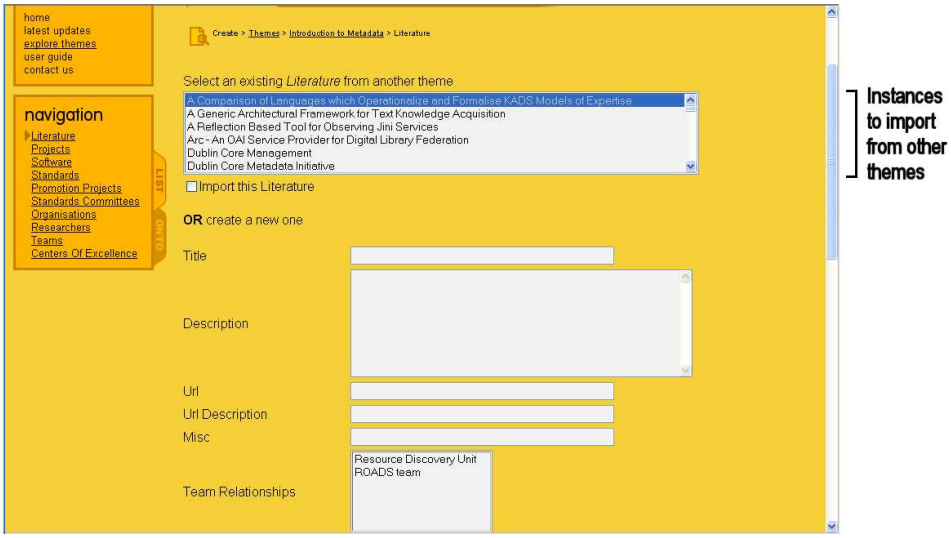


Figure 8.17: OntoPortal import facility

alert users when they are about to add an instance that already exists in another theme.

The authoring task therefore involves identifying concepts and their relationships in the material for which the portal provides information. While this requires an increased authoring overhead when compared to basic approaches of manually constructing scholarly hypertexts (e.g. in a digital library), creating the ontological hypertext manually is likely to be a significant and error prone undertaking, as past studies that evaluated the manual linking of complex sites and scholarly material have demonstrated (Baragar, 1995; Ellis *et al.*, 1996; Theng, 1999).

8.4 Applications of OntoPortal

OntoPortal's generic framework has been used to create four applications for different scenarios. Initially it was used to create the metadata research portal for DERA researchers. This proved to be a useful exercise in the context of this research as it enabled the ontological hypertext principle to be explored on a larger scale and in a real-world scenario. Three further portals have also been created with an emphasis on other activities of research.

The four applications of OntoPortal are:

- MetaPortal: Portal designed for DERA to provide information on metadata.
- TPortal: To support lecturers of computer programming by helping them discover Java teaching resources.
- XPortal: To help students studying XML to navigate relevant material.
- Icon Directory: A catalogue of icon meanings and uses.

The process of creating an ontology for each domain was left to the individual authors, who analysed the material they planned to use to determine the salient concepts and relationships.

8.4.1 MetaPortal

The conceptualisation of the ontology used in MetaPortal was constructed based on the requirements documentation received at the start of the project. As proposed by several ontology construction guidelines (Uschold & Gruninger, 1996; Gomez-Perez, 1996), the process started with an approximate conceptualisation and was then iteratively refined by the members of the OntoPortal team. The conceptualisation illustrated in Figure 8.18 was then derived.

The *Literature* concept is only one of the ten first-class concepts represented. *Literature* discusses *Software* and *Standard*s. It is published by a *Team*, authored by *Researcher*s, and may be the product of a *Project*. A *Team* is part of a *Centre of Excellence*. A *Standard* is defined by a *Standards Committee*, which may be part of an *Organisation* and consists of *Researcher*s. A *Standard* is promoted through a *Promotion Project*.

Experienced metadata authors were employed to use the authoring tools in OntoPortal to populate the MetaPortal application. Each author was in charge of

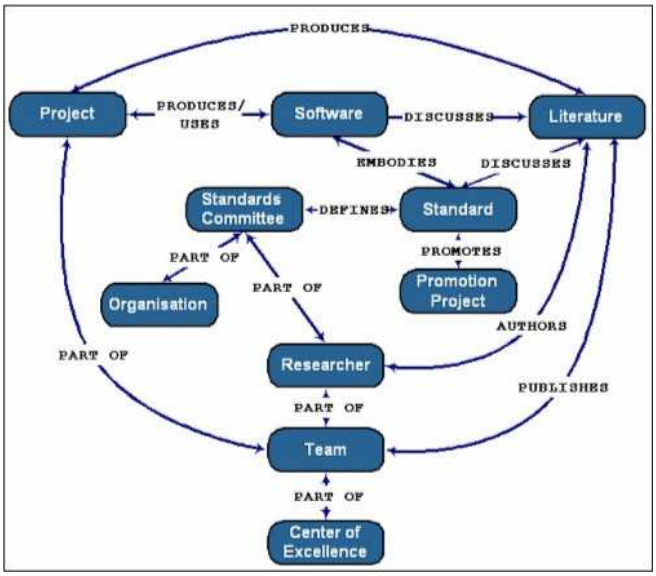


Figure 8.18: MetaPortal ontology

adding resources and providing commentary for one or two themes. The following themes were created and populated with a total of 461 instances and 641 relationships:

- Bibliographic Metadata for Digital Libraries
- Inferences and Metadata
- Introduction to Metadata
- Metadata Acquisition
- Metadata for Expertise Models
- Metadata Schemas
- Middleware and Metadata
- Ontologies and Metadata
- Reasoning and Knowledge
- The Semantic Web

A *static* snapshot of the resulting MetaPortal site is available at `http://www.ontoportal.org.uk/snapshot/explore/themes/`.

8.4.2 TPortal and XPortal

The Teaching Portal (TPortal) and the XML Portal (XPortal) were created by other users within the IAM Group and were not aimed specifically at scholarly research. TPortal is a didactic tool that supports lecturers teaching Java by providing a method to explore resources relating to a course. Its objective was to help lecturers

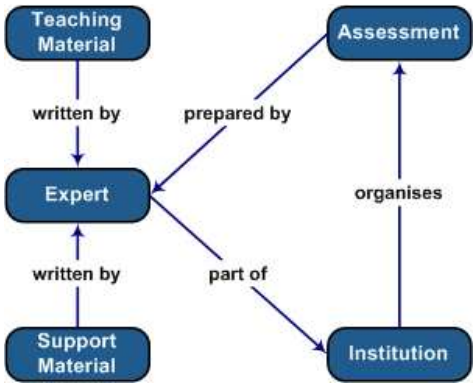


Figure 8.19: TPortal ontology

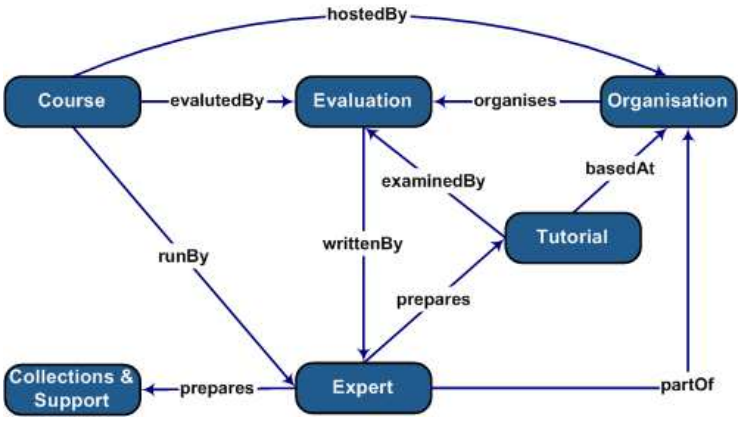


Figure 8.20: XPortal ontology

better understand the available teaching material. A simple ontology was used (Figure 8.19).

XPortal supports students studying XML by enabling them to retrieve course information. The portal assists students in finding all necessary information about any course in an intuitive and comprehensive manner. The XPortal ontology is illustrated in Figure 8.20.

8.4.3 Icon Directory

The Icon Directory is an informal project undertaken within the IAM Group and is based on producing a catalogue of *icons*: visual symbols used to convey a message such as a warning or purpose. The creator has a keen interest in icons found on various objects: hi-fi icons, packaging warning symbols, washing symbols.

Icons may be similar or based on other icons, they may convey the same message, or be used on the same objects (e.g. different washing symbols on articles of

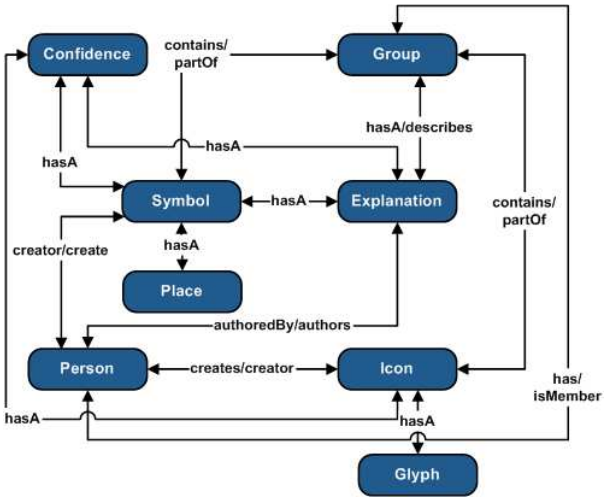


Figure 8.21: Icon Directory ontology

clothing). The creator wanted a method of defining and then exploring these relationships to enable a comprehensive Web site to be constructed that offered a novel way of navigating around an icon library. OntoPortal was used and the following ontological concepts identified (Figure 8.21).

- Icon: The conceptual representation of an icon. No physical characteristics.
- Symbol: A representation denoting some message. This will relate to (several) icons that are used to represent this symbol (e.g. the fragile symbol).
- Group: A collection of symbols (e.g. washing symbols).
- Glyph: The physical representation of an icon (e.g. a GIF or SVG image).
- Place: A location where a symbol appears (e.g. on a video recorder).
- Person: A researcher or author.
- Explanation: A description or clarification.
- Confidence/Provenance: Adds authority to concepts such as symbols. For example, the relative proportions of the stripes on the French flag are not equal and may be labelled as 33%, 37%, and 30%. The authority might be from the French government site, in which case most users will believe the proportions. However, if the percentages had been extracted from an Internet newsgroup, users may be less inclined to believe it.

The work for the Icon Directory is at an elementary stage with several different ontological representations being tested to determine which one delivers the best

browsing experience. The portal is being viewed as one of the methods the creator plans to use to provide access to his icon database.

8.5 Commentary

OntoPortal organises information using an ontological model, unlike existing portals, which usually use a hierarchical structure. For example, Yahoo classifies resources into various topics. To view links on 'electrochemistry', the user drills down from 'Science', then 'Chemistry', and then selects the sub-topic 'electrochemistry'. While this may be an effective search mechanism in circumstances where the user is specific in their search requirement, it is inappropriate when the user is unaware of the relevant category to search or simply wishes to browse between various resources.

Existing research portals also provide no further context or information on the resources being viewed. For example, when perusing 'electrochemistry' resources the user is not made aware of similar resources on electricity and chemical reactions. Additionally, if the user is interested in the material provided by a particular person or organisation, and wishes to view other material provided by them, then the user must use other means than those provided by the portal to find the related information. As OntoPortal also contains information *about* resource (metadata), users do not leave the realms of the portal as soon as they explore a resource. If users then choose to visit the actual resource once they have reviewed the collected information on it, they follow the provided link to the actual resource, which is opened in a separate browser window.

OntoPortal also serves as a useful adjunct to authors (and users). It forces them to focus on the structure of the domain they are involved in. Rather than simply viewing a paper as an isolated document, authors ask themselves questions such as 'What else has the author of this paper written?', 'Was this paper presented at a conference?', 'Are there any other papers that discuss the same project?', 'Has the university where this author works published any similar articles?' This helps authors become familiar with the field and provides a pre-planned information strategy.

However, OntoPortal applications suffer from two problems that are also evident on most Web portals. A large collection of instances can quickly accumulate for a

concept making it potentially laborious to locate particular instances as users are forced to work through long lists. Further sub-dividing a subject area using themes or creating finer grained concepts may alleviate the problem, although this could result in more navigational steps.

Secondly, while OntoPortal provides a mechanism to import concepts from other themes within the same OntoPortal application to reduce the possibility of duplicates, this is entirely the responsibility of the author. The system does not conduct any consistency checking within and across themes so duplicates can still appear. This then causes instances to be incompletely linked, as the links are likely to be divided between the duplicate instances, leading to user confusion.

8.6 Improvements

Through the four applications of OntoPortal, the following improvements have been identified as the key changes that would benefit it.

Adaptive Hypertext: Using the extensibility of OntoPortal and XSLT to produce a personalised view of the data on a per user basis. Currently, the personalisation applies to the authentication status of the user (i.e. not logged in, logged in to participate in discussions, and logged in to author). However, personalisation could be applied to the expertise of a user. This could lead to ways of reducing the number of proposed resources. For example, a novice is only presented with preparatory resources.

Further ontological support: The range of ontological structures supported is limited. Support for subsumption is desirable as this enables concepts to be specialised. For example, the concept 'Literature' can be specialised to 'Thesis' and 'Conference Publication'. If a user is interested in all literature, then the literature concept is selected. However, if only conference publications are of interest, then this can be specified by selecting the 'Conference Publication' concept. Furthermore, applying a blanket ontology across all themes is inappropriate in certain scenarios. For example, in introductory themes some of the concepts and relationships may be overly complex or inappropriate (e.g. 'promotion projects' in the introductory theme in the MetaPortal application). Therefore, authors may wish to apply different ontologies to the themes in an application to hide instances of particular concepts.

Generic stylesheet construction: Although the back-end processes of OntoPortal are generic, in the sense that an ontology can be plugged in to create a new application, the construction of the accompanying style sheets is a manual process requiring a significant time overhead. Naturally, this does have the advantage that the user can fully customise the look and feel of the OntoPortal application, and this ability should not be removed. However, it would also be useful to create a set of standard style sheets based on the application's ontology automatically.

Discussion moderation: OntoPortal only allows authorised users to post a discussion about a resource within the application. This process is unmoderated, relying on the authorised users to conduct appropriate discussions. This approach has two disadvantages: unregistered users are unable to participate and inappropriate discussions can arise. A moderated system on the other hand, solves both problems. Any user is allowed to post discussions and the possibility of inappropriate discussions appearing is eliminated. The downside is the time lag between a user posting a message and the moderator evaluating it and adding it to the site.

8.7 Summary

OntoPortal is an example of a Semantic Web application; it uses ontologies and machine-readable ontological metadata, combined with hypertext, to construct semantically interlinked research material. By having an explicit understanding of the concepts and their relationships evident in resources, OntoPortal is able to accurately control the linking between them and produce a hypertext that resembles a model of a particular domain.

XML is used as the underlying representation format for the ontologies and the authored metadata. However, extensions to OntoPortal that use the upper layers of the Semantic Web architecture, in particular logic support, would enable OntoPortal to reason over the metadata and uncover new information. For example, reasoning could be used in the MetaPortal application to determine which standards are described by a large number of papers and implemented in software, and then propose these as the significant ones that researchers should investigate.

OntoPortal was created using off-the-shelf components and widely used standards to explore and demonstrate the ontological hypertext principle on a large scale and in a real world scenario. It was used to create four portals based around

different ontologies and these demonstrated the possibility of improved interlinking that ontological hypertext affords to complex domains. Users cited the ease of being able to navigate around a particular resource as the main benefit while authors noted that they better understood their subject area as the knowledge structure (i.e. ontology) forced them think about the subject in new and useful ways.

For authors of research portals, OntoPortal shifts the emphasis from making explicit connections between parts of documents (and manually specifying the anchors), to creating an ontology and identifying instances and relationships (ontological metadata). The linking mechanism is then intensional (computed at delivery time) rather than extensional (authored in documents) resulting in more maintainable and flexible linking as modifications to the linking are not carried out in every affected document. The authoring process also serves as a useful adjunct by providing authors with a pre-planned information strategy.

For researchers, OntoPortal provides a principled, consistent, and highly interlinked site that assists them in exploring *all* objects in their research field and therefore becoming better informed. Ontological hypertext aims to reduce disorientation and cognitive overload because links are based on real-life relationships between artifacts in the application's domain, rather than structural links between parts of incongruous documents.

OntoPortal was successfully used to interlink material from complex domains and provided the motivation and confidence to further extend and apply the principle. It also provided further experience and training in constructing ontologies.

The e-Scholar Knowledge Inference Model (ESKIMO), influenced by the ontological hypertext mechanism in OntoPortal, provides a support environment for scholarly research and introduces reasoning services to enable scholarly inquiry. ESKIMO is the focus of the next chapter.

Chapter 9

Supporting e-Scholars with ESKIMO

9.1 Introduction

This chapter introduces a scholarly Semantic Web application, E-Scholar Knowledge Inference Model (ESKIMO), which demonstrates the principles of ontological hypertext and scholarly inquiry to provide comprehensive support for research.

ESKIMO is the third evolution of a scholarly support environment. Originally, WSS explored ontological hypertext and was subsequently extended in OntoPortal to provide a large-scale and robust system for creating research portals. ESKIMO builds on these and adds reasoning services.

The study described in Chapter 6 investigated how well the Web supports scholars in their research tasks; the results indicated problems of locating scholarly material and answering research questions. Therefore, the objective of ESKIMO was to address these issues and promote a new approach to supporting research on the Web.

This chapter begins by describing the scholarly ontology used in the ESKIMO system to model the domain of academic research. The 'ACM Conference on Hypertext and Hypermedia' series was selected to initially populate ESKIMO, this involved process is outlined, and then the architecture of ESKIMO is detailed. Then an overview of how the system is used and an evaluation study are described.

9.2 Scholarly community Ontology

The core of the ESKIMO system (as with any system that supports ontological hypertext) is the ontological model that drives the hypertext and the reasoning services. Chapter 7 introduced a simple scholarly ontology used in WSS and this section extends this model to create the scholarly community ontology and then describes how it is represented in a machine-readable form for use by ESKIMO.

9.2.1 Overview

Modelling a domain is a difficult and iterative task especially when many inter-related concepts are evident (Farquhar *et al.*, 1996). Overly complex ontologies consisting of many fine-grained concepts and multiple relationships can perform poorly and will inherently increase the maintenance overhead as the population and implementation process are intensified. Conversely, basic ontologies may fail to capture the essence of the domain and thereby prove unsuitable in their final application.

While individual research areas have their own characteristics (e.g. ancient manuscripts are important in the study of the classics, but have little significance in the field of telecommunications), these were not included in the interest of being generic. Therefore, the scholarly ontology represented a compromise between detail and simplicity that could be applied to a vast range of different research fields.

Purpose and Scope

Clarifying the purpose and scope of an ontology helps focus the construction process on the relevant concepts and relationships in a domain (Uschold, 1996). Therefore, the purpose of the scholarly ontology was to model the research domain so it could be used to assist scholars in using the Web for information discovery. The initial scope would be the hypertext research domain.

Noy *et al.* (2001) propose also using competency questions to determine the purpose of an ontology and help focus it on the nature of the task. These questions are also used when evaluating an ontology to determine whether it sufficiently captures a domain. Hypothetical questions for the scholarly ontology were identified through discussions with fellow researchers, the survey of professors in the IAM Group discussed in Section 6.4, and the experiment discussed in Chapter 6.

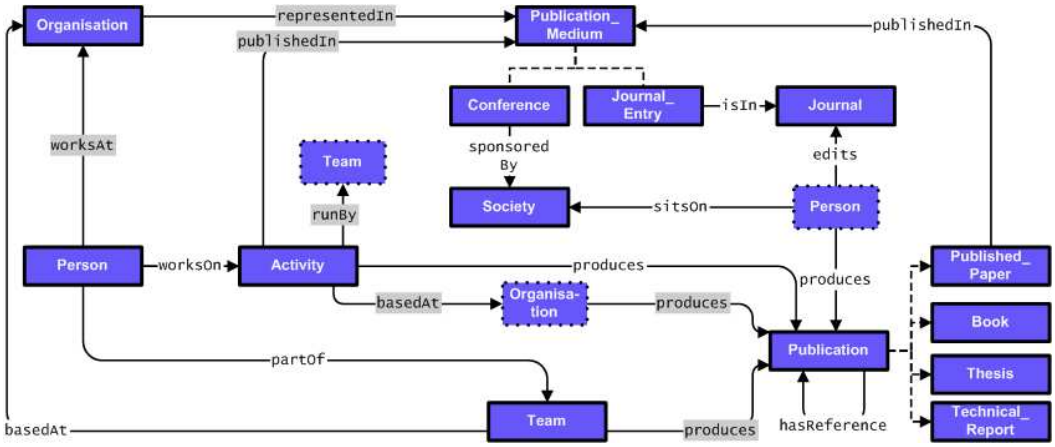


Figure 9.1: Scholarly community ontology

- What project, researchers, and institutes are related to this literature?
- What other papers were presented at this conference?
- Are there any similar journals?
- Who are the experts in knowledge management?
- What are the seminal papers from the IEEE Computer series of journals?
- Which research teams collaborate?

Conceptualisation

The conceptualisation process began by identifying the key concepts and relationships in the academic domain by considering all artifacts and objects evident in scholarly material and activities. Deciding on *borderline* concepts such as *Funding Body* and *Digital Resource* required careful consideration and a compromise between minimal ontological commitment and expressiveness was made. The same approach was considered when specialising concepts. For instance, rather than defining further sub-concepts for *Person* (e.g. researcher, postgraduate student, research fellow, professor) only the single super concept was used to broadly classify all types. Synonym resolution was another issue that became apparent. For example, 'Researcher', 'Academic', 'Scientist', 'Person', and 'Scholar' are all terms that could be used to label the same concept. It is also important to include only the salient relationships as superfluous relations could lead to confusion, especially when used in ontological hypertext.

After several iterations of the conceptualisation process, which included discussions with fellow researchers and experimentation with the WSS system, the

Concept	Slots
Activity	title, description, URI, fundingSource
Journal	title, description, URI, publisher, type
Person	name, description, URI, email, role
Publication	title, description, URI
Published_Paper	
Book	
Thesis	
Technical_Report	
Publication_Medium	title, description, URI
Conference	location, type
Journal_Entry	issue, volume, pages, year
Organisation	title, description, URI, location, type
Team	title, description, URI
Society	title, description, URI

Table 9.1: Slots for each concept in the Scholarly Ontology

ontology illustrated in Figure 9.1 was derived. Researchers, scientists, scholars, and other participating members of the scholarly domain were modelled as *Person*. Four types of *Publication* were modelled (*Published Paper*, *Thesis*, *Book*, *Technical Report*) which cover the main body of scholarly literature, in particular the literature types found in the ACM Hypertext Conference series and the types of papers in their references. Two major types of publication medium were identified: *Conference* and *Journal Entry*. The *Journal Entry* represents a published entity in a *Journal*. A conference is organised by a *Society* that researchers are part of. A project or similar activity is modelled as an *Activity* and is an important entity as it is frequently the focal point of research teams and their publications. Finally, a *Research Team* is part of a larger *Organisation*.

Each concept also has slots (or properties) attached to it which are literal values (Table 9.1).

9.2.2 Representation

There are several approaches for formally representing ontologies and these were discussed in Chapter 4. The WSS prototype used the Southampton Framework for Agent Research (SoFAR) (Moreau *et al.*, 2000) which supports ontologies defined using an XML notation. A similar representation was also used by OntoPortal. However, the scholarly ontology used in ESKIMO is more complex due to its use of hierarchies and self-referential relationships (both of which are not supported by OntoPortal). In addition, support tools, such as Protégé 2000, were required

for testing and evaluation purposes. Naturally, these tools only read ontologies in well-known ontology formats.

Four ontology definition languages were considered due to their technical capability and the acknowledgements they have received from academia and industry: RDFS, OIL, DAML, and DAML+OIL. The latter three have extensive ontology modelling support (e.g. intersections, class equivalence, sub-property, inverse relations, and enumerations) and constraint mechanisms (e.g. cardinality, data types, range). However, these proposals are evolutionary and therefore likely to change. This means that few tools, such as parsers, editors, and visualisers, support them, or support different (possibly incompatible) versions of them.

RDFS has been designed as the underlying schema language for the Semantic Web. It has basic modelling support although lacks facilities for complex statements, such as class intersections and negations (e.g. the statement 'cow is a type of mammal and not carnivore' cannot be modelled in RDFS). Aside from restricting the domain and range of a property or relationship (e.g. *sitsOn* has a domain of *Person* and a range of *Society*), RDFS provides no further constraint mechanism making it impossible to specify, for instance, that a son has to have exactly one father.

The scholarly ontology requires a language to specify concepts (and their hierarchies) and simple binary concept-to-concept relationships. Constraints could be specified (e.g. a publication has at least one author and is published in exactly one publication medium) although for the eventual task of this ontology, navigation and scholarly analysis, constraints are less important as neither tool requires them and therefore the necessary additional authoring overhead is not justified. In addition, the plans were that ESKIMO would be built from public domain tools and so a standard language that is supported by as many tools as possible was necessary.

Therefore, the three main criteria for selecting the ontology language were:

1. Modelling support for concepts, hierarchies, and simple relationships.
2. Constraint support is unnecessary.
3. Standardisation and portability are desirable.

RDFS fulfills these criteria and has the advantage of being supported by the W3C which guarantees a degree of standardisation and integration with other W3C standards.

The ontology was constructed using Protégé 2000 and an RDFS definition was exported and used in ESKIMO. The RDFS representation has been included in Appendix C.

9.2.3 Evaluation and Documentation

The scholarly ontology was informally evaluated by constructing an OntoPortal application based on it and experimenting with the resulting ontological hypertext. As OntoPortal only supports limited ontologies, the scholarly ontology was modified to exclude the hierarchies and self-referential relationships. It was felt that these changes would not significantly alter the overall evaluation process as the general structure could be modelled and this part of the evaluation was only concerned with navigational benefits provided by ontological hypertext. Data from the ACM Hypertext Conference series for 1997, which covered the concepts in the ontology, was then used to populate the application manually. After evaluating the constructed OntoPortal application by exploring its interlinked information, inconsistencies and incorrectly modelled concepts could not be identified.

Protégé 2000 was also used with the ontology and the sample data to test the inference possibilities and to determine whether the ontology was expressive enough to support scholarly questions. For example, the competency question *Who are the experts in knowledge management?* was defined by drawing on the publications of researchers and the citations these have received, the projects researchers work on, and the research teams they are members of. However, the query could only be applied to the entire hypertext field, and not be specialised to 'knowledge management'. Therefore, a secondary theme ontology was designed to extend the scholarly community ontology and is discussed in the next section.

The scholarly ontology was documented using the methodologies proposed by Uschold *et al.* (1996) and Skuce (1996). Each concept is defined with a description detailing the concept, terms by which it is also known by (synonyms), assumptions, its unique system name, and its slots. This is presented in Appendix D.

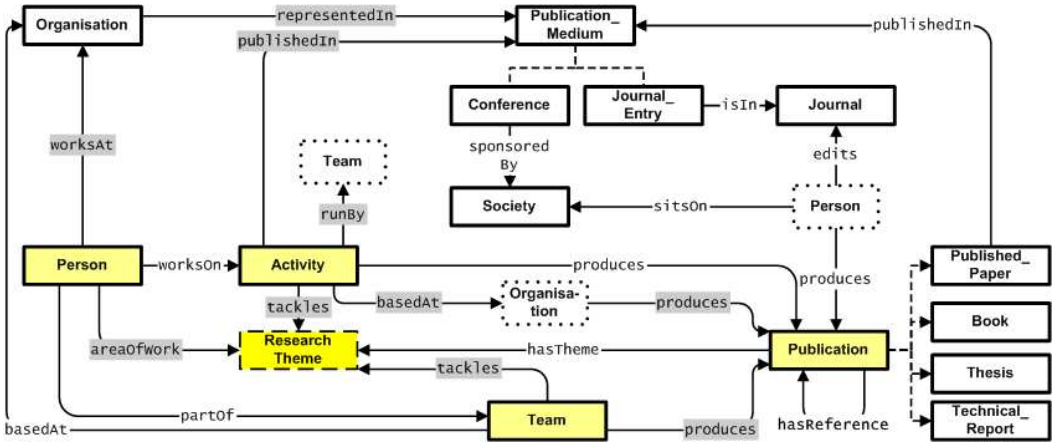


Figure 9.2: Scholarly community ontology with themes

9.3 Hypertext Research Theme Ontology

A secondary ontology was constructed to extend the scholarly ontology and account for the different themes within hypertext research. ESKIMO uses this to enable scholars to restrict their navigation and querying of the ACM Hypertext proceedings to a specific speciality area. The competency questions highlighted that this facility was necessary. For example, instead of querying for the experts in 'Hypertext', the query can be specialised to the experts in 'Hypertext Theory'.

Figure 9.2 illustrates how this research theme concept is combined with the scholarly ontology. The *Person*, *Activity*, *Team*, and *Publication* concepts are about a research theme. This feature also enables different research areas to be modelled (e.g. high performance computing, modern history, oceanography) and 'plugged' into ESKIMO.

The hypertext theme ontology was based on the topic classification used by the ACM (the *Computing Classification System*) for categorising the papers in their digital library. It includes the theme *Hypertext* and four specialisations of it: Architectures, User Issues, Theory, and Navigation. However, this only provided a coarse categorisation and further specialisations were required to enable users to specify accurately the theme in which they are interested.

Therefore, the conference proceedings of the ACM Hypertext series from the last 13 years were inspected. These provided valuable information such as keywords, track and session names, and index terms which were compiled and topics/themes distilled from them and organised into a hierarchical structure (Figure 9.3).

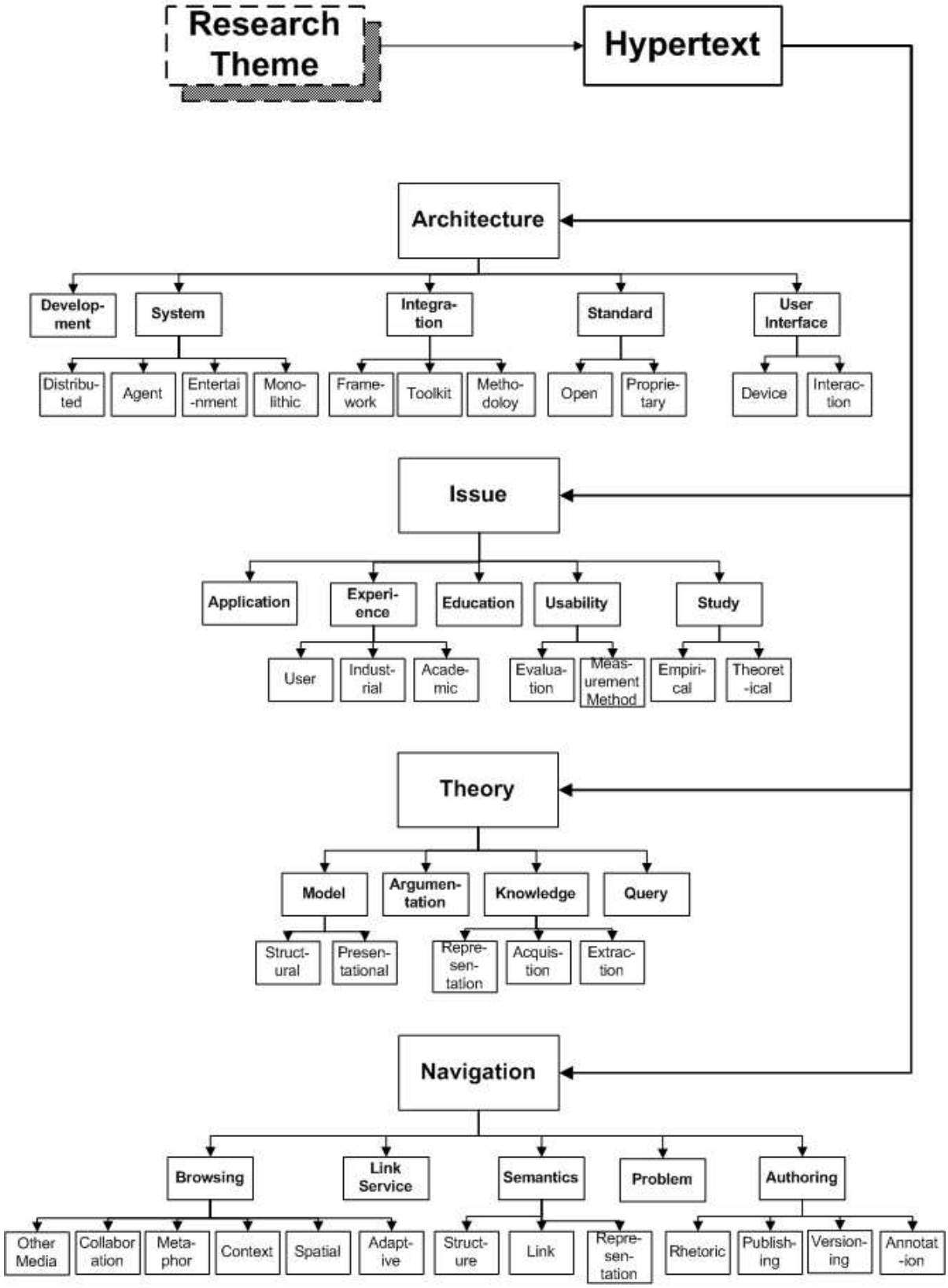


Figure 9.3: Hypertext Theme Ontology

9.4 Populating ESKIMO

A discussion of the process used to populate the ontological knowledge base in ESKIMO is presented as several knowledge acquisition issues are raised. ESKIMO was populated using data from the 'ACM Conference on Hypertext and Hypermedia' series from 1988 to 2000. This conference was chosen because of its appropriateness to this research area and as it represents a long established canonical conference in hypertext where seminal hypertext works have been published. In addition, some of the data was already in a structured and machine processible format from earlier work within the IAM Group.

The following two sections discuss how these data were extracted and converted to knowledge for use by ESKIMO.

9.4.1 Data Acquisition

The data from the conference series was extracted using a semi-automatic approach. First, the HTML pages about the conference proceedings and the PDF files of the actual documents were downloaded. A PDF citation extraction tool was used to parse the papers and extract the citation information (title, author, journal or conference title, volume, issue, year). However, due to the unstructured and inconsistent format of citations, many errors, such as in author names, page numbers, and places of publication, were evident and had to be manually fixed.

The conference details (e.g. year, location), paper details (e.g. title, abstract, author names), and author details (e.g. the research team and organisation they are based at) were automatically extracted from introductory HTML documents provided by the ACM for each paper. These documents had a consistent format so processes could be designed to parse them. However, the publication type (published paper, book, thesis, and technical report), activity information, and a paper's theme had to be manually identified.

Instance resolution proved a major problem at this stage. Many apparently different textual representations of an instance (e.g. an organisation's name) were actually referring to the same entity. This was most notable for researcher appellations. For example, the following versions of the person 'Frank Shipman III' were evident in the captured data.

- Frank Shipman III

- Prof. Frank Shimpan III
- F. Shipman
- F. Shipman III
- Frank Shipman
- Shipman
- Frank M. Shipman
- F. M. Shipman

In addition, there were frequent misspellings (e.g. Shipmann) and errors (e.g. H. Shipman). Therefore a tool was created that used a series of natural language processing tools (fuzzy matching, heuristics, and clustering) to analyse the data and group similar instances together. For example, all the versions of 'Frank Shipman' listed above would be grouped together. Manual inspection then confirmed the similarity and removed the excess versions. This technique was applied to all concept types.

When completed, the data were transformed into an XML format that would enable the knowledge acquisition stage to parse it and extract the knowledge (i.e. instances and relationships). The data acquisition stage required approximately two months effort.

9.4.2 Knowledge Acquisition

Once the data had been collected, the knowledge had to be derived and represented as ontological metadata, ready to be used by the various processes in ESKIMO (Figure 9.4). As the scholarly ontology was defined using RDFS, its complement, RDF, was selected as the metadata representation language. The availability of RDF parsers and inference (e.g. SiLRI) and visualisation (e.g. RDFViz[1]) tools, also made this a viable approach. The knowledge was also translated into logic statements to be used by the inference engines in ESKIMO (Prolog is used as the inference language (Section 9.5.3)).

After several scripts and XSLT style sheets were used to convert the XML data into RDF and logic statements, the logic interpreter SiLRI, was used to reason over the RDF statements and identify further ontological relations not already explicitly

[1] `http://www.rdfviz.org/`

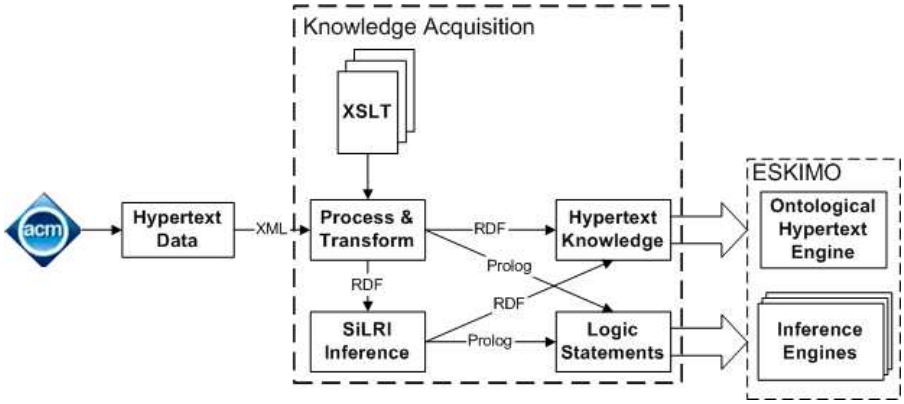


Figure 9.4: Knowledge acquisition for ESKIMO

defined. For example, if all researchers' publications and the teams they are members of are specified, then SiLRI can deduce the publications that were produced by a particular team.

Ten relations in Figure 9.2 have a shaded background. This is to indicate that these relations are deduced by SiLRI, and not manually authored or identified; meaning the authoring overhead (or requirement on automated extraction procedures) is significantly reduced. In ESKIMO, SiLRI was used to deduce over 19,000 additional relationships between instances.

At the end of the knowledge acquisition stage, over 53,000 RDF expressions had been compiled. This includes all the statements required to introduce instances and their properties, as well as the statements to describe the relationships between them. Logic statements are finer grained than RDF expressions, as a separate statement is required for each property and relation. Three files containing logic statements were produced, one for each inference engine. These ranged from containing 54,000 statements to over 84,000 statements.

However, errors, especially duplicates, were still evident once the knowledge acquisition stage had been completed. For example, the journal titles 'Hypermedia And Library Studies' and 'Hypermedia And Literary Studies' could plausibly refer to different journals; some investigation however, concluded that they referred to the same journal. These errors are difficult to identify, as they require manual inspection.

The data and knowledge acquisition stages required a large amount of effort that involved both automated tools and manual intervention. Due to the general

lack of available metadata from digital libraries, such as the ACM, a highly efficient approach is unattainable. Automatically harvesting (scraping) the ontological metadata is problematic due to the many variations and errors evident in the source text (e.g. the differences in reference styles), and as a result issues of referential integrity are common and require complex solutions (Bergmark *et al.*, 2001; Alani *et al.*, 2002).

New Semantic Web initiatives, such as the Academic Metadata Format (AMF) (Brody *et al.*, 2001b) and interoperability facilities like the Open Archives Initiative (OAI) (Lagoze & de Sompel, 2001), could significantly reduce or eliminate this overhead. However, these are only being gradually adopted by smaller initiatives and projects. Furthermore, research into extracting metadata from scientific papers has progressed. Giuffrida *et al.* (2000) propose a method based on spatial and visual knowledge principles for extracting a paper's metadata from postscript files. The approach is able to uncover title, author, and affiliation data with an 80% accuracy. Citeseer (Lawrence *et al.*, 1999a) succeeded in automatically extracting reference data, with reasonable accuracy, from scientific papers in various formats, such as HTML, PDF, and postscript. The Open Citation Project (Harnad & Carr, 2000) also developed tools to locate references within papers.

9.5 Architecture

ESKIMO is influenced by the original OntoPortal design and its ontological hypertext mechanism. However, it differs in the representation and management of the underlying ontology; it is capable of supporting more elaborate models, such as the scholarly ontology proposed earlier. Furthermore, ESKIMO adds scholarly inference facilities that are not available in OntoPortal.

As in OntoPortal, the back end processes use an SQL database, together with an ontology module that reads an RDFS ontology definition, to store and access a knowledge base that contains the ontological metadata. A similar procedure as illustrated in Figure 8.4 to set up OntoPortal applications was used to configure ESKIMO.

Figure 9.5 illustrates the major components of ESKIMO. Requests made by the user through an ESKIMO browser interface are handled by the *Query Engine*. This engine inspects the query, constructs further queries when necessary, and directs

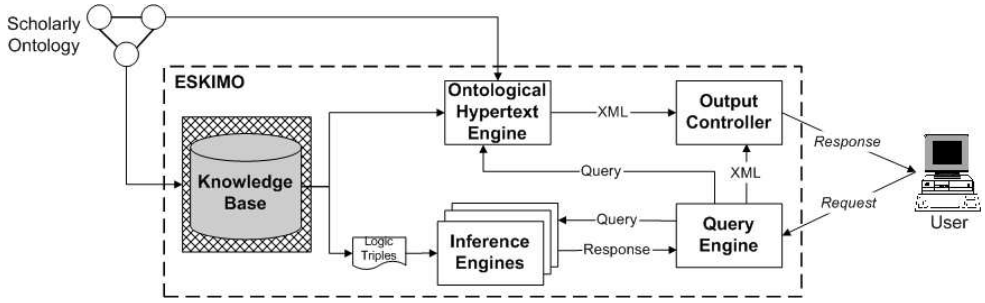


Figure 9.5: Components of ESKIMO

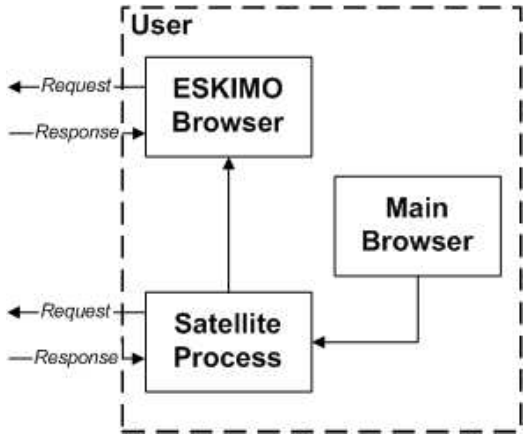


Figure 9.6: ESKIMO user request procedure

these to either the ontological hypertext engine to create the hypertext dynamically or one of the connected inference engines to resolve an analytical query. The ontological hypertext engine retrieves the properties of an instance and its relations to other instances and assembles this into an XML document. Likewise, the inference engines resolve scholarly questions and construct responses in an XML format. The output controller receives the XML, converts it into the designated output format (e.g. HTML), and returns this to the user.

Users either interact with the ESKIMO browser interface directly, or use a satellite program which monitors their activity while browsing the Web, and upon recognising a URL (e.g. that of a known paper), instructs the ESKIMO browser window to update to reflect the new context (Figure 9.6). The latter approach ensures that ESKIMO always presents scholars with information that is relevant to their current task (e.g. browsing papers in a digital library).

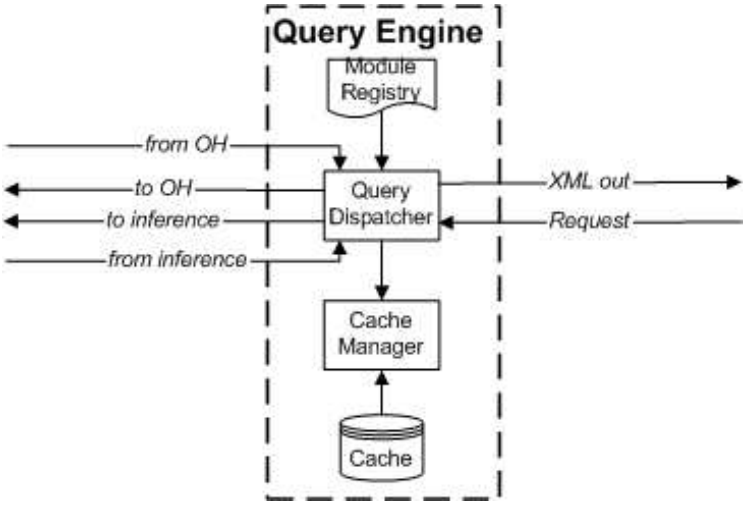


Figure 9.7: ESKIMO query engine

9.5.1 Query Engine

Figure 9.7 illustrates the components of the *Query Engine*. This module receives requests from the user and forwards them to the appropriate module. These requests are either navigational or inquiry. A navigational request is received every time the user navigates the scholarly material and therefore requires ontological hypertext to be dynamically constructed. An inquiry is a request for an analytical scholarly question to be resolved by an appropriate inference engine.

When a navigational query is received, it is forwarded to the ontological hypertext engine which constructs a document containing the properties of the relevant instance together with its relationships (links). Then it queries each registered inference engine for its support for the specified concept type. The module registry stores information on how each of the inference engines is contacted. If there is a match, the inference engine returns details of each question: the query name and label, and any further information required to execute the question (e.g. notification of the current theme). These are then appended to the XML document returned by the ontological hypertext engine.

If an inquiry is received by the *Query Engine*, it first consults the cache manager to determine if this request has already been resolved. Some analytical questions take several minutes to execute so a cache facility is useful. If the query has not been added to the cache, then it is forwarded to the appropriate reasoning engine for execution and upon completion added to the cache.

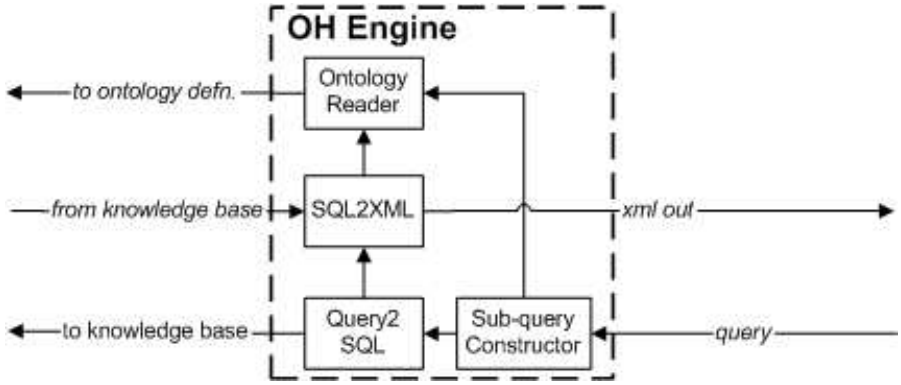


Figure 9.8: ESKIMO Ontological Hypertext engine

9.5.2 Ontological Hypertext Engine

In Figure 9.8 the *Ontological Hypertext Engine* is illustrated. When a query is received, it consults the ontology to determine the concepts the requested instance can be related to. For example, for the 'Society' concept in Figure 9.2, 'Person' and 'Conference' are returned.

This information is then used to construct the necessary sub-queries, which gather the instance properties and relationships from the knowledge base. Each query is translated into an SQL statement and issued against the database to extract the data. The result is converted into an XML document and returned.

9.5.3 Inference Engines

The inference engines define the various modules that provide scholars with an inquiry facility. They reason over the scholarly ontological metadata to uncover facts and patterns, and make plausible suggestions. Rules (e.g. bibliometrics) and heuristics (e.g. the number of publications an expert is expected to have) are also combined with the reasoning facility to support various scholarly questions.

Architecture

The architecture of the inference engines used in ESKIMO is illustrated in Figure 9.9. When the engine receives a query requesting its support for a particular concept, it consults the *Query Registry* and returns the details of all supported questions as an XML document fragment. If the query is an instruction to execute one of its supported inferences, it first ensures that it supports the question and then retrieves the necessary data (i.e. the rules and facts) required to execute it. A logic process then runs the question against the input statements.

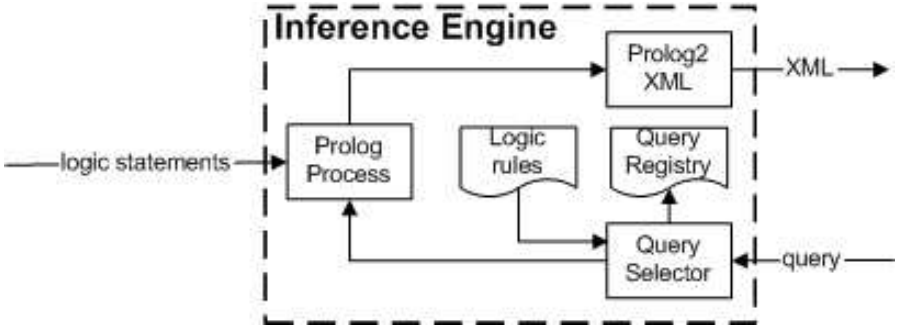


Figure 9.9: ESKIMO inference engine

The inference engines comply to a common interface that ESKIMO uses to communicate with them. This requires them to act as Web servers and respond to two types of queries. Firstly, a request for details of all supported scholarly questions that use a particular concept. Secondly, a synchronous request for resolving a scholarly question that the engine supports. Beyond this, the architecture of the engine is totally at the discretion of the implementor. The main ESKIMO component runs independently of these inference engines, meaning they can be *dynamically* started and halted without interfering with the other services.

Three inference engines were built for ESKIMO: one to apply bibliometric rules, one to use bibliometric rules on scholarly objects other than literature, and one that used the entire scholarly ontology to make general abductions (e.g. determining potential experts).

The declarative logic language Prolog (PROgramming in LOGic) (Warren, 1977) was used as the inference language for the engines. Prolog is a language based on predicate logic and uses Horn clauses. A clause is a disjunction of literals (e.g. $p_1 \vee p_2 \vee q$). A Horn clause is a clause where at most *one* literal evaluates to positive (e.g. $\sim p_1 \vee \sim p_2 \vee q$ or $\sim p_1 \vee \sim p_2 \vee \sim q$). This characteristic makes Horn clauses suitable for computation as they can be stated as a rule in the form of $q \rightarrow p_1 \wedge ... \wedge p_n$. This expression contains a head (q) and a body consisting of a series of claims (p_1 ... p_n). It is sufficient to only prove that the body is true in order for the head (or rule) to be true. Therefore, Horn clauses have the advantage of providing programmers with an instinctive approach to declaratively defining rules.

Prolog was also chosen because of its wide support in academia and the facilities and tools available to integrate with other programming constructs (e.g. I/O, networking, other programming languages).

Each Prolog clause contains the rule name (predicate) and the body of the rule, which consists of other predicates and variables. The following code defines three clauses that can be used to determine the wife of a particular husband.

```
gender(P, G).
spouse(P, S).
wife(M, F) :- spouse(F, M), gender(F, 'female').
```

Facts are specified in a Prolog program by supplying terms for the variables, such as *gender('Sally', 'female')*. The *wife* predicate can then be used to determine all matching wives and husbands as well as particular instances (e.g. *wife('Bob', F)*).

An example inference used by the scholarly community engine to identify possible significant (or seminal) organisations is listed below.

```
findAllSeminalOrganisations(List) :-
  findall(Org, seminalOrganisation(Org), List).

seminalOrganisation(Org) :-
  organisation(OrgID),
  enoughPublications(OrgID, 10),
  workEnoughPeople(OrgID, 5),
  title(OrgID, Title),
  Org = [['id', OrgID], ['title', Title], ['type', 'Organisation']].

workEnoughPeople(Org, N) :-
  findall(Person, worksAt(Person, Org), List),
  length(List, Number),
  Number > N.

enoughPublications(Thing, N) :-
  findall(Publication, produces(Thing, Publication), List),
  length(List, Number),
  Number > N.
```

The main predicate is called *findAllSeminalOrganisations*. Using Prolog's *findall* predicate, all seminal organisations are retrieved with the *seminalOrganisation* predicate. This extracts the organisations that (i) have produced at least 10 publications (as defined by *enoughPublications*) and (ii) have at least 5 members (as defined by *workEnoughPeople*). The numbers used in these heuristics are comparatively small, as they only apply to the objects within the ACM Hypertext Conference series, and not the total number of publications or members in an organisation. For each discerned organisation, its title and ID are constructed into a list and returned.

To add support for a new scholarly question, the Prolog rule is first added to the logic file. Then details of the question (e.g. its label, required data, rule name, and concept it applies to) are added to the inference engine's query registry and the ESKIMO configuration file. The question is then available to users of ESKIMO.

Communication and Interchange

If a user is viewing an instance of type *Person*, the *Query Engine* requests each inference engine for its support for the *Person* concept. The format of this query is an XML fragment and is sent to the engine as an HTTP request:

```
<query type="SUPPORT" class_type="Person"/>
```

The type of the query is specified as *SUPPORT*, which denotes that this is a request for the details of all supported questions. Each inference engine is consulted and the resulting XML fragments are merged and returned to the *Output Controller*. The following listing illustrates the XML received by the *Output Controller* after all inference engines have been queried for their support for the *Person* concept.

```
<module name="AUG" module_name="Augmented Bibliometrics">
  <query type="QUERY" name="mostCoCitedPerson"
   text="Which authors have been regularly co-cited?"
   class_type="Person"/>
</module>

<module name="SC" module_name="Scholarly community Analysis">
  <query type="QUERY" name="findAllPersonCollaborators"
   text="Which researchers collaborate?" class_type="Person"/>
  <query type="QUERY" name="findPossibleCollaborators"
   text="With which fellow researchers does this person collaborate?"
   class_type="Person"/>
  <query type="QUERY" name="findAllExperts"
   text="Who are the experts?" class_type="Person"/>
</module>
```

If the user then decides to resolve the question, *Who are the experts?*, a request is constructed and a request is issued to the appropriate inference engine.

```
<query type="QUERY" class_type="Person" id="56" query="findAllExperts">
  <rthemes join="and">
    <theme>navigation</theme>
    <theme>issue</theme>
  </rthemes>
</query>
```

This query requests that the indicated query ('findAllExperts') is executed. Request also supply the concept type, *Person* in this example, and the ID of the

unique instance being viewed by the user (e.g. '56'). This particular question does not actually require this additional information, although questions such as *Who collaborates with this person?* would. The theme context is also provided, in this case 'Hypertext Navigation' and 'Hypertext Issues'. The join applied to these themes is an 'and'. An 'and' requires returned instances to be about both themes, whereas an 'or' simply requires that the instances are about at least one of the listed themes.

The inference engine runs the query, converts the result into an XML document fragment, and returns it. The fragment contains details about the original question and the result. The following listing illustrates the XML fragment that is returned by the inference engine when the above query is completed.

```
<resolved_query type="QUERY"
                name="findAllExperts"
                text="Who are the experts?"
                class_type="Person"
                class="56" rtheme="navigation,issue"
                rjoin="and">
  <class type="Person" id="126">B. Schneiderman</class>
  <class type="Person" id="135">M. Bieber</class>
  <class type="Person" id="168">U. K. Wiil</class>
  <class type="Person" id="6">J. J. Leggett</class>
  <class type="Person" id="223">G. Salton</class>
  <class type="Person" id="21">C. C. Marshal</class>
  <class type="Person" id="34">R. Furuta</class>
  <class type="Person" id="60">M. J. Bernstein</class>
  <class type="Person" id="85">W. Hall</class>
  <class type="Person" id="87">H. C. Davis</class>
</resolved_query>
```

9.5.4 Output Controller

The *Output Controller* receives XML fragments from the *Ontological Hypertext Engine* and *Inference Engines* and converts them into the requested output format, for example HTML for displaying in a Web browser. As in OntoPortal, the conversion from XML to HTML is completed using several XSLT style sheets. Alternatively, raw XML or RDF (i.e. ontological metadata) could be returned to be used by further services or metadata processes.

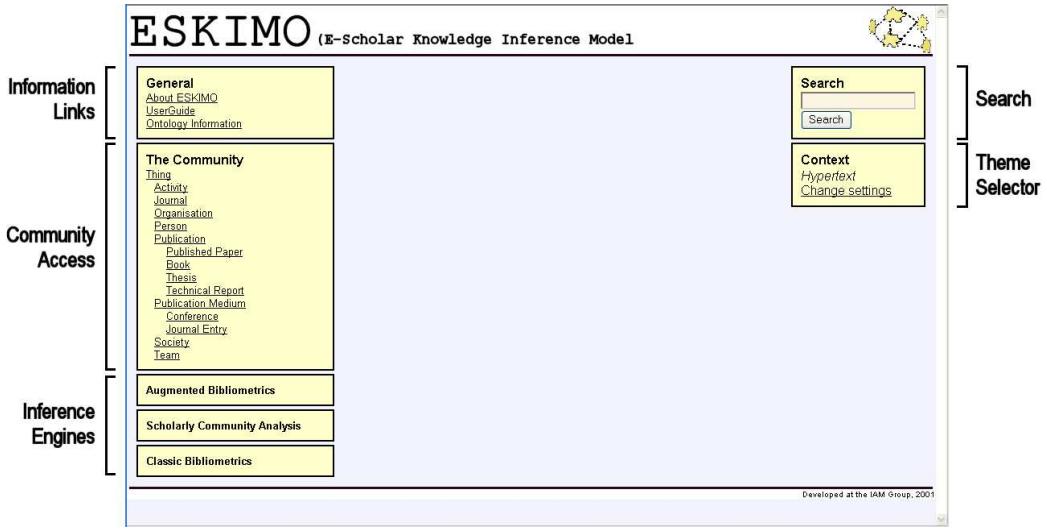


Figure 9.10: Initial Screen (ESKIMO)

9.6 ESKIMO in Practice

While ESKIMO supports a scholar by providing further insight into aspects of research, it does not provide a starting point, or *seed*, such as a table of contents or index page, to initiate this. The scholar is expected to provide the starting point, such as a researcher's name, the title of a project, an organisation's name, or even a theme. ESKIMO is then used to explore this object and discover further related material. This section discusses the two exploratory methods ESKIMO provides to scholars: navigation and inquiry.

9.6.1 Navigating Scholarly Material

Scholars review the material available in ESKIMO by traversing the hypertext produced using the ontological hypertext principle. Figure 9.10 displays the initial screen presented by ESKIMO. The top left box provides general links about the ESKIMO system. The box below this is the community access where a textual outline of the ontology is listed which provides links to view all available instances of a particular concept (Figure 9.11). Below this are three outlined areas for each of the registered inference engines. A search box is provided at the top right and the current theme and link to the theme selector page (Figure 9.12) is provides below this.

By following one of the activity instances from Figure 9.11, the system returns the information about the instance along with its relationships to other instances

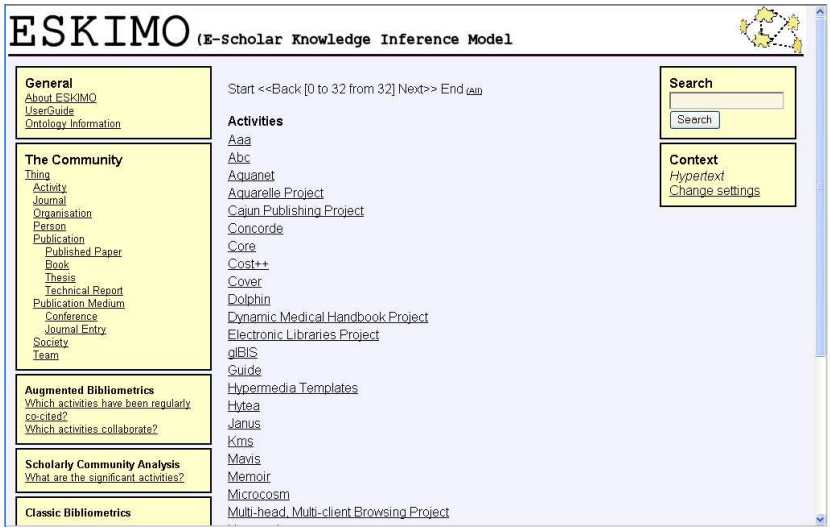


Figure 9.11: All activities (ESKIMO)

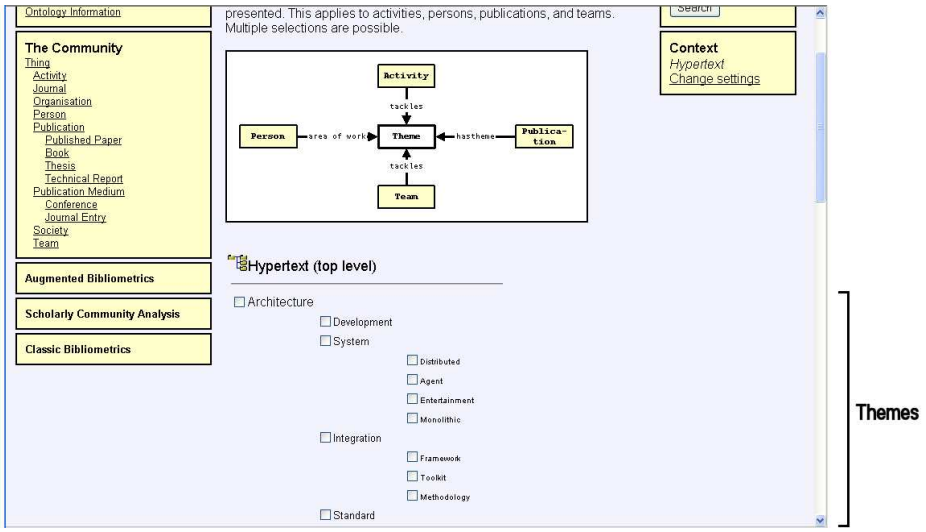


Figure 9.12: Theme selector (ESKIMO)

(Figure 9.13). The central area at the top is used to list its properties and graphically illustrate the relationships the concept can have. The related instances are then listed below this area.

Jonassen (1993) also proposes explicitly and graphically stating the relationships between a current node and related nodes to act as structural cues for users. However, his subsequent study found little evidence of these actually assisting users in their knowledge acquisition tasks. In ESKIMO though, these graphical representations of the underlying relationships form more than structural cues and provide users with vital information on the structure of the ontology that is controlling the

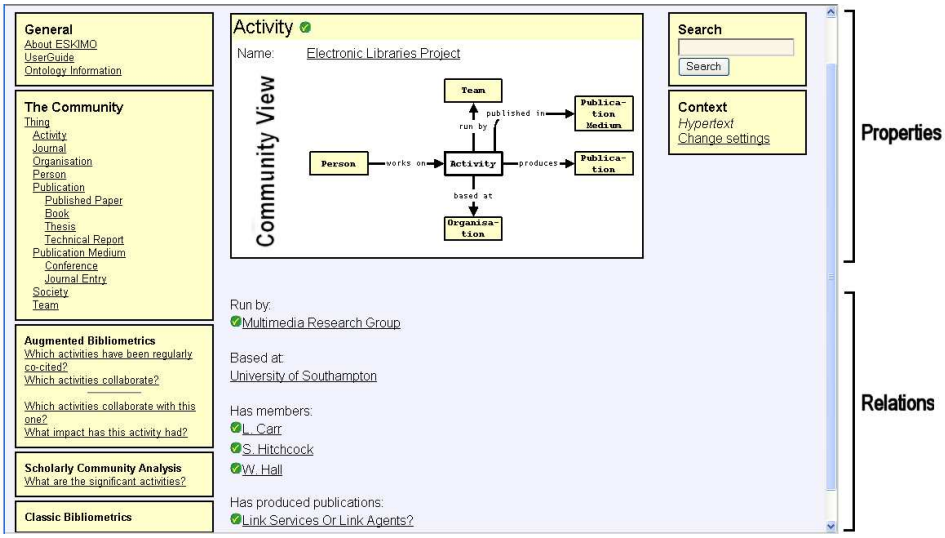


Figure 9.13: Activity instance (ESKIMO)

hypertext.

Figure 9.14 displays an instance of a published paper concept. As with Onto-Portal, scholars browse around an instance to explore its relations to other instances and thereby explore paths and material that they might not otherwise have considered. For example, by following one of the author links, the properties and relationships of that person are returned (Figure 9.15).

The theme selector enables users to specify a more specific context than 'Hypertext'. For example, by changing the context to 'Semantics' the list of activities presented by the community access has been sharply reduced (Figure 9.16).

When viewing instances with a theme context set, related instance links are preceded with a context icon which designates if the related concept is within the current context, out of it, or that the instance has been unclassified (Figure 9.17).

The hierarchies specified in the ontology enable scholars to decide whether to view instances of a general concept or specialisations of it. For example, if information on all publications is desired, then instances of type *Publication* are selected. Alternatively, if the scholar is only interested in theses, then the *Thesis* type is used.

A search facility is provided to enable scholars to locate known material efficiently. Figure 9.18 illustrates the result from searching for 'Microcosm'. Each located instance is presented under its respective concept type, in this case results

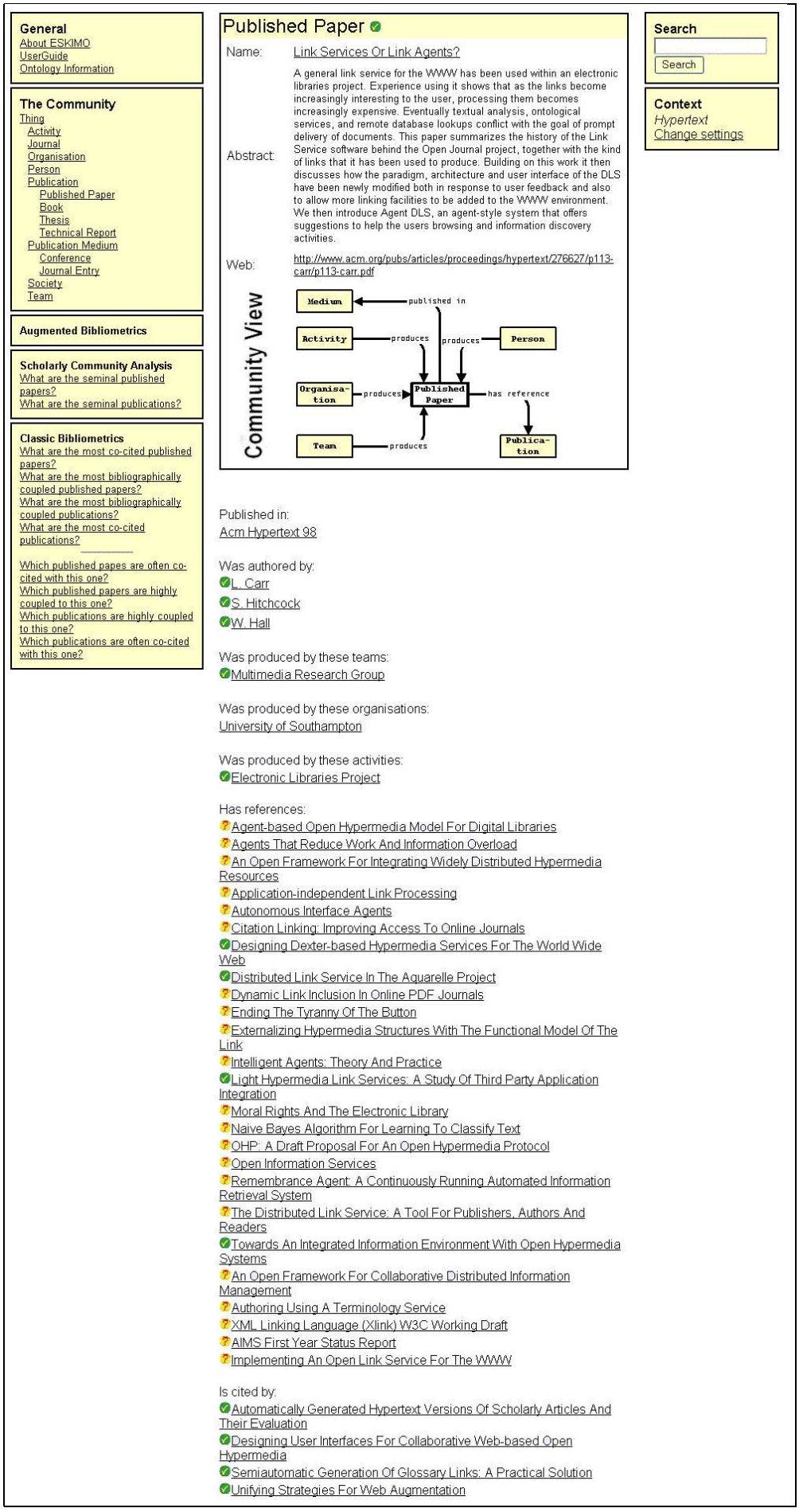


Figure 9.14: Published paper (ESKIMO)

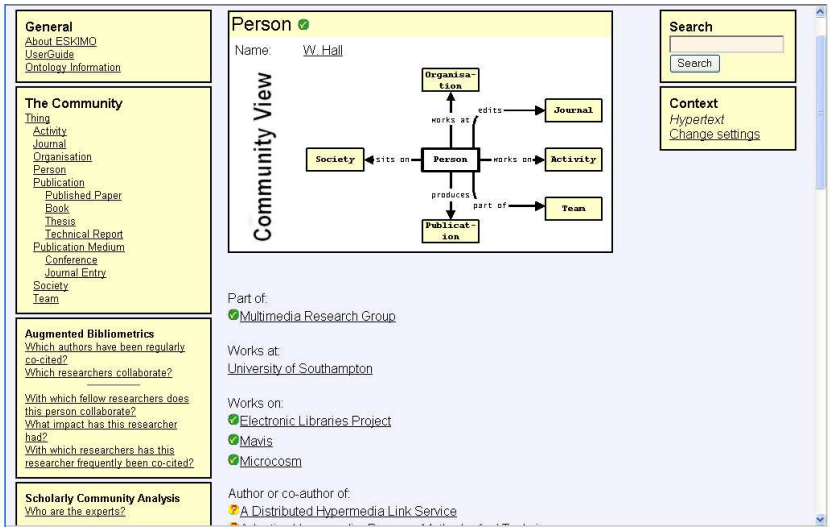


Figure 9.15: Person instance followed from the published paper (ESKIMO)

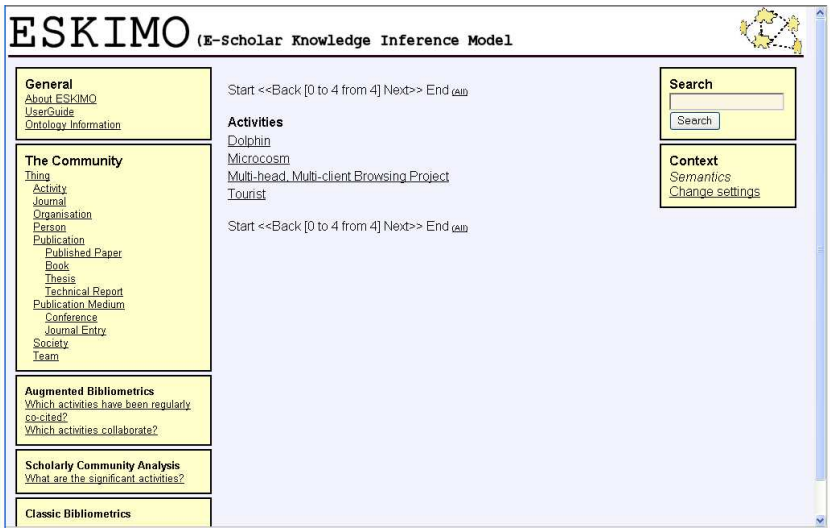


Figure 9.16: All activity instances with the theme context applied (ESKIMO)

were found for *Activity*, *Publication*, *Technical Report*, *Published Paper*, and *Book* concepts.

9.6.2 Scholarly Inquiry

Three inference engines are provided by ESKIMO to enable scholars to make intricate questions about the knowledge in their research field. While the inference engines are able to respond to over 50 scholarly questions, these are only intended to demonstrate the potential of the ESKIMO approach, rather than provide canonical answers to all research questions.

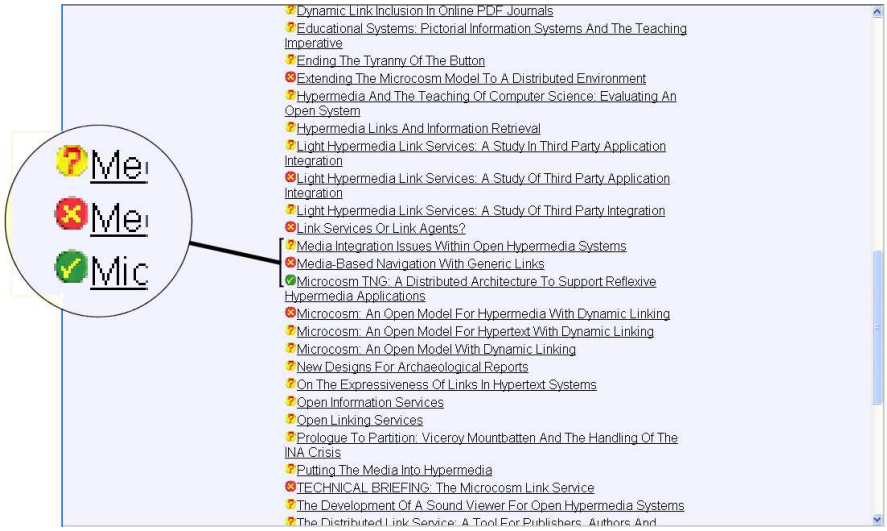


Figure 9.17: Context icons (ESKIMO)

Figure 9.18: Search results (ESKIMO)

Bibliometrics

The bibliometrics engine applies standard bibliometric rules, such as impact factor, co-citation, bibliographic coupling, and collaboration to the scholarly ontological metadata. These rules are only applied to standard bibliographic data: publications (and their references), journals, and authors. The module assists scholars in further understanding their research field from a purely bibliographic viewpoint. For example, they can determine the most highly co-cited or coupled papers, the impact factors of journals, and which authors collaborate (based only on co-authorship).

Concept	Queries
Publication	What are the most co-cited publications? Which publications are often co-cited with this one? What are the most bibliographically coupled publications? Which publications are highly coupled to this one?
Published Paper	What are the most co-cited published papers? Which published papers are often co-cited with this one? What are the most bibliographically coupled published papers? Which published papers are highly coupled to this one?
Book	What are the most co-cited books? Which books are often co-cited with this one? What are the most bibliographically coupled books? Which books are highly coupled to this one?
Thesis	What are the most co-cited theses? Which theses are often co-cited with this one? What are the most bibliographically coupled theses? Which theses are highly coupled to this one?
Technical Report	What are the most co-cited technical reports? Which technical reports are often co-cited with this one? What are the most bibliographically coupled technical reports? Which technical reports are highly coupled to this one?
Journal	What impact has this journal had?
Person	Which researchers collaborate? With whom does this researcher collaborate?

Table 9.2: Questions supported by the classic bibliometrics inference engine

Facts and patterns uncovered are similar to those provided by the Web of Knowledge (ISI, 2002) described earlier. The limitations of depending on citation data alone for scholarly analysis have been highlighted several times throughout this thesis. Therefore, scholars should only use the results from this inference engine as general indicators.

Table 9.2 lists the questions that this engine supports for each concept.

Augmented Bibliometrics

The augmented bibliometrics engine applies standard bibliometric rules and inferences to all other concepts modelled by the scholarly ontology. For example, impact factors are provided for research teams, collaboration measures for activities, and co-citation analysis for organisations. These enable scholars to use common, accepted, and well-understood rules to determine useful facts about other objects in their research field.

Concept	Queries
Research Team	What impact has this team had? Which team(s) collaborates with this one? With which teams has this team frequently been co-cited? What teams have been regularly co-cited? Which teams collaborate the most?
Organisation	What impact has this organisation had? Which organisations collaborate with this one? Which organisations have been regularly co-cited? Which organisations collaborate?
Conference	What impact has this conference had?
Activity	What impact has this activity had? Which activities collaborate with this one? Which activities have been regularly co-cited? Which activities collaborate?
Person	What impact has this researcher had? With which researchers has this researcher frequently been co-cited? With which fellow researchers does this person collaborate? Which authors have been regularly co-cited? Which researchers collaborate?

Table 9.3: Questions supported by the augmented bibliometrics inference engine

In addition, this module augments standard bibliometric rules with additional domain knowledge to improve the accuracy and breadth of results. For example, rather than just using co-authorship as a measure of researcher collaboration, the research teams and activities that researchers work in are also analysed. The different results achieved through this approach, compared to using only standard bibliometrics, are demonstrated in the next section.

Table 9.3 lists all the questions supported by this inference engine.

Scholarly community

The scholarly community inference engine draws from all concepts in the scholarly ontology to make general abductions and inferences. For example, by drawing on the activities that a researcher works on, their publications and the citations these have received, the journals they edit, and the research teams they are members of, significant researchers (or experts) are identified.

Seminal papers are identified through the citations they have attracted. However, this approach could be extended by weighting citations differently if they were made by experts, by other significant papers, or by papers describing a highly

Concept	Queries
Person	Who are the experts?
Research Team	What are the significant research teams?
Publication	What are the seminal publications?
Published Paper	What are the seminal published papers?
Book	What are the seminal books?
Thesis	What are the seminal theses?
Technical Report	What are the seminal technical reports?
Organisation	What are the significant organisations?
Activity	What are the significant activities?
Conference	What are the significant conferences?
Journal	What are the significant journals?

Table 9.4: Questions supported by the scholarly community inference engine

influential project. Unfortunately, a lack of available processing power made this extension infeasible.

Traditionally, experts and other significant objects are identified through citation impact. However, this inherently incurs a time lag as further papers are published and is also tied to the quality and quantity of citations. By drawing on other parts of the domain however, these can be calculated without relying on the citation factor alone and therefore circumvent the time lag.

The scholarly questions provided by this module are especially useful for identifying the key material and developments in a research field. It enables researchers new to a field, such as Ph.D. students, to determine the major publications, activities, and researchers. The informal survey of the professors in the IAM Group (Section 6.4) concluded that these are indeed the types of questions Ph.D. students must determine the answers to. Each of the returned results can then be further analysed to discover additional related information. For example, the collaborators of experts are likely to conduct relevant (and possibly significant) work.

Other scholarly activities, such as reviewing a journal paper or completing a literature survey, are also supported by this module and ESKIMO as scholars are able to quickly identify key work and understand how it relates to and is positioned within the research field in general, and to specific activities, researchers, and publications.

Table 9.4 lists the questions supported by this engine.

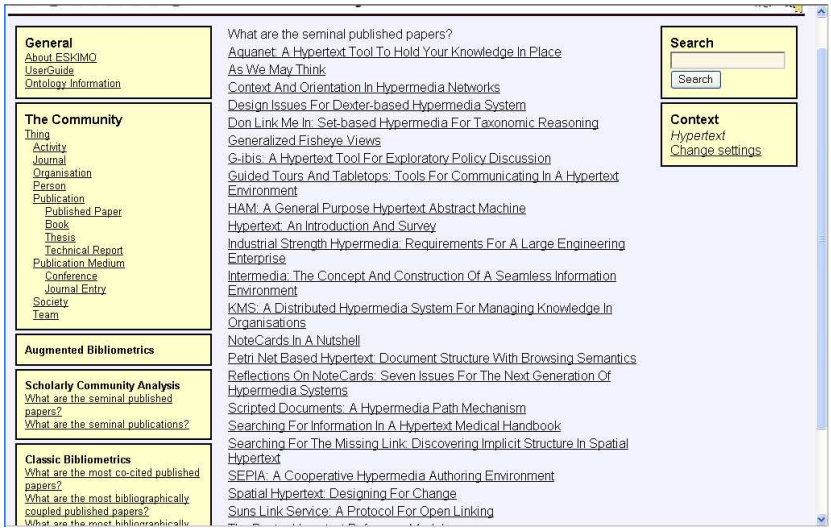


Figure 9.19: Seminal papers (ESKIMO)

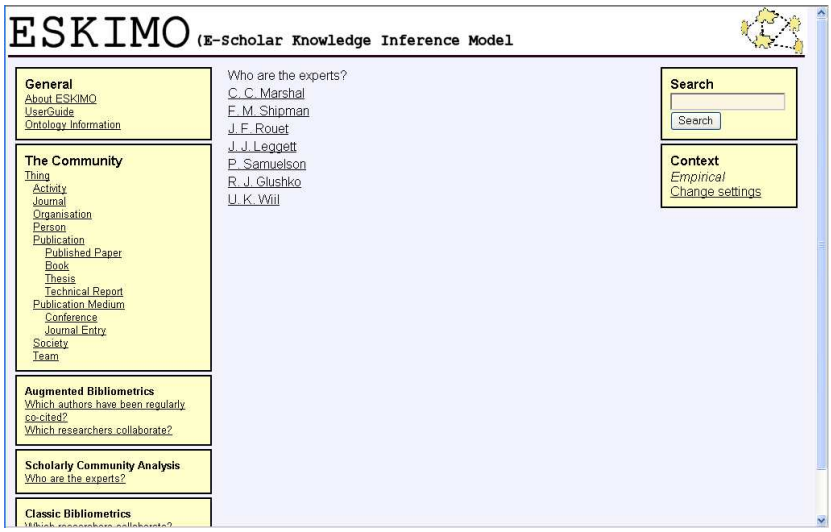


Figure 9.20: Experts of empirical studies in hypertext (ESKIMO)

Scenarios

A researcher new to a field is likely to be initially interested in the significant activities, papers, researchers, and research teams in that area. Figure 9.19 illustrates how the scholarly community inference engine in ESKIMO is used to propose seminal papers in 'Hypertext'. Figure 9.20 displays the results for the experts in empirical hypertext studies.

Collaborating researchers have overlapping research interests and this principle can be exploited to identify further and related work. When viewing a person instance, ESKIMO presents several queries it supports for this concept type (Figure 9.21), including two different methods for identifying collaborators.

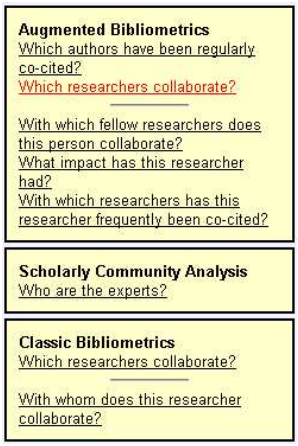


Figure 9.21: Queries available for the person concept (ESKIMO)

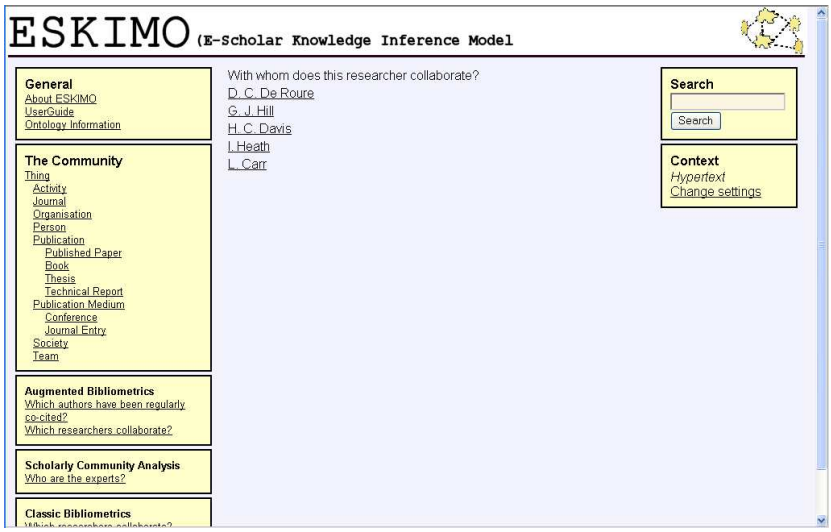


Figure 9.22: Collaborators based on bibliometric measures (ESKIMO)

Firstly, the bibliometric inference engine defines collaboration as those researchers that have co-authored a large number of papers. The determination of this threshold depends on the desired number of results; the higher the number the fewer potential collaborators are returned. However, if the threshold is set too low (e.g. 1 co-authored paper indicates collaboration) then not only will a large number of collaborators be returned, but many are also likely to be incorrect. For the researcher 'W. Hall', five collaborators are proposed by the bibliometric engine (Figure 9.22). From the author's experience, the proposed collaborators are appropriate candidates.

Secondly, the augmented bibliometrics inference engine also uses co-authorship as an indicator of collaboration, but inspects the research teams, organisations, and activities of potential collaborators as well. The results for 'W. Hall' are illustrated

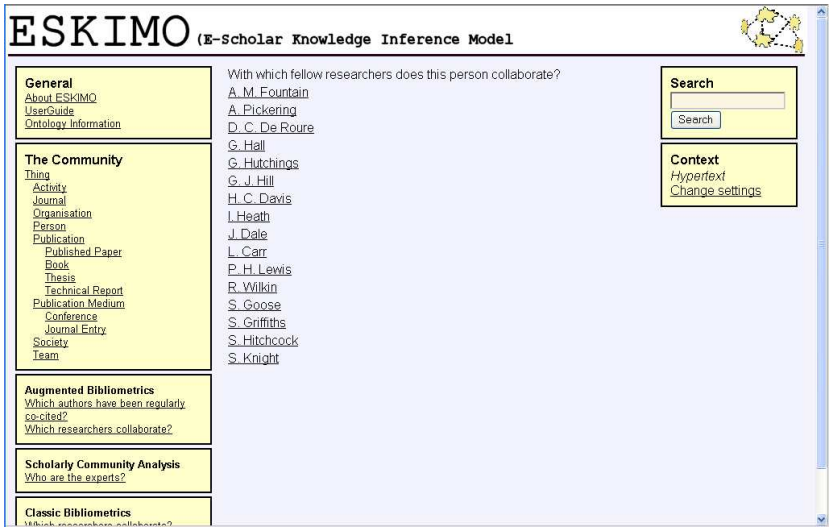


Figure 9.23: Collaborators based on entire scholarly domain (ESKIMO)

in Figure 9.23. 16 researchers are returned using this approach, including the five that the standard measure returned. However, several further important researchers are proposed that the classic approach failed to identify, such as 'S. Hitchcock' and 'P. H. Lewis'. From experience, it is known that these researchers have collaborated on many occasions.

It would also be plausible to select collaborators based on author co-citation patterns, although this may only find researchers who have similar research interests. Figure 9.24 presents the results of researchers that are frequently cited with 'W. Hall'. The list of 50 researchers contains two out of the three researchers identified using the classic collaboration measure, and nine out of the thirteen proposed using the augmented approach.

Organisations and research teams are effective places for locating information. For example, the IAM Group (formerly the 'Multimedia Research Group') conducts hypertext research and scholars in this discipline are likely to view the work published by its members. ESKIMO can then be used to advise on other possibly related teams, for example, by using collaboration measures or team co-citation. Figure 9.25 presents the result when team collaboration analysis has been completed. Four teams are suggested, all of which have had connections with the IAM Group, and in particular the research team at 'Multicosm Ltd.' which was initially set up to commercialise the Microcosm work started by the IAM Group.

Impact factors are useful for assessing an object's influence and significance in

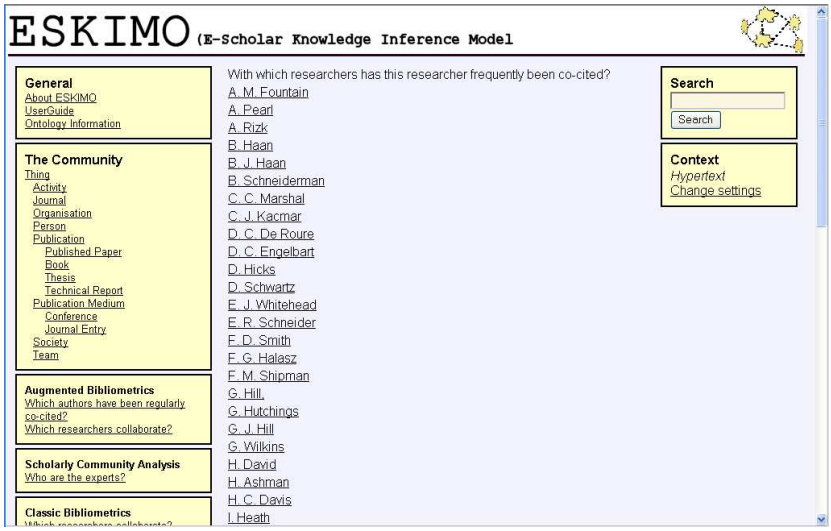


Figure 9.24: Co-cited fellow researchers (ESKIMO)

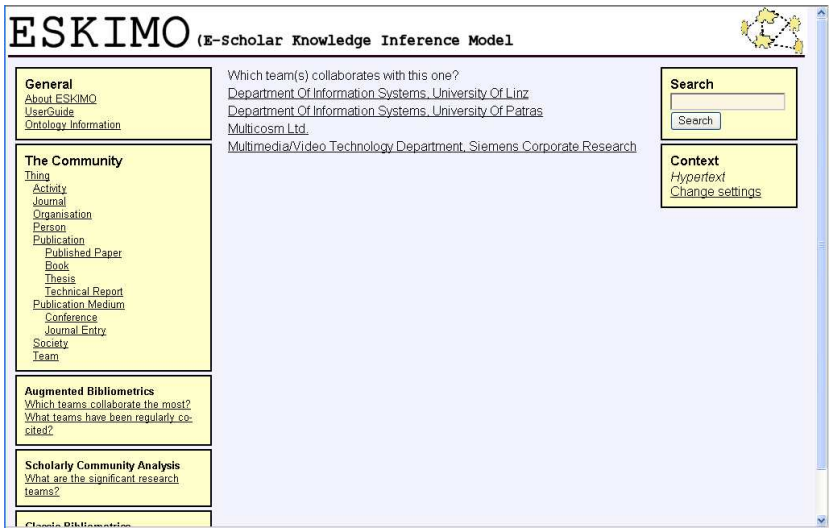


Figure 9.25: Collaborating teams (ESKIMO)

relation to the other instances within a domain. Traditionally, impact factors have been used to assess journals, however the approach can be used to rank other scholarly artifacts. For example, the augmented bibliometrics inference engine enables the determination of the team impact factor for a research team. This is calculated by dividing the total number of citations to papers authored by researchers based at the team, by the total number of papers produced by it.

Figure 9.26 illustrates the impact factor for the 'Multimedia Research Group'. The calculated factor of '2.25' is not informative in isolation. Therefore, a graphic is provided by ESKIMO to indicate the range of impact factors evident for all the teams in the system and thereby assist the scholar in comparing the value. In

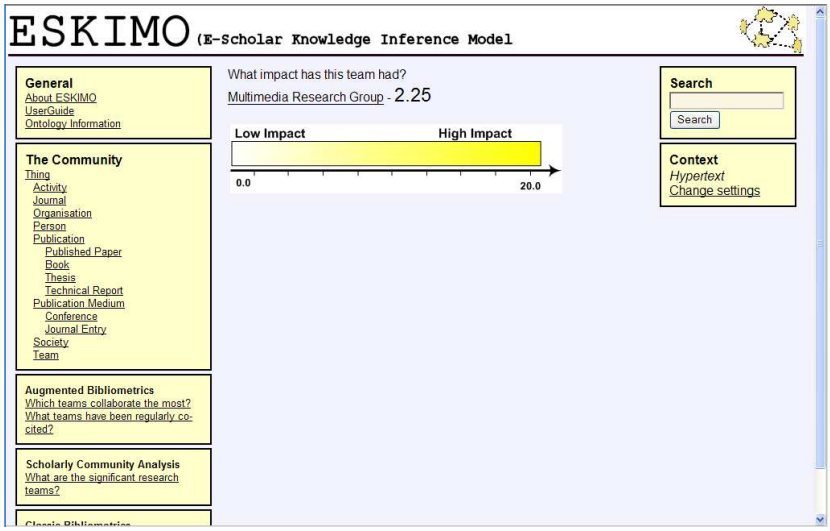


Figure 9.26: Team impact factor (ESKIMO)

addition, by comparing it to other research teams' impact factors a better judgment is possible. For example, the 'C&C Research Laboratories' has an impact factor of '1.66'. It could therefore be argued that based on citation factors, the research group at Southampton has had a greater impact on the hypertext research field (or more precisely, the ACM Hypertext Conference series). This could be explained by the fact that the IAM Group traditionally concentrates its effort in hypertext research, while 'C&C Research Laboratories' focuses more on computer communications.

9.7 Integrating ESKIMO

ESKIMO's objective is to provide scholars with an improved understanding of their research field. However, a less *proactive* approach to using ESKIMO would be advantageous as switching between browser windows, and thereby temporarily losing the focus of a task, incurs a cognitive load.

Using the ESKIMO satellite described earlier, the content of the information provided by ESKIMO can reflect the current task the scholar is engaged in (e.g. viewing a paper within a digital library). Therefore, two different approaches that provide integrated linking are presented but have not been implemented.

First, ontological links and queries are augmented into the material of a scholar's task. The links are attached alongside the scholarly material using a similar approach as in the document discussion tool, D^3E (Sumner & Shum, 1998), where supplementary material is provided in a separate browser frame.

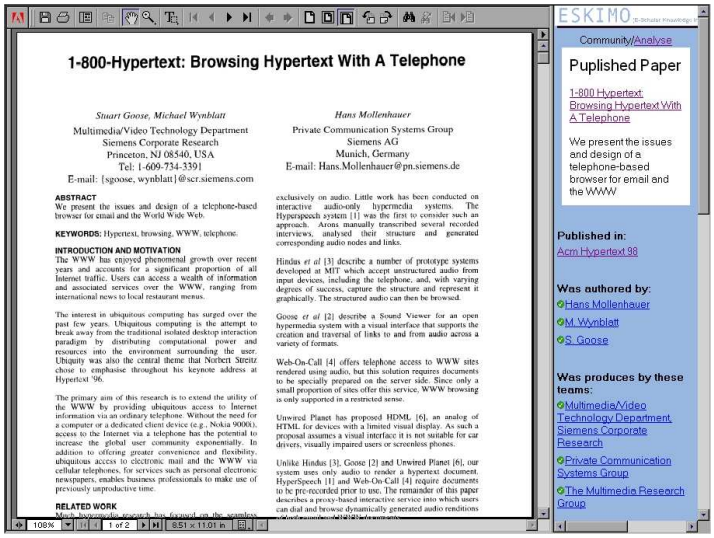


Figure 9.27: Augmenting ESKIMO: community View

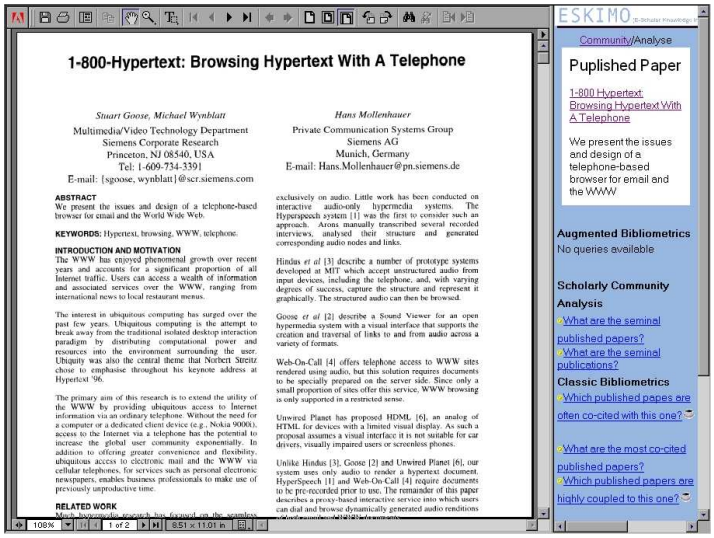
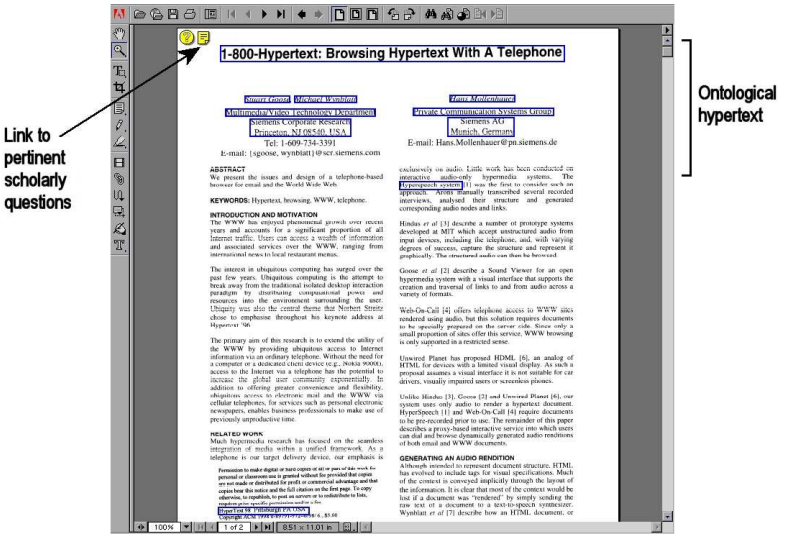


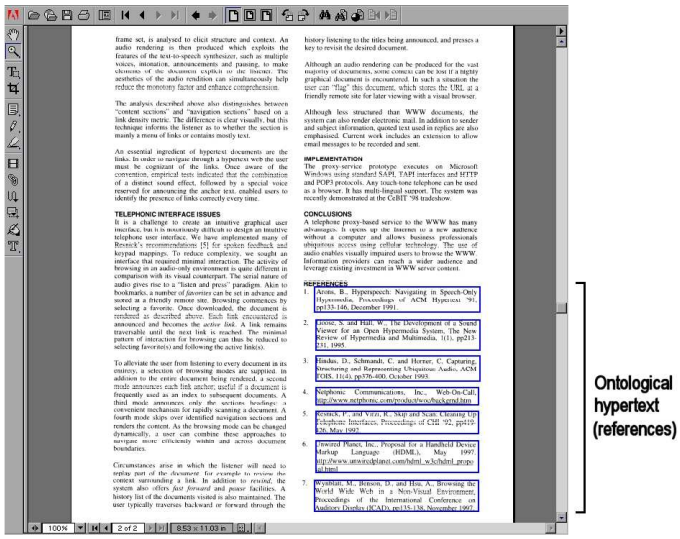
Figure 9.28: Augmenting ESKIMO: Analysis

Figure 9.27 illustrates a paper presented along with the community access information. By selecting the *Analyse* link the list of possible analytical queries is presented (Figure 9.28)

Alternatively, open hypermedia principles could be used to dynamically add links to the literature as they are required, for example by using link services such as the DLS (Carr *et al.*, 1995). This has already been demonstrated in the Open Journal Framework (Hitchcock *et al.*, 1998) which explored ways of combining the DLS with electronic journals to enable articles to contain hypertext links and be 'be released from their isolated state'.



(a) Ontological hypertext and scholarly inquiry



(b) Ontological hypertext

Figure 9.29: ESKIMO facilities within Adobe Acrobat

However, while ESKIMO contains the necessary linking metadata to perform dynamic linking, the predominant paper formats on the Web are PDF and postscript, which traditionally have limited support for Web facilities such as linking. Fortunately, newer versions of PDF support simple annotations and linking, enabling PDF documents to be enhanced with the facilities provided by ESKIMO. Text matching techniques could be used to parse a PDF document and add the links at the correct positions; an approach successfully applied to citations and references in the Open Citation Project (Harnad & Carr, 2000). Figure 9.29 illustrates how a PDF document is decorated with ESKIMO links.

These approaches present a scholar with a more integrated and seamless environment that provide scholarly services directly in the content they are consuming and task they are engaged in, rather than forcing scholars to switch tasks and use ESKIMO as an independent support tool.

9.8 Evaluating ESKIMO

An evaluation of ESKIMO was conducted to collect comments and evidence on its usefulness as a research support tool. The experiment in Chapter 6 assessed how well the Web supported typical scholarly questions. A different type of evaluation was completed for ESKIMO however, for the following reasons:

1. ESKIMO is limited to hypertext material. The original experiment covered material in various disciplines.
2. It is likely that answers to the type of questions posed in the earlier experiment are directly available in ESKIMO, and thus such an experiment would not necessarily contribute useful new evidence.
3. It would be possible to script the experiment to suit the types of analytical questions supported by ESKIMO. For example, the question requesting the identification of seminal papers is directly supported by ESKIMO, and would require minimal effort to resolve.
4. It would be useful to observe how ESKIMO supports 'real' research tasks, rather than just the individual questions that make up these tasks.

Therefore, an evaluation was constructed that required participants to use ESKIMO in a practical research task without a pre-scripted series of questions or tasks to complete. This type of experiment also enabled the intuitiveness of ESKIMO to be considered, as users were given little guidance on how to resolve the specified task and were instead asked to devise their own strategy.

9.8.1 Structure of the evaluation

Five experts in the field of digital libraries were recruited to participate in the study and draw from their experience to provide high quality feedback on ESKIMO and its suitability for a typical research activity. The task of the evaluation required participants to read a short paper from the ACM Hypertext Conference 2001 (a paper not in ESKIMO) and review it to determine its quality and whether it includes

all relevant material. The paper was selected because of its subject area (it is covered by ESKIMO) and its short length (2 pages). A shorter paper is likely to be unable to refer to all relevant material, and therefore provide more scope for the participants to discover uncited work.

The details of the paper selected for the evaluation are outlined below. The paper discusses an approach that uses hypertext to capture issues and events in organisational meetings. The ACM has only classified the paper as being about hypertext 'Theory', although the themes ontology enables this to be more accurately specified as 'Argumentation'. The paper also covers hypertext authoring, as the approach uses hypertext to author conceptual maps.

Title	Facilitated hypertext for collective sensemaking: 15 years on from gIBIS
Authors	Jeff Conklin, Albert Selvin, Simon Buckingham Shum, Maarten Sierhuis
Published	ACM Conference on Hypertext and Hypermedia 2001
ACM Classification	Hypertext/Hypermedia $\mapsto$ Theory

Before using ESKIMO to evaluate the paper, participants were asked to judge their confidence in the subject area of the paper and their ability to provide an accurate review. They were asked to make these judgments again after completing the review. This helped establish if using ESKIMO raised participants' confidence. They were also asked to provide additional feedback and comments on the system in general, and on its suitability for the particular task. The evaluation instructions are provided in Appendix B.

9.8.2 Trial and Hypothesis

The experiment was initially trialled by the author to appreciate the extent to which ESKIMO could be used for evaluating the selected paper. The trial concluded that while it was difficult to use ESKIMO alone to determine a paper's noteworthiness (this requires experience and judgment), ESKIMO could be used to position the work presented in the paper to other work within the research field. ESKIMO also highlighted several projects and papers that while significant and related to

the work in the paper, were not mentioned by it. The following points outline the major conclusions discovered about the paper after using ESKIMO in the trial evaluation.

1. gIBIS is the major project mentioned in the paper. ESKIMO confirms that this is a high impact paper. The paper "What's Eliza Doing In The Chinese Room? Incoherent Hyperdocuments And How To Avoid Them", cites the main gIBIS paper and appears to be a significant related paper that discusses hypertext authoring and structure, but its work is not mentioned by the evaluation paper.
2. The evaluation paper cites the Notecards project. ESKIMO is used to determine that this is a high impact and significant project and that one of the project's members is Frank Halasz. This researcher also works on the Aquanet project (according to ESKIMO a very high impact activity) which is a knowledge tool to capture structured tasks, and is also highly relevant to the evaluation paper.
3. The SEPIA activity is frequently cited by papers on the Notecards and gIBIS work. The abstract of a SEPIA paper notifies that this project is about cooperative hypermedia authoring and is also worth considering in respect to the evaluation paper.
4. As the evaluation paper is about hypertext authoring, the theme is set to 'authoring'. The available activities for this theme are then inspected. The activities ABC (Artifact-Based Collaboration) and AAA (Author's Argumentation Assistant) appear relevant. ESKIMO also concludes that the AAA and SEPIA activities are collaborating projects. Upon further inspection, this appears likely as they are both based at the same university and have an overlapping membership. The evaluation paper does not cite either project.
5. As ESKIMO proposed many experts for the theme 'Authoring', the theme is modified to 'Argumentation' (a sub-theme of the ACM 'Theory' classification) and fewer researchers are proposed. Two new activities are discovered that these experts work on: Phidias and Janus. While these projects are about hypertext structure, they are less relevant than the previously discovered activities. Viewing the available publications within this theme, the

Participant	Confidence before		Confidence after	
	Subject	Review	Subject	Review
P1	1	1	3	3
P2	3	3	4	4
P3	2	2	3	3
P4	3	2	2	2
P5	2	1	1	1
Mean	2.2	1.8	2.6	2.6

Table 9.5: Confidence values for the ESKIMO evaluation

paper entitled 'Arguments In Hypertext: A Rhetorical Approach' appears a highly relevant paper on the subject that the evaluation paper does not cite.

6. With the theme set to 'Authoring', the available research teams are explored. Again the above-mentioned activities appear, in particular the Phidias and AAA projects.

The trial demonstrated that ESKIMO could be used to evaluate the paper and identify further research that it fails to mention. In addition to this, ESKIMO was also useful in further understanding the material in the paper and helped position it within existing research.

Therefore, the hypothesis for this evaluation is that participants will be able to use ESKIMO to evaluate the paper and make several conclusions about its merit. However, it is envisaged that participants will be unsure, at least initially, in their operation of ESKIMO. This is because ESKIMO presents a new method of viewing and exploring scholarly material, which demands time to learn and appreciate. This factor is also inclined to affect confidence ratings, which are likely to remain constant.

9.8.3 Results

Table 9.5 lists the confidence values for each participant before and after the task. The subject area confidence values increased for participants 1, 2, and 3, and decreased for participants 4 and 5. The confidence in reviewing the paper increased or remained the same for all participants. Overall, the confidence values increased.

Participant 1 found ESKIMO effective and noted that it gave a 'useful feel for people, expertise, organisations, etc.' and added that it 'reinforced my rusty knowledge of some of these people and their work'. This is confirmed by both her confidence ratings increasing. The participant combined ESKIMO with occasional

Web searches to find ancillary information not captured by ESKIMO, such as the full-texts of papers. However, the participant did note that to complete a full review of the paper more disciplines than hypertext would have to be examined.

The participant managed to locate several further papers relevant to the topic of the paper. For example, 'Accessing Hyperdocuments Through Interactive Dynamic Maps' and 'Augmenting Human Intellect: A Conceptual Framework'. In addition, the projects DOLPHIN, Guided Tours, and Aquanet were discovered through investigations of related research teams, papers, experts, and impact factors. Although the participant noted the effectiveness of collaboration measures, these could not be used to any great effect in the context of this evaluation.

Participant 2 also noted that their confidence in the subject area and their ability to review the paper increased. The participant noted several relevant projects discovered by further investigating the gIBIS project, such as AAA, ABC, and Sepia. Experts were identified as Conklin, Schuler, and Smith and three further papers on gIBIS were noted, as well as the papers 'Authors Argumentation Assistant: A Hypertext Based Authoring Tool For Argumentative Texts' and 'KMS: A Distributed Hypertext System For Managing Knowledge In Organisations'. After exploring the subject area of the paper, the participant decided that the paper represented novel material as 'it brings it up to date as most of the [existing] work is pre-97'.

Participant 3 found the organisation of material useful and commented that *even* more information 'about/around key concepts' would further improve ESKIMO. This person's confidence ratings also increased. gIBIS was chosen as the main starting point for most explorations. Several systems were discovered: Aquanet, AAA, SEPIA, Phidias. The seminal publications were also retrieved within the topic area of the paper to determine significant papers not mentioned, such as the KMS paper noted by participant 2.

The confidence level in the subject area decreased for participant 4. This participant noted that there was 'no way into ESKIMO' and stated that they preferred an information retrieval approach, such as a search engine, to locate information. Their complaint centres around the fact that ESKIMO does not provide a 'seed' to start an exploration session. ESKIMO requires the scholar to provide some initial

facts (e.g. a person's name, a project title, an institute) and then use these to explore and discover further related information. Nevertheless, the participant used ESKIMO to identify the experts in the field and noted that the projects, Xanadu, CoVer, and Multicard were not mentioned by the paper.

Participant 5 also felt that their confidence in the subject area decreased after using ESKIMO. However, the participant stated that this was only because using ESKIMO demonstrated that much of the subject matter was, contrary to the initial belief, actually *unfamiliar* and 'as a result I feel less confident of my knowledge of this area'. While the participant found ESKIMO an effective tool which enabled 'making connections with aspects of the paper', the lack of up-to-date information in ESKIMO (the conference series used to populate ESKIMO covered 1988 to 2000) and it only containing links to the full-texts of hypertext conference papers, made it difficult for the participant to draw from recent advances.

9.8.4 Discussion of Results

Overall, the participants found ESKIMO to be an effective tool for exploring and further understanding a research field and were therefore able to comment on the paper's merit. The results demonstrated that they managed to identify several projects and papers that the evaluation paper failed to mention, and as a result, the confidence rating for three of the participants increased. Both the navigation and reasoning parts of ESKIMO were extensively used, especially questions to determine the significant researchers, papers, activities, and organisations within a field.

One participant noted the limited application area of ESKIMO as it only covers hypertext material over 12 a year period. A fully deployed version of ESKIMO would address this problem. More significantly, one participant noted that they preferred an information retrieval approach. This is an important issue as it indicates a shift in thinking is required by some scholars to prompt them to use an exploration environment over the conventional search engines many researchers are used to. ESKIMO requires scholars to take a more active role in their research and information gathering, and spot developments and further material and *act* on them, rather than just locating an isolated document through a Web search engine and only consuming its content and not examining directly related material.

The results suggest that further development and evaluation of ESKIMO are worthwhile. The main improvement suggested by the participants, the coverage provided by ESKIMO, could be addressed by making ESKIMO dynamically updateable (c.f. ResearchIndex (Lawrence *et al.*, 1999a), CiteBase (Brody *et al.*, 2001a)) and thereby allowing scholars to quickly spot new developments and reassess the fundamentals of a field. However, such an approach depends on further advances in either the sharing of scholarly metadata or in the facilities to automatically capture it.

9.9 Summary

ESKIMO is a scholarly support tool, which expands on the principles of ontological hypertext introduced in WSS and OntoPortal, and introduces reasoning facilities to enable scholars to explore their research field and make pertinent questions about it, and therefore become more informed about their community. The study described in Chapter 6 indicated that responding to typical research questions was time consuming, required the perusal of multiple resources, and, especially for analytical questions, was difficult. ESKIMO provides a semantic network over scholarly resources to enable researchers to locate related material quickly and efficiently, and analytical questions are supported through the inference engines.

ESKIMO enables scholars to explore their research field and follow paths they would not necessarily have considered, and then ask involved questions about the field. This helps support a scholar's high level cognitive thought processes, enabling them to ask *why* (Why is this document significant?) and *what* (What else is related to this project?). This is in contrast to current Web scholarly services such as offered by digital libraries, which improve access to scholarly material, but remain disconnected and lack the required semantics and metadata to enable further machine processing.

ESKIMO is particularly suitable for scholars beginning their research in a new field, such as a postgraduate student or a researcher involved in an interdisciplinary activity, who are required to quickly grasp a complex field and become proficient in it. For example, ESKIMO enables them to discover the experts and significant publications and projects. These can then be further explored to locate related material and position them with respect to the field. However, ESKIMO is also

suitable for experienced researchers who use it to explore paths and material they have not yet considered and as an instrument to support their review process (as demonstrated by the evaluation).

While ESKIMO could conceivably provide a new framework for the organisation and publication of scholarly material in digital libraries and e-journals, this is not proposed and its main objective is to act as a support service or portal/subject gateway. Section 9.7 proposed methods of seamlessly integrating ESKIMO into a scholar's task, such as browsing a digital library. In this case, ESKIMO provides a meta-layer over the underlying information in the digital library, and provides scholars with an in-depth service to improve their understanding of *all* connected scholarly material and not just the literature.

ESKIMO is an application and demonstration of the Semantic Web; it captures, represents, and publishes knowledge. The five lower layers of the Semantic Web architecture (Figure 3.5) are supported (unicode, both schema layers, ontology, and logic), but ESKIMO does not support the proof and trust layers, nor the digital signature facility. This is partly due to the lack of available documentation on the exact purpose and nature of these layers, and partly because they would not add to the principles described in this research.

In addition, the ontological metadata used by ESKIMO can be made available for other services to use. For example, the *Output Controller* can be switched to output raw XML, which provides other processes with machine-readable metadata. Furthermore, the controller can be modified to return the information as ontological metadata for other knowledge services.

The current lack of available metadata and interoperability facilities in publishing mediums, results in significant authoring and maintenance overhead in constructing ESKIMO. Section 9.4 discussed the laborious task of extracting and cleansing the data from the ACM Hypertext Conference series to produce the ontological knowledge. At the moment, this overhead results in ESKIMO not being readily portable to other research fields or being used in an open and unstructured environment, although the emergence of standards like the OAI (Lagoze & de Sompel, 2001) and AMF (Brody *et al.*, 2001b) are aimed at addressing this issue.

ESKIMO is a scholarly application for the Semantic Web and demonstrates advantages of adopting and publishing ontological metadata to offer scholarly services. It offers a novel approach to viewing, exploring, and understanding a research field that removes the restrictions of access and inquiry in traditional paper-based research, and the constraints and limitations of scholarly interlinking and search facilities on the current Web. ESKIMO also promotes the creation, distribution, and use of scholarly metadata and illustrates the benefits this affords.

Chapter 10

Conclusions and Further Work

10.1 e-Scholars in the Semantic Web

This research has explored and demonstrated a novel approach for supporting scholarly research on the Web; by integrating principles from hypertext and the Semantic Web, scholars can comprehensively explore their research field and ask questions about it.

10.1.1 Summary

Scholarly research involves the exploration of the academic domain and the various artifacts evident in it. These form a network of associated concepts and issues that scholars use to traverse a research field and understand it. Traditionally, scholars have focused on literature and the citations between them to explore their field. With the advent of the Web however, the possibility of providing instantaneous and interconnected access to all scholarly material becomes possible. Unfortunately, current publishing mediums have failed to exploit hypertext and the interlinking of the academic domain to its full potential, as well as providing services to support inquiry to allow scholars to ask pertinent questions about their field. Researchers are therefore required to use detective skills, intuition, and significant effort to locate relevant material and then attempt to position it within the context of the discipline.

Traditionally, scholars relied on slow, and sometimes ineffective, peer-to-peer communication for resolving their questions and conducting research. Indeed, Bush (1945) noted the "growing mountain of research" and realised that the plethora of

publications inhibited scholars making "real" use of it. The Web adds immediacy and the facility to find the answers to their research questions. However, as the experiment in Chapter 6 demonstrated, the Web still does not measure up to the requirements of scholarly research.

The ESKIMO system was implemented to address these issues and provide researchers with a comprehensive research environment that enables them to understand the facts and concepts in their research field and make pertinent questions about them.

Constructing scholarly hypertext is difficult and error prone (Baragar, 1995; Theng, 1999; Mendes *et al.*, 2001) and therefore ESKIMO exploits the *ontological hypertext* principle to structure the scholarly hypertext. Ontological hypertext uses an ontological representation (a explicit conceptualisation of a domain) of the scholarly community in the hypertext layer to *control* and *constrain* the linking between research material and contribute to reducing information and cognitive overload, and disorientation. It is useful in interconnecting complex research domains as it constructs a principled and consistent hypertext based on the real-life relationships specified in the ontology. It was explored on a large scale in OntoPortal to create real-world research portals and in ESKIMO to interlink research material from the ACM Hypertext Conference series.

The scholarly ontology also provides the mechanism for reasoning services to answer typical research questions. By analysing the ontological metadata captured, inferences are realised to uncover new information and patterns, and make plausible suggestions about the research field. For example, the discovery of collaborations, impact factors, and co-citations are possible, in addition to extensions to these that draw from other artifacts in the academic domain (e.g. research team collaboration, organisation impact, project co-citation)

The inference engines also *abduct* (the process of making plausible suggestions) over a research field to determine significant researchers, publications, projects, journals, and organisations. This is particularly useful to new postgraduate students or experienced researchers involved in an interdisciplinary project, as it enables them to quickly identify the salient and prominent work and position it within the context of the research field, resulting in a more informed scholar.

10.1.2 Influencing Work

This research was particularly influenced by earlier work in hypertext and the Semantic Web. Hypertext provides the mechanism to interconnect related material on the Web. Early hypertext systems failed to exploit the obvious semantic nature of hypertext links, however systems such as Textnet (Trigg, 1983), gIBIS (Conklin & Begeman, 1988), and Aquanet (Marshall *et al.*, 1991) introduced link semantics that enabled interlinked documents to be more effectively presented to users who were then able to predict the effect of traversing a link. In fact, Trigg believed that eventually all scientific activity would move on-line and typed hypertext would provide the necessary infrastructure to support it (Trigg, 1983).

Open hypertext systems, such as Microcosm (Fountain *et al.*, 1990) and Chimera (Anderson *et al.*, 1994), demonstrated the use of abstracting the various hypertext constructs to enable the interchange and processing of hypertext data. Link abstraction is useful as it allows links to be considered as independent semantic specifications, a principle exploited in ontological hypertext.

Hypertext also introduces a new paradigm for searching and locating information: navigation. Unlike information retrieval, it enables users to browser a subject area and home in on required information. This method of information discovery is especially suitable for scholars who often explore a research field without details of a specific paper or researcher.

The Semantic Web describes an extension to the Web where machines understand the concepts and relationships described within Web documents. Early metadata standards were too basic to accurately represent and constrain a domain, and were only useful in primitive indexing facilities. Ontologies, on the other hand, provide an explicit conceptualisation of a domain in a rigorous and expressive way. They are an integral part of the Semantic Web, where they are used to accurately describe Web resources. ESKIMO uses ontological metadata to accurately represent the knowledge contained in scholarly material and then provide ontological hypertext and scholarly inquiry.

Finally, the field of digital libraries has been influential for understanding scholarly activities and in identifying current support levels on the Web. The field has introduced methods of analysing scholarly data, such as bibliographies, to provide

scholars with an overall impression of their field. Projects have also explored how scholarly Web facilities, such as discourse (Shum & Sumner, 2001), publication (Harnad, 1995a; Harnad *et al.*, 1999), and interlinking (Harnad & Carr, 2000) can be supported and implemented. These have been influential in understanding how the Web can be adapted for research activities.

10.1.3 Where are we heading?

The principal remaining issue is the *interoperability* between publishing mediums and institutional websites to promote the use and availability of scholarly metadata. While ESKIMO demonstrated the benefits that can be achieved by having access to large quantities of accurate metadata, it also exposed the enormous authoring overhead required to create it in the first place. Therefore, until interoperability and the required infrastructure improve, advanced scholarly services are likely to remain restricted to specific applications and limited domains.

Interoperability also enables scholarly services to dynamically update their metadata and provide up-to-date information and a wide-coverage for a research field. Scholars can then track developments in their field *as they occur* and have an advantage over less informed researchers. However, this also requires scholars to become more proactive and involved in their research field, and identify emerging material and developments as they happen. Rather than simply locating and printing out a paper to read, scholars endeavour to always uncover further relevant material.

In addition, more effective inquiry to answer scholarly questions is required. While bibliometric patterns and scholarly community abductions provide a useful insight into particular aspects of a research field, they still require scholars to filter through and disseminate the returned information, and then decide which trails are worth further investigation. However, more accurately targeted or adaptively controlled scholarly inquiry, would reduce this effort.

10.2 Contributions

10.2.1 Does the Web support scholarly activity?

An experiment to asses how well the Web supports a scholar's information seeking tasks was completed. Participants were observed as they used the Web to answer typical research questions. The results of the experiment indicated that problems

do exist for efficiently and effectively using the Web to answer research questions, especially analytical questions that require scholars to draw from and associate information. It was also evident that scholars rely heavily on search engines, and less on digital libraries and other scholarly tools, indicating that these tools are ineffective, too restricted or limited, or that more training is required.

10.2.2 The Scholarly Community

Scholarly research involves more than reading isolated publications. A research field includes many artifacts that assist scholars in their exploration. However, traditionally these have not been comprehensively exploited, partly because of accessibility issues, and partly because of a lack of infrastructure to interconnect them. ESKIMO demonstrates how the objects can be interlinked to provide researchers with the facility to seamlessly explore *all* artifacts within their research field.

10.2.3 Interlinking Research Material using Ontological Hypertext

Ontological hypertext provides a principled approach to interlinking highly complex research fields and enables scholars to navigate research material comprehensively and intelligently; it was demonstrated in OntoPortal and ESKIMO. It also promotes serendipitous browsing as scholars explore paths and associations that they may otherwise have ignored or not considered.

10.2.4 Supporting Scholarly Inquiry

The scholarly community ontology affords machine analysis to uncover useful facts and patterns. This is used in ESKIMO to allow scholars to examine their research field with common bibliometric rules, and inferences that draw from the entire scholarly community to make plausible suggestions about the prominent researchers, papers, organisations, and research teams. Resolving these types of questions in the traditional paper-based world is either difficult, time consuming, and relies on the knowledge and experience of peers, or is impossible.

10.2.5 Scholarly Research in the Semantic Web

Ontological hypertext and scholarly inquiry provide scholars with a comprehensive research environment that enhances their view on a field and presents them with knowledge not previously available; scholars are more informed.

An evaluation of ESKIMO indicated that researchers benefit from these services, although a shift in scholarly practices may be necessary, as ESKIMO requires researchers to take a more active role in disseminating material.

Furthermore, the work demonstrates the tangible benefits of using scholarly metadata and Semantic Web principles, and therefore promotes the creation, publication, and reuse of such knowledge.

10.2.6 An Integration Exercise

This research has drawn from three distinct areas in exploring methods of supporting scholars: hypertext, Semantic Web and knowledge technologies, and digital libraries. This demanded an understanding of these disparate disciplines to *integrate* them and provide an effective scholarly support service.

10.3 Further Work

The work presented in this thesis has scope for further research and these possibilities are presented.

10.3.1 Who are the experts?

ESKIMO does not provide canonical methods to resolving research questions, rather it introduces a potential framework. Resolving queries, such as *Who are the experts?*, requires investigation and research. It could be defined based on the publications researchers produce and how these are received in the community. However, frequently researchers are named on a paper for political reasons, although they might only be partly knowledgeable on the subject and have little technical experience. Should these researchers still be considered experts? Therefore, many approaches are likely to return potential experts, though whether all peers agree on the result is uncertain. Research is required to understand the exact nature of scholarly questions and how they can be resolved.

10.3.2 The Knowledge Cycle

The management of knowledge involves six stages that a knowledge application has to address: acquisition, modelling, maintenance, retrieval and extraction, reuse, and publishing.

Knowledge *acquisition* is the capture of relevant knowledge for a particular task. OntoPortal provides an authoring interface to add the ontological knowledge to the system. Although this promotes a community driven approach to constructing research portals, the task is laborious and authors must carefully select and recognise the concepts and relationships they wish to add. Alternatively, ESKIMO employs a semi-automatic method that converts raw data about hypertext papers to ontological metadata. Advances in scholarly metadata (e.g. AMF (Brody *et al.*, 2001b)) and better metadata extraction (scraping) tools (Bergmark *et al.*, 2001) are required before knowledge acquisition of scholarly metadata dramatically improves.

Knowledge *modelling* was used to derive the scholarly community ontology. However, changes to this representation are not automatically supported by OntoPortal or ESKIMO. Instead, modifications require transformations to the ontological metadata and the inference rules, meaning serious *maintenance* implications are raised.

The ontological knowledge could also be dynamically *retrieved and extracted* from papers and other scholarly material as they are published, allowing ESKIMO to always present scholars with up-to-date information and enabling them to spot developments as they happen.

Knowledge *reuse* is a vital constituent of any knowledge system as it significantly reduces the authoring overhead. For example, ESKIMO reused part of the ACM classification index for the themes ontology. Furthermore, the captured metadata in ESKIMO is available for other processes to reuse.

ESKIMO *publishes* the ontological knowledge and presents it to scholars in an interconnected way. Inquiry facilities are also provided to reason over the knowledge and present researchers with a particular view of it.

The six knowledge tasks are applicable to ESKIMO and raise issues and implications that demand further research. The AKT project is exploring methods of managing these stages (Shadbolt, 2001a) which will contribute to these issues and the Semantic Web in general.

10.3.3 Supporting other Research Areas

ESKIMO has been populated with hypertext data from the ACM Hypertext Conference series. Due to the significant effort this process demanded and the lack of

available metadata from other disciplines, ESKIMO was not populated with knowledge from further research fields. Therefore, the scholarly ontology could not be evaluated to determine how well it supports other disciplines, especially disciplines that differ significantly from hypertext, such as the classics or sociology. This may have concluded that different scholarly ontologies are indeed necessary for providing advanced services.

10.3.4 Temporal Aspects

A significant exclusion from the scholarly ontology is the aspect of time. The ontology does not account for temporal changes, such as researchers moving to other organisations. When the hypertext data were captured, no inspections or analyses were completed to determine whether the data were conceptually valid. Therefore, situations appear where researchers are listed as being members of several organisations. Related to this, name changes (e.g. 'Multimedia Research Group' to 'IAM Group') are not accounted for, meaning instances and relationships are divided across the different versions, causing confusion to scholars and affecting the results of scholarly analyses.

A solution is to use inferences to identify anomalies either as new metadata is added, or at regular maintenance intervals. For example, if a publication from 1992 states in the affiliation section that the author works at the 'Multimedia Research Group', and a paper published in 2002 states that the same author works at 'Siemens Corporate Research', then it could be inferred that the researcher has moved.

10.3.5 Facts and Issues

ESKIMO provides semantic access to the facts in a research field. However, it does not include the issues and opinions made by researchers, which are evident in the text of scholarly articles and indicate the inter-social climate of a research field. However, by combining the knowledge of this (e.g. ScholOnto (Shum *et al.*, 1999)) together with the facts and contributions of a research field, an in-depth understanding of a discipline is possible. Researchers can ask insightful questions such as *Do the researchers that collaborate with Tim Berners-Lee share his views on the Semantic Web?*

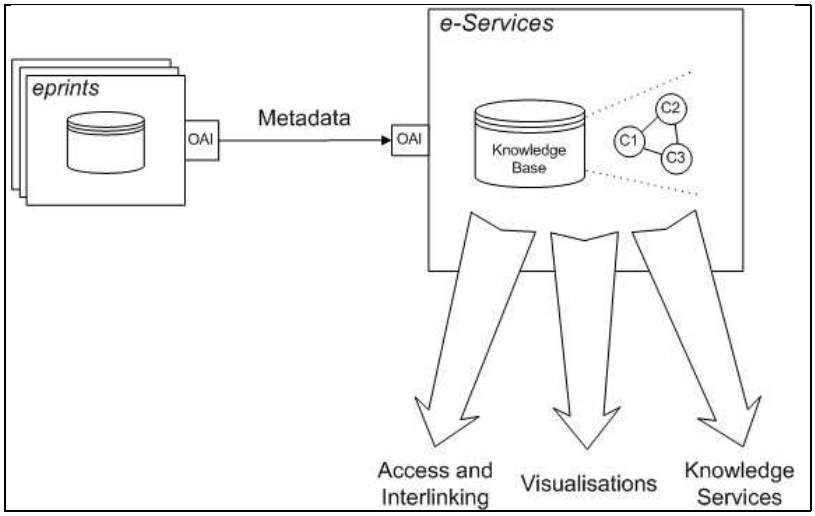


Figure 10.1: Scholarly e-Services

10.3.6 Scholarly Services

Interoperability standards, such as the OAI (Lagoze & de Sompel, 2001), enable large amounts of scholarly metadata to be made publicly available. This is used in the OpCit (Harnad & Carr, 2000) project to provide 'e-Services' to e-prints archives (Figure 10.1). The 'e-Services' connects to document repositories and through the OAI interoperability standard requests scholarly metadata. This is then internally represented in a knowledge representation format and services are exposed to the researcher, such as ontological hypertext, visualisations like co-citation networks, and knowledge services like scholarly inquiry. Scholars simply point the service to a location of an archive, and then use advanced facilities to access and explore the material.

Indeed, the work from ESKIMO is being applied to the 'e-Services' framework in OpCit and in the AKT (Shadbolt, 2001a) project for an advanced research portal.

10.4 Concluding Statement

Scholarly research on the Web has yet to reach its full potential; while accessibility has dramatically improved, comprehensive interconnectivity and inquiry have yet to emerge. The work presented in this thesis proposes a novel approach to providing these services. ESKIMO immerses scholars in knowledge about their research community and allows them to make intricate and pertinent questions. As a result, the e-Scholar is more informed to make better contributions.

Continued advances in interoperability, scholarly metadata, and the adoption of Semantic Web principles, will see dynamic scholarly services emerge to provide unique facilities not available in paper-based research or on the current Web. This will dramatically change the way scholars interact with and disseminate material.

Negroponte (1995) notes a shift from the industrial age governed by atoms (i.e. *physical* things), to an information age governed by bits (i.e. *digital* information). This transformation is evident in research as e-Scholars adopt the Web as their primary research tool and depend on it to discover scholarly knowledge. For the first time, research can be instantaneously accessible, interconnected, and analysed.

> "The web is changing the way that researchers locate and access scientific publications" (Lawrence & Giles, 1999)

At the centre of all research is the consumption and production of knowledge; any tool that improves this process will make research more productive. As the printing press profoundly changed the way papers were distributed and disseminated and marked the beginning of the first scientific revolution (Eisenstein, 1979), the Web has also changed the way scholars "disseminate knowledge; retrieve knowledge; and acquire knowledge" (Dewar, 1998) and has provided the platform for a profound new research environment. *Are we at the beginning of the next Scientific Revolution?*

Appendix A

Experiment 1 Instructions

How do Scholars Use the Web?

This experiment has been designed to get an impression of how scholars use scholarly data on the Web. This does not just include the techniques you employ to find information, but also the type of information you use.

While you are working through the tasks, your browsing activity will be recorded by a proxy. Any comments you wish to make during the experiment should be made using the note applet.

Thank you.

Guidelines

- Use only the web to answer the questions, even if you have prior background knowledge. However, if the latter is the case, make a note of this.
- Try not to spend more than about 5 minutes on any step. Give up and make a note explaining the difficulties you encountered.
- There is no significance to the projects, researcher, etc. chosen in the questions.
- Answer the questions using the note applet and add any thoughts that you have during the process. If you feel the task is impossible to answer or represents a type of question that you would never consider asking, indicate this using a note.

Details

Proxy host	pip.ecs.soton.ac.uk
Proxy port	3000
Proxy traffic	use the proxy only for HTTP traffic, not HTTPS
Note applet	http://pip/ta/message.html

Questions

Task 1

1. Locate a noteworthy paper on the Ontobroker project.
2. Was this paper ever presented at a conference, and if so, which one?
3. Are there any other related papers at this conference?
4. What projects are related/similar to Ontobroker?
5. Who are the researchers that are part of this project, and where do they work?
6. What other project has the institute produced?

Task 2

1. Eugene Garfield is probably the most prominent researcher within the field of citation analysis. Find one of his seminal papers and explain why you believe the paper to be seminal?
2. Which institute participates in significant ontology research?
3. Broadly speaking, how has the perspective of hypertext changed over the last decade?

Appendix B

Experiment 2 Instructions

Hypertext Paper Evaluation

Instructions:

The objective of this evaluation is to determine how the additional scholarly knowledge provided by ESKIMO helps researchers have a better understanding of the issues and research raised in papers.

The evaluation first requires you to read a paper and judge how confident you would feel in providing feedback for it. Then use the ESKIMO system to evaluate the research quality of the paper provided (e.g. is the literature review complete) and attempt to position it with respect to other research in the area to help determine if it represents worthwhile research. When you have completed this, indicate how confident you now feel in your evaluation of the paper.

ESKIMO is a *support* tool so the WWW may also be used as an auxiliary tool. For example, use it to find information about a particular software system. However, please do not use the WWW to directly help evaluate the paper (e.g. by searching for '*systems similar to x*' in Google).

ESKIMO is accessible from within the department at:
http://tractor.ecs.soton.ac.uk/eskimo/cgi-bin/get.cgi

Step 1 – Read the paper

Read the paper and complete the following statements.

1. I feel sufficiently informed about the subject to comment on this paper.

0	Disagree completely
1	Disagree strongly
2	Disagree
3	Agree
4	Agree strongly
5	Agree completely

2. I feel confident about the feedback I can provide on this paper.

0	Disagree completely
1	Disagree strongly
2	Disagree
3	Agree
4	Agree strongly
5	Agree completely

Step 2 - Use ESKIMO to evaluate the paper

Using ESKIMO, evaluate the content of the paper (e.g. does it cite all relevant research) and assess its research quality. While completing the task, please indicate all your reasoning and how you came about it.

Step 3 – Rate the confidence in your evaluation

Complete the following statements.

1. I feel sufficiently informed about the research area of this paper to comment on it.

0	Disagree completely
1	Disagree strongly
2	Disagree
3	Agree
4	Agree strongly
5	Agree completely

2. I feel confident about the accuracy and correctness of my feedback for this paper.

0	Disagree completely
1	Disagree strongly
2	Disagree
3	Agree
4	Agree strongly
5	Agree completely

3. Any other comments? How could your review/feedback have been improved?

Appendix C

Scholarly Community Ontology represented in RDFS

The RDFS used for the representation is based on the March 2000 RDFS 1.0 specification, available at `http://www.w3.org/TR/2000/CR-rdf-schema-20000327/`.

```
<?xml version="1.0"?>

<rdf:RDF xml:lang="en"
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
   xmlns:themes="file://c:/PhD/ontologies/web_page/s_themes#">

<rdf:Description ID="Thing">
 <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
 <rdfs:subClassOf
  rdf:resource="http://www.w3.org/2000/01/rdf-schema#Resource"/>
</rdf:Description>

<rdf:Description ID="Person">
 <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
 <rdfs:subClassOf rdf:resource="#Thing"/>
</rdf:Description>

<rdf:Description ID="Team">
 <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
 <rdfs:subClassOf rdf:resource="#Thing"/>
</rdf:Description>

<rdf:Description ID="Organisation">
 <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
 <rdfs:subClassOf rdf:resource="#Thing"/>
</rdf:Description>

<rdf:Description ID="Activity">
 <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
 <rdfs:subClassOf rdf:resource="#Thing"/>
</rdf:Description>

<rdf:Description ID="Society">
 <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
 <rdfs:subClassOf rdf:resource="#Thing"/>
```

```
</rdf:Description>

<rdf:Description ID="Journal">
 <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
 <rdfs:subClassOf rdf:resource="#Thing"/>
</rdf:Description>

<rdf:Description ID="Publication">
 <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
 <rdfs:subClassOf rdf:resource="#Thing"/>
</rdf:Description>

<rdf:Description ID="Published_Paper">
 <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
 <rdfs:subClassOf rdf:resource="#Publication"/>
</rdf:Description>

<rdf:Description ID="Book">
 <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
 <rdfs:subClassOf rdf:resource="#Publication"/>
</rdf:Description>

<rdf:Description ID="Thesis">
 <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
 <rdfs:subClassOf rdf:resource="#Publication"/>
</rdf:Description>

<rdf:Description ID="Technical_Report">
 <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
 <rdfs:subClassOf rdf:resource="#Publication"/>
</rdf:Description>

<rdf:Description ID="Publication_Medium">
 <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
 <rdfs:subClassOf rdf:resource="#Thing"/>
</rdf:Description>

<rdf:Description ID="Conference">
 <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
 <rdfs:subClassOf rdf:resource="#Publication_Medium"/>
</rdf:Description>

<rdf:Description ID="Journal_Entry">
 <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
 <rdfs:subClassOf rdf:resource="#Publication_Medium"/>
</rdf:Description>

<rdf:Description ID="partOf">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Person"/>
 <rdfs:range rdf:resource="#Team"/>
</rdf:Description>

<rdf:Description ID="representedIn">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Organisation"/>
 <rdfs:range rdf:resource="#Publication_Medium"/>
</rdf:Description>

<rdf:Description ID="produces">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Activity"/>
 <rdfs:domain rdf:resource="#Organisation"/>
 <rdfs:domain rdf:resource="#Person"/>
 <rdfs:domain rdf:resource="#Team"/>
 <rdfs:range rdf:resource="#Publication"/>
```

```
</rdf:Description>

<rdf:Description ID="sponsoredBy">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Conference"/>
 <rdfs:range rdf:resource="#Society"/>
</rdf:Description>

<rdf:Description ID="hasTheme">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Publication"/>
 <rdfs:range rdf:resource="file://c:/PhD/ontologies/web_page/s_themes#Research_Theme"/>
</rdf:Description>

<rdf:Description ID="hasReference">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Publication"/>
 <rdfs:range rdf:resource="#Publication"/>
</rdf:Description>

<rdf:Description ID="publishedIn">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Activity"/>
 <rdfs:domain rdf:resource="#Published_Paper"/>
 <rdfs:range rdf:resource="#Publication_Medium"/>
</rdf:Description>

<rdf:Description ID="isIn">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Journal_Entry"/>
 <rdfs:range rdf:resource="#Journal"/>
</rdf:Description>

<rdf:Description ID="edits">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Person"/>
 <rdfs:range rdf:resource="#Journal"/>
</rdf:Description>

<rdf:Description ID="sitsOn">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Person"/>
 <rdfs:range rdf:resource="#Society"/>
</rdf:Description>

<rdf:Description ID="worksAt">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Person"/>
 <rdfs:range rdf:resource="#Organisation"/>
</rdf:Description>

<rdf:Description ID="worksOn">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Person"/>
 <rdfs:range rdf:resource="#Activity"/>
</rdf:Description>

<rdf:Description ID="areaOfWork">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Person"/>
 <rdfs:range rdf:resource="file://c:/PhD/ontologies/web_page/s_themes#Research_Theme"/>
</rdf:Description>

<rdf:Description ID="runBy">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Activity"/>
```

```
 <rdfs:range rdf:resource="#Team"/>
</rdf:Description>

<rdf:Description ID="tackles">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Team"/>
 <rdfs:domain rdf:resource="#Activity"/>
 <rdfs:range rdf:resource="file://c:/PhD/ontologies/web_page/s_themes#Research_Theme"/>
</rdf:Description>

<rdf:Description ID="basedAt">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Activity"/>
 <rdfs:domain rdf:resource="#Team"/>
 <rdfs:range rdf:resource="#Organisation"/>
</rdf:Description>

<rdf:Description ID="title">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Activity"/>
 <rdfs:domain rdf:resource="#Publication_Medium"/>
 <rdfs:domain rdf:resource="#Journal"/>
 <rdfs:domain rdf:resource="#Publication"/>
 <rdfs:domain rdf:resource="#Organisation"/>
 <rdfs:domain rdf:resource="#Team"/>
 <rdfs:domain rdf:resource="#Society"/>
 <rdfs:domain rdf:resource="#Person"/>
 <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
</rdf:Description>

<rdf:Description ID="abstract">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Publication"/>
 <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
</rdf:Description>

<rdf:Description ID="uri">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Activity"/>
 <rdfs:domain rdf:resource="#Publication_Medium"/>
 <rdfs:domain rdf:resource="#Journal"/>
 <rdfs:domain rdf:resource="#Publication"/>
 <rdfs:domain rdf:resource="#Team"/>
 <rdfs:domain rdf:resource="#Society"/>
 <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
</rdf:Description>

<rdf:Description ID="email">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Person"/>
 <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
</rdf:Description>

<rdf:Description ID="fundingSource">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Activity"/>
 <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
</rdf:Description>

<rdf:Description ID="description">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Activity"/>
 <rdfs:domain rdf:resource="#Publication_Medium"/>
 <rdfs:domain rdf:resource="#Journal"/>
 <rdfs:domain rdf:resource="#Person"/>
 <rdfs:domain rdf:resource="#Team"/>
```

```
 <rdfs:domain rdf:resource="#Society"/>
 <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
</rdf:Description>

<rdf:Description ID="location">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Conference"/>
 <rdfs:domain rdf:resource="#Organisation"/>
 <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
</rdf:Description>

<rdf:Description ID="volume">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Journal_Entry"/>
 <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
</rdf:Description>

<rdf:Description ID="issue">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Journal_Entry"/>
 <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
</rdf:Description>

<rdf:Description ID="page">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Journal_Entry"/>
 <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
</rdf:Description>

<rdf:Description ID="year">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Journal_Entry"/>
 <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
</rdf:Description>

<rdf:Description ID="misc">
 <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#Journal_Entry"/>
 <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
</rdf:Description>

</rdf:RDF>
```

Appendix D

Scholarly Community Ontology Documentation

The convention used within this document to describe the ontology is discussed in (Uschold & Gruninger, 1996) and (Skuce, 1996).

Purpose

The purpose of the ontology is to model the scholarly community/academic domain. It will be used to create ontological hypertext and enable machine analysis of scholarly data.

Scope

In identifying the concepts within the community, a fine balance was drawn between the complexity and expressiveness of the ontology and the population process. The ACM Hypertext Conferences from 1988 to 2000 was used as the basis with which to create the knowledge base. Therefore, the ontology construction was also influenced by the data available.

The initial concepts were identified during group brainstorming sessions.

The concepts initially identified were:

committee, article, journal entry, journal, person, conference, activity, institute

with most having the following literals:

title, description, URI, misc

After several iterations of the ontology and more discussions, a final version was derived. This is described in the remainder of this document.

RDF Schema Representation

See 'RDF Schema Specification 1.0 W3C Candidate Recommendation 27 March 2000'

Scholarly Community - s_community.rdfs

Thing

Class

Info	**Description**
C	There is ONE top category that includes everything.
T	Thing. Sometimes referred to as object or entity but Thing is the preferred term.
D	Anything that can be thought of or referred to by some symbol meaningful to other people.
EX	MIT, Wendy Hall, hypertext, desk, MAVIS, ACM, pencil

Literals

Name	**Description**
none	none

Predicates

Name	**Description**	**Domain**	**Range**	**Cardinality**
none	none	none	none	none

Organisation

Class

Info	**Description**
C	A type of Thing. Also referred to as institute, company, research lab but organisation is the preferred term.
T	Organisation
D	An administrative and functional structure (as a business or a political party).
EX	University of Southampton, MIT, Ford Motor Company, Oxfam

Literals

Name	**Description**
title	The name associated with the organisation.
desc	A brief description.
uri	A web address where more information can be found regarding the organisation. Usually this will be a homepage.
location	The location of the organisation. Actual content is decided by the user, but standard postal address is the recommended minimum.
type	The type of work the organisation participates in. Options: Research, Commercial, Other

Predicates

Name	**Description**	**Domain**	**Range**	**Cardinality**
representedIn	Indicates the journal entry where an organisation is represented (e.g. through an author)	Organisation	Publication_ Medium	1:M
produces	Indicates the deliverables (e.g. reports, software) that the organisation produces.	Organisation	Publication	1:M

Society

Class

Info	**Description**
C	A type of Thing. Also referred to as association or club, but society is the preferred term.
T	Society
D	A voluntary association of individuals for common ends; especially: an organized group working together or periodically meeting because of common interests, beliefs or profession.
EX	ACM, IEEE.

Literals

Name	**Description**
title	The name associated with the society.
desc	A brief description.
uri	A web address where more information can be found regarding the society. Usually this will be a homepage.

Predicates

Name	**Description**	**Domain**	**Range**	**Cardinality**
None	None	None	None	None

Publication Medium

Class

Info	**Description**
C	A type of Thing. Also referred to as workshop, journal, conference, publishing medium, but publication medium is the preferred term.
T	Publication_Medium
D	A physical place or event to publish a scholarly work.
EX	Hypertext 2000, Communications of the ACM, IAM Seminar Series

Literals

Name	**Description**
title	The name associated with the conference desc A brief description
uri	A web address where more information can be found regarding the conference. Usually this will be a homepage

Predicates

Name	**Description**	**Domain**	**Range**	**Cardinality**
none	none	none	none	none

Conference

Class

Info	**Description**
C	A type of Thing. Also referred to as seminar, workshop, meeting, symposium, but conference is the preferred term.
T	Conference
D	A meeting of two or more persons for discussing matters of common concern
EX	Hypertext 2000, WWW10, IAM Seminar Series

Literals

Name	**Description**
location	The location of the conference. Actual content is decided by the user, but standard postal address is the recommended minimum
type	The type of conference.

Predicates

Name	**Description**	**Domain**	**Range**	**Cardinality**
sponsoredBy	Indicates the sponsors of the conference	Conference	Society	1:M

Journal Entry

Class

Info	**Description**
C	A type of Publication Medium
T	JournalEntry
D	An publication entry within a journal.
EX	Issue 43, pages 905-909, 1995

Literals

Name	**Description**
issue	The issue number of the journal containing the publication entry
volume	The volume number of the journal containing the publication entry
pages	The page numbers of the journal containing the publication entry
year	The year of the journal containing the publication entry

Predicates

Name	**Description**	**Domain**	**Range**	**Cardinality**
isIn	Indicates the journal in which this entry appears	Journal_ Entry	Journal	1:1

Publication

Class

Info	**Description**
C	A type of Thing. Also referred to as literature, but publication is the preferred term.
T	Publication
D	A published work
EX	Thesis, Conference Paper, Journal Paper

Literals

Name	**Description**
title	The name associated with publication.
desc	A brief description.
uri	A web address where more information can be found regarding the publication.

Predicates

Name	**Description**	**Domain**	**Range**	**Cardinality**
hasTheme	Indicates the research theme the publication covers	Publication	Research_ Theme	1:M
hasReference	Indicates the references a publication has.	Publication	Publication	1:M

Published Paper

Class

Info	**Description**
C	A type of deliverable.
T	Published_Paper
D	A published work.
EX	Thesis, Conference Paper, Journal Paper

Literals

Name	**Description**
none	none

Predicates

Name	**Description**	**Domain**	**Range**	**Cardinality**
publishedIn	Indicates the journal entry where the publication is published.	Publication	Publication_ Medium	1:M

Book

Class

Info	**Description**
C	A type of Thing. Also referred to as manual, user manual but book is the preferred term.
T	Book
D	A published work.
EX	Literary Machines

Literals

Name	**Description**
none	none

Predicates

Name	**Description**	**Domain**	**Range**	**Cardinality**
none	none	none	none	none

Thesis

Class

Info	Description
C	A type of Thing. Also referred to as dissertation but thesis is the preferred term.
T	Thesis
D	A published work representing research completed for doctoral work.
EX	a thesis

Literals

Name	Description
none	none

Predicates

Name	Description	Domain	Range	Cardinality
none	none	none	none	none

Technical Report

Class

Info	Description
C	A type of Thing. Also referred to as report, working paper, draft, scientific report but technical report is the preferred term.
T	Technical_Report
D	A technical report.
EX	a failed conference paper submission

Literals

Name	Description
none	none

Predicates

Name	Description	Domain	Range	Cardinality
none	none	none	none	none

Journal

Class

Info	Description
C	A type of Thing. Also referred to as magazine, periodial but journal is the preferred term.
T	Journal
D	A publication that appears at regular intervals.
EX	Communications of the ACM, Journal of The Electrochemical Society

Literals

Name	Description
title	The name associated with the journal.
publisher	Details about the publisher of this journal.
desc	A brief description.
uri	A web address where more information can be found regarding the journal. Usually this will be a homepage.
type	The type of journal. Options: Academic, Popular, Professional

Predicates

Name	Description	Domain	Range	Cardinality
none	none	none	none	none

Person

Class

Info	Description
C	A type of Thing. Also referred to as human, researcher, individual, but Person is the preferred term.
T	Person
D	Any living or extinct member of the family (Hominidae) to which the primate belongs.
EX	Bill Gates, Ted Nelson, Vannevar Bush

Literals

Name	Description
name	The full name, including title, associated with this person.
desc	A brief description.
uri	A web address where more information can be found regarding the organisation. Usually this will be a homepage.
email	The full email address of this person.
role	The person's work role. Options: Student, Technical Staff, Academic Staff

Predicates

Name	Description	Domain	Range	Cardinality
partof	Indicates the team that the person is a member of.	Person	Team	1:M
produces	Indicates the deliverables the person has produced.	Person	Publication	1:M
edits	Indicates the journal that the person is an editor of.	Person	Journal	1:M
sitsOn	Indicates the society that the person sits on.	Person	Society	1:M
worksAt	Indicates the organisation where the person works.	Person	Organisation	1:M
worksOn	Indicates the activity(ies) that the person works on.	Person	Activity	1:M
areaOfWork	Indicates the topic of work the person is involved in.	Person	Research_Theme	1:M

Team

Class

Info	Description
C	A type of Thing. Also referred to as group, syndicate, but Team is the preferred term.
T	Team
D	A number of persons associated together in work or activity.
EX	Equator Team, KMi Team

Literals

Name	Description
title	The name associated with the team.
desc	A brief description.
uri	A web address where more information can be found regarding the team. Usually this will be a homepage.

Predicates

Name	Description	Domain	Range	Cardinality
tackles	Indicates the topic of work that the team tackles.	Team	Research Theme	1:M
basedAt	Indicates the organisation where the team is based at.	Team	Organisation	1:M
produces	Indicates the deliverables produced by the team.	Team	Publication	1:M

Activity

Class

Info	Description
C	A type of Thing. Also referred to as project, exercise, but activity is the preferred term.
T	Activity
D	A planned undertaking.
EX	PhD work, AKT project, ScholOnto Project

Literals

Name	Description
title	The name associated with the activity.
desc	A brief description.
uri	A web address where more information can be found regarding the activity. Usually this will be a homepage.
fundingSource	The name associated with the activity's funding source.

Predicates

Name	Description	Domain	Range	Cardinality
produces	Indicates the deliverables produced by the activity.	Activity	Publication	1:M
tackles	Indicates the topic of the activity.	Activity	Research_ Theme	1:M
basedAt	Indicates the organisation where this activity is based.	Activity	Organisation	1:M
runBy	Indicates the team(s) that is responsable for the activity.	Activity	Team	1:M
publishedIn	Indicates where the activity is published.	Activity	Publication_ Medium	1:M

Research Theme

Class

Info	**Description**
C	A type of Thing. Also referred to as subject, theme, point, issue, proposition, but topic is the preferred term.
T	Topic
D	The subject of a discourse or of a section of a discourse.
EX	Hypertext Navigation, User Interfaces, Database Design

Special Case: Merges with Hypertext Research Theme Ontology.

References

Abasolo, J., & Gomez, M. 2000 (September). MELISA: An ontology-based agent for information retrieval in medicine. *In: Proceedings of the ECDL 2000 Workshop on the Semantic Web.* Available from: `http://www.ics.forth.gr/isl/SemWeb/program.html`.

Ackoff, R. 1967. Management Misinformation Systems. *Management Science*, **14**(4), 147–156.

Akscyn, R., McCracken, D., & Yoder, E. 1988. KMS: A distributed hypermedia system for managing knowledge in organizations. *Communications of the ACM*, **31**(7), 820–835.

Alani, H., Dasmahapatra, S., Gibbins, N., Glaser, H., Harris, S., Kalfoglou, Y., O'Hara, K., & Shadbolt, N. 2002. Managing Reference: Ensuring Referential Integrity of Ontologies for the Semantic Web. *In: Proceedings 13th International Conference on Knowledge Engineering and Knowledge Management, Siguenza, Spain.* Springer.

Alexa. 1997. *Alexa Internet Introduces Web Navigation That Learns From People.* Available from: `http://www.alexa.com/press/press_releases/intro.html`.

Anderson, K. 1999. *Internet use among college students: An exploratory study.* Available from: `http://www.rpi.edu/~anderk4/research.html`, Rensselaer Polytechnic Inst., New York, USA, unpublished report.

Anderson, K., Taylor, R., & Whitehead, E. 1994. Chimera: Hypertext for heterogeneous software environments. *Pages 94–107 of: Proceedings of the ACM Hypertext 1994 Conference, Edinburgh, Scotland.* ACM Press.

Andrews, K., & Dieberger, A. 1996. *Reinventing the Wheels. Usability problems on the World Wide Web.* Available from: `http://www.mindspring.com/~juggle5/Writings/NotesAndReports/Usability_and_Web.html`.

Andrews, K., Kappe, F., & Maurer, H. 1995. Hyper-G and harmony: Towards the next generation of networked information technology. *Pages 33–34 of: Proceedings of Conference companion on Human factors in computing systems, Denver, CO, USA.* ACM Press.

Bailey, C., El-Beltagy, S., & Hall, W. 2001. Link Augmentation: A Context-Based Approach to Support Adaptive Hypermedia. *Page in print of: Proceedings of the Third Workshop on Adaptive Hypertext and Hypermedia at the ACM Conference on Hypertext and Hypermedia 2001, Aarhus, Denmark.* ACM Press.

Baird, P., & Percival, M. 1989. Glasgow On-Line: database development using Apples HyperCard. *In:* McAleese, R. (ed), *Hypertext: Theory into Practice.* Oxford: Intellect.

Baragar, G. 1995 (May). *Editing a Scholarly Journal Article for the Electronic Medium.* M.Phil. thesis, Faculty of General Studies, University of Calgary, Calgary, Alberta.

Baron, L. 1994. *The Effectiveness of Labelled, Typed Links as Cues in Hypertext Systems.* Ph.D. thesis, The University of Western Ontario.

Beaver, D., & Rosen, R. 1978. Studies in Scientific Collaboration: Part I - The Professional Origins of Scientific Co-authorship. *Scientometrics*, **1**, 65–84.

Bechhofer, S., & Horrocks, I. 2000 (August). Driving User Interfaces from FaCT. *In: Proceedings of International Workshop on Description Logics (DL 2000), RWTH Aachen, Germany.* Available from: `http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/`.

Bechhofer, S., Horrocks, I., Patel-Schneider, P., & Tessaris, S. 1999a. A Proposal for a Description Logic Interface. *Pages 33–36 of: Proceedings of the International Workshop on Description Logics.* CEUR.

Bechhofer, S., Stevens, R., Ng, G., Jacoby, A., & Goble, C. 1999b. Guiding the User: An Ontology Driven Interface. *Pages 158–161 of: Workshop on User Interfaces to Data Intensive Systems, Edinburgh, UK.* IEEE Computer Society.

Bechhofer, S., Broekstra, J., Decker, S., Erdmann, M., Fensel, D., Goble, C., van Harmelen, F., Horrocks, I., Klein, M., McGuinness, D., Motta, E., Patel-Schneider, P., Staab, S., & Studer, R. 2000 (November). *An informal description of Standard OIL and Instance OIL.* Available from: `http://www.ontoknowledge.org/oil/downl/oil-whitepaper.pdf`. white paper.

Bechhofer, S., Goble, C., & Horrocks, I. 2001 (July-August). DAML+OIL is not Enough. *In: The First Semantic Web Working Symposium.* Available from: `http://www.semanticweb.org/SWWS/program/full/paper40.pdf`.

Beeman, W., Anderson, K., Bader, G., Larkin, J., McClard, A., McQuillan, P., & Shields, M. 1987. Hypertext and Pluralism: From Lineal to Non-Lineal Thinking. *Pages 67–88 of: Proceedings of the Hypertext 1987 Workshop, University of North Carolina, USA.* ACM Press.

Benjamins, R. V., Fensel, D., & Perez, A. G. 1998 (October). Knowledge Management through Ontologies. *In: Proceedings of the 2nd International Conference on Practical Aspects of Knowledge Management.* Available from: `http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-13/`.

Bergmark, D., Phempoonpanich, P., & Zhao, S. 2001. Scraping the ACM Digital Library. *ACM SIGPLAN Notices*, **35**(2), 1–7.

Berners-Lee, T. 1991 (December). *Hypertext Transfer Protocol (HTTP).* Available from: `ftp://ftp.w3.org/pub/www/doc/http-spec.txt`.

Berners-Lee, T. 1992. *Hypertext Markup Language (HTML).* Available from: `http://www.w3c.org/History/19921103-hypertext/hypertext/WWW/MarkUp/MarkUp.html`.

Berners-Lee, T. 1998. *Semantic Web Road map.* Available from: `http://www.w3.org/DesignIssues/Semantic.html`.

Berners-Lee, T. 2000 (December). XML 2000 Keynote. *In: Proceedings of XML 2000*. Available from: `http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide1-0.html`.

Berners-Lee, T., Cailliau, R., Groff, J., & Pollerman, B. 1993. World Wide Web: The Information Universe. *Electronic Networking: Research, Applications and Policy*, **1**(2).

Berners-Lee, T., Masinter, L., & McCahill, M. 1994 (December). *Uniform Resource Locators (URL)*. Available from: `http://www.w3c.org/Addressing/rfc1738.txt`. Proposed Standard (RFC 1738).

Berners-Lee, T., Hendler, J., & Lassila, O. 2001 (May). *The Semantic Web*. Scientific American. Available from: `http://www.sciam.com/2001/0501issue/0501berners-lee.html`.

Bernstein, M. 1990. An Apprentice That Discovers Hypertext Links. *Pages 212–223 of: Hypertext: Concepts, Systems and Applications, Proceedings of the Hypertext '90 Conference, INRIA, France*. ACM Press.

Bernstein, M. 1991. Position statement for Panel on Structure, Navigation, and Hypertext: The Status of the Navigation Problem. *Pages 365–366 of: Proceedings of the ACM Hypertext 1991 Conference, San Antonio, Texas, USA*. ACM Press.

Bieber, M., Vitali, F., Ashman, H., Balasubramanian, V., & Oinas-Kukkonen, H. 1997a. Fourth Generation Hypermedia: Some Missing Links for the World Wide Web. *International Journal on Human Computer Studues*, **47**, 31–65.

Bieber, M., Vitali, F., Ashman, H., Balasubramanian, V., & Oinas-Kukkonen, H. 1997b. Some Hypermedia Ideas for the WWW. *Pages 309–319 of: Proceedings of the 30th Annual Hawaii International Conference on System Sciences, Wailea, Hawai, USA*. IEEE Press.

Bishop, A. 1998. Digital Libraries and Knowledge Disaggregation: The Use of Journal Article Components. *Pages 29–39 of: Proceedings of the ACM Conference on Digital libraries 1998, Pittsburgh, PA, USA*. ACM Press.

Boulos, M. Kamel, Roudsari, A., & Carson, E. 2001. Health Geomatics: An Enabling Suite of Technologies in Health and Healthcare. *Journal of Biomedical Informatics*, **34**(3), 195–219.

Bowman, C., Danzig, P., Hardy, D., Manber, U., & Schwartz, M. 1995. The Harvest Information Discovery and Access System. *Computer Networks and ISDN Systems*, **28**, 119–125.

Bra, P. De, & Calvi, L. 1998. 2L670: a flexible adaptive hypertext courseware system. *Pages 283–284 of: Proceedings of the ACM Hypertext 1998 Conference, Pitsburgh PA, USA*. ACM Press.

Bradford, J. 1934. The distribution of papers in a given science journal. *Engineering*, **137**, 85–86.

Bradford, J. 1981. Empirical and Theoretical Bases of Zipf's Law. *Library Trends*, **30**(1), 53–64.

Brody, T., Jiao, Z., Hitchcock, S., Carr, L., & Harnad, S. 2001a (September). Enhancing OAI Metadata for Eprint Services: two proposals. *In: Proceedings of the OAI Workshop at the European Conference on Digital Libraries, Darmstadt, Germany.* Available from: `http://notesmail.cs.odu.edu/faculty/zubair/workshop.nsf/OaiEcdlWorkshop?OpenForm`.

Brody, T., Jiao, Z., Krichel, T., & Warner, S. 2001b (September). *Syntax and Vocabulary of the Academic Metadata Format.* Available from: `http://amf.openlib.org/doc/ebisu.html`.

Brown, J., & Duguid, P. 1996. The Social Life of Documents. *First Monday*, **1**(1). Electronic journal. Available from: `http://www.firstmonday.dk/issues/issue1/documents/`.

Bush, Vannevar. 1945. As We May Think. *The Atlantic Monthly*, **176**(1), 101–108.

Calvi, L., & Bra, P. De. 1997. Improving the usability of hypertext courseware through adaptive linking. *Pages 224–225 of: Proceedings of the ACM Hypertext 1997 Conference, Southampton, UK.* ACM Press.

Cameron, R. 1997. A Universal Citation Database as a Catalyst for Reform in Scholarly Communication. *First Monday*, **2**(4). Electronic journal. Available from: `http://www.firstmonday.dk/issues/issue2_4/cameron/index.html`.

Cane, G., Chavez, R., Mahoney, A., Milbank, T., Rydberg-Cox, J., Smith, D., & Wulfman, C. 2001. Drudgery and Deep Thought. *CACM*, **44**(5), 35–40.

Carr, L., Roure, D. De, Hall, W., & Hill, G. 1995. The Distributed Link Service: A Tool for Publishers, Authors and Readers. *World Wide Web Journal*, **1**(1), 647–656.

Carr, L., Hall, W., Bechhofer, S., & Goble, G. 2001. Conceptual Linking: Ontology-based Open Hypermedia. *Pages 334–342 of: Proceedings of the Tenth World Wide Web Conference, Hong Kong, China.* ACM Press.

Carr, L., Kampa, S., & Miles-Board, T. 2001 (May). *MetaPortal Final Report: Building Ontological Hypermedia with the OntoPortal Framework.* IAM Group, University of Southampton, Final Project Report. Available from: `http://www.ontoportal.org.uk/papers/fr.pdf`.

Carr, L., Kampa, S., Roure, D. De, Hall, W., Bechhofer, S., Goble, C., & Horan, B. 2002 (March). *Ontological Linking: Motivation and Analysis.* Available from: `http://cohse.semanticweb.org/papers/cikm.doc`, IAM Group, University of Southampton, report.

Chen, C., & Carr, L. 1999a. A Semantic-Centric Approach to Information Visualisation. *Pages 18–23 of: Proceedings of Information Visualisation 99.* IEEE Press.

Chen, C., & Carr, L. 1999b. Trailblazing the Literature of Hypertext: Author Co-Citation Analysis (1989-1998). *Pages 51–60 of: Proceedings of the ACM Hypertext 1999 Conference, Darmstadt, Germany.* ACM Press.

Chen, C., & Paul, R. 2001. Visualizing a Knowledge Domain's Intellectual Structure. *IEEE Computer*, **34**(3), 65–71.

Clarke, B. 1964. Multiple Authorship Trends in Scientific Papers. *Scientific*, **143**, 822–824.

Cockburn, A., & Jones, S. 1996. Which way now? Analysing and easing inadequacies in WWW navigation. *International Journal of Human-Computer Studies*, **45**, 105–129.

Collier, G. 1987. Thoth-II: Hypertext with Explicit Semantics. *Pages 269–289 of: Proceedings of the Hypertext 1987 Workshop, University of North Carolina, USA*. ACM Press.

Conklin, E., & Begeman, M. 1989. gIBIS: A Tool for All Reasons. *Journal of the American Society for Information Science*, **40**(3), 200–213.

Conklin, J. 1987. Hypertext: An Introduction and Survey. *IEEE Computer*, **2**(9), 17–41.

Conklin, J., & Begeman, M. 1988. gIBIS: A hypertext tool for exploratory policy discussion. *ACM Transaction on Office Information Systems*, **6**(4), 303–331.

Conklin, J., Selvin, A., Shum, S. Buckingham, & Sierhuis, M. 2001. Facilitated Hypertext for Collective Sensemaking: 15 years on from gIBIS. *Pages 123–124 of: Proceedings of the ACM Hypertext 2001 Conference, Aarhus, Denmark*. ACM Press.

Cui, Z., Jones, D., & O'Brien, P. 2001. Issues in Ontology-based Information Integration. *In: Proceedings of Workshop on Ontologies and Information Sharing at the International Joint Conference on Artificial Intelligence 2001, Washington, USA*. Morgan Kaufmann Publishers: San Francisco, USA.

Cycorp. 2002. *Cycorp Incorporated*. Available from: `http://www.cyc.com/`.

da Silva, D. Pilar, Durm, R. Van, Duval, E., & Olivi, H. 1998. Adaptive navigational facilities in educational hypermedia. *Pages 291–292 of: Proceedings of the ACM Hypertext 1998 Conference, Pitsburgh PA, USA*. ACM Press.

DARPA. 2000 (October). *DAML-ONT Initial Release*. Available from: `http://www.daml.org/2000/10/daml-ont.html`.

Davis, H., Hall, W., Pickering, A., & Wilkins, R. 1993. Microcosm: an open hypermedia system. *Page 526 of: Proceedings of Conference on Human factors in computing systems*. ACM Press.

Davis, Hugh C. 1999. Hypertext link integrity. *ACM Computing Surveys (CSUR)*, **31**(4es), 28.

DCMI. 1999. *Dublin Core Metadata Element Set, Version 1.1: Reference Description*. Available from: `http://dublincore.org/documents/1999/07/02/dces/`.

Decker, S., Brickley, D., Saarela, J., & Angele, J. 1998. A Query and Inference Service for RDF. *In: Proceedings of the The Query Languages Workshop*. W3C. Available from: `http://www.w3.org/TandS/QL/QL98/pp/queryservice.html`, Position Paper.

Decker, S., van Harmelen, F., Broekstra, J., Erdmann, M., Fensel, D., Horrocks, I., Klein, M., & Melnik, S. 2001. *The Semantic Web - on the respective Roles of XML and RDF*. Available from: `http://www.ontoknowledge.org/oil/downl/IEEE00.pdf`.

DeRose, S. 1989. Expanding the Notion of Links. *Pages 249–257 of: Proceedings of the Hypertext 1989 Conference on Hypertext, Pittsburgh, Pennsylvania, USA*. ACM Press.

Deutsch, A., Fernandez, M., Florescu, D., Levy, A., & Suciu, D. 1998. *XML-QL: A Query Language for XML*. Available from: `http://www.w3.org/TR/NOTE-xml-ql/`.

Deutsch, P., Emtage, A., Bunyip, Koster, M., Nexor, & Stumpf, M. 1995. *Publishing Information on the Internet with Anonymous FTP.* Available from: `http://www.ifla.org/documents/libraries/cataloging/metadata/iafa.txt`.

Dewar, J. 1998. *The Information Age and the Printing Press: Looking Backward to See Ahead.* Available from: `http://www.rand.org/publications/P/P8014/`.

Dillon, A., Richardson, J., & McKnight, C. 1989. The human factors of journal usage and the design of electronic text. *Interacting with Computers*, **1**(2), 183–189.

Domingue, J. 1998 (April). Tadzebao and WebOnto: Discussing, Browsing, and Editing Ontologies on the Web. *In: Proceedings of the 11th Banff Knowledge Acquisition Workshop, Banff, Alberta, Canada.* Available from: `http://ksi.cpsc.ucalgary.ca/KAW/KAW98/KAW98Proc.html`.

Eastman, Caroline M. 1999. 30,000 Hits May be Better than 300: Precision Anomalies in Internet Searches. *Pages 313–314 of: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Posters/Late Breaking Results, no. C.IR.99.313. ACM Press.

Edge, D. 1977. Why I am not a co-citationist? *Society for Social Studies of Science Newsletter*, **2**, 13–19.

Eisenstein, E. 1979. *The Printing Press as an Agent of Change.* Cambridge University Press, New York.

El-Beltagy, S., Hall, W., Roure, D. De, & Carr, L. 2001. Linking in context. *Pages 151–160 of: Proceedings of the twelfth ACM conference on Hypertext and Hypermedia.* ACM Press.

Ellis, D., Furner, J., & Willett, P. 1996. On the creation of hypertext links in full text documents - measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, **47**(4), 287–300.

Empolis. 2001. *X2X.* Available from: `http://www.empolis.co.uk/products/prod_X2X.asp`.

Engelbart, D. C., Watson, R. W., & Norton, J. C. 1973. The Augmented Knowledge Workshop. *Pages 9–21 of: Proceedings of AFIPS Conference*, vol. 42. National Computer Conference.

Falasconi, S., Lanzola, G., & Stefanelli, M. 1996 (November). Using Ontologies in Multi-Agent Systems. *In: Proceedings of the 10th Banff Knowledge Acquisition For Knowledge Based Systems Workshop.* Available from: `http://ksi.cpsc.ucalgary.ca/KAW/KAW96/KAW96Proc.html`.

Farquhar, A., Fickas, R., & Rice, J. 1996 (November). The Ontolingua Server: a Tool for Collaborative Ontology Construction. *In: Proceedings of the 10th Banff Knowledge Acquisition for Knowledge Based System Workshop.* Available from: `http://ksi.cpsc.ucalgary.ca/KAW/KAW96/KAW96Proc.html`.

Febvre, L., & Martin, H-J. 1984. *The Coming of the Book: The Impact of Printing, 1450-1800.* (Translated by D. Gerrard) edn. London: Verso.

Fensel, D. 2000 (October). *Ontoknowledge: Intelligent Information Brokering in Intranets.* Available from: `http://www.ontoknowledge.org/countd/countdown.cgi?1review-otk.pdf`.

Fensel, D., & Musen, M. 2001. The Semantic Web: A Brain for Humankind. *IEEE Intelligent Systems*, **16**(2), 24–25.

Fensel, D., Decker, S., Erdmann, M., & Studer, R. 1998. Ontobroker in a Nutshell. *Pages 663–664 of: European Conference on Digital Libraries, Crete, Greece*. Springer.

Fensel, D., Horrocks, I., Harmelen, F. Van, Decker, S., Erdmann, M., & Klein, M. 2000. OIL in a nutshell. *In:* Dieng, R. (ed), *Proceedings of the European Knowledge Acquisition Conference*. Springer-Verlag.

Fiderio, J. 1988. A Grand Vision. *Byte*, October, 237–244.

Fountain, A., Hall, W., Heath, I., & Davis, H. 1990. Microcosm: An Open Model for Hypermedia with Dynamic Linking. *Pages 298–311 of:* Rizk, A., Streitz, N., & Andre, J. (eds), *Hypertext: Concepts, Systems and Applications, Proceedings of the Hypertext '90 Conference, INRIA, France*. ACM Press.

Fox, E., & Marchionini, G. 1998. Toward a Worldwide Digital Library. *CACM*, **41**(4), 29–32.

Freese, E. 2000. Topic Maps vs. RDF. *Pages 79–86 of: Proceedings of Extreme Markup Languages 2000*. IDEAlliance.

French, J., Powell, A., & Creighton, W. 1998. Efficient searching in distributed digital libraries. *Pages 283–284 of: Proceedings of the ACM Conference on Digital libraries 1998, Pittsburgh, PA, USA*. ACM Press.

Fujitsu. 2001. *XLink Processor(XLiP)*. http://www.labs.fujitsu.com/free/xlip/en/.

Furnas, G. W. 1986. Generalized Fisheye Views. *Pages 16–23 of: Proceedings of CHI 1986 Human Factors in Computing Systems, Boston, Massachusetts, USA*. ACM Press.

Furner, J., Ellis, D., & Willett, P. 1999. Inter-Linker Consistency in the Manual Construction of Hypertext Documents. *ACM Computing Surveys*, **31**(4es).

Furuta, R., III, F. M. Shipman, Marshall, C. C., Brenner, D., & Hsieh, H. 1997. Hypertext Paths and the World Wide Web: Experiences with Walden's Paths. *Pages 167–176 of: Proceedings of the ACM Hypertext 1997 Conference, Southampton, UK*. ACM Press.

Garfield, E. 1955. Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science*, **122**(3159), 108–11.

Garfield, E. 1979a. *Citation Indexing - Its Theory and Application in Science, Technologies, and Humanities*. John Wiley & Sons, New York.

Garfield, E. 1979b. Is citation analysis a legitimate evalutation tool? *Scientometrics*, **1**, 359–375.

Garfield, E. 1983. How to use the Science Citation Index (SCI). *Essays of an Information Scientist*, **6**(February), 5–14.

Garfield, E. 1993. Scientists should understand the limitations as well as the virtues of citation analysis. *The Scientist*, **7**(14), 12.

Garfield, E. 1994. The impact factor. *Current Contents*, **25**, 3–7.

Garfield, E. 1996 (October). *Citation Indexers for Retrieval and Research Evaluation.* Presentation at the Consensus Conference on the Theory and Practice of Research Assessment, Capri, Italy.

Garfield, E., Sher, I., & Torpie, R. 1964 (December). *The use of citation data in writing the history of science.* Available from: `http://www.garfield.library.upenn.edu/papers/useofcitdatawritinghistofsci.pdf`. Institute for Scientific Information, Philadelphia.

Genesereth, M. 1990. *The Epikit Manual.* Epistemics, Inc. Palo Alto, CA.

Genesereth, M. 1998. *Knowledge Interchange Format (KIF).* Available from: `http://logic.stanford.edu/kif/dpans.html`.

Genesereth, M., & Fikes, R. 1992. *Knowledge Interchange Format, Version 3.0 Reference Manual.* Tech. rept. Logic-92-1. Computer Science Department, Stanford University.

GILS. 1997. *Global Information Locator Service.* Available from: `http://www.gils.net/`.

Giuffrida, G., Shek, E., & Yang, J. 2000. Knowledge-based metadata extraction from PostScript files. *Pages 77–84 of: Proceedings of the fifth ACM conference on Digital libraries.* ACM Press.

Gloor, A. 1991. CYBERMAP Yet Another Way of Navigating in Hyperspace. *Pages 107–121 of: Proceedings of the ACM Hypertext 1991 Conference, San Antonio, Texas, USA.* ACM Press.

Golovchinsky, G. 1997. Queries? Links? Is There a Difference? *Pages 407–414 of: Proceedings of the CHI 1997 Human Factors in Computing Systems, Atlanta, Georgia, USA.* ACM Press.

Golovchinsky, G., & Chignell, M. H. 1996 (April). *The Newspaper as an Information Exploration Metaphor.* Working Paper 96-05. Dept. of Mechanical and Industrial Engineering, University of Toronto.

Gomez-Perez, A. 1995. Some ideas and examples to evaluate ontologies. *In: Proceedings of the 11th Conference on Artificial Intelligence Applications.* Available from: `ftp://ftp.ksl.stanford.edu/pub/KSL_Reports/KSL-94-65.ps`.

Gomez-Perez, A. 1996. A framework to verify knowledge sharing technology. *Expert Systems with Application*, **11**(4), 519–529.

Gordon, S., & Lewis, V. 1992. Enhancing hypertext documents to support learning from text. *Technical Communication*, **39**(2), 305–308.

Grønbæk, K., & Trigg, R. 1994. Design issues for a Dexter-based hypermedia system. *Communications of the ACM*, **37**(2), 40–49.

Grønbæk, K., Bouvin, N. Olof, & Sloth, L. 1997. Designing Dexter-based hypermedia services for the World Wide Web. *Pages 146–156 of: Proceedings of the ACM Hypertext 1997 Conference, Southampton, UK.* ACM Press.

Gruber, T. 1993. A translation approach to portable ontology specification. *Knowledge Acquisition*, **5**, 199–220.

Guarino, N. 1998. Formal Ontology in Information Systems. *Frontiers in Artificial Intelligence and Applications*, **46**, 347.

Halasz, F. 1988. Reflections on NoteCards: Seven issues for the next generation of hypermedia systems. *CACM*, **31**(7), 836–852.

Halasz, F., & Schwartz, M. 1994. The Dexter hypertext reference model. *CACM*, **37**(2), 30–39.

Halasz, F., Moran, T., & Trigg, R. 1987. NoteCards in a nutshell. *Pages 45–52 of: Proceedings of Human Factors in Computing Systems and Graphics Interface*. ACM Press.

Hall, W., Hill, G., & Davis, H. 1993. The microcosm link service. *Pages 256–259 of: Proceedings of the ACM Hypertext 1993 Conference, Seattle, Washington, USA*. ACM Press.

Halsey, B., & Anderson, K. M. 2000. XLink and open hypermedia systems: a preliminary investigation. *Pages 212–213 of: Proceedings of the ACM Hypertext 2000 Conference, San Antonio, Texas, USA*. ACM Press.

Handschuh, S., Maedche, A., Stojanovic, L., & Volz, R. 2001 (October). *KAON - The KArlsruhe ONtology and Semantic Web Tool Suite*. Tech. rept. University of Karlsruhe.

Harnad, S. 1991. Scholarly Skywriting and the Prepublication Continuum of Scientific Inquiry. *Current Contents*, **45**, 9–13.

Harnad, S. 1995a. Electronic Scholarly Publication: Quo Vadis? *Serials Review*, **21**(1), 70–72.

Harnad, S. 1995b. The postgutenberg galaxy: How to get there from here. *Information society*, **11**(4), 285–292.

Harnad, S., & Carr, L. 2000. Integrating, Navigating and Analyzing Eprint Archives Through Open Citation Linking (the OpCit Project). *Current Science*, **79**(September), 629–638. (Special issue in honour of Eugene Garfield).

Harnad, S., Varian, H., & Parks, R. 1999 (November). *Academic publishing in the online era: What Will Be For-Fee And What Will Be For-Free?* Available from: `http://culturemachine.tees.ac.uk/Cmach/Backissues/j002/Articles/art_harn.htm`. on-line discussion.

Harter, S. 1996. The Impact of Electronic Journals on Scholarly Communication: A Citation Analysis. *The Public-Access Computer Systems Review*, **7**(5). Electronic Journal. Available from: `http://info.lib.uh.edu/pr/v7/n5/hart7n5.html`.

Hatch, R. 1998 (February). *The Scientific Revolution: Definition, Concept, History*. Available from: `http://web.clas.ufl.edu/users/rhatch/pages/03-Sci-Rev/SCI-REV-Teaching/03sr-definition-concept.htm`.

Haustein, S. 2001. Semantic Web Languages: RDF vs. SOAP Serialisation. *In: Proceedings of the Second International Workshop on the Semantic Web*. Available from: `http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-40/`.

Haustein, S., & Pleumann, J. 2002. Is Participating in the Semantic Web too Difficult? *In: Proceedings of the International Semantic Web Conference (ISWC), Sardinia, Italia*. Springer.

Heflin, J., Hendler, J., & Luke, S. 1998 (July). Reading Between the Lines: Using SHOE to Discover Implicit Knowledge from the Web. *In: Proceedings of the AAAI-98 Workshop on AI and Information Integration*. Available from: `http://www.isi.edu/ariadne/aiii98-wkshp/proceedings.html`.

Heflin, J., Volz, R., & Dale, J. 2002 (March). *Requirements for a Web Ontology Language*. Available from: `http://www.w3.org/TR/webont-req/`.

Hellman, E. 1999. The S-Link-S Framework for Reference Linking: Architecture and Implementation. *Pages 68–73 of: Proceedings of the ICCC/IFIP Conference on Electronic Publishing, Ronneby, Sweden*. ICCC Press.

Hitchcock, S., Carr, L., Harris, S., Hey, J., & Hall, W. 1997a. Citation Linking: Improving Access to Online Journals. *Pages 115–122 of: Proceedings of the ACM Conference on Digital Libraries 1997, Philadelphia, PA, USA*. ACM Press.

Hitchcock, S., Carr, L., Quek, F., Witbrock, A., Tarr, I., & Hall, W. 1997b (April). Linking Everything to Everything: Journal Publishing Myth or Reality? *In: Proceedings of ICCC/IFIP Conference on Electronic Publishing: Electronic Publishing - New Models and Opportunities*. Available from: `http://journals.ecs.soton.ac.uk/IFIP-ICCC97.html`.

Hitchcock, S., Carr, L., Hall, W., Harris, S., Probets, S., Evans, D., & Brailsford, D. 1998. Linking electronic journals: Lessons from the Open Journal project. *D-Lib Magazine*, December. Available from: `http://www.dlib.org/dlib/december98/12hitchcock.html`.

Holmes, A., & Oppenheim, C. 2001. Use of citation analysis to predict the outcome of the 2001 Research Assessment Exercise for Unit of Assessment (UoA) 61: Library and Information Management. *Information Research*, **6**(2).

Hughes, G., & Carr, L. 2002. Microsoft Smart Tags: support, ignore or condemn them? *Page 80 of: Proceedings of the ACM Hypertext 2002 Conference, Maryland, USA*. ACM Press.

Ichimura, S., & Matsushita, Y. 1993. Another Dimension to Hypermedia Access. *Pages 63–72 of: Proceedings of the ACM Hypertext 1993 Conference, Seattle, Washington, USA*. ACM Press.

International Organization for Standardization. 1986. *Standard Generalized Markup Language (SGML)*. ISO 8879:1986(E).

ISI. 2002. *ISI Web of Knowledge Fact Sheet*. Available from: `http://www.isiwebofknowledge.com/isiwokfactsheet.pdf`. Rev 01/02.

Jacobs, D. 2001. A bibliometric study of the publication patterns of scientists in South Africa 1992-96, with particular reference to status and funding. *Information Research*, **6**(3).

Jennings, N., Woghiren, K., & Osborn, S. 2000. *Interacting Agents - the way forward for Agent-Mediated Electronic Commerce*. Available from: `http://www.lostwax.com`. Lostwax white paper.

Jonassen, D. 1993. Effects Of Semantically Structured Hypertext Knowledge Bases on Users Knowledge Structures. *Chap. 7 of:* McKnight, C., Dillon, A., & Richardson, J. (eds), *Hypertext: A Psychological Perspective*. Ellis Horwood, New York: IEEE Press.

Kampa, S., & Carr, L. 2000 (May). Web Scholars. *Pages 44–45 of: In Poster Proceedings of the 9th International WWW Conference, Amsterdam, Netherlands.* Available from: `http://www9.org/w9cdrom/index.html`.

Kampa, S., Miles-Board, T., Carr, L., & Hall, W. 2001a. Hypertext in the Semantic Web. *Pages 237–238 of: Proceedings of the twelfth ACM conference on Hypertext and Hypermedia.* ACM Press.

Kampa, S., Miles-Board, T., Carr, L., & Hall, W. 2001b (March). *Linking with Meaning: Ontological Hypertext for Scholars.* Tech. rept. ECSTR-IAM01-005. IAM Group, University of Southampton, Southampton, UK.

Kappe, F., Maurer, H., & Sherbakov, N. 1993. Hyper-G: A Universal Hypermedia System . *Journal of Educational Multimedia and Hypermedia*, **2**(1), 39–66.

Katz, J., & Martin, B. 1997. What is Research Collaboration? *Research Policy*, **26**, 1–18.

Kessler, M. 1963. Bibliographic Coupling between Scientific Papers. *American Documentation*, **14**, 10–25.

King, J. 1987. A review of bibliometric and other science indicators and their role in research evaluation. *Journal of Information Science*, **13**(5), 261–276.

Kobayashi, M., & Takeda, K. 2000. Information retrieval on the Web. *ACM Computing Surveys (CSUR)*, **32**(2), 144–173.

Koivunen, M., & Swick, R. 2001. Metadata Based Annotation Infrastructure offers Flexibility and Extensibility for Collaborative Applications and Beyond. *In: Proceedings of K-Cap Workshop on Knowledge Markup and Semantic Annotation.* Available from: `http://semannot2001.aifb.uni-karlsruhe.de/schedule_new.html`.

Kopak, R. 1999. Functional link typing in hypertext. *ACM Computing Surveys (CSUR)*, **31**(4), 16.

Kopmanis, J., & Wirzenius, L. 1994. *Linux Software Map Entry Template.* Available from: `ftp://sunsite.unc.edu/pub/Linux/docs/LSM/lsm-template`.

Kramer, S. 1963. *The Sumerians.* Chicago: The University Chicago Press.

Kwok, C., Etzioni, O., & Weld, D. 2001. Scaling question answering to the Web. *ACM Transactions on Information Systems (TOIS)*, **19**(3), 242–262.

Lacher, M., & Decker, S. 2001 (July-August). On the Integration of Topic Maps and RDF Data. *In: The First Semantic Web Working Symposium.* Available from: `http://www.semanticweb.org/SWWS/program/full/paper53.pdf`.

Ladd, B., Capps, M., & Stotts, P. 1997. The World Wide Web: what cost simplicity? *Pages 210–211 of: Proceedings of the ACM Hypertext 1997 Conference, Southampton, UK.* ACM Press.

Lagoze, C., & de Sompel, H. Van. 2001. The open archives initiative: building a low-barrier interoperability framework. *Pages 54–62 of: Proceedings of the first ACM/IEEE-CS joint conference on Digital libraries.* ACM Press.

Landow, G. 1990. Popular fallacies about hypertext. Berlin: Springer-Verlag.

Landow, G. 1997. *Hypertext 2.0: The Convergence of contemporary critical theory and technology.* Baltimore: John Hopkins University Press.

Lang, D. 1996. Mining for gems in an information overload. *Pages 167–178 of: Proceedings of the 14th annual international conference on Marshaling new technological forces : building a corporate, academic, and user-oriented triangle.* ACM Press.

Lawrence, S. 2001. Online or Invisible? *Nature*, **411**(6837), 521.

Lawrence, S., & Giles, C. Lee. 1999. Searching the Web: General and Scientific Information Access. *IEEE Communications*, **37**(1), 116–122.

Lawrence, S., Giles, C., & Bollacker, K. 1999a. Digital Libraries and Autonomous Citation Indexing. *IEEE Computer*, **32**(6), 67–71.

Lawrence, S., Bollacker, K., & Giles, C. 1999b. Indexing and Retrieval of Scientific Literature. *Pages 139–146 of: Eighth International Conference on Information and Knowledge Management.* Kansas City, Missouri: ACM Press.

Lee, J., & Malone, T. 1990. Partially shared views: A Scheme for communicating among groups that use different type hierarchies. *ACM Transactions on Information Systems*, **8**(1), 1–26.

Lee, Y., Yoo, S., Yoon, K., & Berra, P. 1996. Index structures for structured documents. *Pages 91–99 of: Proceedings of the ACM Conference on Digital Libraries 1996, Bethesda, MD, USA.* ACM Press.

Liew, C., Foo, S., & Chennupati, K. 2001a. A user study of the design issues of PROPIE: a novel environment for enhanced interaction and value adding of electronic documents. *Journal of Documentation*, **57**(3), 377–426.

Liew, C., Foo, S., & Chennupati, K. 2001b. Towards a new generation of information environment for the user of e-documents. *Journal of Information Science*, **27**(5), 327–342.

Lotka, A. 1926. The frequency distribution of scientific productivity. *Journal of the Washington Acad. of Science*, **16**, 317.

Luke, S., Spector, L., & Rager, D. 1996. Ontology-Based Knowledge Discovery on the World Wide Web. *Pages 96–102 of: Proceedings of the AAAI-98 Workshop on Internet-based Information Systems.* AAAI Press.

MacGregor, R. 1990. *LOOM Users Manual.* ISI/WP-22. USC/Information Sciences Institute.

MacRoberts, M., & MacRoberts, B. 1989. Problems of citation analysis: a critical review. *Journal of the American Society for Information Science*, **40**(5), 342–349.

Maedche, A., Staab, S., Stojanovic, N., Studer, R., & Sure, Y. 2001. SEAL - A Framework for Developing SEmantic portALs. *Pages 663–664 of: BNCOD 2001 - 18th British National Conference on Databases, Oxford, UK.* Springer.

Mahalingam, K., & Huhns, M. 1997. An Ontology Tool for Query Formulation in an Agent-Based Context. *Pages 170–179 of: Proceedings of the 2nd IFCIS International Conference on Cooperative Information Systems.* IEEE Press.

Mar. 2000 (January). *MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media.* Available from: `http://www.loc.gov/marc/specifications/spechome.html`. Library of Congress.

Marshall, C., Halasz, F., Rogers, R., & Janssen, W. 1991. Aquanet: a hypertext tool to hold your knowledge in place. *Pages 261–275 of: Proceedings of the ACM Hypertext 1991 Conference, San Antonio, Texas, USA*. ACM Press.

Marshall, C. C., Shipman, F. M., & Coombs, J. H. 1994. VIKI: spatial hypertext supporting emergent structure. *Pages 13–23 of: Proceedings of the ACM Hypertext 1994 Conference, Edinburgh, Scotland*. ACM Press.

McCray, A., & Gallagher, M. 2001. Principles for digital library development. *CACM*, **44**(5), 49–54.

McDonald, S., & Stevenson, R. 1998. The effects of text structure and prior knowledge of the learning on navigation in hypertext. *Human Factors*, **40**(1), 18–27.

McIlraith, S., Son, T. Cao, & Zeng, H. 2001. Mobilizing the Semantic Web with DAML-Enabled Web Services. *Pages 82–93 of: Proceedings of the Second International Workshop on the Semantic Web*. Available from: `http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-40/`.

McKnight, C. 1997. Designing the Electronic Journal: Why Bother? *Serials*, **10**(2), 184–188.

McKnight, C., Dillon, A., Richardson, J., Haraldsson, H., & Spinks, R. 1992. Information Access in Different Media: An Experimental Comparison. *Contemporary Ergonomics*, 515–519.

Megginson, David. 2000. *SAX 2.0 - The Simple API for XML*. Available from: `http://www.megginson.com/SAX/index.html`.

Mendes, E., Mosley, N., & Counsell, S. 2001. Web Metrics - Estimating Design and Authoring Effort. *IEEE MultiMedia*, **8**(1), 50–57.

Middleton, S., Alani, H., Shadbolt, N., & Roure, D. De. 2002. Exploiting Synergy Between Ontologies and Recommender Systems. *In: Proceedings of The Eleventh International World Wide Web Conference, Haiwai, USA*. Sementic Web Workshop 2002. Available from: `http://www2002.org/CDROM/alternate/index.html`.

Millard, D., Moreau, L., Davis, H., & Reich, S. 2000. FOHM: a fundamental open hypertext model for investigating interoperability between hypertext domains. *Pages 93–102 of: Proceedings of the ACM Hypertext 2000 Conference, San Antonio, Texas, USA*. ACM Press.

Miller, E., Swick, R., Brickley, D., & McBride, B. 2001. *Semantic Web Activity at the W3C*. Available from: `http://www.w3.org/2001/sw/`.

Miller, Jim. 1996. *Rating Services and Rating Systems*. Available from: `http://www.w3.org/TR/REC-PICS-services`.

Moreau, L. 2000. *Southampton Agent Framework (SoFAR)*. Available from: `http://www.iam.ecs.soton.ac.uk/software/sofar/`.

Moreau, L., Gibbens, N., Roure, D. De, El-Beltagy, S., Hall, W., Hughes, G., Joyce, D., Kim, S., Michaelides, D., Millard, D., Reich, S., Tansley, R., & Weal, M. 2000. SoFAR with DIM Agents. *Pages 369–388 of: Proceedings of the 5th International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology (PAAM 2000), Manchester, UK*. Springer.

Motta, E. 1998 (January). An Overview of the OCML Modeling Language. *In: Proceedings of the 8th Workshop on Knowledge Engineering Methods and Languages*. Available from: `http://www.aifb.uni-karlsruhe.de/WBS/dfe/keml98/proceedings.html`.

Motta, E., Shum, S. Buckingham, & Domingue, J. 1999 (October). Case Studies in Ontology-Driven Document Enrichment. *In: Proceedings of the Twelfth Workshop on Knowledge Acquisition, Modeling and Management*. Available from: `http://ksi.cpsc.ucalgary.ca/KAW/KAW99/KAW99Proc.html`.

Mukherjea, S., & Hara, Y. 1997. Focus+Context Views of World Wide Web Nodes. *Pages 187–196 of: Proceedings of the ACM Hypertext 1997 Conference, Southampton, UK.* ACM Press.

Murugesan, P., & Moravcsik, M. 1978. Variations of the nature of citation measures with journals and scientific specialities. *Journal of the American Society for Information Science*, **29**, 141–147.

Nanard, J., & Nanard, M. 1991. Using structured types to incorporate knowledge in hypertext. *Pages 329–343 of: Proceedings of the ACM Hypertext 1991 Conference, San Antonio, Texas, USA*. ACM Press.

Nanard, J., & Nanard, M. 1993. Should anchors be typed too?: An experiment with MacWeb. *Pages 51–62 of: Proceedings of the ACM Hypertext 1993 Conference, Seattle, Washington, USA*. ACM Press.

Nanard, J., & Nanard, M. 1995. Hypertext Design Environments and the Hypertext Design Process. *CACM*, **38**(8), 49–56.

Needleman, M. 1999 (February). *Meeting Report of the NISO Linking Workshop, Washington D.C, USA*. Available from: `http://www.niso.org/linkrpt.html`.

Negroponte, N. 1995. *Being Digital*. Alfred A. Knopf, New York.

Nelson, M. 1994. We have the information you want, but getting it will cost you! *ACM Crossroads*, **1**(1). Available from: `http://info.acm.org/crossroads/xrds1-1/mnelson.html`.

Nelson, T. 1980. Replacing the printed word : a complete literary system. *Pages 1013–1023 of: Proceedings of IFIP Congress 80*. North-Holland/IFIP.

Nelson, T. 1987. *Literary Machines*. 87.1 edn. Computer Books.

Nelson, T. 1999. The unfinished revolution and Xanadu. *ACM Computing Surveys*, **31**(4es), 37.

Network Inference. 2002. *Network Inference Ltd.* Available from: `http://www.networkinference.com`.

Nevill-Manning, C., Witten, I., & Paynter, G. 1997. Browsing in digital libraries: a phrase-based approach. *Pages 230–236 of: Proceedings of the second ACM international conference on Digital libraries*. ACM Press.

Newcomb, S. R., Kipp, N. A., & Newcomb, V. T. 1991. The HyTime hypermedia / time-based document structuring language. *CACM*, **34**(11), 67–83.

Nielsen, J. 1990. The art of navigating through hypertext. *Communications of the ACM*, **33**(1), 296–310.

Noy, N., & McGuinness, D. L. 2001. *Ontology Development 101: A Guide to Creating Your First Ontology.* Report SMI-2001-0880. Dept. of Mechanical and Industrial Engineering, University of Toronto.

Noy, N., Sintek, M., Decker, S., Crubezy, M., Fergerson, R., & Musen, M. 2001. Creating Semantic Web Contents with Protégé-2000. *IEEE Intelligent Systems*, **16**(2), 60–71.

O'Hara, K., Alani, H., & Shadbolt, N. 2002. Identifying Communities of Practice: Analysing Ontologies as Networks to Support Community Recognition. *In: Proceedings of 17th IFIP World Computer Congress.* Montreal, Canada: Kluwer.

Ontoprise. 2002. *Ontoprise Limited.* Available from: `http://www.ontoprise.de/`.

Pam, A. 1995 (September). Where World Wide Web Went Wrong. *In: Proceedings of Asia-Pacific World Wide Web Conference, Hong Kong.* Available from: `http://www.csu.edu.au/special/conference/apwww95/sept-all.html`.

Pearl, A. 1987. Sun's Link Service: A Protocol for Open Linking. *Pages 137–146 of: Proceedings of the Hypertext 1989 Conference on Hypertext, Pittsburgh, Pennsylvania, USA.* ACM Press.

Pepper, S., & Moore, G. 2001. *XML Topic Maps (XTM) 1.0.* Available from: `http://www.topicmaps.org/xtm/1.0/`.

Peters, J. 1995. *Electronic Peer Review.* Keynote at the Electronic Peer Review Internet Conference.

Peters, J. 1996. The hundred years war started today: an exploration of electronic peer review. *Journal of Electronic Publishing.* Electronic journal. Available from: `http://www.press.umich.edu/jep/works/PeterHundr.html`.

Pikrakis, A., Bitsikas, T., Sfakianakis, S., Hatzopoulos, M., DeRoure, D. C., Hall, W., Reich, S., Hill, G. J., & Stairmand, M. 1998. MEMOIR - Software Agents for Finding Similar Users by Trails. *Pages 453–466 of: Proceedings of the Third International Conference and Exhibition on The Practical Application of Intelligent Agents and Multi-Agents. London, UK.* Springer.

Planet, Bright. 2000 (July). *The Deep Web: Surfacing Hidden Value.* Available from: `http://www.brightplanet.com/deepcontent/tutorials/DeepWeb/deepwebwhitepaper.pdf`. White Paper.

Price, D. 1965. Networks of scientific papers. *Science*, **149**, 510–515.

Rath, H. Holger, & Pepper, S. 2000. *Topic Maps - Introduction and Allegro.* Available from: `http://www.empolis.com/englisch/pdf/Rath-introduction.pdf`.

Rittle, H. 1972. On the Planning Crisis: Systems Analysis of the 'First and Second Generations'. *Bedriftsokonomen*, **8**, 390–396.

Rizk, A., & Sauter, L. 1992. Multicard: an open hypermedia system. *Pages 4–10 of: Proceedings of the ACM Hypertext 1992 Conference, Milan, Italy.* ACM Press.

Robert, L., & Lecolinet, E. 1998. Browsing Hyperdocuments with Multiple Focus+Context Views. *Pages 293–294 of: Proceedings of the ACM Hypertext 1998 Conference, Pitsburgh PA, USA.* ACM Press.

Robie, J., Chamberlin, D., & Florescu, D. 2000. *Quilt: an XML Query Language*. Available from: `http://www.almaden.ibm.com/cs/people/chamberlin/quilt_euro.html`.

Salton, G. 1989. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley: Reading, Mass, USA.

Schulman, J. 2000 (May). *Using Medical Subject Headings (MeSH) to examine patterns in American medicine*. Course paper STS 5206. Falls Church, Virginia Polytechnic Institute and State University, Northern Virginia Center.

Senders, J. 1977. An on-line scientific journal. *The Information Scientist*, **11**(1), 3–9.

Shackel, B. 1982. Plans and initial progress with BLEND - an electronic network communication experiment. *International Journal of Man-Machine Studies*, **17**, 225–233.

Shadbolt, N. 2001a (October). *AKT Manifesto*. Available from: `http://www.aktors.org/publications/Manifesto.doc`.

Shadbolt, N. 2001b. Knowledge Technologies. *Ingenia, The Royal Academy of Engineering*, **8**(May), 58–61.

Shneiderman, B. 1987. User interface design for the Hyperties electronic encyclopedia. *Pages 199–204 of: Proceedings of the Hypertext 1987 Workshop, University of North Carolina, USA*. ACM Press.

Shneiderman, B., & Kearsley, G. 1989. *Hypertext Hands-On! An Introduction to a New Way of Organizing and Accessing Information*. Reading, MA, USA: Addison-Wesley.

Shum, S. Buckingham, & Sumner, T. 2001. JIME: An Interactive Journal for Interactive Media. *First Monday*, **6**(2). Electronic Journal. Available from: `http://firstmonday.org/issues/issue6_2/buckingham_shum/index.html`.

Shum, S. Buckingham, Motta, E., & Domingue, J. 1999. Representing Scholarly Claims in Internet Digital Libraries: A Knowledge Modeling Approach. *Pages 423–442 of: Proceedings of 3rd European Conference on Research and Advanced Technology for Digital Libraries*. Edusite.

Simpson, R. 2001. ConceptLab: An Information Structures Spatial Hypermedia Environment. *Page 6 of: Poster Proceedings of the ACM Hypertext 2001 Conference, Aarhus, Denmark*. ACM Press.

Skuce, D. 1996. Conventions for Reaching Agreement on Shared Ontologies. *In: Proceedings of 9th BANFF Knowledge Acquisition for Knowledge-Based Systems Workshop, Alberta, Canada*. Available from: `http://ksi.cpsc.ucalgary.ca/KAW/KAW96/KAW96Proc.html`.

Small, H. 1973. Co-Citation in the Scientific Literature: A New Measure of the Relationships Between Two Documents. *Journal of the American Society for Information Science*, **24**, 265–269.

Smith, M. 1958. The Trend Toward Multiple Authorship in Psychology. *American Psychologist*, **13**, 596–599.

Smith, T., & Bernhardt, S. 1988. Expectations and experiences with HyperCard: a pilot study. *Pages 47–56 of: Proceedings of the 6th international conference on Systems documentation proceedings*. ACM Press.

Staab, S., Angele, J., Decker, S., Erdmann, M., Hotho, A., Maedche, A., Schnurr, H., Studer, R., & Sure, Y. 2000 (May). Semantic community Web portals. *In: Proceedings of the Tenth World Wide Web Conference, Hong Kong, China*. Available from: `http://www9.org/w9cdrom/index.html`.

Stackpole, L., & Atkinson, R. 1998. *The National Research Library Alliance: A Federal Consortium Formed to Provide Inter Agency Access to Scientific Information*. Available from: `http://www.library.ucsb.edu/istl/98-spring/article6.html`.

Stevens, R., Goble, C., & Bechhofer, S. 2000. Ontology-based knowledge representation for bioinformatics. *Briefings in Bioinformatics*, **1**(4), 398–414.

Sumner, T., & Shum, S. Buckingham. 1998. From Documents to Discourse: Shifting Conceptions of Scholarly Publishing. *Pages 95–102 of: Proceedings of the CHI 1998 Human Factors in Computing Systems, Los Angeles, California, USA*. ACM Press.

Tennison, J., & Shadbolt, N. 1998 (April). APECKS: A Tool to Support Living Ontologies. *In: Proceedings of 11th Knowledge Acquisition Workshop*. Available from: `http://ksi.cpsc.ucalgary.ca/KAW/KAW98/KAW98Proc.html`.

The Economist. 1997. Methodical progress: Applying the scientific method to the processes of science can be illuminating. *The Economist*, **344**(September), 89–90.

Theng, Y. 1999. 'Lostness' and digital libraries. *Pages 250–251 of: Proceedings of the fourth ACM conference on Digital libraries*. ACM Press.

Toulmin, S. 1958. *The Uses of Argument*. Cambridge University Press: Cambridge, UK.

Trigg, R. 1983 (November). *A Network-Based Approach to Text Handling for the Online Scientific Community*. Ph.D. thesis, Department of Computer Science, University of Maryland, College Park, MD, USA. Technical Report TR-1346.

Trigg, R., & Weiser, M. 1986. TEXTNET: A Network-Based Approach to Text Handling. *ACM Transactions on Office Information Systems*, **4**(1), 1–23.

Uschold, M. 2001 (May). Where are the Semantics in the Semantic Web? *In: Workshop on Ontologies in Agent Systems held at the 5th International Conference on Autonomous Agents, Montreal, Canada*. unpublished.

Uschold, Mike. 1996. Building Ontologies: Towards a Unified Methodology. *In: Proceedings of the 16th Annual Conference of the British Computer Society Specialist Group on Expert Systems*.

Uschold, Mike, & Gruninger, Michael. 1996. Ontologies: Principles, Methods, and Applications. *The Knowledge Engineering Review*, **11**(2), 93–136.

Valauskas, E. 1997. Waiting for Thomas Kuhn: First Monday and the Evolution of Electronic Journals. *First Monday*, **2**(12). Electronic journal. Available from: `http://www.firstmonday.dk/issues/issue2_12/valauskas/index.html`.

van der Vet, P., & Mars, N. 1998. Bottom Up Construction of Ontologies. *IEEE Transaction on Knowledge and Data Engineering*, **10**(4), 513–526.

van Elst, L., & Abecker, A. 2001. Domain Ontology Agents in Distributed Organizational Memories. *Pages 39–48 of: Proceedings of International Joint Conference on Artificial Intelligence 2001, Washington, USA*. Morgan Kaufmann Publishers: San Francisco, USA.

van Harmelen, F., Patel-Schneider, P., & Horrocks, I. 2001 (March). *DAML+OIL*. Available from: `http://www.daml.org/2001/03/reference.html`. Revision 4.2.

Vargas-vera, M., Domingue, J., Kalfoglou, Y., Motta, E., & Shum, S. Buckingham. 2001. Template-driven information extraction for populating ontologies. *In: Proceedings IJCAI 2001 workshop on Ontologies Learning*. Available from: `http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-31/`.

W3C. 1997. *Meta Content Framework Using XML*. Available from: `http://www.w3.org/TR/NOTE-MCF-XML/`.

W3C. 1999a. *Resource Description Framework (RDF) Model and Syntax Specification*. Available from: `http://www.w3c.org/TR/REC-rdf-syntax/`.

W3C. 1999b. *XML Path Language (XPath) Version 1.0*. Available from: `http://www.w3c.org/TR/xptr`.

W3C. 1999c. *XSL Transformations (XSLT)*. Available from: `http://www.w3c.org/TR/xslt`. W3C Recommendation.

W3C. 2000a. *Extensible Markup Language (XML) 1.0 (Second Edition)*. Available from: `http://www.w3c.org/TR/REC-xml`.

W3C. 2000b. *Resource Description Framework (RDF) Schema Specification 1.0*. Available from: `http://www.w3c.org/TR/rdf-schema/`.

W3C. 2000c. *XML Linking Language (XLink) Version 1.0*. Available from: `http://www.w3c.org/TR/xlink/`.

W3C. 2001a. *SiRPAC - Simple RDF Parser & Compiler*. Available from: `http://www.w3.org/RDF/Implementations/SiRPAC/`.

W3C. 2001b. *XML Pointer Language (XPointer) Version 1.0*. Available from: `http://www.w3c.org/TR/xpath`.

W3C. 2001c. *XML Schema Part 0: Primer*. Available from: `http://www.w3c.org/TR/xmlschema-0/`.

Warren, D. 1977. *Implementing Prolog - compiling logic programs (1 and 2)*. Tech. rept. 39 and 40. DAI Research, University of Edinburgh.

Waterworth, J., & Chignell, M. 1991. A Model for Information Exploration. *Hypermedia*, **3**(1), 35–38.

Wells, H. 1938. *World Brain*. New York: Doubleday.

Wessels, Duane. 1996. *The Summary Object Interchange Format (SOIF)*. Available from: `http://kvtr.elte.hu/harvest/node42.html`.

Whalley, P. 1990. Models of hypertext structure and learning. *In:* Jonassen, D., & Heinz, M. (eds), *Designing Hypermedia for Learning*. New York: Springer-Verlag.

White, H., Buzydlowski, J., & Lin, X. 2000. Co-Cited Author Maps as Interfaces to Digital Libraries: Designing Pathfinder Networks in the Humanities. *Pages 25–30 of: Proceedings of the International Conference on Information Visualisation (IV2000)*. IEEE Press.

Wills, G. 1995. Embracing Electronic Publishing. *The Learning Organisation*, **2**(4).

Woodward, H., McKnight, C., Pritchett, C., & Rowland, F. 1997. Use of Electronic Journals by Academic Staff and Postgraduate Students in an Information-literate University. *New Book Economy*, 274–281.

Yankelovich, N., Haan, B., Meyrowitz, N., & Drucker, S. 1988. Intermedia: The Concept and the Construction of a Seamless Information Environment. *IEEE Computer*, **21**(1), 81–96.

Young, L. De. 1990. Linking considered harmful. *Pages 238–249 of: Hypertext: Concepts, Systems and Applications, Proceedings of the Hypertext '90 Conference, INRIA, France*. ACM Press.

Zellweger, P. 1991. Structure, Navigation and Hypertext: The Status of the Navigation Problem. *Pages 363–366 of: Proceedings of the ACM Hypertext 1991 Conference, San Antonio, Texas, USA*. ACM Press.

Index