

1997

A hypermedia database to manage World-Wide-Web documents


Schubert Shou Boon FOO

Ee Peng LIM

Singapore Management University, eplim@smu.edu.sg

DOI: [https://doi.org/10.1016/S0378-7206\(96\)01088-9](https://doi.org/10.1016/S0378-7206(96)01088-9)

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 Part of the [Databases and Information Systems Commons](#), and the [Numerical Analysis and Scientific Computing Commons](#)

Citation

FOO, Schubert Shou Boon and LIM, Ee Peng. A hypermedia database to manage World-Wide-Web documents. (1997). *Information & Management*. 31, (5), 235-249. Research Collection School Of Information Systems.

Available at: https://ink.library.smu.edu.sg/sis_research/65

This Journal Article is brought to you for free and open access by the School of Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email libIR@smu.edu.sg.

Research

A hypermedia database to manage World-Wide-Web documents

Schubert Foo^{*}, Ee-Peng Lim

Division of Software Systems, School of Applied Science, Nanyang Technological University, Nanyang Avenue, Nanyang, Singapore 639798

Abstract

The surge of interest in the World-Wide-Web (WWW) with its potential commercial payoff has resulted in an explosion of information as organisations join in to publish and do business on the Internet. A related development, Intranet, which basically uses the same technology to build private corporate WWW-based networks, has emerged to provide cost-effective and efficient groupware and information management solutions for organisations. As a result, the Hyper Text Markup Language (HTML), used for constructing WWW documents, has become a contender for authoring future office documents. With this scenario, this study examines WWW practices and highlights the inadequacy and drawbacks of current publishing on the WWW. This paper demonstrates the need as well as the advantages in having a hypermedia database system to maintain and manage office or WWW documents and describes the design and prototyping of a hypermedia database system. © 1997 Elsevier Science B.V.

Keywords: Data modelling; Hypermedia database; Hyper Text Markup Language; Information retrieval; Internet; Intranet; Query; World-Wide-Web

1. Introduction

Internet, the world network inter-connecting other networks across continents, is already the IT phenomenon of the nineties. Originally conceived in 1973 by the U.S. Advanced Research Projects Agency (ARPA) as a research program to develop a suite of communication protocols to allow networked computers to be connected transparently across multiple, linked packet networks, it has grown substantially to become a world-wide network. This rapid growth stems from continual support from the US Federal Government, international adoption of its Transmission Control Protocol/Internet Protocol (TCP/IP)[1] communica-

tion protocols and introduction of the World Wide Web (WWW) and commercial facilities. The bulk of the system today is made up of private networking facilities in educational and research institutions, businesses and governmental organisations across the world. It is estimated that up to one billion users will have access to the Internet by the turn of the century.

Internet exhibits a host of features and facilities whose contents are contributed from many different people and parties around the world. The more popular uses of Internet include electronic mailing, **USENET** Newsgroup (a world-wide distributed discussion system), **WWW** (a hypermedia system for browsing and retrieving Internet resources), **File Transfer Protocol** (FTP – a file transfer facility for uploading and downloading files around the world), **Gopher** (a globally searchable collection of menu-based hierarchical

^{*}Corresponding author. Tel.: 7994854; fax: 7926559; E-mail: {assfoo,aseplim}@ntu.edu.sg

information resources) and Internet Relay Chat (IRC – a global chat line).

The popularity of Internet is derived from a number of factors:

- **Easy accessibility.** Individuals and organisations without computing resources can gain easy access to the Internet by acquiring an off-the-shelf computer, a low-cost modem and an account with a commercial Internet service provider (ISP) at competitive rates;
- **Ease of use.** The development of better human-computer interfaces and improvement in the level of computer literacy among the people of many developing nations around the world has resulted in more people finding the Internet technology accessible and easy to use.
- **Global connectivity.** By virtue of Internet's global inter-connectivity, all that is required is the knowledge of another user's 'Internet address' before communication can take place between parties almost anywhere in the world. Participation in the WWW and USENET Newsgroups guarantees a world-wide audience.
- **Speed of access to Internet facilities.** Access to facilities is almost instantaneous. Electronic mail sent from anywhere reaches its destination within minutes. Downloading freeware or shareware programs from an ftp site and installing is achievable quickly. Likewise, browsing and searching for information at WWW sites can be very fast.
- **Commercialisation opportunities.** The advent of the WWW has transformed the Internet into an environment that can support commercial interests and form support business-oriented networks. From an organisation's standpoint, the Internet offers a novel and attractive feature (unlike normal mass media tools, such as television or newspapers) to allow customers to 'interact' with multimedia information provided by the organisation and, in some cases to carry out business transactions on the spot.
- **Scale and distance independence.** Given the technical know-how of the technology, it takes only a little extra effort and costs to reach out to thousands of users than to reach out to an individual. Additionally, the cost to reach someone at the other side of the world is typically the same as for someone in the same country.

- **Low barriers to entry and equal opportunities.** As a result of the ease of access and an open and extremely democratic environment, equal opportunities exists for all in this world-wide business community. Large organisations investing more financial resources into making more impressive WWW pages may not necessarily have any advantage over an individual or a small team who may focus on better communications and support for its niche market customers.

Internet has indeed revolutionised business, in that it redefines the methods used in traditional business practice and offers another important channel for mass communication that is likely to grow even further in the future. Depending on the type and nature of the business, Internet provides a platform to carry out a host of potential business applications:

- as a public relations tool in establishing a global presence and heighten public interest;
- as a marketing tool in advertising goods and services and in opening up international markets;
- as a marketplace, in selling goods and services;
- as an information kiosk, in providing up-to-date business information, answering frequently-asked-questions (FAQs) and releasing time-sensitive information;
- as an alternative support tool, in answering customers' queries and soliciting feedback;
- as a research or information gathering tool for market surveys, product launches, and even in obtaining solutions to problems;
- as a human resource tool for staffing and recruiting purposes;
- as a support tool, in serving mobile employees or telecommuters;
- as a computer supported co-operative work (CSCW) tool in facilitating groupwork and communication in an organisation or across organisations, both locally and globally.

The WWW in its current form, has already made many such applications realisable. WWW-based business computing is already seen as a new competitive business weapon. With the move towards the era of global information technology, regionalisation and globalisation of businesses, and the ongoing enhancement of Internet's technology (such as system and data

security, encryption techniques, and communication protocols), the WWW is set to expand rapidly in the commercial dimension. This will be especially true when the pay-offs of organisations investing in this technology are revealed.

Furthermore, **Intranets** are proving to be an exceptionally cost-effective way to distribute reports, track assets, improve employee communications, provide access to diverse corporate databases, distribute and run applications and enhance collaborations at a fraction of a cost of groupware solutions.

2. World Wide Web

2.1. Basic concepts

The WWW is a wide area hypermedia information retrieval system aimed at giving universal access to a large universe of documents. It is organised as a set of Hyper Text Transfer Protocol (HTTP)[17] servers designed specially for rapid distribution of hypermedia documents. Hypermedia is a way of representing and accessing information. It views the information space as a graph whose nodes store information and whose arcs (links) represent semantic relationships between nodes. The links are usually associated with words or image icons within the document that describe the meaning of the links. A set of nodes and the corresponding links makes up a hypergraph (or hyperstructure) that can be represented as a network. This is shown in Figure 1.

Node and node contents are independent objects; the association is made through a general referencing mechanism that allows the referencing of an entire

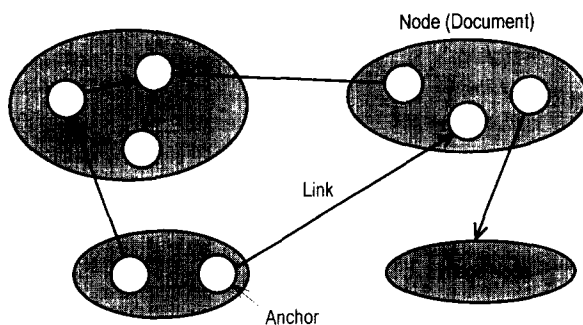


Fig. 1. A hyperstructure of a hyperdocument.

document. As hypermedia systems are often used to structure information contained in pre-existing documents, the same document can be referenced by different nodes. Thus, the hypermedia model allows the sharing of node contents. When a hyperlink is activated (usually by double clicking the mouse button), the system extracts its destination node and correspondingly loads the document as the currently referenced document. Users navigate around the hyperspace by moving from link to link, or by specifying direct links.

The Hyper Text Markup Language (HTML)[18] is a hypermedia language used to construct WWW documents. It is designed to specify the logical organisation and formatting of general text documents, with extensions to include inline images, audio, video clips, fill-in forms and hyperlinks to other HTML documents and other Internet resources (such as files, ftp, USENET, telnet). As a result, HTML is not only applicable to WWW documents alone, but also to the environment of office documents within an organisation. It is expected that HTML will emerge as the main contender for office documentation and that future office documents will be authored using authoring tools developed for the WWW. Hyperlinks will present a powerful means for navigation and to cross reference information. Each WWW resource has a unique address known as the Uniform Resource Locator (URL)[19].

In order to publish and read information on the WWW, three components are required:

1. *Access to a HTTP (Web) server.* This is normally part of the package offered when an individual or organisation subscribes to an information service provider. For organisations with computing resources, they can set up their own HTTP server and get it connected to the Internet via a commercial network provider.
2. *An HTML authoring tool.* This is used to compose the WWW documents (also known as Web pages). Since an HTML document is basically an ASCII text file that contains embedded HTML tags, any text editor can be used for this purpose. However, this is not the normal way to create HTML documents, since many existing commercial and free-ware HTML editors are available for this purpose. Most editors are user-friendly and easy to use.

Some uses the WYSIWYG (What-You-See-Is-What-You-Get) paradigm for authoring. Most editors have a browser so that the end result can be reviewed immediately after composition. Corel WebDesigner [2] and HoTMetal Pro 3.0 [12] are examples of MS-DOS based editors. Phoenix [16], Emacs hhm [8] and HoTMetal 1.0 are examples of UNIX based editors (Laviolette[5] keeps a recent compilation of HTML editors and related tools)

- 3. *A Web browser.* This is the user client software used for navigating and reading the huge volume of Web pages stored on various Web servers on the Internet. Although both forms of text and graphical-based browsers are available, the latter are much more popular, as they can display graphics and icons that are commonly found in most Web documents. Netscape Navigator [11], NCSA Mosaic for Windows[10] and Windows Internet Explorer [9] are among the three most popular Web browsers currently in use.

2.2. Case example: modelling an university environment for publication

The academic environment of Nanyang Technological University (NTU) is used as an example of an

information profile (i.e., information that is to be presented on the WWW) modelled to a form suitable for publication on the WWW. Its structure is shown in Figure 2.

By using a hierarchical tree structure, the various components of the organisation can be defined and cross-linked to each other. The links are hyperlinks that connect one piece of information to another. Cross-hierarchical tree structures can easily be achieved by linking hierarchical tree structures together. Breaking up of the information profile in a systematic and logical manner will result in tree structures which are used as the framework for publication. This can also be applied in any business organisation.

At the top of the tree is the organisation's home page. In WWW terminology, this refers to the document intended to be viewed first. It contains introductory information and/or a master menu of documents within the publication. It is generally associated with a particular Web site, person, named collection, or business organisation. This master menu contains the hyperlinks to subsequent home pages or documents.

Using the NTU example, the first level menu contains the main focus and is separated into a number of different general categories:

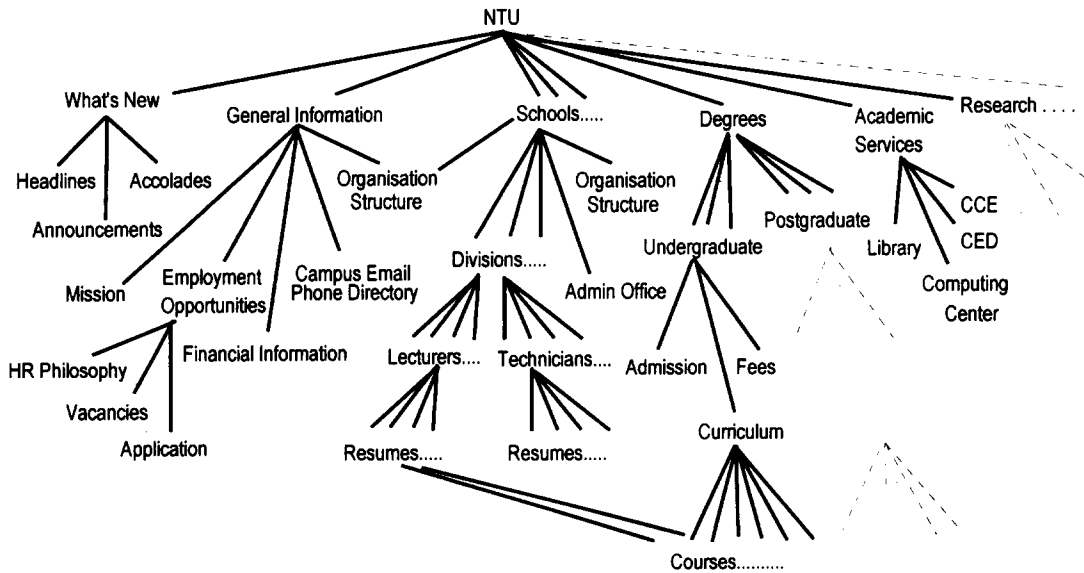


Fig. 2. Information profile of Nanyang Technological University on the WWW.

- *What's new!* This is a public relations area; it contains announcements and headlines (such as new product launches and accolades).
- *General information.* This serves to provide general information about the organisation: its mission, objectives, organisation structure, financial information, human resource (HR), and other pertinent information. For a business organisation, it may contain the company's financial highlights, balance sheets, profit and loss accounts, funds flow statement, excerpts from the company's annual report, human resource philosophy, employment opportunities, vacancies, duties, requirements and application procedures.
- *Schools and degrees.* These are the general categories that provides information on the main thrust of the organisation: the products and services offered by the organisation. For large organisations this may be replaced by the various strategic business units (SBUs) with further breakdowns to define their products and services. In the example, each school (equivalent to a product division) is organised into a number of groups with a supporting administrative office; each division comprises lecturers and technicians, with their personal resumes. Resumes contain information on the courses taught, research interests, and industrial consultancy expertise. Each school offers different undergraduate and postgraduate degree programs (products) together with the associated information of admission requirements, fees, financial assistance, curriculum structure, and detailed course descriptions. It should be apparent that hyperlinks are used to link related information (e.g., resume is linked to course). Links can have a one-to-one, one-to-many, and many-to-many relationships. Links also play the important role of eliminating repetitive information.
- *Academic services.* This plays the supporting role of the organisation. In the example, the services include the Library, Computer Centre, Centre for Continuing Education, Centre for Educational Development, etc. In a business organisation, this may take on the role of customer support and enquiries. It may be used by customers to request product servicing, checking for availability of spare parts, and to pose other queries.

With the detailed information available for each component, it becomes an easy task to compose the Web documents and to cross-link them using HTML conventions (i.e., anchors and URLs). As the information is contributed from various sources within the organisation, there is generally an overall administrator (or Chief Information Officer) who is in-charge of the overall project, with various centres set up for actually defining and maintaining the information. Guidelines and rules of thumb are distributed to the network service personnel to ensure homogeneity and conformity in presenting information. A schedule of regular updates and maintenance and a system to handle proposed changes to existing information is also needed to ensure overall data integrity and accuracy of information in the system.

2.3. *Current problems with WWW practices*

Organisations throughout various industries have jumped onto the band-wagon and started to use Web technology. Companies without computing resources will have to rely on commercial vendors to put up and maintain the information on their behalf. However, publishing in the WWW and using it in the present manner can lead to a number of future difficulties:

- *Maintainability of publication.* As organisations realise the ease with which they can put information on the WWW, there will be a tendency to enter more rather than less information; this can result in a rapid increase in the number of Web pages. Ultimately, it will reach a point where maintaining and ensuring the accuracy of information becomes difficult.
- *Integrity of data.* The chances of duplication or inaccuracy will arise as different people contribute towards publishing. The integrity of the data can be compromised. Information should be cross-linked to that provided by other authors. However, deletion of other people's documents will cause integrity problems. In addition, information updates are a problem if it is shared among different domains. For example, a change in the pricing of goods or services may require corresponding changes to a number of Web pages that are maintained by different people. Thus, any loss of data integrity of the system can have potentially severe business con-

sequences, especially if the system is used for direct business transactions. In order to avoid such problems, there must be a structure, a change mechanism, and manpower resources for the overall management of the system. Such a scenario may not be feasible for organisations with limited financial and manpower resources. Additional costs is thus incurred by having such a system in place.

- *Inadequate search facilities on the WWW.* There is currently no high level query language for locating, filtering, and presenting WWW information. Searching is achieved by using any of the many existing Internet search engines. Search engines work in many different ways: some search titles of headers of documents, other search the documents themselves, and still others search indexes of other directories. Some are specially dedicated to the WWW (e.g. Lycos [6], Infoseek Guide [4], etc.), while others support searching of other Internet facilities, such as USENET, FTP and Gopher (e.g. Magellan [13], Alta Vista [3], etc.). However, searching in this manner usually results in a long list of matches (with possible duplicates) if indexing is used; these must then be explored one at a time. Furthermore, the home page of the document is usually presented so that navigating around the hypermedia space via hyperlinks is still required until the desired information is located. Such a form of searching is suitable in “Net Surfing” for looking up general information from several sources. However it is inefficient, time consuming, and unsuitable for cases when a specific and detailed query is known.
- *Degradation of performance with number of users.* As the number of WWW users grow, it will eventually overload the underlying network capabilities and result in a degradation of performance. Such observations are already apparent when accessing the more popular sites with extremely high traffic. Depending on the severity of the problem and expectations of the users, this could lead to frustration and subsequent loss of interest in using the WWW. For organisations that have invested heavily on doing business via the Internet, this is a real threat about which they can do little.

Bearing such factors in mind, we propose the use of an experimental hypermedia database system for the

management and publication of WWW documents for organisations intending to use the Internet for business.

3. A Hypermedia database system for WWW documents

D4W3 is a hypermedia database system being developed at NTU. It is designed to address the management of HTML-based documents (including WWW documents) within organisations. We now define the **system requirements** satisfied by the D4W3 design.

- **Hypertext editing.** As the HTML language becomes a well-recognised standard, a wide range of tools has been made available as public domain software. Tools include WWW browsers, editors, parsers and translators. Office documents authored in HTML allow the inclusion of text, multimedia objects, and hyperlinks that provide the means for document navigation. Some of the HTML documents can be made available on the Internet for public viewing, while others may be kept for internal use only.
- **Multiple user support in a networked environment.** Being an office document management system, D4W3 must be able to support concurrent requests initiated by users working on different personal computers. All office documents have to be collected and stored at a centralised database that can be accessed remotely. Figure 3 depicts an office environment in which a number of users’ machines and a central repository of office documents are linked together by a local area network. A gateway can be provided to allow remote access from the outside world. To prevent one user from tampering or accidentally modifying other users’ documents, some form of security must be included.
- **Flexible search and retrieval capabilities.** To support content-based queries on the office documents, D4W3 must allow users to specify queries with keywords as well as other important attributes of the documents, such as author, subject, title, etc. In the D4W3 database, every document is assigned a unique document identifier that is the URL of the document. The D4W3 query engine can therefore

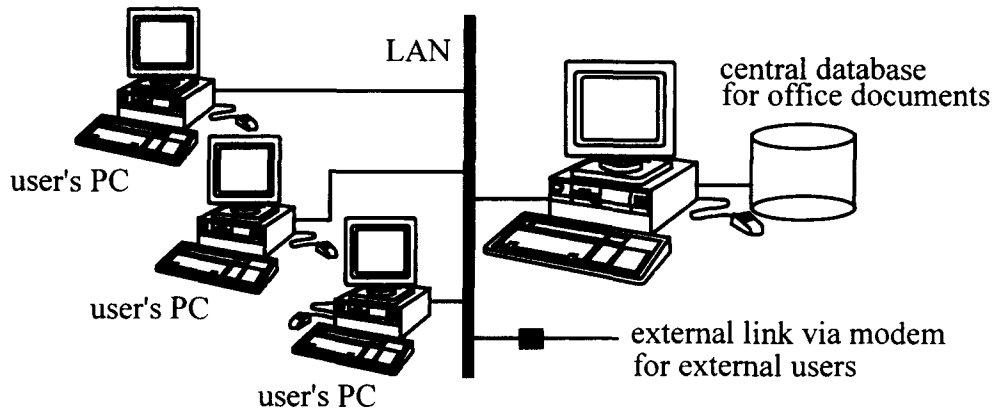


Fig. 3. Document management in a networked environment.

support retrieval by URLs. Such a search and retrieval facility surpasses the commercial search engines available on Internet. For instance, a user can specify a search to locate documents of a subject area authored after a particular date.

- **Easy document import and authoring functions.** Importing a bulk of documents and authoring them are two important consideration in D4W3. The former simplifies the migration of existing office documents into the D4W3 database. The latter facilitates the creation of new documents and updates on existing documents. Existing documents should be appended to the D4W3 database in batch mode. To benefit from a wide variety of WWW authorware, D4W3 must adopt a flexible design to interface with different WWW authoring tools.

3.1. The D4W3 architecture

A general overview of the architecture of D4W3 is shown in Figure 4. An HTML-based document application can be constructed by using the following modules:

3.1.1. User modules

D4W3 is designed to support hypermedia applications including three important applications: a **Querying and Browsing Tool**, an **Import Tool** and an **Authoring Tool**. Other hypermedia applications if necessary can be added in future. The Querying and Browsing Tool allows users to formulate content-based searches and to view hypermedia documents. Using the Import Tool, collections of hypermedia documents can be inserted into the D4W3 database

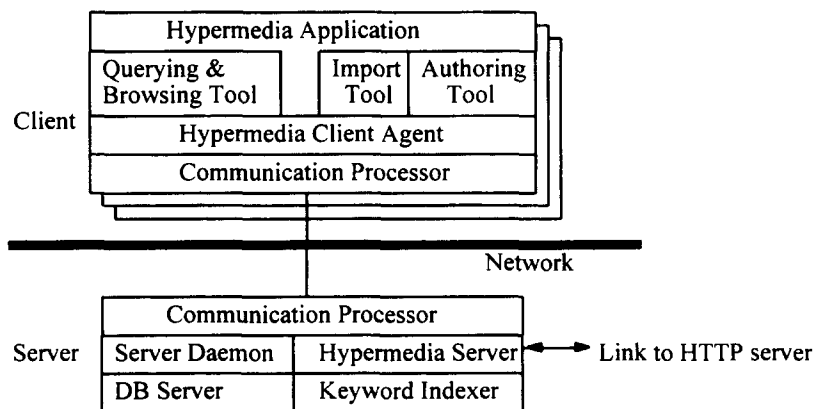


Fig. 4. D4W3 architecture.

in batch mode. The Authoring Tool facilitates the creation of new documents and modification of existing documents.

The **Hypermedia Client Agent** provides a common client machine interface to the D4W3 database server modules which reside on a server host. Based on the client agent's library of application programming interface (API) functions, different types of hypermedia applications can be developed. A call to any API function is evaluated by sending a request to the D4W3 server and receiving its response. The **Communication Processor** on the client machine is responsible for sending and receiving messages to the D4W3 server. Before messages are sent, they are encoded according to the request type and the related parameters. Messages received are decoded in the reverse manner.

3.1.2. Server modules

Like the client machines, the D4W3 server requires a **Communication Processor** to interpret incoming messages and to construct outgoing messages. To handle concurrent multiple client-server sessions, a **Server Daemon** must listen to incoming server connection requests. For each client-server connection, the daemon creates a **Hypermedia Server** process, that handles all queries from the client within the session established by the connection. To support flexible queries, some important attributes of hypermedia documents are extracted and stored in a **Database Server**. For practical reasons, the database server used by D4W3 is a relational database system. The hypermedia documents are stored as files fully indexed using a **Keyword Indexer**. The concept of distributed servers can be utilised, if necessary, to even out the load on the database server and improve efficiency in searches.

3.2. Hypermedia database modeling and database schema

In D4W3, hypermedia documents are represented as a **directed graph (digraph)**. A node of the digraph represents a document which can be either a HTML document or image document. At present, we have restricted D4W3 to handle only text and image data. A **directed link** from one node to another denotes a hyperlink; this allows users to navigate from one

location in a document to an image document or another location in the same or different HTML document. Given a directed link from node A to node B, we call node A the **source node** and node B the **destination node**. To conform with the HTML language, we do not allow a directed link to be defined from an image document. On the other hand, zero or more directed links can be defined between two HTML documents or within the same HTML document.

The overall D4W3 database model is depicted in Figure 5, which is an *entity-relationship diagram* consisting of entity types represented by boxes, their attributes represented by ovals, and relationship types between entities represented by diamonds. As shown in the database model, every hypermedia **Document** consists of a document id, a URL, a title and a filename. Documents can be further classified into either HTML documents or image documents. While the attributes of HTML documents include author, subject, header, and date of creation/update, there are only two attributes for image documents: image format (such as GIF, TIFF, etc.), and description (text describing the content of the image document).

The **has_link** relationship captures the directed links between HTML documents, with attributes: anchor_name, link parameter, text description, and location (offset) in the source document. The **has_image** relationship captures the directed links between HTML documents and image documents. Only its offset is kept by the D4W3 database.

To further organise hypermedia documents, D4W3 allows users to group a number of related documents together to form a **Publication**. For example, the top nodes of the page may be a publication consisting of many different documents. Every publication is owned by some user who is responsible for authoring and updating the documents in the publication. Therefore, we have an **Owner** entity type maintaining an account name and password for every owner.

Since most database systems are not designed to represent and manipulate large text objects, we have chosen to store the original hypermedia documents as ordinary files. However, the filenames of documents are maintained within the D4W3 database.

3.2.1. Database Relations

From the entity-relationship diagram, we derive the following relations to be stored in the database. In our

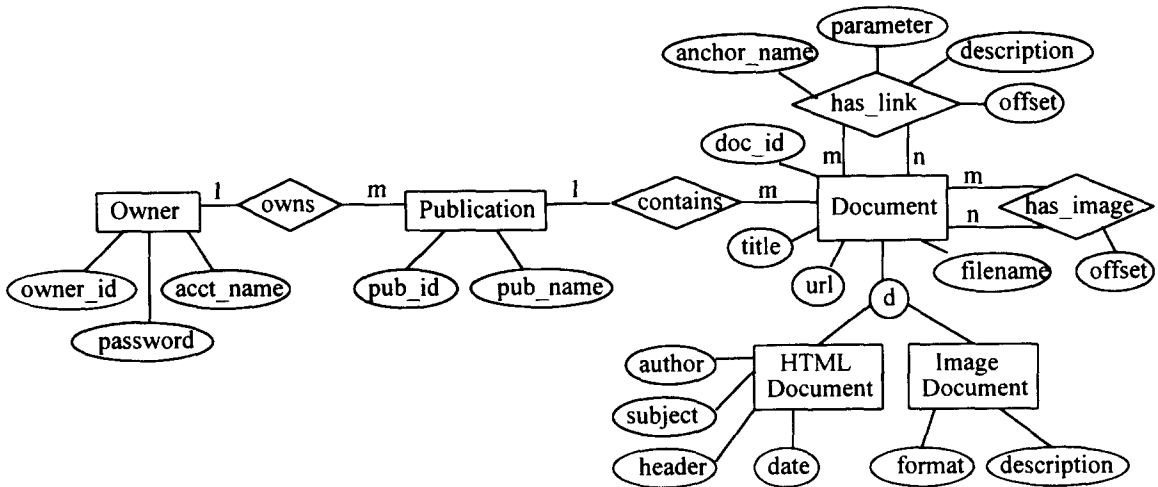


Fig. 5. Entity-relationship modelling of D4W3 database.

implementation, we have chosen **POSTGRES** [14, 15], an extended relational database management system. It allows generalisation relationship to be defined between two relations so that it can inherit attributes from its superclass; tuples in a relation are also tuples of its superclass relations. Every tuple in POSTGRES is assigned a unique tuple id. We have therefore used it in place of *owner_id* and *pub_id*.

- **OWNER**(*acct_name*, *password*)
- **PUBLICATION**(*pub_name*, *owner_id*)
- **DOCUMENT**(*url*, *title*, *pub_id*, *filename*)
- **HTML_DOC** is a subclass of **DOCUMENT**
- **HTML_DOC**(*author*, *subject*, *header*, *date*)
- **IMAGE_DOC** is a subclass of **DOCUMENT**
- **IMAGE_DOC** (*format*, *description*)
- **HAS_LINK**(*src_doc_id*, *dest_doc_id*, *anchor_name*, *parameter*, *description*, *offset*)
- **HAS_IMAGE**(*src_doc_id*, *dest_doc_id*, *offset*)

3.3. D4W3 Query facility

Users can query hypermedia documents by specifying search criteria on selected attributes and keywords. The attributes that can be queried include *author*, *title*, *subject*, and *header*. A complex search criterion can be constructed by a conjunction of multiple simple search criteria. To handle potentially large query result sets for any kind of search queries, D4W3 is designed to

keep the result sets temporally at the server so that subsets of result sets can be returned to the user. Figure 6 illustrates the steps a user has to perform in order to search and retrieve documents from the server.

A query session must be established before any query can be submitted to the server. The user then specifies search criteria on the documents' attributes and keywords. Since every document is assigned a unique URL, the query result of the search request will be represented as a set of document URLs and attribute values. The user can choose to perform one or more retrieval requests in order to obtain subsets of result from the server. Based on the summarised information about the documents that satisfy the search request, the user may wish to browse some of these documents and continue to navigate for other documents. This can be achieved by performing a **breadth-first search** on a user-selected document, and retrieving its related documents: those that can be reached by direct or indirect links. Finally, the user-selected document and its related documents are transferred to the client machine for browsing. For example, let A in Figure 7 be one of those documents which satisfy the user's search criteria. A has links to B and C which in turn have links to other documents. If the user wish to browse A and 5 of its closely related documents, a breadth-first search on the digraph will return the set of documents: {A,B,C,D,E,F}.

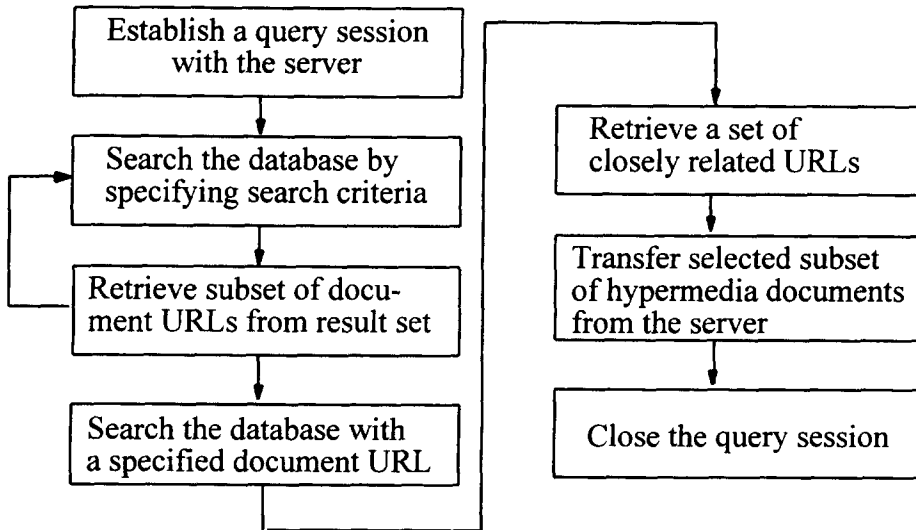


Fig. 6. Query formulation steps.

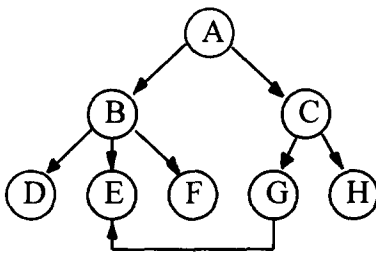


Fig. 7. Breadth-first search.

3.4. Authoring and import facility

In order to benefit and support a wide range of existing authorware that currently exist, a flexible authoring tool utilising the "plug-and-play" paradigm is chosen. With this, users can use their own preferred authorware to create and modify HTML documents. Flexibility is achieved as users can switch and upgrade authorware without affecting the functionality of the overall authoring tool.

4. Implementation Issues

The D4W3 server, authoring and import tools have been implemented on the SUN SparcStation running

SunOS 4. We have developed a D4W3 query front-end on the IBM PC running Windows 3.1. All software has been developed using C or C++. The client-server communication is implemented using TCP/IP.

To implement the query, authoring, and import facilities on the D4W3 clients, we have structured the client-server integration by designing a comprehensive set of application programming interface (API) functions. The same set can later be used to realise new D4W3 applications or tools.

4.1. User interfaces

The two main D4W3 user interfaces that have been developed are the document query front-end and the authoring tool.

4.1.1. D4W3 query front-end

As most D4W3 users are expected to be PC users, a graphical query front-end has been developed on the PC client to allow users to perform query and browsing tasks. The essential features include:

- **Integrated query and browsing:** Other than querying the documents at the D4W3 server, our query front-end allows documents to be transferred from the server to the PC client and invoked using

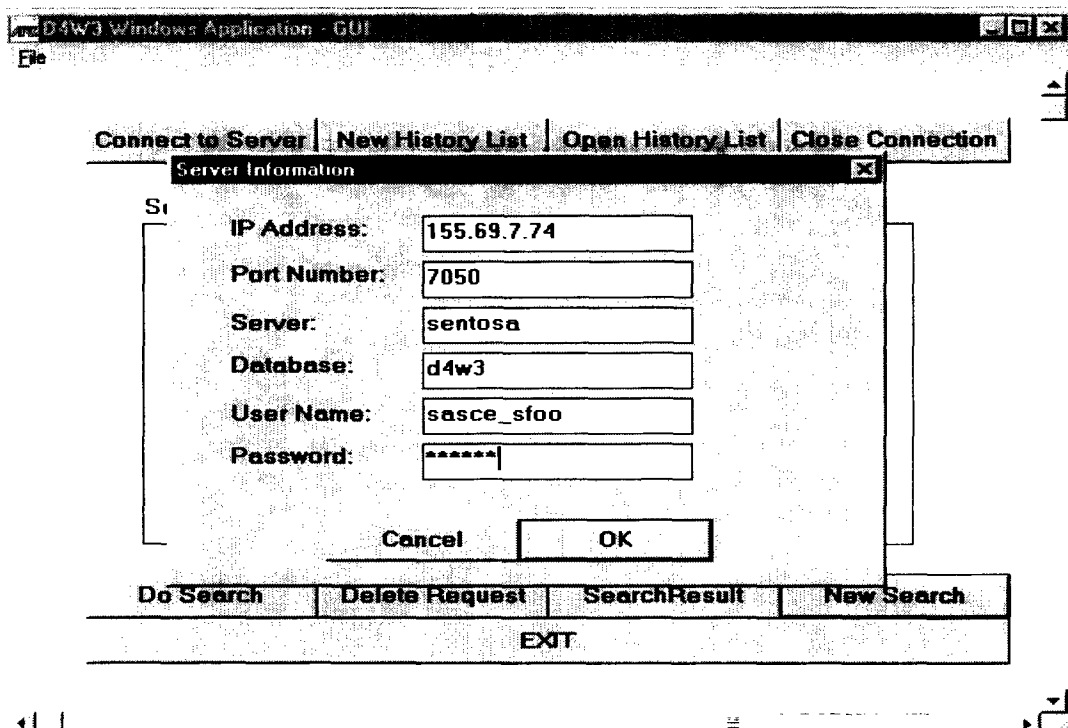


Fig. 8. D4W3 client opening screen with request for server connection.

any user-selected HTML browser to read the documents.

- Search history maintenance:** The search criteria, together with their results, are kept within a search history maintained by our query front-end. This allows users to recall the searches they have previously performed. By modifying previous search criteria, they can easily derive new ones. In addition, only full result sets of searches are stored in the search history. Partial result sets can arise because the user has not brought the complete result set from the server to the client machine. As such, these result sets are not stored since they cause integrity problems and can lead to confusion in future.

Following the query formulation steps, a user first starts D4W3 and asks for a connection to a D4W3 server, as shown in Figure 8. Subsequently, a search criteria is formulated and submitted for query, as shown in Figure 9. A list of search results is returned

to the user, as shown in Figure 10. Having identified the result set, the user selects the "Retrieve Document" option and specifies the number of levels of related-documents to be retrieved. When all the documents have been transferred to the client machine, the user can call up any Web browser to display the HTML document using the "Load Browser" option. A browser setup dialog box showing the default browser (e.g. Netscape's Navigator 2.0) is presented to the user (this is not shown here). Figure 11 shows the retrieved document been displayed using the Internet Explorer browser. Upon terminating from the browser, the application returns to the Query result display screen.

4.2. Keyword indexing and searching

Though the D4W3 server keeps extracted document attributes in POSTGRES relations, the documents themselves are stored as UNIX files. To support keyword search on these documents, a keyword index has

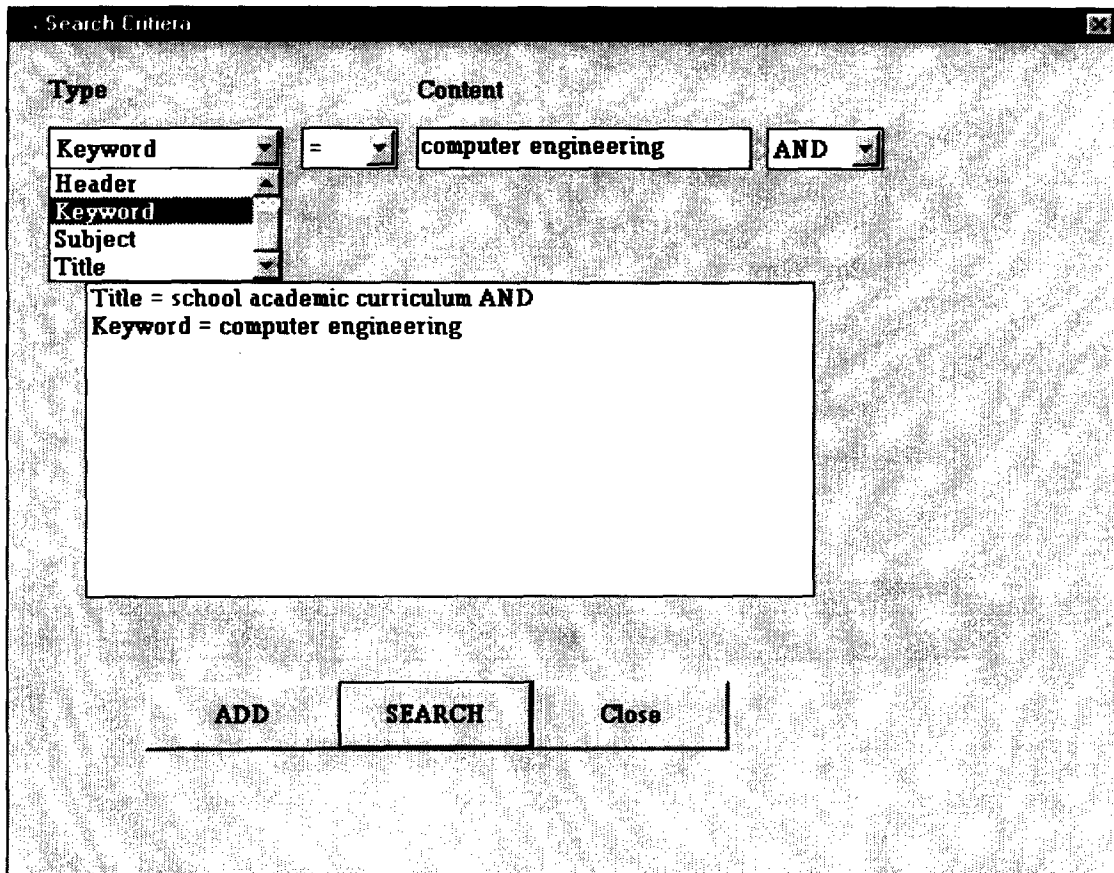


Fig. 9. Search screen for query formulation.

to be established. **GLIMPSE** (**GL**obal **IM**PLICIT **SE**arch), a search tool developed by Manber and Wu at the University of Arizona, has been used to provide the indexing features required by D4W3[7]. Its novelty is that it uses a very small index but allows very flexible full-text retrieval, including Boolean queries, approximate matching, and even searching for regular expressions. Its index builder, **glimpseindex**, is used to index HTML documents for the D4W3 server.

In D4W3, all HTML documents are stored as files in a designated directory. To allow us to identify which documents satisfy a user's keyword search criteria, we use the POSTGRES ids as the filenames for the HTML files. The keyword index of GLIMPSE allows us to perform keyword searches and obtain the filenames of HTML documents containing the keywords using the

following UNIX command line: **glimpse -c -I -l -y keywords > result_file**

4.3. System features and advantages

Using D4W3 as a means for WWW publication and office documents management offers a distinct number of features and advantages over traditional WWW practice. It:

- provides an integrated system for the management of HTML documents from authoring to publication;
- allows users to login into the system directly to carry out query and retrieval operations. It offers an enhanced query and retrieval facility not normally found in WWW search engines. Users can prescribe

The screenshot shows a window titled "Query Result Display" with the following fields and controls:

- Total number of records found :**
- Total number of records retrieved :**
- URL to be retrieved :**
- Search Result List** (scrollable area):
 - Author =sas_go
Title =SAS Course Curriculum Index
Subject =Curriculum|Course Information
URL =http://www.ntu.ac.sg/sas/curri.html
 - Author =ro_a0
Title =NTU SCHOOLS
Subject =General|School Information
URL =http://www.ntu.ac.sg/general/schools.html
 - Author =eee_go
- Retrieve More Records** (button)
- Retrieve Document** (button)
- Save Search** (button)
- Load Browser** (button)
- Close** (button)

Fig. 10. Query result display.

the amount of search results to be brought back to the client machines, as well as the amount of browsing information. As the central repository of information is smaller, stand-alone, and specific to the organisation, it results in superior efficiency of searches;

- can be configured to act as a mini-Web site to store an exact copy of the organisation's publication on the WWW, thereby allowing remote users to login and visit the local site and use normal hyperlink navigation to search for information;
- enhances quality of service (e.g. speed of access), as the organisation provides the direct computing resources over which they have full control;
- is easy to define an access control structure to control ownership of documents and access the information domain within the information profile.

Such a structure provides a framework to control the information resources of the organisation;

- preserves data integrity while editing and updating of documents. It eliminates the chances of deleting linked documents by mistake. The locking mechanism of the database system ensures document integrity during updates;
- ensures integrity of system as the database is responsible for the management of the overall system. Each time information is updated on the system, it can be triggered to automatically update the same information in the actual Web and local sites;
- provides flexibility, configurability and upgradability as users select their own HTML authorware and browsers;
- minimises duplicate information as the query engine is a convenient facility for checking prior to publication.

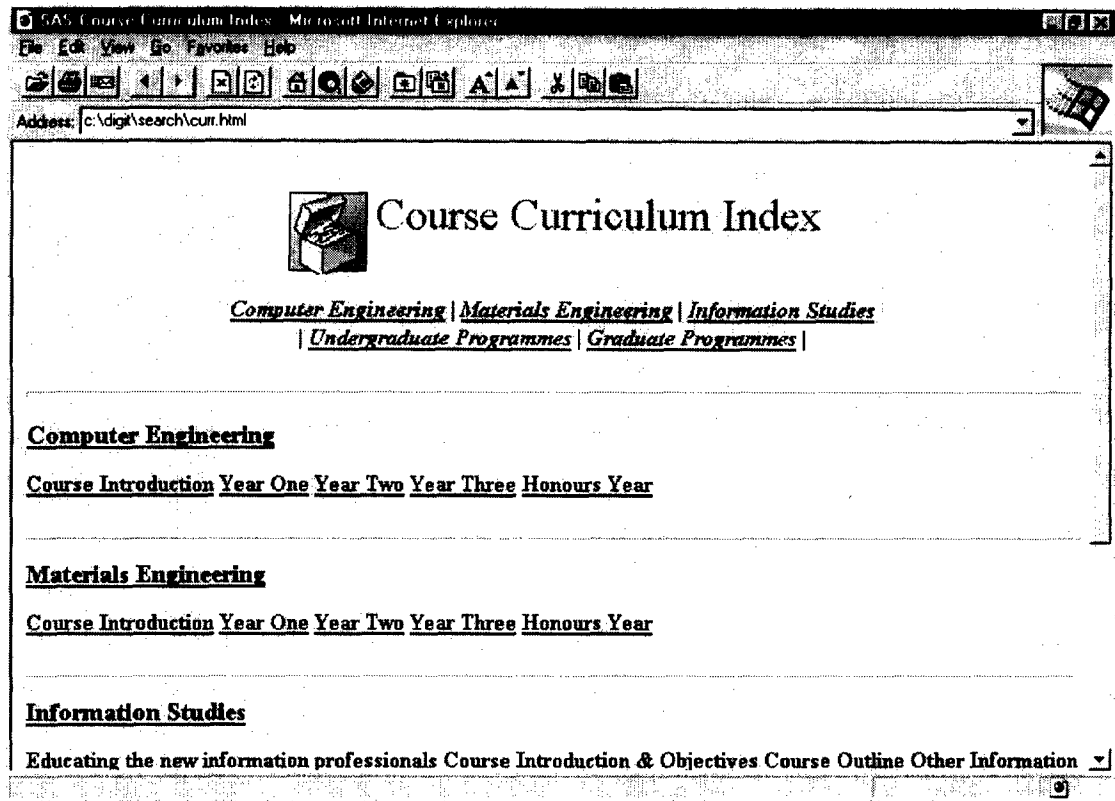


Fig. 11. Browsing the retrieved home document using Microsoft's Internet Explorer.

5. Conclusions

This research has demonstrated the need, as well as the advantages, in having a hypermedia database system to maintain and manage office or WWW documents in HTML format. A prototype, D4W3, had been implemented to demonstrate the research idea and it allows a structured and controlled means to publish, maintain and query HTML documents; this is lacking today. The design is generic and employs a plug-and-play paradigm so that users can choose their own HTML editors and browsers for authoring, etc. The database search and retrieval engine is interfaced with a set of API functions. The same set of functions can be used later to realise new D4W3 applications or tools, etc.

References

- [1] Comer, D. and Stevens, D. *Internetworking with TCP/IP*, Prentice Hall, Englewood Cliffs, NJ, 1991.
- [2] Corel Corporation, *Corel WebDesigner* <URL: <http://www.corel.com/corelweb/webdesigner/index.htm>>.
- [3] Digital Equipment Corporation, *Alta Vista* <URL: <http://www.altavista.digital.com/>>.
- [4] Infoseek Corporation, *Infoseek Guide* <URL: <http://www.infoseek.com/Home>>.
- [5] Lavolette, N., *The Web Designer* <URL: <http://www.kosone.com/people/nelson/nl.htm>>.
- [6] Lycos and Carnegie Mellon University, *Lycos* <URL: <http://lycos.cs.cmu.edu/>>.
- [7] Manber, U. and Wu, S., "GLIMPSE: A Tool to Search Through Entire File Systems", USENIX Technical Conference, Winter, 1994.
- [8] Minar N. *Emacs hhm (html-helper-mode)*, <URL: <http://www.santafe.edu/~nelson/tools/>>.
- [9] Microsoft Corporation, *Internet Explorer*, <URL: <http://www.microsoft.com/ie/ie.html>>.
- [10] NCSA, *NCSA Mosaic WWW Browser*, <URL: <http://www.ccs.org/winsock/mosaic.html>>.
- [11] Netscape Communications, *Netscape Navigator 2.02 WWW Browser*, <URL: http://home.netscape.com/comprod/products/navigator/version_2.0/index.html>.
- [12] SoftQuad, *HotMetal Pro 3.0/HotMetal 1.0*, <URL: <http://www.sq.com/products/hotmetal/hmp-org.htm>>.

- [13] The McKinley Group, *Magellan*, <URL: <http://www.mckinley.com/>>.
- [14] The POSTGRES User Manual, University of California, Berkeley, 1994.
- [15] The POSTGRES Reference Manual, University of California, Berkeley, 1994.
- [16] University of Chicago, IL, *Phoenix HTML Editor*, <URL: <http://www.bsd.uchicago.edu/>>.
- [17] World Wide Web Consortium, *HyperText Transfer Protocol*, <URL: <http://www.w3.org/hypertext/WWW/Protocols>>.
- [18] World Wide Web Consortium, *HyperText Markup Language*, <URL: <http://www.w3.org/hypertext/WWW/Markup>>.
- [19] World Wide Web Consortium, *Uniform Resource Locators*, <URL: <http://www.w3.org/hypertext/Addressing/URL>>.



Schubert Foo received his BSc in Mechanical Engineering, PhD in Materials Engineering, and MBA from the University of Strathclyde, U.K in 1982, 1985 and 1989 respectively. He joined Babcock's Research Centre in Renfrew, U.K as a Project Engineer in 1985. He subsequently became Systems Consultant specialising in developing real-time software for the process and power industries. He joined School of Applied

Science, Nanyang Technological University as Lecturer in 1990. He was Sub-Dean of the School from 1991 to 1996. His research interests include multimedia technology, Internet technology, CSCW systems, digital libraries and project management.



Ee-Peng Lim received his BS (Honours) in information systems and computer science from the National University of Singapore, in 1989, and Ph.D in computer science from the University of Minnesota, Minneapolis, in 1994. Since 1994, he has been on the faculty of the School of Applied Science at the Nanyang Technological University, where he is currently a

lecturer. His current research interests include database integration, multi-database systems, and digital libraries. Dr. Lim is a member of the *IEEE Computer Society* and the *ACM*. His professional activities have included being on the program committees of the *IEEE Conference on Tools of Artificial Intelligence 1995(USA)* and the *ACM Digital Library Conference 1997(USA)*, and refereeing for journals and conferences.