URL's CONSIDERED HARMFUL: LARGE SCALE INFORMATION MANAGEMENT FOR THE WEB

W Hall

Rapporteur: Michael Elphick

URL's CONSIDERED HARMFUL: LARGE SCALE INFORMATION MANAGEMENT FOR THE WEB

Wendy Hall
Multimedia Research Group
Department of Electronics and Computer Science
University of Southampton
SO17 1BJ
wh@ecs.soton.ac.uk

Abstract
The concept of hypertext has existed since the 1960's and its potential for providing access to information has never been in doubt by its proponents. But it was not until the advent of the World Wide Web that hypertext became a household word and now anyone with a computer, a modem and telephone can have access to a global network of information through the use of hypertext links. However current common usage embeds Web links in documents, link management is generally non-existent and spaghetti hypertext authoring is rife. Many individuals and organisations do not realise until it is too late how impossible it is to maintain and up-date information created in such an anarchic fashion. This comes as no surprise to software engineers!

However, there is nothing inherent in the design features of the Web that says this has to be the case. Links don't have to be embedded, link management tools can and are being implemented and more sophisticated link authoring tools can and are becoming available. This paper will consider the problems that application developers face in dealing with large-scale information management in the Web, or any other distributed hypertext system, and present some solutions based on work being undertaken at Southampton.

Introduction
We all think by making associations between different pieces of information stored in the seemingly infinite database of our brains. Sometimes we don't even remember how we originally made these links but they enable us to retrieve both related and seemingly unrelated pieces of information. Ever since people have been able to read and write, they have wanted to be able to capture such cross-references across the written word. The example that would be most well known to all of us is the amazing amount of scholarly work represented in cross-referencing the ancient scriptures such as the Bible and the Talmud.

In the twentieth century, the prospect of digital computers to help us make and use such cross-references became apparent. One of the early visionaries in this field was Vannevar Bush, who was the scientific advisor to Roosevelt in the second world war. He worked on the Manhattan project and foresaw the explosion of scientific information which makes it impossible even for specialists to follow developments in a single discipline, let alone the multitude of disciplines involved in a large scientific or engineering project.

In his seminal paper "As We May Think", which was published in the Atlantic Monthly in 1945 [1], Bush proposed the Memex (or memory extender) system which he described as a sort of mechanized private file and library. His ideas were based on a mechanical machine using microfilm or something similar, but they are obviously even more applicable

applicable to the world of digital computers. Bush described the Memex in theory and it was never implemented in its original form, but the ideas live on in today's Internet society.

His design, incorporated the idea of associative indexing, or "As we may think", and the building and following of "trails" of information. The Memex had a scanner that would scan in all information of interest to the user. Here we have the concept of a machine that photographs everything the scientist looks at and stores for future use. This may sound like the realms of science fiction, but people are just beginning to use such devices in everyday life today.

Bush even forecast a new breed of information scientists called "trail blazers" whose job it would be to create useful or relevant trails through vast information spaces. Bush likens the idea of creating a trail to that of cross-referencing or linking. To quote from his paper:-

> "The Memex affords an immediate step to associative indexing, the basic idea of which is a provision whereby an item may be caused at will to select another immediately and automatically. This is the essential feature of the Memex. The process of tying two items together is the important thing."

The idea of making and following links between pieces of digital information is now known as *hypertext* and although Bush never really returned to the idea after the publication of "As We May Think", he is known as the "grandfather of hypertext" because of the inspiration that his paper gave to others who followed in his footsteps. These ideas are as relevant today as they were in 1945 - maybe even more so because now we can see that the implementation of such a system is actually possible, if still extremely complex.

Hypermedia systems in the 1990's

The terms hypertext and hypermedia were coined by Ted Nelson in the 1960's, his ideas inspired by the writings of Bush. It was thirty years however before these ideas became widely appreciated in the computing community and beyond. The first ACM conference on hypertext was held in North Carolina in 1987. This marked a watershed in the development of hypermedia systems - a coming of age. By the 1990's there were three main research themes in the development of hypermedia systems. Firstly the standards community was looking for common methods of modelling and languages such as Dexter [2] and HyTime [3] Another community was concentrating on developing tools to provide access to information available on the Internet such as the World Wide Web work at CERN [4] and Hyper-G at the University of Graz in Austria [5]. And a third community, of which the research group at Southampton is very much part, was developing so-called open hypermedia systems and link services, but more of that later.

Let us consider the phenomenon that became one of the major driving forces of the information revolution - the World Wide Web. Everybody reading this paper will have either used the World Wide Web or at the very least have read about it in the newspapers, more probably the former. It was developed by an Englishman, Tim Berners Lee, working at CERN with a Belgian physicist, Robert Cailliau, from about 1989 onwards. They developed it in order to help the scientists at CERN organise the enormous amounts of information they generated and share that information with other groups around the world. It was really part of building and maintaining the organisational memory of the work at

CERN, very much following in the footsteps of Bush and his vision of the Memex machine.

It remained a largely obscure research based system until it was popularised through the development of the Mosaic interface, which was released by the University of Illinois in 1993. This easy to use interface, whilst not realising the full potential of the system developed by Tim Berners Lee, became almost overnight the easy to use hypertext interface to the information on the Internet. The developer s of Mosaic left Illinois to establish Netscape Inc. and the rest as they say is history.

The World Wide Web essentially created a universal hypertext system in three years, when Nelson hadn't succeeded with Xanadu in 30 years. The main reasons were that the Web and its browsers were free to use, it was pioneered by academics on the (essentially free) Internet, used open protocols so anyone could develop tools that used the system, provided a distributed cross-platform environment, was easy to use and the time was right!

The Web has shown us that global hypertext is possible, but it has also shown us that is easier to put rubbish on the net than anything of real and lasting value. In its current form it also encourages the development of unstructured, unmaintainable, unreusable and uncustomisable hypertexts. It is clear that authoring effort and the management of links are major issues in the development of large hypertexts. This has led to the design of systems which separate the link data from the document data thus enabling the information about links to be processed and maintained like any other data rather than being embedded in the document data. Research effort has also been concentrated on the development of link services that enable hypermedia functionality to be integrated into the general computing environment and allow linking from all tools on the desktop. The hypertext management system then becomes much more of a back-end process than a user interface technology. Such systems are usually referred to as open hypermedia systems although much use and abuse has been made of the term "open" in this context There are those who use the term to mean that the application runs on an open system such as UNIX: in this sense the World Wide Web is an open system. However, this is not what is meant in the context of open hypermedia.

There is not room to give a detailed description of what characterises such systems here but their capabilities will be discussed in later sections in this paper. The interested reader should look at Davis et al [6], Wiil and Osterbye [7] and Hall et al [8] for details. In the remainder of this section, we shall discuss the Microcosm open hypermedia system [6, 8] as an example, including the motivations behind its development and its application to digital libraries.

Motivations behind the development of Microcosm

One of the main motivations behind the development of the Microcosm open hypermedia system was the problem of providing hypertext support in large scale digital libraries. This was sparked when in 1987 the University of Southampton acquired the archive of the Earl Mountbatten of Burma from the Broadlands Trust. The archive contains about 250,000 text documents, 50,000 photographs, many of his speeches recorded on 78 rpm records, and a large collection of film and video. The archive spans his entire lifetime (1900-1978) and essentially mirrors British history in the twentieth century. It is also very multimedia in nature, composed as it is of much unstructured information in the form of free text, photographs, audio and video.

The multimedia nature of the archive makes publication and access impossible by traditional methods. For the foreseeable future, there are no developments in technology that can reduce the enormous effort required to catalogue all the items in the archive, this requires specialist knowledge, but developments in information systems, both hardware and software can make it easier to store and organise the catalogues. Also, we can envisage a time when the information in such archives is available in digital form for anyone to make use of.

Creating a digital version of the archive is not something that fits neatly into a database. There is no linear sequence to the material other than that of chronological date, and users will want to move seamlessly from one document to other related documents. Every user is going to approach the material from a different perspective, so there is a need to create different 'views' for different users. from a hypermedia perspective this means the creation of different sets of hypertext links. Hence the need for an open hypermedia system/link service that stores links separately, allows those links to be applied across data of any media type and format across a network of heterogeneous platforms, and supports multiple users, allowing each user, or group of users, to maintain their own private view of the objects in the system.

We started in 1988 with the period when Mountbatten was in India, and we quickly learnt that the biggest issue was that of copyright. However, working with the archive helped us learn what was required of the next generation of hypermedia systems and indeed what was required of a multimedia information system in general. We needed to integrate hypermedia technology with database management and information retrieval systems to even begin to tackle the issues.

Structured access to the information through the use of document and database management systems is needed to answer questions like

"List the photographs taken by ...."

To support similarity matching across indexed documents to answer questions like

"Find photographs similar to ...."

we need information retrieval techniques, which is of course pretty hard when you are dealing with anything other than textual information. And you need hypermedia techniques to enable sophisticated browsing and cross-referencing capabilities to answer questions like

"Tell me about the people in that photograph ...."

The philosophy behind the design of the Microcosm model was aimed at creating a system that not only provided an open hypermedia link service, but that seamlessly integrated that service with database and information retrieval functionality, and in fact any other information processing capability that was required for a particular application area.

The Microcosm model

The three-layered model of the Microcosm system [6, 8] has not really changed since its conception in 1989 although of course the way it is implemented has. The three layers essentially represent the way the user interacts with the system, the hypermedia functionality of the system, which is where the links are created and stored, and the information storage layer of the system. One of the key tenets of the design was that links could be defined on the basis of the content of an object in a document and/or its context, and not just on the basis of the position of a button within a document [9]

For example, suppose in working on the digital version of the Mountbatten archive we want to define a link that tells readers who Ghandi was. The word Ghandi appears all over the documents in the archive, so we don't want to have to manually define a link from the word Ghandi to say his biography or a picture of him, we want that link to automatically be available if the reader needs it. From this sort of idea the concept of the generic link evolved. This is a link that is available at any occurrence of a particular object, such as the text string Ghandi, wherever it appears. The scope of the link can be restricted as the author requires: for example we would probably put the generic link that associates the text string Ghandi with his biography into a link database designed for novice users, rather than one designed for expert researchers. A natural extension of this type of concept is that people will exchange and publish link databases in the same way that they exchange and publish linear texts at the moment.

The other main contribution that Microcosm has made to hypertext research, is the provision of a set of communication protocols that allows its links to be accessed through any viewer or application program that can communicate with the link service. This enables the integration of filters, or processes, that provide a framework for the dynamic generation of links through database, information retrieval or knowledge based techniques. In today's terminology such filters might be known as agents, an idea that we shall discuss in more detail later.

An example of a filter that dynamically generates "links" in the current version of Microcosm is the compute links filter (see [8] for details). The user is able to query the system to "compute links" based on a particular text selection, and the response is in essence a list of (pre-indexed documents) that contain that text selection. Because the filter architecture is highly configurable, this in-built text retrieval process can be replaced by any third party text retrieval process that can talk to Microcosm through its API. This facility, very much equivalent to the use of a search engine in a WWW context, is useful for authors as well as users to help them make links. But of course not all links can be made in this way, for example if we wanted to make a link from the text string Ghandi to a picture of Ghandi, that picture could not have been found by a standard text retrieval process. The author must either be using a more sophisticated information retrieval package or know where such a picture can be found.

Generally speaking the associative hypermedia links, i.e. links that associate content in the same or different documents, that we store in databases, or indeed that might be embedded in the documents in a different hypermedia system, represent some "knowledge" about the content of the documents on the part of an author. The destinations might have been system-computed, such as through the use of a text retrieval process, but the author, or indeed in some cases the system, has decided that the association thus generated is worth storing in a database for future reference.

Link services and digital libraries for the Web

The growth of the Web has been phenomenal and it could be argued to be the killer technology of the information revolution let alone of hypertext. Suddenly everyone -

journalists, industrialists, politicians, school children - want to have their own home page on the net. Most people who use computers at work or at home are now aware of the concept of hypertext as it appears in the Web. Due largely to the influence of the design of the early Web browsers, such as Mosaic and Netscape, common usage embeds Web links in the documents using html, and only allows the presentation of one document at a time. Link management is generally non-existent and spaghetti hypertext authoring is rife. Many individuals and organisations do not realise until it is too late how impossible it is to maintain and up-date information created in such an anarchic fashion. This comes as no surprise to software engineers!

However, there is nothing inherent in the design features of the Web that says this has to be the case. It is actually possible to implement almost any hypertext model in the Web environment. Links don't have to be embedded, Web document and link management tools can and are being implemented, and new browsers could allow multiple documents to be presented at one time, more sophisticated navigation features and integrated editing facilities. Indeed the use of frames in Web browsers has already led to the development of Web sites that are much more user friendly in their interface design. See the book by West & Norris for a discussion of such techniques [10].

There has been much progress in the creation and management of large scale Web sites through the use of relational databases to store the pages and tables that represent the relationships between them, so that for example the intelligence concerning which page should be loaded next can be built into the database. However in the case of loosely structured digital libraries, there is often no concept of the next or previous page. To enable users to make full use of such information sources we need to provide the rich cross-referencing and linking capabilities described above as well as the integration of other information processing tools such as text retrieval systems. This means we need to be able to support the large-scale creation and maintenance of associative links between the content of documents, as well as any navigational or structural links that are inherent in the design of the information system. Just as we separate the structural links from the data using a relational database, we need to store the associative (sometimes called unstructured) links separately from the data as well, as described above. As a result a number of groups in the hypermedia community are working on the development of link services for the Web to support applications such as digital libraries [11, 12, 13, 14, 15].

One of the most well-known of these is the Hyper-G system, now available commercially as HyperWave [11]. Hyper-G is described by its creators as a second generation Web system because it represents an advance over the standard Web and supports tools for structuring, maintaining and serving multimedia data. In addition it guarantees automatic link consistency and supports links among multimedia documents, full-text retrieval, a UNIX-like security system, and client gateways to Gopher and Web browsers such as Netscape, Mosaic and MacWeb.

Like the Web, Hyper-G is based on a client server architecture but unlike most Web environments, it stores its link separately from the documents in an object-oriented database. In this way it can provide the document and link management tools to ensure consistency and integrity of links, as well as allowing different sets of links to be applied to the same set of documents. If the Web in its current form is like HyperCard on the Internet, then Hyper-G could be described as Intermedia on the Internet since although it supports standard Web browsers, its full functionality can only be accessed through dedicated Hyper-G viewers - the main current ones being Harmony for the X-Windows environment and Amadeus for Microsoft Windows. Also as in Intermedia, links are generally point to point. In text documents they can be created, as is common in HTML, by using an HREF reference to a URL. Using the Hyper-G authoring tools, Harmony and Amadeus, they can be created interactively by point-and-click methods with source and destination anchors in multimedia documents such as videos, PostScript and VRML documents. One of the main strengths of Hyper-G are the tools it provides for visualising the structure of the documents and the links in Hyper-G applications through graphical browsers. Like Microcosm, it also contains in-built search facilities, both structured (database query access to documents) and unstructured (full-text retrieval of indexed documents).

Microcosm and the Web: the Distributed Link Service

The original aim of the Microcosm project was to take the fully 'open' system approach and to provide a light-weight link service that is available to the user whatever information they are dealing with, whether it is information under their own or their organisations control, or external information over which they have no control, or indeed legacy data within their own environment. The Microcosm architecture also provides methods for alleviating the link management problem by enabling the development of links that do not simply link a point in one document to a point in another document, but can be dynamically created by whatever algorithm is most suitable for the information environment.

The Distributed Link Service (DLS) is a development of the Microcosm philosophy applied to the distributed environment of the World-Wide Web that has been developed in Southampton [15, 16]. In the same way that a client connects to a remote Web server to access a document, the DLS allows the client to connect to a link server to request a set of links to apply to the data in a document. There are several different link database categories supported by the system, at the most general level are server databases, which apply whenever the system is queried. Link databases may also be provided for a group of documents, or a particular document. In addition, a variety of 'context' link databases are available which the user may select from. By choosing a different context, the user may adjust the available link set to best suit their current information requirements. Like Microcosm, the system supports the use of generic links, which allows links to be applicable beyond the scope in which they were originally created and considerably reduces link authoring effort.

The link server facilities of the DLS were implemented first as CGI scripts invoked by a standard WWW server. The user could make a selection in a Web document, and choose an option such as Follow Link from a menu just as in Microcosm. However a major problem with this interactive client approach is the engineering requirements of producing and maintaining software that applies the available link services to a range of different viewing applications using a variety of WWW browsers on a range of different host operating systems. Hence an alternative, 'interfaceless' approach was investigated: to make the link service transparent to its users by embedding it in the Web's document transport system, compiling links into documents as they are delivered to the user by a specially adapted WWW proxy server.

This approach requires no extra client software for the user, which is an immediate practical benefit, but it does suffer from a number of disadvantages. Firstly, the loss of interaction makes it impossible to create a link by the usual method (in Microcosm) of making a selection and choosing Start Link from the menu. It also changes (perhaps for the worse) the browsing paradigm from 'reader-directed enquiry' to 'click on a predefined choice' [9]. Secondly, this behind-the-scenes link compilation is applicable only to documents which are delivered via the WWW and which are coded in well-understood document formats that can themselves support some form of hypertext link. These requirements abandon some of the advantages of the open system previously described, since there are relatively few document formats which can have links embedded. However, because of its engineering advantages, it is this version of the DLS that is currently being commercialised as Webcosm. A link control panel provides the user with configuration possibilities to overcome some of the disadvantages described above and interfaces that provide the full interactive capabilities of Microcosm are being investigated.

The DLS is being used in a digital libraries project concerned with electronic journals funded by the Electronic Libraries Programme (eLIB) of the UK Higher Education Councils' Joint Information Systems Committee (JISC). The aim of this project is to produce an open journal framework (OJF) which integrates collections of electronic journals with networked information resources [17].

The seamless integration of journals that are available electronically over the network with other journals and information resources is not possible with traditional paper publishing but has long been sought by the academic community. The emergence of the World-Wide Web and its freely available user, or 'browser', interfaces has dramatically simplified access to such resources and has raised awareness and expectations of electronic distribution of information and data, and has particular significance for journal development. In its present form, however, the Web is limited in its capabilities for page presentation, hypertext linking, user authentication and charging, but it defines a flexible framework into which more advanced technologies can be slotted. It is in these areas that the Open Journal Framework project intends to provide mechanisms that realize the wider potential for networked journals that has been created by the Web.

The philosophy of the project, at the most basic level, is to provide immediate access to electronic versions of existing quality journals, the information content of which includes scientific formulae, tables, diagrams and high-resolution colour photographs. Beyond that the aim is to provide powerful hypermedia linking techniques to allow naive users direct access to secondary information resources, instead of requiring them to use these resources independently. The links are created dynamically at the users' request and do not need to be explicitly embedded in the journal papers when they are authored, thus realizing the concept of the 'open' journal.

This is achieved by making use of current hypermedia technologies, for example Adobe Acrobat for presentation, and Microcosm, in the guise of the DLS, for information linking, as well as the World Wide Web. In addition, through the development of subject-expert agents implemented within the DLS framework, the user will be offered a greater range of resources than he or she alone would normally be aware of.

Making links intelligently: filters become agents

During the initial design for Microcosm, it was expected that the combination of the generic link facility and other dynamic link generation processes would create a significant problem of information overload for Microcosm users. Hence the link processors were called *filters* because it was expected that their main task would be to reduce the number of links offered to users according to some appropriate algorithm. In fact, most of the intervening years have been spent designing and building filters to create links, such as the 'Compute Link' filter, so that the term became rather a misnomer for current usage of the system. However, things are now coming round full circle because the system has been developed to such a point that it is possible in practice rather than just in principle to use it to build large-scale applications with very large numbers of links, and processes are needed that truly act as filters to decrease the information overload for users.

In today's terminology, filters would probably be called *agents* because Microcosm can essentially be considered as a group of processes working together to perform tasks on behalf of the user, in other words a community of software agents.

The current Microcosm communication model allows isolated agent processes to be constructed and both Prolog and neural networks have been used to construct experimental agents within the current design. However the limitations in the design of the current Windows version of Microcosm mean that no direct communication between processes is possible. Experiments with a direct model for communication between processes have shown that this model supports co-operative agent processes. The extension of this model to construct a heterogeneous fully distributed version of Microcosm as part of the Next Generation (TNG) project, opens the way for extensive research into the implementation and use of co-operative agents within the new architecture fully integrated with other information systems available via the Internet [18]. A large subset of these agents will be doing fairly straightforward information processing tasks such as library management. These agents need not be intelligent and we tend to refer to them as "distributed information management" or DIM agents. However, the aim is to enable the user to customise the environment to suit their own purposes and to "train" the supervisor agents to this end. It is here that the intelligence of the system will lie.

The results of the TNG project are being utilised in an EU funded project, MEMOIR [19], which is aiming to build a corporate information system, where documents, links and user trails are stored in an object-oriented database, accessible to users via standard Web browsers and a distributed hypermedia system brokered by agent services. As users access the information in the system, and indeed external to it, they can keep a record of the trails of information they have followed. The MEMOIR system helps to match people with trails, trails with trails and people with people to try and overcome the problems of finding information in large corporations.

A Vision for the Future

We have all seem the science fiction films and TV series where the crew of the spaceship find the information they need through communication with an "intelligent" agent. The most famous example is probably HAL in the film *2001: A Space Odyssey*. It is possible to see now the sort of building blocks that are required to create and information processing environment such as HAL would need access to. Just as we think by making associative links between disparate pieces of information, so will the intelligent information processing agents of the future. They will draw on information from vast banks of distributed databases and document management systems. They will access this information by both structured and unstructured means using combinations of database, information retrieval and hypermedia techniques. In effect, all these technologies are about making links.

If we ask the question "Tell me about Ghandi" of an intelligent agent such as HAL, then first the agent must have some understanding of the context in which I am asking the question in, in order to answer it with some degree of confidence. The next step maybe a search of a database of significant figures in history. If none is available then a general search of relevant documents may be made or brokered using search engine techniques. Or alternatively, someone may have asked this question before and the agent will "remember" the link and produce the information that successfully answered the question previously, or that is flagged as being a suitable answer to the question in an appropriate link database. The possibilities are endless, and of course all of them require the agent to intelligently prioritise and process the available information sources. Still very much the realms of science fiction!

References

1. Bush, V., As We May Think, Atlantic Monthly, pp 101-108, July 1945

2. Halasz, F. & Schwartz, M. "The Dexter Hypertext Reference Model". Communications of the ACM, Vol. 37, No. 2, pp. 30 - 39, (Feb 1994).

3. DeRose, S.J. & Durand, D.G., Making Hypermedia Work: A User's Guide to HyTime. Kluwer Academic Press, 1994.

4. Berners-Lee, T.J., Cailliau, R. & Groff, J-F., The World Wide Web. Computer Networks and ISDN Systems, Vol. 24(4-5), pp 454 - 459, 1992.

5. Andrews, K., Kappe, F. & Maurer, H. "Hyper-G: Towards the Next Generation of Network Information Technology". Journal of Universal Computer Science, April 1995.

6. Davis, H.C., Hall, W,. Heath, I., Hill, G.J. and Wilkins, R.J. "Towards an Integrated Information Enivornment with Open Hypermedia Systems". In the Proceedings of the ACM Conferenc e on Hyerptext, ECHT'92, pp. 181 - 190. ACM Press, 1992.

7. Wiil, U.K. & Osterbye, K (eds.) "The Proceedings of the ECHT'94 Workshop on Open Hypermedia Systems, Edinburgh, 1994". Technical Report R-94-2038, Aalborg University, October 1994.

8. Hall, W., Davis, H.C. & Hutchings, G.A. "Rethinking Hypermedia: the Microcosm Approach". Kluwer Academic Press, Boston, 1996.

9. Hall, W. "Ending the Tyranny of the Button". IEEE Multimedia, Vol.1, No.1. pp. 60 - 68, 1994

10. West, S. & Norris, M. "Media Engineering: a Guide to Developing Information Products". Wiley, 1997

11. Maurer, H. "Hyper-G now HyperWave: the Next Generation Web Solution". Addison-Wesley, 1996.

12. Rizk, A. & Sutcliffe, D. "Distributed Link Service in the Aquarelle Project". In the Proceedings of the ACM Conference on Hypertext, HT 97, pp 208 - 210, ACM Press (1997)

13. Gronbak, K., Bouvin, N. & Sloth, L. "Designing Dexter-based Hypermedia Services for the World Wide Web". In the Proceedings of the ACM Conference on Hypertext, HT 97, pp 146 - 156, ACM Press (1997)

14. Anderson, K. "Integrating Open Hypermedia Systems with the World Wide Web". In the Proceedings of the ACM Conference on Hypertext, HT 97, pp 157 - 166, ACM Press (1997)

15. Carr, L.A., De Roure, D.C., Hall, W. & Hill, G.J. "The Distributed Link Service: A Tool for Publishers, Authors and Readers". Proceedings of the Fourth International World Wide Web Conference: The Web Revolution, Boston, MA, December 1995.

16. Carr, L.A., Davis, H.C., De Roure, D.C., Hall, W. and Hill, G.J. "Open Information Services". Computer Networks and ISDN Systems, 28, 1027-1036, Elsevier 1996

17. Hitchcock, S., Carr, L., Harris, S., Hey, J. & Hall, W. "Citation linking: improving access to online journals". Proceedings of the Second ACM Conference on Digital Libraries, Philadelphia, pp. 115 - 122 (1997).

18. Goose, S., Dale, J., Hill, G.J., DeRoure, D. & Hall, W. "An Open Framework for Integrating Widely Distributed Hypermedia Resources". In the Proceedings of the IEEE International Conference on Multimedia and Computing Systems (ICMCS'96), Hiroshima, June 1996, pp 364-371, IEEE Press, 1996.

19. Hill, G.J., Hutchings, G.A., James, R., Loades, S., Hale, J. & Hatzopulous, M. "Exploiting Serendipity Amongst Users to Provide Support for Hypertext Navigation". In the Proceedings of the ACm Conference on Hypertext (HT'97), Southampton. ACM Press, pp 212 - 214 (1997)

DISCUSSION

Rapporteur: Michael Elphick

In reply to Mr Kerr, who had asked whether links were global or document-specific, Professor Hall said that this could and often did vary; however, dictionary links were usually global. Mr Kerr then said he had in mind a recent occasion, on which he had been trying to introduce some Korean children to the English language. Although he had not attempted to use the Web, it seemed to him that the ideas presented here could provide just what was needed; however, one would need to be very selective on the provision of links, to avoid the danger of information overload. The speaker pointed out that this was precisely the reason for the use of filters; initially, they had been used mainly for adding links, but were now being applied to control the number of links.

Professor Farber asked about the speaker's views on the book mentioned by a previous speaker ("Snow Crash" by Neal Stephenson). Professor Hall said that although she found some of the ideas and concepts presented were fascinating (in particular the Virtual Librarian), she had to say that the book presented a very macho - indeed sexist - image, and summed up what much of the male-dominated world of computing seems (unfortunately) to be about.

Mr Butler commented that there had been no mention of metadata; the speaker agreed that this was probably crucial in making such systems work on a global scale.

Mr Hodgson asked if this system could be used to produce versions of materials suitable for primary school children; if so, would one supply both documents and links, or just the links? Professor Hall said that this was an obvious example of its use, in delivering different interfaces to existing documents for different audiences. At an elementary level one would publish the combined document and link data, while for an expert audience, just the links alone might be sufficient.

Professor Randell commented that he was intrigued by the ideas presented, but was concerned over the apparently simple dichotomy between text and links. His predilection for advocating recursiveness in structure was well-known to these audiences, but surely there were more subtle distinctions to be made in hypermedia? He also raised the question of extending the notion of similarity to other types of object than text.

The speaker said that this was easier to do with text, but the concept could be extended to other data types; with an object-oriented database, one could use this to perform the appropriate matching operations for such queries. Professor Randell said that he had used "similar" in one sense; one could imagine several different meanings for the term. In reply, the speaker pointed out that one could create agents that would choose the appropriate search engines or queries.