Linking Everything to Everything: Journal Publishing Myth or Reality?

S. Hitchcock, F. Quek*, L. Carr, W. Hall, A. Witbrock* and I. Tarr*

Open Journal Project †
Multimedia Research Group
Department of Electronics and Computer Science
University of Southampton SO17 1BJ, UK
*Electronic Press Ltd, London
Contact for correspondence: Steve Hitchcock sh94r@ecs.soton.ac.uk

† The Open Journal project is funded in the UK by JISC's Electronic Libraries (eLib) Programme award ELP2/35.

This version of the paper was presented at the ICCC/IFIP conference on Electronic Publishing '97: New Models and Opportunities held in Canterbury, UK, during April 1997, and was made available to delegates at that meeting as a conference preprint.
This version posted to the Web June 1997. Revised August 1997.

A more sustantially revised version of the paper, Towards Universal Linking for Electronic Journals, was published in *Serials Review*, Vol. 24, No. 1 (Spring 1998) 21-33, and posted to the Web in November 1998.

Abstract

Reference lists are an important facet of the modern academic journal. This form of 'hyperlinking' becomes enormously more powerful when translated to the World Wide Web, both in terms of the speed of link following and in the number of linked documents that can be made accessible. Electronic Press Ltd (EP), one of the first commercial publishers to commit to electronic publishing on the Web, plans to extend the practice of citation linking, aiming to link a document not just to a cited source but to all other documents that contain relevant information. Relevance in this case is defined as all referring, or referred to, documents. The paper discusses EP's approach to link creation on this scale, which is based on an internalised system. One way of extending this approach is to support link creation as well as link following on a distributed network such as the Web. The Open Journal Project has built some first demonstrations, which are outlined in the paper. The convergence between these two approaches suggests some important new motivations for online journal publishing. Some of these features will eventually transform journal usage.

Keywords: *electronic journals, Web publishing, hypertext linking, link services*

Overview of the paper

1 Introduction

Commercially-produced online journals are entering a new phase. In a little over a year those publishers that were among the first to make substantial journal programmes available online have begun to add features which are not directly available in corresponding print editions. The agenda has moved on from how you put journals online to how those journals can be enhanced.

Prior to the wider availability of online journals, talk of enhancements tended to focus on the idea of 'multimedia' content. Although some fields such as medicine and biology may be in a position to build and use such materials, the widespread realisation of audio-visual support for essentially text-based journals, as well as an adequate network infrastructure to distribute such materials, is still some way off.

Instead, attention has focussed on the hypertext link. In fact, links are a vital component of integrating multimedia content created in widely differing formats, but as far as online journals are concerned the first application of links on a large scale extends a convention that is fundamental to the modern academic journal: the use of citations.

Through reference lists within primary journal articles and the proliferation of secondary information sources - indexing and abstracting services, reviews, etc. - the journal literature is intrinsically 'hyperlinked', but the online medium is a more natural environment for this feature. The essence of the online environment is speed of access to the linked materials (Hitchcock 1996). In principle electronic links can be followed in an instant and will prove to be orders of magnitude more productive for the user than hyperlinks in print. In the electronic domain links can give the user access to a cited resource, possibly in its full form or to further information about that resource. Online journals that demonstrate citation linking are appearing in the areas of biology, physics and astronomy (Hitchcock *et al.* 1997).

This paper will examine two approaches being used to create links - citation links, but also authored links to non-traditional sources such as databases and other reference sources. The objective in both cases is to create and maintain large numbers of links, potentially linking everything to everything. One approach is that currently used by Electronic Press Ltd, a commercial publisher and producer of online journals, for the BioMedNet online club for those working in biology and medicine. Also examined is the Open Journal research project, which involves the novel approach of storing the link information separately from the authored documents, thereby potentially improving the flexibility of linking and creating a new, reusable and possibly valuable resource. By comparing the two approaches it may be possible to identify which features will be important in providing the most practical, efficient and cost-effective method for creating and maintaining links, a service that will become a vital component of online publishing.

2 Why links?

In the context of academic journals, the idea of providing 'clickable' links from citations is simple and obvious and on a small scale the implementation can be technically straightforward. In practice our ability to apply links is fraught with social, commercial and legal difficulties. For journals this goes beyond the fact that as long as the electronic archive is small it will be difficult to link to full-text articles: secondary abstracting services, preferably those with citations from the abstracted papers or other indexing terms, are a good substitute. So before looking at the implementations it may be worth considering briefly some of the reasons for using links.

The Web has become a massively popular Internet service in just 2-3 years since the introduction of graphical browsers such as Mosaic and Netscape. During these early years much of the content of the Web has been text-based, and the single dominating feature of text on the Web is the ‘blue button’ link. So it may be possible to conclude, not only intuitively as did Bush (1945), that links between properly connected pieces of information are important. The popularity of the Web shows there is a real demand for links when they are simple to create and to use.

2.1 Content integration will drive Web publishing

Another aspect of this period of Web development is its 'openness' and this has significant implications for online publishing. The Web can be considered an open system because it is based on open published standards and is not tied to any hardware or software platform. Since the Web is as accessible to individuals as it is to large organisations, much of the content has been freely contributed. Consequently some have questioned the quality of content on the Web and others assert that users will not pay for content on the Web. This is to misread the situation. Another interpretation is that, currently, distinctive Web content is rare. It will be hard to place a commercial value on content, especially text, that was designed for another medium, *unless* that content

- can be adapted or enhanced with features that the new medium supports
- has an established identity that can be brought to its new form.

In other words, the former demands links and multimedia enhancements. Looked at generically, the critical features that will enhance Web content are access to a wide range of information and speed of access, and this suggests an environment that is rich in what might be called 'data integration' (European Commission 1996). This will involve a range of services such as personalised alert, customised information, enhanced search engines and advanced database querying, software agents and the inclusion of standardised metadata, as well as links.

In the latter case it is the publishing framework that finds or creates value in something that may not inherently contain that value outside the framework. Through services such as BioMedNet (Quek and Tarr 1996), this framework is beginning to develop on the Web. The BioMedNet Library contains 100 full-text publications and a further 200 will be added in 1997. The club has 60,000 active users (growing at a rate of 2000 per week). Club members now access 200,000 pages a week, and have downloaded a total of a quarter of a million full-text articles.

According to Tim O'Reilly, books and online publisher, early online products have added searchability and multimedia: 'But the Web shows a third key advantage of an online product: the creation of information interfaces'. (O'Reilly 1996) From the user perspective, the need will be to be able to find information on demand quickly, accurately and reliably, whatever the data type (audio, video, graphical, etc.), often starting from an imperfectly formed query. The challenge for online publishers seeking to add value to Web content is to develop interfaces that will locate and deliver the requested information based on the optimum data type.

In this scenario links, and the distinction with which links are applied, will become one of the principal determinants in establishing the competitive position of the information seller.

3 Electronic Press: a publisher's perspective on linking

From a commercial viewpoint, Vitek Tracz, Chairman of the Current Science Group, sees the virtue of links as providing value-added information, with publishers competing to provide better quality links to entice customers to their online products and services. In the early 1990s, before the explosion of the Internet, Tracz saw the potential of links as a commercial entity in future electronic publishing. He set the following brief for Electronic Press Ltd, the electronic publishing arm of the Group:

- To create and manage bidirectional links, e.g. between any citation (reference) and:
  - the full-text of the cited resource (if available)
  - a bibliographic database, e.g. a Medline entry for the citation
  - any other document containing the citation as a reference.
- To support the creation of arbitrary links between documents (i.e. link anything to anything).
- To support automatic and manual creation of links.

A major problem with the Web is that links are unidirectional. There is no simple way of knowing whether a link has been made to your document – evaluating an Alta Vista search to determine this can be time-consuming - and comments cannot be attached to documents you read on the Web. Today, there are forward links but no back links, which creates the problem of 'dead end' documents. If a link is bidirectional, a link always exists in the reverse direction, i.e. bidirectional linking allows the system to deduce the inverse relationship, that if A includes B, for example, that B is part of A. This effectively adds information for free. One compromise is that links are one-way in the data model, but that a reverse link is created when any link is made, so long as this can be done without infringing protection. An alternative is for reverse links to be gathered by a background process operating on a basically monodirectionally linked web (Berners-Lee 1990). In EP's case, bidirectional linking is actually two unidirectional links.

To achieve EP's goal of linking everything to everything, as well as providing bidirectional linking, can only be achieved if a given document links to all other documents that contain relevant information. For the BioMedNet library a level of relevance has, for practical reasons, been defined as all referring, or referred to, documents. To achieve bidirectional linking, there must be control over both source and destination of links to provide back links. In a commercial environment, however, there are restrictions preventing this. Data suppliers do not allow modifications to their databases yet users expect value to be added by data providers.

This section presents an overview of EP's linking system ('Clinky') and the external link database ('BundledLinks') that facilitates commercial competition with database suppliers while maintaining scalability, and discusses how EP achieves its goals with this implementation.

3.1 EP's approach: Medline linking

EP's first major linking initiative is to provide citation linking to Medline (MEDlars onLINE), the National Library of Medicine's (NLM) bibliographic database covering the fields of medicine, nursing, dentistry, veterinary medicine, the health care system and the preclinical sciences. In the biomedical field this database is an important resource for researchers, clinicians, scientists and medical students. The database contains bibliographic citations and author abstracts from over 3800 biomedical journals published in the USA and 70 other countries for the last four years. It also contains over 8.6 million records dating back to 1966. Coverage is worldwide, but 87% of the records in current Medline are from English-language sources and 72% have English abstracts. Medline is usually updated weekly, and about 33 000 new citations are added each month. It includes basic citation information and abstracts, as well as MeSH terms, publication types, GenBank accession numbers and other indexing data.

EP's approach to Medline linking on a commercial scale, involving linking of at least 500k documents from a 16 GB of data, is based on a number of requirements:

- rapid access to the database
- formattable records
- bidirectional linking
- flexible linking from other databases
- link descriptors containing e.g. abstract and file sizes

Taking these decisions in the early 1990s meant that the only feasible solution then was to undertake the development of an in-house linking system. EP obtained a Medline database license for all records, including backfiles since 1966, as well as the weekly updates. Known as Evaluated Medline, this service is offered to all publishers of electronic publications held in BioMedNet. There are now at least eleven Web-based Medline implementations: NCBI PubMed, Medscape, Healthgate, Community of Science, Kfinder, Infotrieve, WebSPIRS, OVID, Internet Grateful Med, Paperchase and BioMedNet (Electronic Press Ltd 1996). EP's aim is for Evaluated Medline to be the most powerful.

3.2 Mechanics of EP's Medline linking

One approach is to generate 'computed links'. Computed links remove the need to manually create large sets of links but, as the name suggests, only data that can be computed can benefit from such a strategy. There are various definitions of computed links (Zhouxun *et al.* 1992, DeRose 1989, Kibby and Mayes 1989). Computed links are links generated (in EP's case *a priori*) by a repeatable process, and EP's linking mechanism uses this approach to create bidirectional Medline links from references within journal articles held in the BioMedNet Library.

This approach discards the node/link hypertext model in favour of a set-based model of direct node intersections (Parunak 1991) or 'complex relations' in which the hypertext takes on many properties of a sophisticated database (Marshall *et al.* 1991).

There is little human involvement in creating the links, which are generated programmatically by querying the bibliographic database with data from SGML tagged references. This query is similar to the common notion of a Standard or Structured Query Language (SQL) used in relational databases. The query is constructed from the last name of the first-named author, the longest word in the title, the volume, issue and start page numbers, the date of publication and a 'mangled' journal identifier. Generation of standard matching keys can also be used. An example of this is the use of standard Serial Item and Contribution Identifier (SICI) codes for matching articles published in journals or, in the case of Medline-aware documents, the use of Medline accession numbers.

The 'mangling' process involves looking up the cited journal title in a list of journals known to be in Medline (compiled from a published list of abbreviated names for journals called Index Medicus), and converting it into the three-letter Medline journal code. If a code exists for the reference in question, then there should be a Medline record to this reference. In practice there will be references that do not belong to the Medline set (such as theses, references to non-medical journal articles), and on average only about 60% of references in a typical medical paper are contained in the Medline data.

Of the references that can be found in Medline, it is still impossible to achieve complete linking. Both ends of the link are human typed and error prone: Medline records are input at the NLM, and the references by, or on behalf of, the author(s). In addition errors can be introduced in the tagging process either manually or programmatically.

The linking program is therefore forced to make a 'guesstimate' based on the available information, and this can achieve a success rate higher than 85%. For EP, given the scale of the database processing task, this is an economical and acceptable threshold. Currently EP only performs a single attempt at a match. If this fails, there is no attempt to resubmit the reference. A log of which records do not match and the reasons are made available so that, in theory, someone could go through the logs and hand-enter the mismatches, but this is not economical (Figures 1 and 2).

Error in Medline Match on citation ref02 in file CA9214.SGM.
Searched for:
Surname : knudson.
Title : mortality.
Source : 8IA.
Year : 1994.
Volume : 129.
Pages : 448.
First & second hit exceed or both equal relevancy threshold!
First Relevancy :6
Second Relevancy :5
I chose 94206201 and disregarded 94175757!

Figure 1. Sample error log of an unmatched reference

Analysis

Missing Titles : 124.
Missing Journal Code : 330.
Links Added : 5061 (links added to the data).
Missed links : 771 (Correctly formed but not in Medline).
Pre-existing links : 0.

Stopped at : Sat Feb 01 19:21:31 1997

Figure 2. Sample analysis of a Medline linking session of an article

Those references for which the linking program can find matches can generate more than one match. The query generated for each reference will return a set of results ranked by relevance based on the seven information fields identified in the reference. The most relevant, i.e. the 'best fit', record is chosen so long as no less than five of the seven fields match and there is no more than one record with the highest ranking.

Another type of linking EP performs is to embed Medline accession numbers (unique identification numbers) into the full-text journal articles. This type of linking implies that by using the same linking mechanism, Medline links can potentially be derived from any Medline database by constructing a query based on the accession number. By storing only the accession number and not a URL the link can be generated by a program, allowing the destination to be changed 'under' the link. One problem with embedding accession numbers is that as the database grows (e.g. adding Embase, Analytical Abstracts) the number of accession numbers needed grows.

3.3 Storing links as a bundle

EP's database of links, BundledLinks, is so called because of the way the links are organised into bundles pointing to related documents or pages (this is synonymous with a 'web'). This concept is based on documents or pages becoming part of a collection of inter-related entities, each perhaps created by a separate commercial enterprise or governing body which also imposes its own trademark or identity.

For example, once a biological research paper has been published on paper or in electronic form, it can always be referred to by its details or an address. If the same paper is then converted to an HTML document and is made available via the Web, it could take on a second address. The same paper might be abstracted for Medline as well as other similar database services such as Embase and Current Contents. Each database publisher may choose to augment the original reference with keywords, evaluations or abstractions, e.g. MeSH terms. The original publisher may choose to hold many different representations of the same document. Many documents published by the Current Science Group appear in BioMedNet as HTML and PDF documents.

For these reasons, within the BioMedNet Library some documents may be represented in four or more different ways, each of which is published by a different publisher which enforces its own rules on augmenting its own data. Each of the links is stored as a record akin to a database record, and are collected and stored in a bundle. A bundle is effectively a collection of links of representations of the same document. Each record can be thought of as an instance of an information unit, that unit being the sum of all its parts. It is not enough to think of the full-text article as the parent of

these records as there may be more information held in annotations and abstracts of the original document than in the original document (Figures 3 and 4).

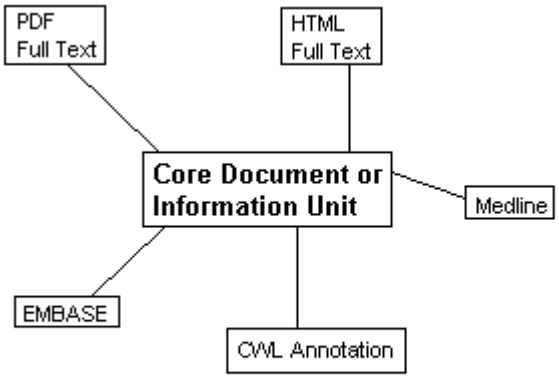

Figure 3. Document Inter-relationships

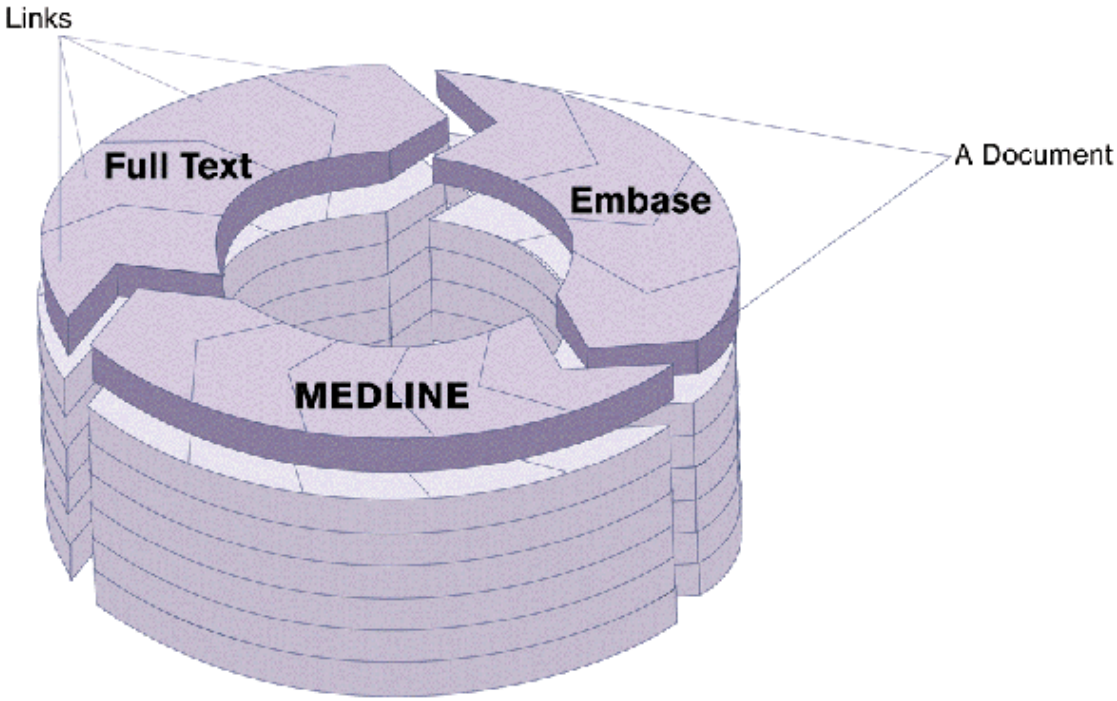

Figure 4. Conceptual view of bundles, documents and links

In many cases the bundle can be generated using data that appears in the records. Many commercial databases store references to the corresponding record in other databases. Medline accession numbers are often used for this purpose. Utilising data directly from the records shifts the linking burden to the database owner. Once two records from different databases are identified as being instances of the same information unit, it becomes possible to use either database identifier as a means of referring to the bundle.

3.4 Retrieving from a bundle

Once the document instances have been collected into a bundle, there are some immediate benefits for users. The BioMedNet library provides free-text searching through all records. This is a cross-database search. Many of the databases impose rules on how the records can be amended (often no changes are permitted).

By collecting the instances of the record into a bundle then for each search a user performs the information units found can be identified and presented to the user with summary information about that unit instead of the usual "found records" display. In effect, if the user searches for a MeSH heading and finds a Medline record, the full-text version and user annotations can be indicated, where they exist, with very little overhead.

By using the bundle as a general storage area for holding meta-details of the information unit, these details can be made available as annotations to whatever document instance the user is viewing. If a user is viewing an abstract from an abstracting service they can also be provided with links to citations.

When a record is displayed in a Web browser, it becomes possible to provide details of which other representations are also available (Figure 5). A link to the PDF version of the full-text can be provided for printing.

Figure 5. Links to other representations of one document source

The bundle is itself a record of the BundledLinks. In this way, not only are the links made available to all instances, but the problem of keeping all records synchronised is reduced to keeping a single entity synchronised. By representing documents as a source / identifier pair it is possible to change the data supplier without changing the destination of the link. So EP could, for example, for financial reasons choose to change document supplier for Medline without having to alter all links in the documents or in the BundledLinks. The hypertext link is generated on-the-fly from the bundle much as an HTML document is generated on-the-fly from the SGML document. For documents that exist only on the Web, the source can be represented as WWW and the identifier becomes the URL of the document.

With this structure it is possible to extract all the links regardless of the record they actually appear in and present them to the user regardless of the record they are viewing.

3.5 Lessons learnt

By identifying the multiple instances in the information unit, it becomes possible not only to present the user with details about the existence of those records, but with links out from the other instances. The links in the bundles are thus 'smarter' than conventional unidirectional links. As a consequence, the bundles begin to acquire a commercial value. The Web has no way of representing how the nodes are related, other than what is contained in the text used to identify the link. Link information on the Web is limited to the URL of the current node and the text of the link. With EP's BundledLinks, each link potentially can contain and provide a lot more information.

Being able to change the way the link is resolved (i.e. to change database provider) has commercial advantage not only for BioMedNet but users could configure their preferred data provider to avoid having to pay a new provider for data they already get from their own supplier. There are many 'free' Medline suppliers on the Web. Users should be able to choose which they prefer.

The commercial value of a link database was recognised before the Web. The Science Citation Index is an example of such a database. A bundle representing a research paper could be used either electronically or on paper for the same purposes.

EP's internalised Medline linking may not be the best approach for everyone, especially since there are now Medline services on the Web, but it has enabled EP to achieve its self-imposed requirements highlighted above. Part of the cost of innovation is that the solution sought may not be there initially. Consequently, EP has a history of building its own technologies such as Evaluated Medline. The important point, however, is that the conceptual model must be workable, flexible and extensible, and the linking mechanism is designed so that it does not depend on a closed system to support external linking. The plan is to be able to implement the same linking mechanism with other large-scale databases such as Embase. In addition, EP is now seeking to extend its linking framework by participating in the Open Journal project which is also applying link publishing software as described below..

4 Enhancing link publishing on the Web

Bush's (1945) crucial insight was to recognise that, despite the long-established and elaborate ways of indexing information that had developed, 'the human mind ... operates by association. With one item in its grasp, it snaps instantly to the next'. The Web, though, has not yet proved flexible enough to support Bush's vision, in effect linking everything to everything.

As the EP example shows, creating many thousands of links to support users in a specific knowledge domain is going to be a significant part of any works published online. Courseware developers have also discovered that creating and maintaining high-quality links is a major economic decision and not an after-thought at the end of the publishing process.

In contrast to EP's internalised linking, one way of widening the use of links is to extend the way in which the Web itself supports the creation and implementation of links. Web links are authored rather than published links, the link type used is limited, and the Web is a closed hypertext environment. The key to exploiting the Web from a publishing perspective is to create a more open environment which supports linking as a publishing and not just an authoring activity, and which reduces the link authoring effort. More powerful ways of developing links are being developed. The use of link services, linkbases and generic links in the Open Journal model are one way in which this can be achieved. Putting this in context requires some understanding of the philosophy underlying the Web and of open hypertext, or hypermedia, systems.

4.1 Open hypertext: publishing implications

In computing, the term 'open systems' describes a framework in which applications or otherwise incompatible software might be allowed to interoperate. This became an important concept in the mid-1980s with the proliferation of Unix systems. Today the same issue applies to the integration of multimedia components, usually developed using packages optimised for a particular format - text, sound, video, etc. - but which were not necessarily designed to work together.

In this sense the Web is a classic open system. With its standard protocol for information transfer and universal addresses, anything that can be displayed can be interconnected. Simply, if the relationship between two works changes the information 'could smoothly reshape to represent the new state of knowledge' (Berners-Lee *et al.* 1994).

The Web though is not an open *hypertext* system. A generally accepted requirement of open hypertext systems is that they do not differentiate between authors and readers. (Malcolm 1991) Each should be offered the same set of functions, that is, a reader should have the same facility for altering a version of a text, say, as the original author. By encoding links within HTML markup, the Web does not conform to this view, because it reduces linking to an author-only task. According to Berners-Lee *et al.* (1994): 'The Web does not yet meet its design goal as being a pool of knowledge that is as easy to update as to read.'

For ‘readers’ read ‘publishers’, because publishers have a greater need to manage content, and on the Web this will not always be content that the publisher ‘owns’, exactly as seen above. O’Reilly (1996) recognised the fundamental shift in publishing that the Web motivates: ‘In the old model, the information product is a container. In the new model, it is a core. One bounds a body of content, the other centers it’. According to Fillmore (1993), a founder of Open Book Systems: 'The successful online publisher will most likely license access to other people's content to supplement or enhance his own, whether that content is online books, databases, bulletin boards, graphic image repositories, or online advice columns. What's ‘for sale’ might be the interactive links, the thought structure the publisher puts around the distributed content area.'

The implications are enormous, but from a technical viewpoint commercial open hypertext systems which give publishers and users the option to make links as well as follow links *from* third-party materials on the Web, for example, are now available.

4.2 Open hypermedia systems and distributed links

The question an open hypertext system raises is how different applications can interact with the hypertext system to share and display links equally. Pearl (1989) introduced a general approach called a ‘link service’. The link service is effectively a database look-up service where the data items are interpreted as links between other data items. The link data are stored in a link database, or linkbase. A link service can potentially integrate a wide range of third-party applications, although this requires that the applications are link-aware, that is, able to communicate with the link service. There are drawbacks - maintaining valid links when documents change, for example, is more difficult in a system mediated by a link service - but Pearl envisaged link services giving links the same utility as computer cut-and-paste operations.

A number of open hypermedia systems reported in the early 1990s adopted a link service approach, including Microcosm (Fountain *et al.* 1990), Hyper-G (Kappe *et al.* 1992), and Multicard (Rizk and Sauter 1992). Two of these systems, Microcosm and Hyper-G, are now commercialised. There are other examples of open hypermedia systems in research (Wiil and Leggett 1997), but current interest centres on extending these linking models to augment the Web, such as the Distributed Link Service (Carr *et al.* 1995) and HyperWave (Maurer 1996).

As the Web is increasingly used to display documents created in common applications such as word processors and spreadsheets, which may or may not support HTML and thus hypertext capabilities, or which may or may not have authored links, the potential for a link service on the Web becomes apparent. 'Without a link service, Web users can follow links from HTML documents or pictures into dead-end media such as spreadsheets, CAD documents or text; with a link service they can also follow links out of these media again' (Carr *et al.* 1995).

It may be obvious to state that the effectiveness of a link service is predicated on the effectiveness of the links that it serves. The links served by the Distributed Link Service (DLS) and its commercial version Webcosm have semantics that are derived from the Microcosm system (Fountain *et al.* 1990). For example, each link is a pattern that can be matched against many potential documents to instantiate an actual link between two actual documents. This allows a link to be parametrised against the identity of a document, against the position of the anchor within a document and against the data contents that the anchor selects.

Thus, each link has the following format, and states the existence of a link from a source to a destination.

```
<link type=local>
     <src><doc>http://diana.ecs.soton.ac.uk/~lac/cv.html
          <offset>
```

```
      <sel>Microcosm
<dest><doc>http://bedrock.ecs.soton.ac.uk
      <offset>
      <sel>The Microcosm Home Page
<owner>Les@holly
<time-stamp>Fri Mar 31 13:32:34 GMT 1995
<title>Hypermedia Research at the University of Southampton
```

Both the source and destination are described as a triple: the document URL, the offset within the document, and the selected object within the document. The system pinpoints the link anchors either by measuring from the beginning of a document (using the offset), or by matching a selection, or both.

Links are of the following types:

- *Specific* if its source anchor is constructed from a complete triple (i.e. a specific occurrence of a selected object in the named document).
- *Local* if its anchor ignores the offset component of the triple (i.e. any occurrence of the selected object in the named document).
- *Generic* if only the selection is used (i.e. any occurrence of the selected object anywhere in any document).

System offsets are frequently used in Microcosm, but due to a lack of integration with the various viewing programs, the DLS usually ignores the offset. Hence only local and generic links can be created and manipulated by the various user interfaces, although it is possible to process specific links programmatically.

Note how the DLS provides flexibility in specifying the source anchor: this means that a single link to a destination may appear in many places at once.

Links supported by the Web are specific, or 'button', links. Each link has to be individually authored.

5 Creating links for the Open Journal project

The project is developing three Open Journals in the areas of biology, cognitive science and computer science. Since the cultures and practices of each field tend to be reflected in the respective literatures, these in turn determine linking strategies. The characteristics of each Open Journal are already markedly different. The important feature here is how the system used for creating links adapts to the different requirements and copes with the formats in which the original materials are presented, the principal formats in this case being those popular for online journals, HTML and PDF. (Figure 6) Link inclusion, from a linkbase, in PDF documents is supported in the project by applications developed in the Electronic Publishing Research Group (EPRG) at Nottingham University, and is an extension of that group's CAJUN (CD-ROM Acrobat Journals Using Networks) project.

The spectrum of link creation options supported by the DLS includes highly pertinent, hand-crafted links such as might take a user from a biology journal page to a graphical molecular database. Links can also be created *en masse* by a batch computational process, for example, in citation linking or linking complex terms to a definition in a specialised dictionary.

On the Web each link must be individually identified and specified, making it a cumbersome process especially if, as today, the navigation facilities of the native Web environment are not particularly advanced and do not aid the link creator sufficiently in browsing relevant resources. (O'Leary 1997) For this reason, the majority of the 12 000 links that are currently being demonstrated in the Biology Open Journal were created by computational methods and demonstrate the power of the *once-only authored* generic link. A cheap way of providing a database of links for a journal archive is to create a link FROM any occurrence of a specific word or phrase within a literary corpus TO any paper with that keyword specified. This approach can also be used for terms in an online dictionary, so that the occurrence of a key dictionary term anywhere in any document is automatically linked directly to its definition. (Figure 6) In fact, this is a variation on the generic link, where the document context is constrained to be 'inside' the boundaries of the Open Journal. So, any mention of the word 'embryo' may link to an entry in the online *Dictionary of Cell Biology* if it is found in the Biology Open Journal, but may not if the instance is inside, say, the Open Journal of Computing.

Figure 6 (98 kb). Sequence of linked biology pages: a, unlinked PDF page: b, same page with links added by the link service via the project's PDF plug-ins; c, a single link can point at multiple destinations; d, following a link on 'embryo' to the *Dictionary of Cell Biology*, a remote resource (note how this resource also has links added to it by the link service, so although it is remote it can still be an integral element of the linked Open Journal)

To create a database of these links requires a source of metadata for the articles of interest: extracting the keyword fields is a small programming effort which leverages links out of another's authorial or editorial effort. In practice, many of the project's resources are in PDF and have not been provided with metadata records, so the programming effort required to extract the keywords from an encoded document display format has been a lot higher - but still less than creating the links manually.

Although the project philosophy is to give users access to links distributed across a network, it has found, as has EP with its Medline linking, that access to localised, formattable data is necessary for creating large linkbases computationally. The application of citation linking in the Cognitive Science Open Journal is an example. Selected abstracts data made available to the project by the Institute for Scientific Information (ISI) required extensive reindexing, but the resulting links from journals such as *Psycoloquy* are proving reliable and relatively complete (Hitchcock *et al.* 1997). The real power of this approach, however, is that once created, this referencing linkbase could be applied to other cognitive science journals from which reference data can be parsed wherever they are on the Web and wherever the abstracts database is held. All that is required is that user is able to access the resources, e.g. as a subscriber, as well as the link service.

5.1 Further development of the link publishing tools

Plans to enhance the link publishing tools include giving the user more control over link selection and link following by exploiting the linkbase structure, and supplementing methods for creating links with network-based software agents.

Having created a set of links, the author can store them either in a single linkbase which must be chosen explicitly by an end-user, or amalgamated into a database of linkbases which can be chosen as part of a larger context. The former option allows for highly specific links, tailored for a small user population, but inevitably results in a large number of these collections. The latter option results in a more generally applicable database, but one that is of lesser relevance to any particular user.

As the project has developed, more of the linkbases have tended to be of the second type. Consequently, the DLS is being revised to give the user explicit control over the kinds of link that they would like to see. Links could be differentiated by colour (or simply pruned according to particular thresholds) according to whether they are hand-authored specific links, machine-generated general links, recently created links or links belonging to (for example) a course tutor. This adds to the user's control over the view of the document's connectivity, which currently exists only at the macro-level by including or excluding whole linkbases.

While the most successful strategies for link creation have depended on data processed locally, the project aims to develop tools to make more use of remote Web resources. One possible approach is to augment an author's resource discovery strategies by piggybacking an existing Web search service. For example, an ActiveX-based interface could allow the author to request that the results of keyword searches automatically be turned into linkbases. This emulates the cheap link creation described above, but also allows the user to view the results of the search from a browser and prune the search results into the most pertinent set of keyword links.

6 Conclusion

The examples given above demonstrate that linking everything to everything is feasible technically. Ultimately it is not possible to represent links from a document to *all* relevant information because the definition of relevance lies in the mind of the reader. It is possible to make some well-founded decisions about what constitutes relevance, however. In electronic journal publishing the application of citation linking presumes from the outset that ultimately every citation will be linked to the cited source in some form and builds on conventions long established in print. EP's Medline linking strategy goes further, defining relevance as all representations of the same document, all documents that refer to them and any documents they refer to.

The application of links must be led, or constrained, by user expectations, although first impressions suggest these should not be underestimated, especially with regard to citation linking. Once established this feature is so powerful that it will be almost mandatory and will transform journal usage.

In other respects the ability to link everything to everything is not always desirable. The practice of citations is well established. In contrast the overlaying of keyword data as links on text is not, so first reactions to such links have been less positive. Simply, not enough is known about the effect on the literature for this type of linking to be applied indiscriminately on a large scale. A better understanding of when and where to use this approach allied to greater

precision in the use of linkbases will make this approach more attractive if, in this case, the demand for universality of links is tempered.

EP's approach enables it to represent what users expect from a commercial system while maintaining source databases as supplied by their owners. The data can change 'underneath' or 'on-top-of' the links without affecting the validity of the link. Effectively this is an open hypermedia approach, similar in principle to that being used to build Open Journals in biology and cognitive science.

What is now required in link publishing are tools such as the DLS to provide more explicit editorial control over the quality of links. By applying distributed open hypertext services such the DLS, we can begin to see that such an approach offers the flexibility and cost-effectiveness for large-scale link creation and maintenance that is not possible with the Web alone.

Most importantly perhaps, as EP shows, links stored in bundles or linkbases have additional commercial value because the bundle becomes a piece of information itself, quite distinct from the underlying text. Exploiting this value is an area of investigation for the future.

Acknowledgement

PDF linking applications were developed by Professor David Brailsford, Steve Probets and David Evans in the EPRG at Nottingham University, which, with the MMRG at Southampton University, are the two research centres for the Open Journal project.

References

Berners-Lee, T (1990) Topology: Should the links be monodirectional or bidirectional? *W3C 1990 archive document* http://www.w3.org/pub/WWW/DesignIssues/Topology.html#14

Berners-Lee, T, Cailliau, R, Luotonen, A, Nielsen, H F, and Secret, A (1994) The World Wide Web. *Communications of the ACM*, Vol. 37, No. 8, August, 76-82

Bush, V. (1945) As we may think. *Atlantic Monthly*, July, 101-108 http://www.isg.sfu.ca/~duchier/misc/vbush/

Carr, L, De Roure, D, Hill, G, and Hall, W (1995) The Distributed Link Service: a tool for publishers, authors and readers. *World Wide Web J., Proceedings of the Fourth World Wide Web conference*, Boston, MA, USA, pp 647-656 http://www.w3.org/pub/Conferences/WWW4/Papers/178/

DeRose, S J (1989) Expanding the notion of links. *Hypertext '89 Proceedings* (New York: ACM), pp. 249-59

Electronic Press Ltd (1996) Medline Services Survey, internal document

European Commission DG XIII/E (1996) *Strategic Developments for the European Publishing Industry Towards the Year 2000* (Brussels, Belgium: European Commission). An information note on this study is available at http://www2.echo.lu/elpub2/en/infonote.html

Fillmore, L (1993) Internet publishing: how we must think. *Meckler's Internet World Conference*, New York, Dec. http://www.press.umich.edu/jep/works/fillmore.think.html

Fountain, A M, Hall, W, Heath, I, and Davis, H C (1990) Microcosm: an open model for hypermedia with dynamic linking. *Proceedings of the European Conference on Hypertext ECHT '90*, France (Cambridge, UK: Cambridge University Press), pp. 298-311

Hitchcock, S, Carr, L, Harris, S, Hey, J M N, and Hall, W (1997) Citation linking: improving access to online journals. *Proceedings of the 2nd ACM international conference on Digital Libraries*, edited by Robert B. Allen and Edie Rasmussen (New York, USA: Association for Computing Machinery), pp. 115-122 http://journals.ecs.soton.ac.uk/acmdl97.htm

Hitchcock, S (1996) Web publishing: Speed changes everything. *IEEE Computer*, Vol. 29, No. 8, August, 91-93 http://www.computer.org/pubs/computer/kiosk/r80091.htm

Kappe, F, Maurer, H, and Sherbakov, N (1992) Hyper-G - a universal hypermedia system. *HCM technical report, Graz University of Technology, Austria*

Kibby, M R and Mayes, J T (1989) Towards intelligent hypertext. *Hypertext: from theory to practice*, edited by R. McAleese

(Norwood, NJ: Ablex)

Malcolm, K C, Poltrock, S E, and Schuler, D (1991) Industrial strength hypermedia: requirements for a large engineering enterprise. *Hypertext '91 Proceedings* (New York: ACM), pp. 13-24

Marshall, C, Halasz, F, Rogers, R, and Janssen, W (1991) Aquanet: a hypertext tool to hold your ideas in place. *Hypertext '91 Proceedings* (New York: ACM), pp. 261-277

Maurer, H (Ed.) (1996) *HyperWave - The Next Generation Web Solution* (Harlow, UK: Addison-Wesley)

O'Leary, D E (1997) The Internet, intranets, and the AI renaissance. *IEEE Computer*, Vol. 30, No. 1, January, 71-78

O'Reilly, T (1996) Publishing models for Internet commerce. *Communications of the ACM*, Vol. 39, No. 6, June, 79-86 http://www.ora.com/oracom/inet/pubmod.html

Pearl, A (1989) Sun's Link Service: a protocol for open linking. *Hypertext '89 Proceedings* (New York: ACM), pp. 137-146

Parunak, H. (1991) Don't link me in: set based hypermedia for taxonomic reasoning. *Hypertext '91 Proceedings* (New York: ACM), pp. 233-243

Quek, F, and Tarr, I (1996) An example of the use of the WWW as a tool and environment for research collaboration. In *Information Systems and Technology in the International Office of the Future*, first edition, edited by B.C. Glasson, D.R. Vogel, P.W. Bots and J. Nunamaker (London: Chapman & Hall)

Rizk, A, and Sauter, L (1992) Multicard: an open hypermedia system. *Proceedings of the European conference on Hypertext, ECHT '92* (New York: ACM), pp. 4-10

Wiil, U K, and Leggett, J J (1997) Workspaces: the HyperDisco approach to Internet distribution. *Hypertext '97 Proceedings* (New York: ACM), pp. 13-23

Zhuoxun Li, Davis, H, and Hall, W (1992) Hypermedia links and information retrieval. *BCS Information Retrieval Colloquium*, London, UK

This page `http://journals.ecs.soton.ac.uk/IFIP-ICCC97.html`

[Top of article] [Other Open Journal papers] [Open Journal Project]