Chapter 2

Hypertext and Associative Writing

This aim of this chapter is to introduce the focus of this work by first providing a brief historical perspective of hypertext systems research, from the pioneering work of early visionaries to more recent work in open hypertext. This perspective then helps to inform a detailed description of the concept of *Associative Writing* within a hypertext docuverse. The specific focus of the work reported over the course of this thesis is to examine the issues and problems surrounding Associative Writing in the context of the World-Wide Web's global information repository (and to develop a solution which helps support such an activity), an investigation which has presented many challenges. In this chapter, a review of the evolution of the Web introduces the first two of these challenges — the controversial "lost in hyperspace" problem, and the recent copyright infringement claims arising from "deep linking" in the Web — the implications of which are discussed in the context of this work.

2.1 Hypertext Foundations

In 1945, Bush envisioned a system which would utilise the computing power that was beginning to appear at that time to manage and store the increasing volume of scientific literature being produced (Bush, 1945). When it came to retrieving information from this storage, he argued that traditional alphabetical or numerical indexes contravened the associative workings of the human mind:

> The human mind does not work that way. It operates by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance with some intricate web of trails carried by the cells of the brain. It has other characteristics, of course;

> trails that are not frequently followed are prone to fade, items are not fully permanent, memory is transitory. Yet the speed of action, the intricacy of trails, the detail of mental pictures, is awe-inspiring beyond all else in nature (Bush, 1945).

The mechanisation of selection by association forms an essential part of Bush's hypothetical Memex device — a personal library station holding a vast corpus of reference materials on microfiche. The Memex may have supported an early form of Associative Writing: users would build associative trails through the library by tying together related frames of microfilm. Thereafter, whenever one of the frames was in view, the other could be instantly recalled. When numerous frames have been tied together to form a trail, they can be reviewed in turn, as if the information had been gathered together to form a new book. Associative trails through the Memex do not fade in the same way as thoughts, and can be passed to other users to add to their own Memex devices. Indeed, Bush imagined a whole range of new products and services arising in a world in which Memex-like technology was realised:

> Wholly new forms of encyclopedias will appear, ready-made with a mesh of associative trails running through them, ready to be dropped into the Memex and there amplified ... There is a new profession of trail blazers, those who find delight in the task of establishing useful trails through the enormous mass of the common record.

The hypothetical Memex machine is one of the foundations of hypertext research (Kahn et al., 1991). Bush's idea of a machine that would help people think was extended in the 1960s by Engelbart and Nelson. Engelbart developed the oNLine System (NLS) (Engelbart, 1963), a complete collaborative work environment which aimed to enhance and augment human abilities through computer technology. Files in NLS were structured as a hierarchy of statements, and links could be created between any two files or statements (also enabling a form of integrated writing) — users followed links through the text by clicking on a link tag and then clicking the window in which they wanted the target text to be displayed.

Memex and NLS are now referred to as *hypertext* systems - a term coined by Nelson (Nelson, 1980). Nelson penned hypertext as "a body of written or pictorial material interconnected in such a complex way that it could not conveniently be presented or represented on paper", or more succinctly, "non-sequential writing" .

Nelson's own visionary hypertext system, Xanadu, represented digitally the literary forms of connection between texts (Nelson, 1987). Nelson envisioned a docuverse where all the world's literary texts would be instantly available and connected (Nelson, 1980). *Xanological* connection constitutes two kinds of literary connection - *content links* and *transclusion links*. Content links are arbitrary connections between texts. Transclusion links show the origins and context of quotations, excerpts and anthologised materials, connecting instances *which are the same* but which appear in different contexts, literally *integrated* writing. Every quotation would have a transclusion link back to its source, allowing original authors to be compensated by a very small amount (a micro-payment) each time the quotation was read. A complete Xanadu system has yet to materialise, although some implemented parts have been recently released into open source[1].

To best support the creation of trials of thought on the fly, hypertext systems must allow users to create any connection between documents (and possibly modify the document content), rendering any readers *simultaneous* authors (Jackson, 1997).

Every implementation of the hypertext vision requires a means for the user to select the specific information to be viewed at any one time, therefore hypertext systems employ the *associative* link in various guises (for example, Xanadu's content and transclusion links) to allow the user to jump from one specific location to another. DeRose provides the following description of associative links, noting that they "are the usual stock in trade of hypertext systems":

> Since associative links attach arbitrary pieces of document, they cannot be replaced by retrieval algorithms, or even by unilateral creation on the part of an author. Rather, every user must be able to create them on the fly and so to organise them in whatever ways seem appropriate...Because these links serve may purposes, they are usually labelled according to type (DeRose, 1989).

(Lowe and Hall, 1999) extend this definition: "the instantiation of a semantic relationship between information elements...The semantic relationship can be something as simple as 'definition of' or something far more complex." The remainder of this section describes some prominent hypertext systems that have emerged since the 1960s, highlighting the implementation of associative links.

2.1.1 Classical Hypertext

Several important hypertext systems were developed in the 1970s and 1980s. KMS was the first system to support the management of large hypertexts in organisations (Akscyn et al., 1988). Based on a simple interaction paradigm, information was displayed a node at a time (although two nodes could be viewed side by side for comparison). A node typically comprised a screen-full of information — quick navigation between short nodes was deemed preferable to scrolling through larger nodes. KMS distinguished hierarchical or structural relationships (tree links) from associative links (called annotation links).

[1] `http://www.udanax.com/`

NoteCards was based on a metaphor of 3×5 index cards, designed to provide a "general purpose idea processing environment" for individuals or small workgroups (Halasz et al., 1987). Like KMS, each node (or *notecard*) was intended to provide a screenful of information to the user, although many notecards could be displayed at once. Each associative link in NoteCards was a typed, directional connection between two cards, anchored at a specific location in the source card. The destination of a link was always an entire card. Links could be labelled arbitrarily; labels were usually employed to describe the properties of the relationship between notecards. Users could organise ideas by producing an interlinked web of notecards, which could then be abstracted as an editable graphical overview (or map) of the network. A special *GuidedTour* link was later added to NoteCards, which allowed users to create guided tours through the main points of the hypertext (perhaps for benefit of novice readers).

Guide was a hypertext system aimed at naive users (Brown, 1987), which disguised the nature of the underlying data structures by presenting hypertexts as a single scrollable document in a single window. Guide supported several different kinds of document link, including replacement, note, glossary, reference, and usage links. When a user selected a replacement link, the link anchor (or an author specified text region) was replaced in situ, with the linked material (which could contain further replacement links); an author therefore typically presented a document in summary form with replacement links to allow the reader to 'unfold' the parts of interest (possibly 'folding' the document back up afterwards). Selecting a note link caused the linked material to be displayed in a separate window alongside the document, but only whilst the mouse button was held. Selecting a glossary link caused the main window to be split between the document and linked glossary definition. A reference link caused a jump to a different point in the current (or possibly a different) document, whereas a usage link displayed cross-referenced material in situ like a reference link (transclusion).

In a classic survey of the field, (Conklin, 1987) maintains that links are the essential feature of a hypertext system, and that other common features (such as text processing facilities and window views) are merely an extension of this basic concept. Conklin also acknowledges the extension of hypertext to other media (for which the term *hypermedia* is introduced), in which the nodes that are interconnected may be text, graphics, sound, video, animation, or any other type of data. Intermedia and Neptune are examples of hypermedia systems.

Intermedia, developed for educational purposes, consisted of a suite of applications including viewers and editors for text, graphics, 3D models, and video (Meyrowitz, 1986; Yankelovich et al., 1988). *Webs* of associative, bi-directional links (links that could be followed in both directions, cf. NoteCards directional links) could be created across these media, which were then stored separately from the media sources. This allowed each user to have their own web of associations (cf. Memex's trails), or explore a web created by another user (for example a student exploring a web that a teacher has created across

Hypertext System	Domain
KMS	Organisation (collaborative work, project management)
NoteCards	Idea processing
Guide	Reading/writing for naive users
Intermedia	Education
Neptune/HAM	Computer-Aided Design

Table 2.1: Application domains of classical hypertext systems.

a set of source materials). Webs were combined with the media sources at runtime to construct the appropriate user view.

Neptune was a serious attempt to apply hypertext to the world of Computer-Aided Design (CAD), concentrating in particular on versioning and configuration management (Delisle and Schwartz, 1986). Designed as a layered architecture based on the Hypertext Abstract Machine (HAM), Neptune maintained a complete version history of the evolving hypermedia network. Associative links could be made between specific versions of nodes within the network, or made to always refer to the latest version.

2.1.2 Open Hypertext

Although impressive in their own right, classical hypertext systems suffered from some problems. They were 'closed' systems, tightly integrating a fixed set of encapsulated applications with hypertext linking mechanisms. The systems worked well in the settings for which they were designed (Table 2.1) but did not often become widely used in other settings.

Meyrowitz identified that the lack of support for third-party editing and viewing tools was a major obstacle to the widespread acceptance of these early hypertext systems (Meyrowitz, 1989). Most organisations relied on standard tools (such as word processors, spreadsheets, and CAD systems) for creating documents, but classical hypertext systems could only provide hypertext functionality to their own integrated editors and file formats. Meyrowitz suggested that rather than try and provide their own content generating tools, hypertext systems should embrace third-party tools and augment their functionality with an underlying hypertext service. This adoption of this approach gave rise to *open hypertext* systems.

A defining feature of open hypertext systems is the separation of service elements. Given the need to integrate third party applications with different (often proprietary or read-only) document formats, it is generally impossible to embed link information inside documents. Open hypermedia systems thus treat links as informational entities in their own right, which can be stored, processed and used independently of the documents to which they refer. Storage of links is usually handled by a hypertext server or *link service*. An open hypertext system can integrate a wide variety of clients using a range of different integration techniques (Davis et al., 1994).

2.1.2.1 Sun's Link Service

Sun's Link Service (Pearl, 1991) is often considered the first example of a truly open hypertext system. The system comprised a link database service which managed both shared and private collections of links. Applications could access links by registering and communicating with the service using a well-defined protocol. The success of the system however, relied on the modification of third-party applications to be made "link service aware". Sun's Link Service did demonstrate, however, that a common link service and standard communication protocol are essential components of an open hypertext system.

2.1.2.2 Microcosm

Amongst other issues concerning open hypertext systems of the time, Microcosm (Fountain et al., 1990; Davis et al., 1992) addressed the problem of the effort required to author associative links. Microcosm's innovative approach was the notion of a *generic link*, an associative link which could be applied from *any* point in *any* document in the system which matched the link anchor. This meant that, once created, a generic link could be made immediately available in any document containing the source anchor, including those added to the collection *after* the creation of the link. Generic links greatly improved authoring efficiency (a link only needed to be created once) and maintenance (only one link had to be changed if the destination changed). A trivial example of the usability of generic links is in a dictionary or glossary lookup - every appearance of a term in any document could be linked to the corresponding definition. Four other types of link could also be created in Microcosm:

1. *Specific links* were analogous to associative links, connecting specific points in specific documents.
2. *Local links* were similar to generic links but once created would only be applied in the scope of a specific document.
3. *Dynamic links* connected a source anchor to a number of dynamically computed destinations (for example, by initiating a text retrieval search across the document database).
4. By pre-indexing the nodes in the database, it was possible to cluster nodes according to their relevance to each other. *Relevance links* connected a document to all the other documents in its cluster.

Messages to perform hypermedia actions (such as *make* or *follow* a link) were sent from Microcosm-aware applications to the Document Control System (DCS) and acted upon by a chain of filters. Non-aware applications could be integrated with Microcosm through the use of a shared clipboard. This "filter model" was a particular feature of the Microcosm architecture. Each filter manipulated messages by blocking, deleting, or adding to it. A special type of filter is a linkbase, which upon finding the source of a link in the message attaches the available destination(s) to it. Messages emerged from filter chain to a *link dispatcher* process, which examined the modified messages and offered actions to the user (such as available links to follow). The filter model allowed the behaviour of the system to be extended by adding new filters.

Any Microcosm user could create their own links, which would be stored in a private linkbase or shared within a user group. Users could also add and remove filters from the chain to control the context in which a particular set of documents was to be investigated (for example, selecting linkbases which offer beginner, intermediate, or advanced dictionary lookup, depending on the user's experience of a topic area).

2.1.2.3 Multicard

Multicard was an attempt to unify apparently diverging directions in open hypermedia system development. Systems at that time focused either on the Runtime and Storage layers or on the Within-component layer of the Dexter model (see Section 2.1.3). Multicard's solution was to provide a clear specification and implementation of the Anchoring interface of the Dexter model in the form of the M2000 model (Rizk and Sauter, 1992). Links in Multicard are viewed as event or message channels between nodes. A variety of messages can be sent "through" the link, including an *activate* request which causes the link to be "followed" and the target node displayed. The Multicard designers felt that this approach would allow scripting languages to easily configure the behaviour of the system.

2.1.2.4 DeVise Hypermedia

DeVise Hypermedia (DHM) (Grønbæk and Trigg, 1994) was an open hypermedia system based firmly on the Dexter Reference Model (Halasz and Schwartz, 1994), in an attempt to both prove the validity of the model and demonstrate its flaws. DHM's major departure from the Dexter model is the inclusion of *dangling links* (links with zero or one endpoint). The DHM designers argue that this is a natural extension of the Dexter model, allowing incremental construction of links within the system and lazy updating and garbage collection when nodes are deleted from the system. DHM also explored the notion of *link directionality*, concluding that an associative link may have at least three notions of direction:

Semantic direction A semantic link between one endpoint and another (for example A *Refutes* B).

Creation direction The order of creation of the endpoints of a link defines the *creation direction* (for example, if A is older than B then A becomes the source and B the destination).

Traversal Direction The user, in creating a link, may explicitly define the *traversal direction*, and the link can only be traversed in this direction.

The Dexter model allows for the specification of direction in a link, but does not explain which of these different notions of direction is actually assumed. The solution in DHM was to specify all links as bi-directional (traversable in both directions) by default and allow the user to modify this where necessary.

2.1.2.5 Chimera

Chimera combined hypertext technology with software development environments, allowing software developers to create links between software objects beyond the restrictions of a type hierarchy (Anderson et al., 1994). Software development environments typically contain a diverse range of development tools and management software, supporting multiple views of the software objects. Chimera therefore enables links to be anchored with respect to interactive views of objects (rather than the objects themselves), allowing n-ary links to be established across heterogeneous object managers.

2.1.2.6 HyperDisco

The HyperDisco project focused on design, development, deployment and assessment of open hypertext systems (Wiil and Leggett, 1996). HyperDisco provides two types of components: workspaces (structure-aware databases that manage core hypermedia abstractions and collaboration services) and tool integrators (integrate services provided by workspaces with participating applications). HyperDisco itself, and the integrated applications can run at different machines on a local area network. Each workspace therefore serves as a "gateway" to a set of multimedia documents residing in the underlying file system, which can have anchors and n-ary, bi-directional, links attached to them. Links can also span workspace boundaries.

2.1.3 Hypertext Models

Hypertext researchers realised towards the end of the 1980s that a major failing of hypertext systems of the time was their inability to inter-operate at a system level or exchange data easily. As a result, a number of hypermedia models were proposed, which attempted to provide a common vocabulary for the comparison of hypertext systems, and thus work towards interoperability standards. Perhaps the best known model is the Dexter Reference Model (Halasz and Schwartz, 1994), which describes hypertext systems as three interacting layers, encapsulating such hypertextual features as associative *multi-headed* links (links with many endpoints), and composite nodes:

1. *Runtime Layer*: facilities for constructing and browsing a hypertext.
2. *Storage Layer*: represents actual hypertext structures (i.e. links and nodes).
3. *Within-Component Layer* : represents the content of nodes.

(Østerbye and Wiil, 1996) presented the Flag Taxonomy of Open Hypermedia Systems, a framework within which to compare and classify open hypermedia systems. The taxonomy described each system in terms of the Storage Manager, Data Model Manager, Session Manager, Viewer, and the interfaces between them. The Open Hypermedia Protocol (OHP) (Davis et al., 1996) focused on providing a standard interface for communication between components of existing open hypermedia systems (so that new services could be used by existing clients, and new clients could take advantage of existing services). OHP software would ‘translate’ server communications to OHP and then to the client format, and vice versa. However, problems with the OHP definition (Anderson et al., 1997), led to its evolution to a component-based model (Reich et al., 2000). More recently, the Fundamental Open Hypermedia Model (FOHM) has attempted to provide a common model which embodies each of OHP’s components (Millard et al., 2000).

2.2 Associative Writing

Hypertext systems, from classical to open and model-based design, represented a new approach to writing, which Nelson described as “non-linear writing” (Nelson, 1987). The opposition between the nonlinearity of hypertext and the linearity of other texts is probably the most discussed theme in hypertext theory (Fagerjord, 2001). Moulthrop suggests that *nondeterminate* writing may be a better term (a “non-sequential” experience being “inconceivable in anything but mystical terms”), since the text Nelson envisions allows writers and readers to substitute multiple alternative sequences for the fixed page order of books (Moulthrop, 1992) — after all a hypertext, although network-like in structure, is still viewed in a linear fashion by the reader through following a series of links (Marshall and Irish, 1989). More recently, (Aarseth, 1997) used the term “multicursal” to describe the several possible courses through a hypertext. However, as Moulthrop succinctly points out, even a nondeterminate (or multicursal) text must retain some kind of intelligible succession, or *local coherence*, if it is to communicate successfully (Moulthrop, 1992).

Bernstein describes several "patterns" in an attempt to develop a vocabulary of structures observed in numerous hypertexts (Bernstein, 1998b). Some patterns, such as *Tree* and *Sequence*, are found in almost any hypertext, and are well-described by the hypertext literature (Brown, 1989; Parunak, 1989). Other patterns, such as *Joyce's Cycle*, *Counterpoint*, and *MirrorWorld* seem to be the domain of more specialised hypertext fiction works such as *afternoon, a story* (Joyce, 1990), *Victory Garden* (Moulthrop, 1991), *Six Sex Scenes* (Eisen, 1996), and *A Dream with Demons* (Falco, 1997); Bernstein provides few examples of patterns in non-fiction works.

In contrast, this work focuses on *integrated writing*; the creation of new hypertexts which have not only local coherence but also *global coherence*, integrating the writer's new contributions with existing ideas, structures, concepts, data, examples, descriptions, experiences, claims, theories, suggestions, reports (*etc.*) that have already been published within the *docuverse* of a hypertext system. As such, this process may be more suited (although by no means restricted, as the dance hypertext work presented in Chapter 9 will demonstrate) to non-fiction writing, which forms the focus of this work. This thesis refers to this process using the term *Associative Writing*, since associative links that are the "glue" that writers use to bind together new contributions with existing global context[2]. In short, Associative Writing is *the process by which a writer creates an integrated hypertext.* The hypertext literature indicates some of the potential benefits of exploring Associative Writing:

Literary and rhetorical theorists have always argued that all documents are interconnected, and that the real power of documents is their ability to refer to, refute, and elaborate on each other (Hill and Mehlenbacher, 1996). Writers are not autonomous, but rather "borrow and sew together [text] to create new discourse" (Porter, 1986). In Associative Writing, there is no need for writers to reiterate material that has already been written and published in the docuverse: new contributions can be linked directly to existing context rather than using hundreds or thousands of words to establish that context, as would be expected in a linear medium (Drexler, 1991). Links can demonstrate the reliability of the conceptual foundation being built on (perhaps showing existing ideas in a new light), and show the innovation and significance of new ideas (Buckingham-Shum et al., 2000). If these links are bi-directional, newer contributions become immediately reachable from older texts.

In hypertext, users not only benefit from the information they read, but also from the richness of associations supported by the network of nodes and links (Theng et al., 1996). Readers can follow links from writers' new contributions through to explanations and

[2] This terminology has (independently) been used elsewhere to describe similar processes; for example, Wideroos presented the "Associative Writing Toolkit" to the 2001 hypertext conference (Wideroos, 2001).

elaborations in existing materials, finding their own path through the docuverse: for readers, freedom of access within an hypertext structure provides a rich environment for understanding the information they find (Thüring et al., 1995).

It is claimed that writers often come to understand information better through the process of structuring that information as an hypertext network, since hypertext represents knowledge in a form close to the cognitive organisational structures of the human mind (Bieber et al., 1997). VanLehn reported his experience with expressing and developing theories in cognitive psychology using NoteCards (VanLehn, 1985): this medium enabled him to experiment with organisations of facts and theories in ways that revealed (and helped correct) serious flaws.

2.2.1 Relation to Scholarly Hypertext

Scholarly works are one (fundamental) example of Associative Writing, using a convention that is fundamental to the modern academic journal — citations — to ground new ideas in existing literature. In fact, citations have played a prominent role in scholarly debate long before the advent of hypertext and electronic publishing; they are “the way researchers have been interconnecting their writings all along” (Harnad and Carr, 2000). All academic writing is thus implicitly hypertextual; Landow describes traditional features such as citations and footnotes as “proto-hypertextual” (Landow and Delany, 1991).

(Dalgaard, 2001) argues that scholarly literature on the Web is composed of increasingly linked archives, describing such archives as an evolving network of texts commenting on, citing, classifying, abstracting, listing and revising other texts. As part of their work on the OpCit project (a project which attempts to provide interlinking and metatextual services for scholarly archives), (Brody et al., 2002) report on evidence of hypertext in scholarly archives in an attempt to quantitatively substantiate these observations.

This work focuses on Associative Writing in general, rather than specifically on the integration of scholarly works.

2.2.2 Relation to Information Retrieval

Associative Writing is the process of integrating new contributions with existing work in a hypertext docuverse. It could be argued that, with the emergence of the Web as the dominant hypertext structure, “hyperlinks” as such are unnecessary since readers can find any document they need using a search engine (Lewis et al., 1999). However, this runs against the importance of the relationships or associations in Associative Writing that are not born of similarity but through the external knowledge of the writer.

The difference between the Information Retrieval and Associative Writing approaches is that retrieval typically answers the user request "find me documents containing something like this query" by associating the keyword or phrase with a document containing something similar (either through pre-indexing or on-the-fly analysis). By contrast, Associative Writing is concerned with user navigation across links or associations which do not necessarily require similarity between the source and destination, but represent some meaningful higher level association that is identified through the mind of the writer (Lewis et al., 1999). (Brown, 2002) explores the spectrum of possibilities between these two opposite "retrieval extremes" of complete freedom (IR) and high levels of constraint (Associative Writing). Approaches on this spectrum include *context-aware retrieval*, in which documents are delivered to the user only if they relate to the user's current context — for example, the user might be a tourist and the documents delivered might describe suitable attractions nearby (Brown, 1998) — and *dynamic hypertext linking*, in which the source and/or destination of a link may be calculated by a function, as demonstrated by Microcosm. This work focuses on the most constrained end of the IR spectrum, *Web links*, perceived by the writer, as a means for the reader to gain a deeper understanding of the writers contributions and their significance in the context of existing work.

2.2.3 Relation to Computed Links

Information Retrieval techniques have also been applied to hypertext, in the use of automatic techniques to discover associative links between documents: (Agosti and Allan, 1997) provide a comprehensive overview of previous work. Automatic hypertext construction techniques are often considered when the effort required to author associative links between a collection of documents becomes significant (Westland, 1991; Mendes and Hall, 1999, 2000; Mendes et al., 2001). As large collections of (unlinked) documents are published in hypertext form, the size of the collection may be simply too large to allow the maintainers to manually integrate them using associative links. Furthermore, if the document collection is larger than can be managed by a single person, there can be problems with link consistency (Ellis et al., 1994; Green, 1998; Furner et al., 1999).

If the documents in the collection are well described structurally (for example using SGML, or XML), it is comparatively easy to automatically turn each individual text into hypertext (Wilkinson and Smeaton, 1999). Automatically discovering associative links between documents however is more difficult, particularly when there are no explicit clues to the existence of a link (for example, a citation — see Section 2.2.1). The similarity between all pairs of documents must be computed, and links inserted between those that are most similar (Green, 1999b). An early attempt to automatically discover associative links between documents reached a pessimistic conclusion:

> When we first began working in hypertext several years ago, we expected that

> it would soon be possible to extract these implicit links automatically with natural language processing or clever indexing techniques...but we have been disappointed so far and we are starting to conclude that implicit document links are best identified by the hypertext reader (Glushko, 1989).

Calculating document similarity based on common words inevitably suffers from problems relating to word meanings: *polysemy* and *synonymy.* Polysemy is when a single word has several meanings; synonymy is when different words have the same meanings. This may lead automatic linkers to infer that there is a relationship between two unrelated documents representing different concepts expressed in similar words when there is none (polysemy), or that there is no relationship between two related documents that represent similar concepts expressed in different words (synonymy). Automatically computed links may therefore vary in quality considerably. To overcome this, some approaches have generated links automatically and then allowed them to be manually vetted (Bernstein, 1990; Chignell et al., 1991). (Green, 1999b) describes the process of identifying links within and between texts using *lexical chaining* — the process of extracting chains of lexically related words from a document with reference to a lexical resource such as WordNet (Beckwith et al., 1991) which relates words by their meaning. The effects of polysemy and synonymy are reduced since related documents representing the same concepts expressed in different words should produce similar chains. The limitations of WordNet's expressiveness, however, has so far limited the success of this approach (Green, 1999a).

Glushko wrote of associative links: "these are the links that are closest to the vision of hypertext, namely links that are not explicit between related documents but that can be extracted by careful and creative analysis of the two texts and the relationship between them" (Glushko, 1989). A question that arises in the context of this work is therefore whether this "careful and creative analysis" can be carried out automatically. The techniques outlined here have met with limited success and in fact, lexical methods may inevitably prove inadequate in every case, as Bernstein noted "even a fluent semantic interpretation of the text cannot suffice — link creation may depend on understanding the knowledge and intentions of both reader and author"' (Bernstein, 1990), concluding that "an accurate automatic linker must be able to recognise and interpret humour, metaphor, euphemism, and irony; clearly this is too ambitious a goal for immediate realization".

Although the benefits, in terms of reducing authoring effort, of automatically identifying associative links in large collections of documents are apparent (and may outweigh the fact that the produced links may be of variable quality), this work takes the view that automatic linking techniques are no substitute for human wisdom and intuition in smaller integrated hypertexts. Automatic linking techniques are not disregarded altogether however; an automatic link (or *knowledge*) discovery process aiding (rather than replacing) human understanding is a theme revisited in the context of the Semantic Web (Chapter 5) and in Chapter 11 as a proposed future direction for this work.

2.3 Associative Writing in the World-Wide Web

This section introduces the specific focus of this work: supporting Associative Writing in the World-Wide Web, the most far-reaching and successful hypertext system to date. A brief historical perspective of the Web, from its conception to the present day is provided, and the first two of a series of core challenges posed by the problem of supporting Associative Writing in the Web are introduced. The first is the controversial "lost in hyperspace" problem which concerns user disorientation in hypertexts (Conklin, 1987). The second involves a recent series of legal cases where content providers have successfully prevented other Web sites linking to their material. Both challenges are similar in that they attempt to restrict linking in the Web, and hence may have significant implications for this work.

2.3.1 Brief History of the Web

The World-Wide Web emerged in the early 1990s (Berners-Lee et al., 1992), and has become the most popular hypertext system in use. Developed by Tim Berners-Lee at CERN[3], the technology was originally conceived as a means of recording connections among the various people, computers and projects at CERN, allowing visiting physicists working in different universities and institutes all over the world to remain "connected" to their work. In fact, the communication of shared knowledge was a major driving force when the Web was first proposed. The intent was that by building a hypertext web, a group of people would be able to easily express themselves, quickly acquiring and conveying knowledge, overcoming misunderstandings and avoiding duplication of effort — the Web would allow ever larger, more interconnected groups of people to act as if they "shared a larger intuitive brain" (Berners-Lee, 1999).

The main cornerstones of the Web were the Uniform Resource Locator (URL), HyperText Markup Language (HTML), and HyperText Transfer Protocol (HTTP). URLs made it possible to uniquely identify anything accessible on the Internet, HTML allowed users to "mark up" the structure of their documents and create URL-based links to other resources, and HTTP made it possible to easily transfer HTML pages across the Web.

Berners-Lee's first graphical Web browser allowed users to create and describe (by adding arbitrary labels) bi-directional associative links. However, the NCSA Mosaic "browse-only" browser (which only let users *view* Web pages and follow associative links in one direction — pages and links had to be created using a separate editing application)

[3] CERN is now the European Organisation for Nuclear Research.

proved more popular, and became the foundation of today's popular Netscape/Mozilla and Microsoft browsers. The Mosaic browser was free, easy to use, and ran on many platforms.

The Web (although somewhat restricted compared to Berners-Lee's original concept) suddenly became usable for people other than academics and specialists, and the amount of information available on the Web has been increasing exponentially ever since: the Web has become a truly global information resource. Researchers now release their results to the Web before they appear in print; corporations list their URLs alongside their toll-free numbers; news media and entertainment companies vie for the attention of a browsing audience. The Web has become the most visible manifestation of a new medium: a global, populist hypertext (Kleinberg, 1999b). A survey carried out by Bright Planet estimated the 'surface' Web to contain over 1 billion documents, and estimated that up to 550 billion documents exist in the 'deep' Web (Bergman, 2001). Pennock asserted that the Web is a reflection of human culture — a massive social network encoding associative links among almost 10^9 documents (Pennock et al., 2002). As Moulthrop & Kaplan have noted:

> If success is measured by numbers of users, documents, and links, then the Web is an overwhelmingly successful implementation of hypertext. (Moulthrop and Kaplan, 1995).

The Web unifies several concepts into a single idea, the *page*:

- The user's view of the information on the screen.
- The unit of navigation (what you get when you click a link or activate a navigation action like a bookmark).
- A textual address used to retrieve information over the net (the URL).
- The storage of the information on the server and the author's editing unit.

The fundamental design of the Web is based on having the page as the atomic unit of information, and the notion of the page permeates all aspects of the Web (Nielsen, 1996b). Several pages collected together form a *site*, which typically has an entry (home) page, and are under the control or ownership of a single designer or design team. Exceptions may exist in large corporate or organisational sites that are complex and involve many hundreds of pages.

2.3.2 On Identifying Core Challenges

As indicated in Chapter 1, this work attempts to address five challenges facing Associative Writing in the Web docuverse: the "lost in hyperspace" problem, legal issues over "deep linking", the Web's restricted hypermedia model, link integrity, and the limited support for Associative Writing provided by popular writing tools. These challenges represent a subset of the many challenges facing hypermedia research in the Web in general, as reported in the literature. As well as the five challenges listed above, other researchers have put forward the challenges of supporting mirroring, versioning, transclusions, 'live' interaction between users browsing the Web, and providing an underlying distributed file system (Pam, 1995), link attributes for structure based query, personalised links, trails and guided tours, and backtracking and history based navigation (Bieber et al., 1997). Although some of these issues will be touched upon in this thesis (for example, XLink as a mechanism for specifying link attributes, Section 4.3; WebDAV as a means of providing versioning support, Section 10.5; Adaptive Hypermedia as an approach to personalised links, Section 11.2.6), this work considers the five challenges listed above as 'core' to supporting Associative Writing in the Web — they are the most immediate problems facing writers of integrated hypertexts and as such represent the minimal subset of challenges which must be addressed in order to help writers achieve the vision of an integrated Web. The other challenges listed here are no less significant (or problematic), but this work considers the core subset as a 'first hurdle', with the anticipation that work towards addressing the remaining challenges will augment the work on supporting Associative Writing described here.

2.3.3 Lost in Hyperspace or Lost in Controversy?

In 1987 Conklin identified two related problems which he felt were endemic to hypertext: *disorientation* and *cognitive overhead* (Conklin, 1987). Cognitive overhead results from "the additional effort and concentration necessary to maintain several tasks or trails at one time." This refers to the reader's ability to follow associative links related indirectly to the current reading task (on a purposeful tangent or detour, or perhaps by accident), as well as the need to follow several interconnected paths to visit as much of the hypertext network as necessary. Disorientation is "the tendency to lose one's sense of location and direction in a nonlinear document", for which Conklin famously coined the phrase "lost in hyperspace".

Since then, the lost in hyperspace problem has given rise to much controversy (Landow, 1990). Some researchers think that the problem is one of the most difficult issues in hypertext research and that there is yet more to be done to address it; others believe that it is not a serious problem and that efforts should be channelled to address other more pressing needs (Thimbleby et al., 1997). For example, the GVU 4th WWW User Survey (Kehoe and Pitkow, 1995)[4] reported from a sample size of approximately 14,500 responses that users were not "lost" and that "lost in hyperspace" was not a problem (6.5%) compared to the most widely cited problem of slow download responses (69.1%).

[4] Earlier GVU surveys (see `http://www.gvu.gatech.edu/user_surveys/`) did not ask specific questions about user problems in the Web.

Problem	Reported (%)	
	1995	1999
Lost in hyperspace	6.5	3.7
Slow download responses	69.1	61.4
Not being able to find known page	34.5	30.0
Not being able to find previously visited page	23.7	16.6
Not being able to visualise location in the hypertext network	14.3	7.4

Table 2.2: Comparing reported user problems in 4th (1995) and 10th (1999) GVU WWW User Surveys.

However, some researchers point out that in the same survey, other responses which may be related to the lost in hypertext problem constitute a potentially enormous number of Web users who might not specifically report that they were "lost", but experienced different forms and degress of "lostness": *not being able to find a known page* (34.5%), *not being able to find a previously visited page* (23.7%), and *not being able to visualise location in the hypertext network* (14.3%). The latest (10th) GVU WWW User Survey (Kehoe et al., 1999) reports similar findings (Table 2.2), and a survey carried out by (Davies et al., 1998) show similar trends.

Lynch and Horton, the authors of the widely-read *Yale C/AIM Web Style Guide*, describe two kinds of link found on Web sites (Lynch and Horton, 1997). Navigational links are the backbone of a site's user interface, connecting pages within the site. Associative links offer parenthetical material, footnotes, digressions, or parallel themes that the author believes will enrich the main content. They argue that the most common difficulties in site design result from the overuse or poor placement of associative links: associative links disrupt the narrative flow by inviting readers to go elsewhere, and instead of enhancing the reader's understanding of a subject, an associative link may "send them to a foreign land without a guide" in an attempt to discern the link's significance. Lynch and Horton argue that most links do not belong in the "middle" of Web pages — they ""aren't important enough to justify the potential distraction". Similarly, a number of hypermedia (and later Web) design models, such as Hypermedia Templates (Catlin et al., 1991), HDM (Garzotto et al., 1993, 1995), and WebML (Ceri et al., 2000), focus on the connections between "atomic" information units (local coherence) rather than associative linking from within the content of each unit (a problem revisited in Chapter 11) More recently, in describing an *Information Architecture for the World-Wide Web*, (Rosenfeld and Morville, 2002) call for writers to systematically follow a simple linking practice, with emphasis on clear, sparse linking. Other researchers made analogies between links in hypertexts and 'go-tos' in programming (Brown, 1987), concluding that go-tos (and therefore links) were "harmful" (De Young, 1990).

Khan and Locatis investigated information retrieval from hypertexts on the Web, and demonstrated that hypertexts with low link densities[5] displayed in a list format produced the best overall search performance (Khan and Locatis, 1998). Users performed six searches of varying difficulty on one of four versions of a hypertext, organised hierarchically with either low or high density links (three or six links respectively) displayed in either lists or paragraphs. Khan and Locatis argued that when fewer links are displayed, cognitive load is reduced because searching and exploration are more focused, leading to the better performance times. However, as the authors themselves note, this study focused only on information finding, not on its extraction or interpretation. However, in a similar experiment (Hitchcock, 2002) found that although users presented with linked content completed tasks more quickly and efficiently than those with unlinked content, they were less satisfied with the results they found.

Bernstein validates widespread concerns of confusing the reader through associative linking, but disagrees with the strategy of restricting the role of links in documents (Bernstein, 1991). Bernstein's view is that disorientation arises not from the fact that information is associatively linked, but the fact that it is linked *badly*, stating that, as with any medium, "hypertext may prove unwieldy and inexpressive when used without care and thought". In addition, Bernstein claims that no convincing evidence exists that associatively linked information necessarily disorients the reader or that a sequential presentation prevents readers from getting lost. In his 1998 Web hypertext *Hypertext Gardens*, he concludes that "hypertext disorientation most often arises from muddled writing, or from the complexity of the subject" and "rigid hypertext makes a large hypertext seem smaller.. Complex and intricate structure makes a small hypertext seem larger, inviting deeper and more thoughtful exploration" (Bernstein, 1998a). Very likely, Bernstein would consider the recommendations of a sparse linking approach outlined above to be examples of a "monotonous web", denying readers "the rich environment and sense of freedom which made hypertext desirable in the first place" (Bernstein, 1991).

In the face of this obvious controversy, (Theng et al., 1996; Thimbleby et al., 1997) re-examine the lost in hyperspace problem and question whether the problem is significant enough to warrant the continued attention of the hypertext research community. They note that in most cases, the lost in hyperspace problem is regarded as a *user's problem*, resulting in improvements being sought in the presentation of information such as graphical browsers and query/search mechanisms — for example, navigational metaphors (Ichimura and Matsushita, 1993; Golovchinsky and Chignell, 1997), sophisticated overview maps (Gloor, 1991; Bieber and Wan, 1994; Robert and Lecolinet, 1998; Kreutz et al., 1999), link reduction algorithms (Furuta et al., 1997), and recommender systems (Pikrakis et al., 1998). They conclude that user disorientation may in fact stem from an *engineering problem*: "Bad" system design causes hypertext authors themselves to become "lost" in the process of designing and authoring hypertexts, and thus they inadvertently contribute to poorly designed hypertexts, which in turn may

[5] Link density is the ratio of linked to unlinked content; hypertexts with low link densities thus have fewer links.

cause difficulties to the reader.

2.3.3.1 Implications for this Work

The lost in hyperspace problem could potentially have important implications for Associative Writing since integrated hypertexts embrace the kind of unrestricted linking so staunchly rejected by those in favour of addressing the problem. However, this thesis takes Bernstein's view that restricting the role of associative links in hypertexts is detrimental to the reader's experience, particularly as the positive aspects of Associative Writing enumerated earlier in this chapter require the reader to have the freedom to explore the global context of new contributions. The challenge therefore, is to avoid hypertext disorientation arising from "muddled writing" (which Bernstein agrees may cause user confusion) which itself may stem from 'bad' system design.

2.3.4 Associative Writing in the Web: Copyright Infringement?

In popular Web parlance, the term *deep linking* refers to the practice of a writer creating an associative link from their web page directly to an interior page of another site[6]. Deep linking is an essential facet of Associative Writing — deep (associative) links integrate new hypertext contributions with existing global context. Nielsen, the Web usability guru, encourages deep linking on the Web, stating that "deep linking enhances usability because it is more likely to satisfy users' needs" (Nielsen, 2002). As an example, linking from an allusion to an event to an online newspaper article describing the event in more detail is far more useful (and more in line with the Associative Writing ethos) than linking to the *home page* of the online newspaper which carries the article.

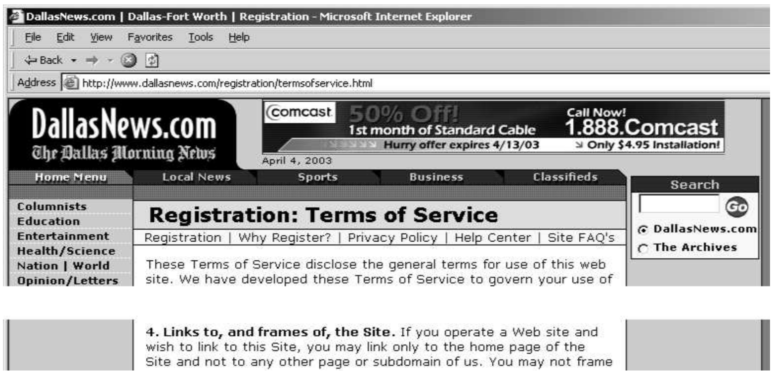
However, some online organisations have claimed that such deep links violate the copyright over their material, and allow users to bypass the advertisements on their home pages. In 2000, a ruling in the United States stated that "Hyperlinking does not itself involve a violation of the Copyright Act..There is no deception in what is happening. This is analogous to using a library's card index to get reference to particular items, albeit faster and more efficiently" (Kennedy, 2000). However, more recently (July 2002) in Europe, a Copenhagen court ruled in favour of the Danish Newspaper Publishers Association[7] which claimed that Danish company Newsbooster[8] violated copyright laws by deep linking to newspaper articles on several Danish newspapers' Web sites[9]. An injunction against Newsbooster forbidding the service to deep link to any association-owned content was granted because the Newsbooster service was deemed to be in direct

[6] No link in the Web is technically or physically "deeper" or "lower" than any other — the idea of deep linking is therefore, in itself, an artificial construct.

[7] `http://www.pressenshus.dk/`

[8] `http://www.newsbooster.com/`

[9] Reported at Wired News — `http://www.wired.com/`



Registration: Terms of Service

These Terms of Service disclose the general terms for use of this web site. We have developed these Terms of Service to govern your use of

4. Links to, and frames of, the Site. If you operate a Web site and wish to link to this Site, you may link only to the home page of the Site and not to any other page or subdomain of us. You may not frame

FIGURE 2.1: Extract from `dallasnews.com` terms of service.

competition with the newspapers to whose content the service linked, a violation of the Danish Copyright Act and the Danish Marketing Act, which forbids profiting by use of other companies' products and/or services.

The fact that the Newsbooster case centres specifically on Danish law may imply that similar cases may be less relevant to the law of other countries. However, in the case of German newspaper Mainpost[10] and German search service NewsClub[11] (which searches through and links directly to Mainpost content), a court in Munich ruled that NewsClub is in violation of European Union Law.

The number of media websites that attempt to ban linking to their material by other sites is growing. The Rodale Press[12], Dallas Morning News[13], National Public Radio[14] (a nonprofit organisation), and Bloomberg[15] have recently tried to enforce rules on deep linking to their content (Figure 2.1).

Similar controversy surrounded Microsoft's recent announcement that "Smart Tags" technology would be included in the release of Windows XP (Hughes and Carr, 2002). Smart Tags allow software plug-ins to Microsoft applications (including the Internet Explorer Web Browser) to identify regions of content in documents that are suitable for "link annotation" and to control the actions that take place when the user clicks on a link annotation (for example, linking a product name to a Web site selling that product). These links annotations are synonymous with Microcosm's generic links (Section 2.1.2.2). Legal (and moral) issues were raised in that original content could therefore be altered by the application without the permission of the author, raising copyright issues over the creation of derivative works. Even if the work is freely published, the moral issue of whether the author has given the right to have their content altered still remains. Microsoft subsequently deactivated Smart Tags in Internet Explorer, although they are

[10] `http://www.mainpost.de`

[11] `http://www.newsclub.de`

[12] `http://www.rodale.com/`

[13] `http://www.dallasnews.com/registration/termsofservice.html`

[14] `http://www.npr.org/about/termsofuse.html`

[15] `http://www.bloomberg.com/tos.html#linking`

available in Office XP applications.

2.3.4.1 Implications for this Work

If legal decisions reached in recent court cases reflect a growing trend towards restricting Web linking, this may have significant implications for this work. For Associative Writing to be effective, and bring the reported benefits to writers and readers, writers need to be able to freely link their contribution to arbitrary existing content, regardless of whether that link constitutes a "deep link" to an organisations interior content — deep linking is the very essence of writing globally coherent hypertexts. If more and more Web sites start to control the content which they provide, Associative Writing becomes difficult, forcing writers to summarise or replicate existing content rather than provide associative links to the (integrated) original sources.

It was posited earlier that if associative links connecting a new work to its global context were bi-directional, newer contributions would become immediately reachable from older texts. The controversy surrounding Microsoft's Smart Tags may therefore also have ramifications for this work, since it demonstrates the problems of adding links to existing Web content.

Nelson's *transcopyright* concept, part of the Xanadu project since the 1960s, demonstrates how the issue of copyright could be resolved in a Xanadu docuverse (Nelson, 1997). Transcopyright preserves the integrity, copyright, and royalty for digital materials, yet allows everyone to freely re-use these materials, which retain their identity at all times. Under this arrangement, writers are free to republish digital materials virtually (as *transcluded* quotations, anthologies, and collages) — each copy of the existing content is separately purchased from the original publisher at the time of delivery. However, it is unlikely that such a mechanism will be widely available any time soon. The challenge therefore, is to demonstrate the advantages of "deep linking" in the context of the Associative Writing process — after all, if no-one links to (integrates) a news article, fewer people will read it!

2.4 Summary

This chapter has introduced the focus of the work which will be reported over the course of this thesis from a historical standpoint, reviewing the evolution of hypertext systems research from its foundations in visionary work by Bush, Engelbart and Nelson, through classical hypertext systems such as NoteCards and Intermedia, to more recent open hypertext efforts which attempt to integrate hypertext services with existing tools on the desktop. This perspective then informed a description of the concept of integrated writing: the process of using associative links — the "stock in trade of hypertext systems" (DeRose, 1989) — to integrate new (typically non-fiction) contributions with existing related work published in the hypertext docuverse, which is termed *Associative Writing* in this thesis. This description also highlighted that this work focuses on Associative Writing in general, rather than specifically emphasising the integration of scholarly works, and on human-created associative links rather those those inferred by a machine comparison of documents.

A review of the evolution of the Web, from its conception to the present day, introduced the first two challenges posed by Associative Writing in this medium: the "lost in hyperspace problem" and concerns over copyright infringement. Both challenges are similar in that they attempt to restrict linking in the Web. The lost in hyperspace problem has given rise to much controversy in that many researchers do not agree on its severity. Some solutions call for writers to systematically follow a simple linking practice, with emphasis on clear, sparse linking, whereas others argue that disorientation only arises from badly linked hypertexts or from bad system design. This thesis takes the view that restricting the role of associative links in integrated hypertexts is detrimental to the reader experience in exploring the contributions of the new work. The challenge therefore, is to avoid avoid badly designed hypertexts by adopting an "engineering approach" (Thimbleby et al., 1997) to the problem.

Recent court rulings that "deep linking" (the practice of an author creating an associative link from their Web page directly to an interior page of another site) infringes the copyright of the owner of the linked content have also been outlined. If these rulings reflect a growing trend towards restricting Web linking, this may have significant implications for this work — for Associative Writing to be effective writers need to be able to freely link their contributions to arbitrary existing content. The challenge in this instance therefore, is perhaps to try and demonstrate the advantages of deep linking (in the context of Associative Writing) to those strongly opposed.