

# Metadocuments supporting digital library information discovery

Unmil P. Karadkar, Luis Francisco-Revilla, Richard Furuta, Frank Shipman, Avital Arora, Suwendu Dash, Pratik Dave, Emily Luke

Center for the Study of Digital Libraries and Department of Computer Science, Texas A&M University, College Station, TX 77843-3112, USA  
e-mail: walden@csdl.tamu.edu

Published online: 27 July 2004 – © Springer-Verlag 2004

**Abstract.** The World Wide Web is a decentralized, unmanaged, dynamically changing repository of digital documents. Walden's Paths provides tools that enable authors to collect, organize, annotate, and present Web-based information to reader communities via a linear metadocument called a path. Walden's Paths includes path authoring and reading interfaces supporting the contextualization of included materials to match authors' goals as well as enabling browsing off the path to match readers' personal interests. It also provides tools to manage these paths of transient Web materials based on the identification and evaluation of changes to the component pages. Experience with Walden's Paths in educational settings and changes to Web technology as well as the Web-savviness of users have led to a variety of changes to earlier designs. Current directions of work include the development of methods for evaluating readers' understanding via quizzes associated with paths and richer path structures.

**Keywords:** Walden's Paths – Metadocuments – Hypertext – Path maintenance – Digital library services

## 1 Introduction

Libraries, traditional as well as digital, act as storehouses of information and also provide tools to find needed information. Patrons often use libraries for information discovery, that is, to acquire information that is needed to fulfill their goals. Most information-location tools are either search-based or index-based. Libraries rarely provide tools to store, manage, and pass on to others the information that patrons have painstakingly discovered.

We examine the use of metadocuments as a means of contextualizing and communicating the information gleaned from collections of digital documents. Metadocuments are documents that contain (or point to) other

documents. Authors of documents organize information elements and impose structure that relates them implicitly by their organization or by explicitly linking them. Similarly, creators of metadocuments provide metastructure, an additional level of structure that holds together independent (possibly disparate) documents in a wider context without altering either the documents they point to or the organizational structure of the digital library that contains them. We focus primarily on linear metastructures that we call *paths*.

Vannevar Bush first proposed the use of paths or *trails* as a means to join related resources from diverse, widely separated source documents [3]. Bush's trails served two purposes: they provided a personal means to remember and organize found information and to communicate this information to colleagues. Paths have since been used in a variety of customized settings, for example, in Note-cards [9], as Guided Tours and Table Tops [19], and in Zellweger's *directed paths* [22, 23].

Walden's Paths uses the World Wide Web as the information substrate over which the paths are created. Being an enormous, decentralized, dynamically changing, unregulated information repository, the Web is possibly the worst-case scenario for a digital library. It lacks crucial library services like cataloging and collection development [12]. In the absence of editors, librarians, and publishers, users of the Web are deprived of guidance to trustworthy materials [20]. Paths lay a metastructure over the Web. Creators of paths evaluate Web-based materials and contextualize the information by including it in paths. Authors may also present their opinion regarding the validity of the information. In some sense, they provide an editorial service that helps reformulate Web-based information in a digital library context.

The rest of the paper is organized as follows. Section 2 describes the design and use of software tools for creating, viewing, and maintaining paths. This section also

discusses the factors that influenced the evolution of these tools. Section 3 discusses the current research directions. Section 4 summarizes and concludes the paper.

## 2 Walden's paths

Paths are annotated lists of URLs (Uniform Resource Locators) or pointers to Web documents to be traversed in the order of their appearance. In general, path authors are not the authors of Web pages contained in the paths. Walden's Paths facilitates reuse of information already available on the Web in a possibly different context to create information structures and present them to audiences with minimal effort. Readers of paths can view these information structures in the author's context without losing the ability to explore the links that emanate from them to gain additional related information, which is a crucial feature of the Web. Creators of Walden's Paths provide a guided environment for exploratory learning where readers are assured to obtain a minimal amount of information upon traversal of the path and may explore the information space for increasing gains depending upon their inclination or the time available.

We have explored the use of paths as tools for K-12 and undergraduate educators to achieve their curricular objectives [5, 17]. Most content on the Web is oriented toward casual readers and not authored with a focus on young audiences or with an educational setting in mind. In the following subsections we provide a brief overview of various Walden's Paths system components.

### 2.1 Creation of paths

Authoring paths is a complex information task. Authors typically start with a concept or topic for a path. They must then locate promising Web pages, browse and evaluate this material, select the information elements to be used in the path, develop an outline for the path, and, finally, add the URLs and annotations to create the path. We have experimented with a variety of tools to support the path authoring process.

The earliest of these provided integrated support for searching the Web, authoring paths, and saving them on the server for user access [6]. User studies with this tool revealed that teachers usually have little time to create paths in school due to high workload at school and few could find the time to author new paths [17]. VIKI [14], a spatial hypermedia system, was tapped into as a possible host for supporting a longer-term authoring process [18]. Authors could save the search results generated and the paths authored as VIKI collections until they were ready to be presented to users.

More recently we have developed a new set of authoring tools to reflect the changes in society, technology, and work practices. Since the days of the first authoring tools, path authors and people in general have grown familiar

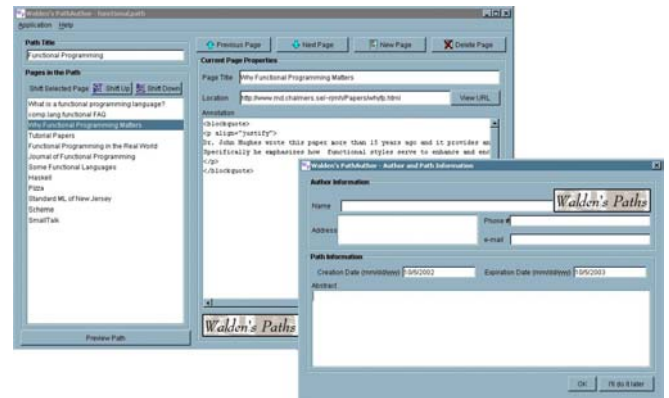


Fig. 1. Walden's PathAuthor

with the use of the Web and with searching for information on the Web. Further, the interfaces to Web search engines change often and updating the authoring tool's search interface is a maintainer's nightmare. With the use of laptop computers by increasingly mobile users, an Internet connection is not always available. The new set of tools lets authors search for information using their favorite search engine and incorporates a two-step process that separates the creation of the path from making it available to the users. The tool for creating the paths is called PathAuthor while the one for copying the path to a public location for presentation is labeled the PathPublisher [11].

PathAuthor is a network-independent Java application that supports the creation of paths without requiring a network connection. Authors may select the information elements to be used and save the links locally while at school where they may have a faster Internet connection. They may then add the annotations and create the rhetorical structure for the path in the absence of a network connection, presumably at home or while traveling. Figure 1 illustrates the interface of PathAuthor. The left side of the interface displays information about the entire path, while the right half displays information about the current page. Author information and path abstract are provided in the Author Information dialog box. The paths thus created must be transferred to a Web server via PathPublisher, a Java Servlet-based Web application, to make them available to their intended audience.

### 2.2 Viewing paths

In 1996, we had the opportunity to observe four classes of sixth graders surfing the Web [13]. The students started browsing from an initial page designed by the instructors for an educational setting. The page included links of various online places like museums, public institutions (for example, the White House), collections of pictures, peer-related Web activities, and entertainment-oriented sites. Five phenomena stood out: the difficulties associated with reaching navigational dead ends, the sociabil-

ity of Web use and collaborative nature of exploration, the success of simple navigational modes, the confusion in interpreting the signs and signals of network performance, and the compelling quality of participation, rather than just interaction [6]. These suggested the design elements in the construction of a path presentation mechanism that would meet the needs of students.

A Web application called Path Server presents paths to the readers. Readers can access Path Server from any standard Web browser that supports frames. Typically, readers begin a Walden's Paths session by selecting a path from a list of paths available on the server. The interface for viewing paths is shown in Fig. 2. The bottom frame displays the Web page that the page points to, as it would be displayed without the mediation by Walden's Paths. The annotation or contextualizing text added by the author of the path is displayed in the top right part. The top left portion contains the controls for viewing the path. Users may view the pages on the path serially by clicking on the "Next" and "Back" images or they may jump to any page on the path by scrolling to and clicking on the "paw" corresponding to the position of the page. While on the path, readers are free to follow links on pages in the path to freely examine the information space. While the reader is browsing the information space off the path, the control widgets described above are replaced by a single image that links back to the last page visited on the path, thus providing readers with an easy way to return to the path once they are done exploring.

In our interactions with teachers and students, we have observed that Walden's Paths has a low technological barrier for use [6]. At workshops organized in 1996, teachers grew accustomed to the technology over

the course of the workshop and soon focused on the information contents of the paths rather than on the technical artifacts and elements of Path Server. Comments received from teachers about our initial prototypes also focused mostly on the contents of the paths rather than the technology used. The technology had disappeared and become an element of the landscape allowing the authors as well as readers to focus on the information communicated by the paths.

Paths depend upon the referred Web pages for their coherence and completeness. As the authors exercise no control over how or when the underlying pages may change, the system must provide support for monitoring the Web pages for change and notifying authors of any changes to ensure the validity of the paths.

### 2.3 Path maintenance

Authors of Web pages change them often, if only to provide a different look and feel for their pages. The median Web page age is about 100 days [2]. Given that about half the Web pages change every 3 months or so, path authors must constantly keep track of how the Web pages of interest change, move, or vanish. Determination of relevance of change can only be done in the use context of the document. Some changes may be trivial and require no action on part of the path authors. Others may be significant and threaten the integrity of the paths. In its simplest form, comparing the Web page periodically with a locally cached copy and returning a Boolean result can track changes. At the other end of the spectrum, an intelligent page tracker could monitor the Web pages for syntactic and semantic changes and report only the changes that significantly affect the fabric of the path. While the simplest system does little to help us, the most complicated one must understand the path's concept as well as the author's intentions at a semantic level. We have settled for a middle path, one that predicts semantic changes based on syntactic changes detected in Web pages [4].

We classify changes into four categories. Content or semantic changes refer to modifications to a page from the reader's viewpoint. Presentation changes modify the display characteristics of the information. Changes to HTML tags (for example, changes in the font face or size of a paragraph) or relocation of information elements within a page are examples of presentation changes. Changes to the underlying connection of the current page to other pages, for example, modification of links, comprise structural changes. Finally, modifications to active components of a document, for example scripts, plug-ins, and applets, refer to behavioral changes. Our change-tracking tool, PathManager, currently monitors for changes in the structure, content, and presentation of the pages. Behavioral changes are difficult to predict reliably and we do not report these yet.

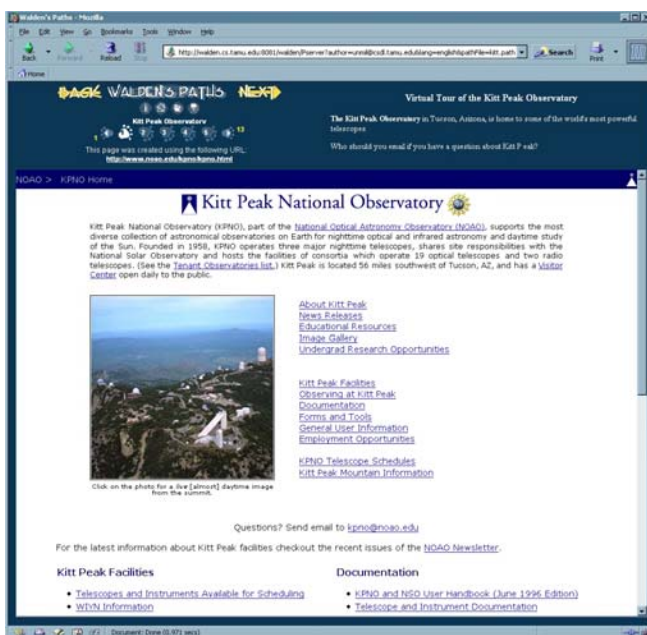


Fig. 2. Path Server interface

We conducted a study to understand how humans perceived changes in Web pages and to estimate the types of changes that the path authors would expect to be notified of, the design and results of which are reported in [5]. Test users were shown pairs of pages in which the experimenters had introduced controlled changes for one of the three types that we monitor. In the first phase they were told what changes had been introduced and asked if they would consider such changes crucial enough that they should be notified of them. In the next two phases, they were allowed to view these pairs for 15s (shorter duration) and 1 min (longer duration) without being informed if the pages had differences in them. For all the phases, they were asked to rate the level of change on a 7-point scale for each of the types of changes. In general, the users' desire for notification increased with the perceived level of change. As the quantity of change increased, the users perceived the change to be of all types (we had modified each page for one kind of change only). Besides the cases where presentation changes were drastic, they were not perceived to be significant. Users reported content changes with equal accuracy in both the short and long observation period presumably because these changes were visually observable. In contrast, they were more successful in identifying structural changes when they had more time to view the pages. The users also indicated that they wished to be notified of structural changes, implying that they care what other pages are accessible from the current page included in the path.

The results of this study helped formulate the design of the path maintenance tool. PathManager helps authors in tracking changes in Web pages and assesses the level as well as relevance of the changes. A snapshot of PathManager in action is shown in Fig. 3.

PathManager stores document signatures that include a document checksum, paragraph checksums, headings, links, and keywords. PathManager supports two algorithms for calculating changes based on these attributes. The first is a variation of Johnson's algorithm [10]. Johnson proposed his algorithm to support decisions about displaying pages in Web-based tutorials based on the level of change. Johnson's algorithm does not include the changes in links. We added these to account for the users' need to know about structural changes. The other, a proportional algorithm, provides a normalized and symmetric distance that is easier to use for different sets of Web pages. The stored signature of a document is compared with the current document to detect syntactic changes. PathManager assesses the level of change. Changes are reported numerically and are color coded to report the level: no change, low, medium, and high degrees of change. PathManager reports levels of changes for each type of change as well as for the page as a whole.

### 3 Research directions

Walden's Paths continues to evolve to cater to our audience, driven by user needs and technological changes and to reflect the advances in our understanding and our ability to provide better tools. Advances in technology coupled with targeted user studies enable us to better address societal and legal issues and support evolving trends in work practices of authors as well as readers of paths.

#### 3.1 Enhancements to system components

Path Server, the authoring components, and PathManager are periodically upgraded to enhance the interface and to reflect the changes in the knowledge and skill sets of authors and readers. Path Server has evolved to match the improvement in Web browser functionality and the enhanced expressiveness of HTML. PathManager is still a relatively new tool and has potential for enhancement to better classify changes in Web pages in general as well as with respect to the paths that contain them. For improving the general understanding of changes in Web pages, PathManager must address several issues like inconsistent use of tags on the part of the Web page creators, use of images that masquerade as text, and customized reminders of nonavailability of pages that return unexpected values that are inconsistent with the HTTP protocol [21].

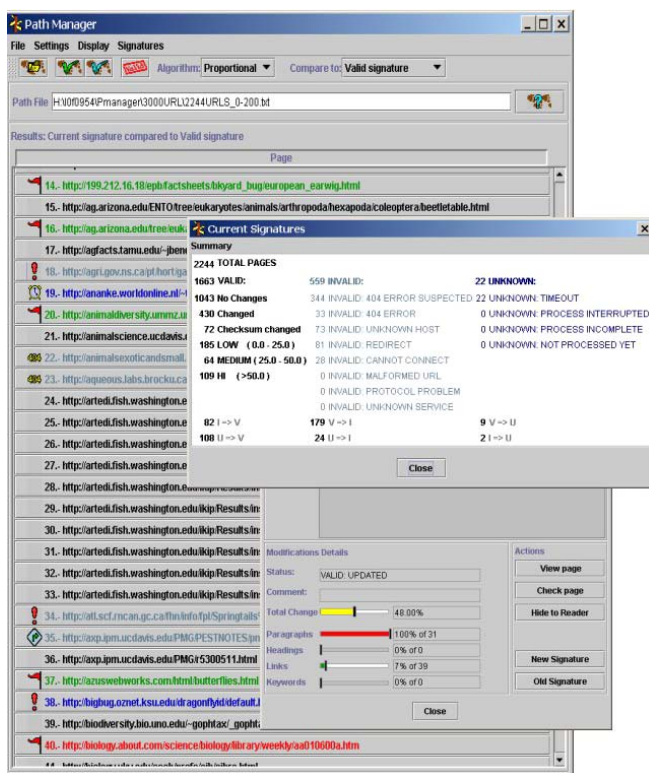


Fig. 3. PathManager

### 3.2 Intellectual property issues

Reuse of materials, possibly in contexts unexpected or unforeseen by the content creators, raises issues regarding intellectual property rights. Path authors have asked for more control than is provided by complete inclusion of Web pages in their paths, for example, inclusion of some information elements on a page and changes to the visual appearance of the included material. These requests conflict with the views and expectations of creators of the Web pages who expect to have control over the use of their materials in content as well as in presentation [17]. We have strived to correctly attribute the credits for the materials used. The enhancement of the presentation interface to highlight the difference between authors' contributions from those of the content providers' has partly been in response to the need for correct attribution of materials to the right sources. The path authors' primary contribution is the path structure and the annotations. This sometimes requires reformulation and reinterpretation of others' work. Such adaptations, though accepted and common in the print form, are a contentious issue on the Web due to the use of the source materials and not copies thereof. While the legal questions regarding fair use are yet to be crystallized, we make every possible effort to attribute the authorship of materials to their sources. In Fig. 2, the link below the paws in the top left frame displays the original URL of the page displayed in the bottom frame, thus enabling users to view this page without the mediating Walden's Paths interface. The frames also clearly separate the content created by the path authors from that created by the authors of pages included in the paths.

### 3.3 Richer path structures

Over the years, creators of paths have used them for a variety of purposes and in various roles. Teachers have created paths that act as extended bookmark lists, as teaching tools that contain a rhetorical structure, and as extended guided tours and presentations [6]. Teachers with more programming-related experience have expressed the need for support for more complex structures. We have experimented with collaborative authoring of paths, sub-path mechanisms that incorporate hierarchy within the path structures, and creation of ephemeral paths on an as-needed basis [7]. We continue to look for various (possibly context-specific) structures that help increase the expressivity of paths.

### 3.4 Evaluation of understanding

Web-based instruction is changing the traditional instructor-learner dichotomy of face-to-face learning. The new paradigm shifts, at least partially, the responsibility for

learning from the instructor to the learner [15, 16]. Goodwin proposes the use of technology to foster a deeper approach to learning by actively engaging students in the learning process and transferring the focus and responsibility for learning to the students [8].

The Walden's Paths Quiz system helps motivate learners by allowing them to evaluate the knowledge gained and retained while browsing the paths [1]. The interaction can be user initiated or enforced by the creator of the path. Users may invoke a quiz session at any point while browsing a path. The system generates a short quiz based on the materials viewed by the user in the current session. The path author may enforce quiz questions at certain points on the path, which must be answered correctly by the readers before they can proceed further. These mechanisms are targeted to providing real-time feedback to the readers about their level of understanding of the information included in the paths.

The Quiz system also helps instructors generate and administer online tests and quizzes for grading. Instructors specify properties of the quiz to be generated by providing the paths over which the quiz is to be created, the types of questions (multiple choice, long answers, etc.), the number of questions, level of difficulty, and other parameters (for a complete description see [1]). The system returns a URL for the generated quiz, which the teacher can then forward to the students. During the time that the quiz must be taken, Path Server may be rendered unavailable to the students. The teachers may further impose additional restrictions to ensure fairness and validity of the tests. For example, they may require that the quizzes be taken in the controlled environment of a laboratory or that they be taken from environments that prohibit access to the Internet. Further, the URLs for the quizzes may be made available only for specific periods of time. There are a multitude of possibilities and scenarios, and teachers may adapt these to suit their level of comfort and trust.

## 4 Conclusion

Walden's Paths provides the tools to create, present, and manage Web-based contextualized discourses in a linear, directed structure in the metaphor of a path. Creators of these presentations may reuse information that is already available on the Web. In doing so, they implicitly or explicitly express their opinions regarding the Web pages used. The exploration issues regarding Web-based information presentation, authoring, and maintenance are ongoing. Although many issues remain open for further study, our experience with Walden's Paths suggests that metadocuments created and mediated by experts are a natural, reliable, and effective means for organizing, communicating, and contextualizing information in digital libraries.

Further information about Walden's Paths may be found on our Web page: <http://www.csdl.tamu.edu/walden/>.

*Acknowledgements.* In addition to the authors of this paper, the Walden's Paths project has benefitted from the efforts of other participants over the years. We gratefully acknowledge their contributions. This material is based on work supported by the National Science Foundation under grant numbers IIS-9812040, DUE-0085798, DUE-0121527, and IIS-0219540.

## References

- Arora A (2002) Walden's Paths quiz: system design and implementation. Masters Thesis, Department of Computer Science, Texas A&M University, College Station, TX
- Brewington B, Cybenko G (2000) How dynamic is the Web? In: Proceedings of WWW9 – the 9th international World Wide Web conference, IW3C2, Amsterdam, May 15–19 2000, pp 264–296
- Bush V (1945) As we may think. *Atlantic Monthly*, August 1945, pp 101–108
- Francisco-Revilla L, Shipman F, Furuta R, Karadkar U, Arora A (2001) Managing change on the Web. In: Proceedings of the ACM/IEEE joint conference on digital libraries, Roanoke, VA, 24–28 June 2001, pp 67–76
- Francisco-Revilla L, Shipman F, Furuta R, Karadkar U, Arora A (2001) Perception of content, structure, and presentation changes in Web-based hypertext. In: Proceedings of Hypertext 2001, Aarhus, Denmark, 14–18 August 2001, pp 205–214
- Furuta R, Shipman F, Marshall C, Brenner D, Hsieh H (1997) Hypertext Paths and the World Wide Web: experiences with Walden's Paths. In: Proceedings of Hypertext '97: the 8th ACM conference on hypertext, Southampton, UK, 6–11 April 1997, pp 167–176
- Furuta R, Shipman F, Francisco-Revilla L, Hsieh H, Karadkar U, Hu S (1999) Ephemeral paths on the WWW: The Walden's Paths lightweight path mechanism. In: Proceedings of WebNet 99 – world conference on the WWW and Internet, Honolulu, 24–30 October 1999, pp 409–414
- Goodwin C, Graham M, Scarborough H (2001) Developing an asynchronous learning network. *J Int Forum Educat Technol Soc IEEE Learn Technol Task Force*, 4(4):39–47. [http://ifets.massey.ac.nz/periodical/vol\\_4\\_2001/scarbo-rough.html](http://ifets.massey.ac.nz/periodical/vol_4_2001/scarbo-rough.html)
- Halasz F, Moran T, Trigg R (1987) Notecards in a nutshell. In: Proceedings of the ACM CHI+GI conference 1987, Toronto, 5–9 April 1987, pp 45–52
- Johnson D (1997) Enabling the Reuse of World Wide Web documents in tutorials. PhD Dissertation, Department of Computer Science and Engineering, University of Washington, Seattle
- Karadkar U, Francisco-Revilla L, Furuta R, Hsieh H, Shipman F (2000) Evolution of the Walden's Paths authoring tools. In: Proceedings of WebNet 2000, San Antonio, TX, 30 October–4 November 2000, pp 299–304
- Levy D, Marshall C (1995) Going digital: a look at assumptions underlying digital libraries. *Commun ACM* 38(4):77–84
- Marshall CC (1996) Observations of sixth graders exploring the World-Wide Web. <http://www.csdl.tamu.edu/~marshall/caeti/observations.html>
- Marshall CC, Shipman F, Coombs JH (1994) VIKI: Spatial hypertext supporting emergent structure. In: Proceedings of the European conference on hypermedia technologies, Edinburgh, UK, 18–23 September 1994, pp 13–23
- Martinez M, Bunderson VC (2000) Foundations for personalized web learning environments. *J Asynchron Learn Netw* 4(2). <http://www.aln.org/publications/magazine/v4n2/burdenson.asp>
- Nichols M (2002) Principles of best practice for 21st century education. *J Int Forum Educat Technol Soc IEEE Learn Technol Task Force* 5(2):7–9. [http://ifets.ieee.org/periodical/vol\\_2\\_2002/discuss\\_summary\\_april2002.html](http://ifets.ieee.org/periodical/vol_2_2002/discuss_summary_april2002.html)
- Shipman F, Furuta R, Brenner D, Chung C, Hsieh H (1998) Using Paths in the classroom: experiences and adaptations. In: Proceedings of Hypertext '98, Pittsburgh, 20–24 June 1998, pp 267–276
- Shipman FM, Furuta R, Marshall CC (1997) Generating Web-based presentations in spatial hypertext. In: Proceedings of the 1997 international conference on intelligent user interfaces, Orlando, FL, 6–9 January 1997, pp 71–78
- Trigg R (1988) Guided Tours and Tabletops: tools for communicating in a hypertext environment. *ACM Trans Office Inf Sys* 6(4):398–414
- Wiederhold G (1995) Digital libraries, value, and productivity. *Commun ACM* 38(4):85–96
- W3C (The World Wide Web Consortium) (1999) Hypertext Transfer Protocol – HTTP/1.1. <http://www.w3.org/Protocols/rfc2616/rfc2616.html>
- Zellweger P (1988) Active paths through multimedia documents. In: van Vliet JC (ed) Proceedings of the international conference on electronic publishing, document manipulation, and typography, Nice, France, 20–22 April 1988. Cambridge University Press, Cambridge, pp 19–34
- Zellweger P (1989) Scripted documents: a hypertext path mechanism. In: Proceedings of Hypertext '89, Pittsburgh, 5–9 November 1989, pp 1–26