

# Structuring Access to a Dynamic Collection of Digital Documents: The Walden's Paths Virtual Directories

Unmil P. Karadkar, Luis Francisco-Revilla, Richard Furuta, Frank M. Shipman III

Center for the Study of Digital Libraries and Dept. of Computer Science  
Texas A&M University  
College Station, TX 77843-3112  
{unmil, l0f0954, furuta, shipman}@csdl.tamu.edu

## Abstract

The Walden's Paths project facilitates incorporation of Web-based documents into the K-12 classroom environment. Currently Walden's Paths uses a static presentation mechanism, based on the authors of the paths, to display the available paths to the readers. As the number of authors and readers increases, it becomes increasingly difficult for readers to find the paths from a list that is based solely on the authors. The most suitable organization of paths varies with the task at hand and the reader's environment.

The use of virtual directories has been proposed for managing dynamic collections of digital documents. These directories are not physically present in the file system; only a user query is stored. At access time, the database of files is queried to select files of interest and these are included in the directory. This mechanism allows readers to organize paths according to their needs while maintaining a hierarchy and preserving the context. This paper proposes that inclusion of the virtual directory mechanism in Walden's Paths will enable the readers to organize data to better suit their conceptual model and presents a working prototype of this feature.

## 1 Introduction

Traditional file systems provide access to and help organize collections of documents. The structure of most file systems looks identical to all those who access it. These file systems do not take into account the difference in users' conceptual models and needs. Users can access documents in these file systems by referring to them only by their location in the document hierarchy. Most file systems are created and managed by trained system administrators, who have a good sense of user expectations from the file system and tailor their systems to obtain optimal performance and to satisfy user requirements. These file systems have well formed document hierarchies that are either intuitive to users in a certain domain or can be remembered, due to frequent use over time. However, these file systems do not scale well from the user perspective. While small collections of documents are manageable, users find it difficult to remember long paths to documents that are stored in intricate hierarchies [5].

---

This material is based upon work supported by the National Science Foundation under Grant Nos. IIS-9812040 and DUE-0085798.

This scenario seems to be changing with the emergence of the World-Wide Web (henceforth referred to as the Web or WWW), which may be viewed as a distributed hierarchical file system. While individual Web sites still follow the traditional model of file systems, the Web, as a whole, has no discernable structure that is either intuitive or documented. Also, it is unlikely that any two Web sites have identical structure. Hence, while most Web sites are managed by knowledgeable system administrators, a casual reader can easily get the impression of chaos as the lessons learnt on any Web site do not apply to another. It is not practical to expect readers to know or learn the entire document hierarchy of the Web due to the sheer volume of data and the number of Web sites involved. Typical users tend to browse many Web sites, but visit only a few of them frequently. Users perceive the Web sites they visit often to be organized, while the ones they seldom visit to be idiosyncratic. The Web also differs from the traditional file systems as it imposes a clear distinction between the authors and readers of documents. While authors of files in traditional file systems can grant modification privileges to readers, the Web does not permit modification of documents by readers. However, some Web services do allow readers to personalize and tailor the presentation to their liking and mental model.

Mostly, users organize their files based on their functional requirements. However, the organization varies between users, and for a given user, between tasks. Users tend to reorganize their files based on the current task at hand. File systems that provide a single view do little to help users reorganize files dynamically, without changing the physical hierarchy of the file system. Removing hierarchy from file systems is not a solution as hierarchy provides a context to the collection of documents. Search results returned by commercial Web search engines is an example of a non-hierarchical collection of documents. These collections are built dynamically but they lack context; an essential classifier for documents.

Users often share documents by passing document locations between them directly or indirectly. Sharing a document by passing its location via e-mail is an example of direct communication, where both readers must know the document location in order to access it. A reader may also create links to the documents of interest and allow other readers to access them via the link. The readers may not know the actual document location in this case, resulting in an indirect access to the document. Readers also share documents by creating copies in their personal space and allowing other readers to access them. With respect to the original document, this is another method of indirect sharing. The documents may be shared privately or publicly. Documents can be shared privately via email to a clique or via direct contact. Documents may be shared publicly by publishing pointers at well-known locations for public viewing.

The issues involved in presentation and sharing of documents in a file system are further compounded when the documents and the file system change often. Systems that provide access to dynamic documents must address the issues mentioned above in addition to reflecting the changes in the document collection as they happen. These systems should display only the documents that are available at the time of access.

Thus, there is a need for systems that provide context-based display and sharing of documents without requiring the users to modify the physical file hierarchy. This helps users build contexts dynamically on a collection of documents. Thus various users may model a given document collection differently for various purposes to suit their individual needs and preferences. In this paper we present the Walden's Paths virtual directories, a mechanism to address some of the issues outlined earlier, in the context of a Web-based collection of meta-documents, for use in K-12 schools.

We present a survey of the earlier research regarding virtual directories in section 2. Sections 3 and 4 provide a brief overview of the Walden's Paths path server. Section 5 describes the features provided by Walden's Paths virtual directories. Section 6 is a discussion of the features and comparison with a leading Web portal. Section 7 outlines the directions for the future and section 8 concludes the paper.

## **2 Virtual Directories**

Managing and using large file systems is a daunting task. Users find it difficult to remember exact path names and may often confuse similar path names. Over the years, attempts have been made to provide content-based access to file systems in order to help users find the right information in a short time without having to remember elaborate file locations.

Semantic File System (SFS) [4] was one of the early efforts in this direction. SFS provides access to file systems through queries. Each query is represented as a *virtual directory*, which points to a set of files that satisfy the query. The system provides associative access to data by extracting attributes from files via the use of file type-specific transducers. Transducers are filters that accept files as input and return the files entities and their corresponding attribute values. A more recent implementation of a semantic access system provides simultaneous access using the hierarchical file-structure as well as a content-based access. This system is called HAC (Hierarchy And Content) [5]. HAC treats queries as files or directories and calls these *semantic directories*. While SFS attempts to treat queries as files, HAC extends a hierarchical file system to support queries. The Essence system [7] associates wrapper objects with ordinary files in order to support a unified data extraction mechanism over multiple file types. The Essence system also incorporates an understanding of Unix file semantics and context to generate representative summaries of large amounts of data to aid faster resource discovery in large data collections.

## **3 Walden's Paths**

Vannevar Bush, in 1945, proposed hypertext paths as a means to associate two items that were conceptually related, but placed physically apart in an information web [1]. He referred to them as trails. Paths have subsequently been incorporated into many hypertext systems, most notably in NoteCards [6] as Guided Tours and TableTops [10] and in Scripted Paths [12].

Walden's Paths is a Web-based implementation of the hypertext path mechanism. Paths lay a meta-structure over the Web. The paths refer to and elucidate documents on the Web and hence can be termed as meta-documents. The Walden's Paths project aims at supporting the incorporation of Web-based documents into K-12 classrooms in order to help teachers achieve their curricular goals. Most of the material available on the WWW is not oriented towards young audience and is tailored to suit casual readers. Hence, teachers who author the paths may add HTML annotations to the paths and to individual pages on the paths to contextualize the information for easier comprehension by the students. A detailed description of various issues and experiences with Walden's Paths may be found in [2, 8 and 9].

Readers can access the paths from any standard Web browser that supports frames. Typically, readers begin a Walden's Paths session by selecting a path from the list of paths (shown in figure 1) that is displayed at startup. Figure 2 illustrates the Walden's Paths interface when viewing a path. The browser window contains three frames. The bottom frame is called the "Content Frame". It displays the Web-based document that the page refers to, as it would appear if viewed in the browser without the Walden's Paths interface. The top-left frame, called the "Control Frame", displays widgets for navigating along the path. The reader may navigate along the path by clicking on the left right arrows to move along the path. The reader may also click on any of the numbered images to view that page on the path. Clicking on the image labeled as "Walden's Paths" takes the reader back to the start page. This frame displays the Web-

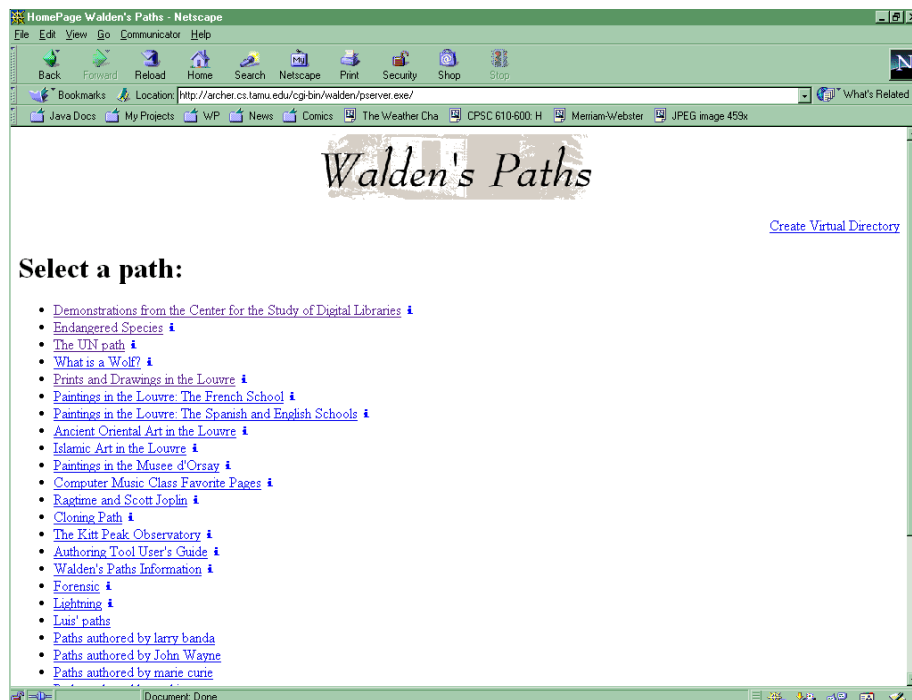


Figure 1: The Walden's Paths Start Page

location of the document displayed in the bottom frame. Selecting the image labeled with the letter "i" brings up an overview of the path via an information icon denoted. The top-right frame, called the "Annotation Frame", contains the additional information added by the creator of the path for the page displayed below. The annotation may contextualize, explain or analyze the information in the Content Frame. It may direct users to focus on certain issues or aspects of the information. It may point the reader towards related interesting issues or pose questions with respect to the document displayed below [9].

Along with the annotations and WWW document references, the paths also store metadata associated with the path. This information includes the name and contact information of the authors, a brief abstract and the dates of creation and expiration for the path.

#### 4 Ephemeral Paths and Author Directories

The Walden's Paths system permits readers to customize the authored paths [3]. Readers can create a new path by selecting a subset of pages from an existing path. The paths thus created are stored in a special directory and are accessible only via a handle that the system provides when creating them. As these paths are expected to have a short life span, they are termed as "ephemeral" paths.

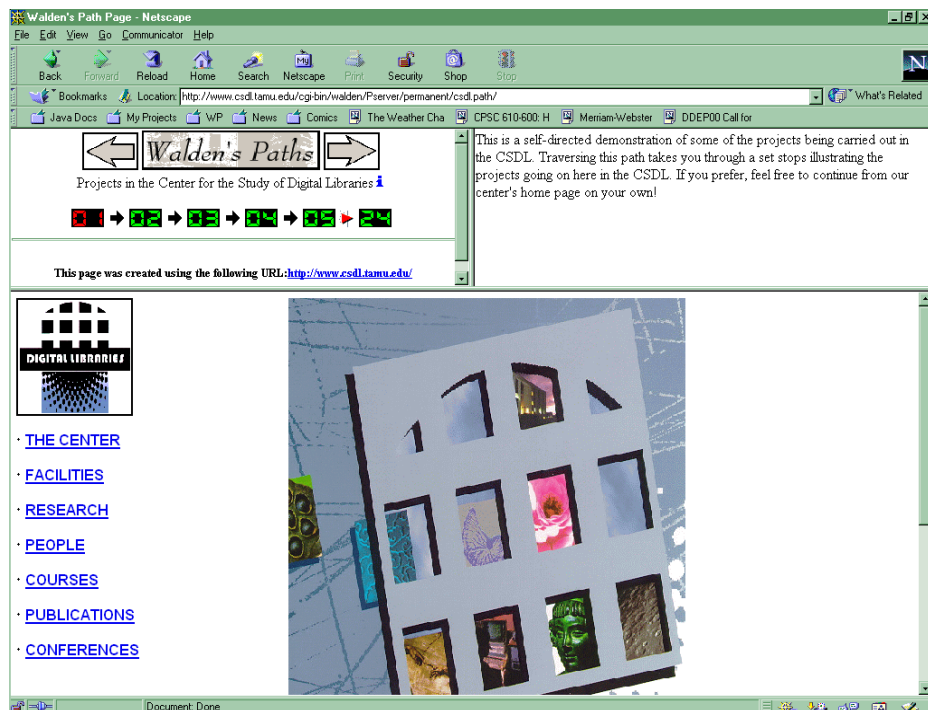


Figure 2: The Walden's Paths Interface

The Walden's Paths system allows registered authors to compile and publish paths. The published paths are accessible to all readers of the system. These paths are saved in the authors' respective directories. An author may choose to provide a link to his home directory to help users easily locate his paths. The main list of paths displays links to paths as well as to author directories. In figure 1, the links that lead to author directories are indicated visually by the absence of the information icon, as well as verbally. The paths in these directories are displayed in an interface identical to figure 1 when a reader selects to view them.

While creation of ephemeral paths allows readers to customize the individual paths, the author directories impose a hierarchical structure on the paths in the system. Creation of virtual directories allows readers to create customized directories of paths based on reader-specified attribute values.

### 5 Virtual Directories in Walden's Paths

As the number of published paths increases, organization of paths based solely on their authors may be too constraining to be meaningful to the readers. Also, readers may like to organize the available paths along various dimensions to best suit their goals. It is possible that teachers from two or more schools may share a path server to reuse resources and/or to split the costs. In such cases, the paths authored by teachers in one school may be available to teachers in the other schools if the authors choose to do so. The domains are less clearly defined and may easily overlap across

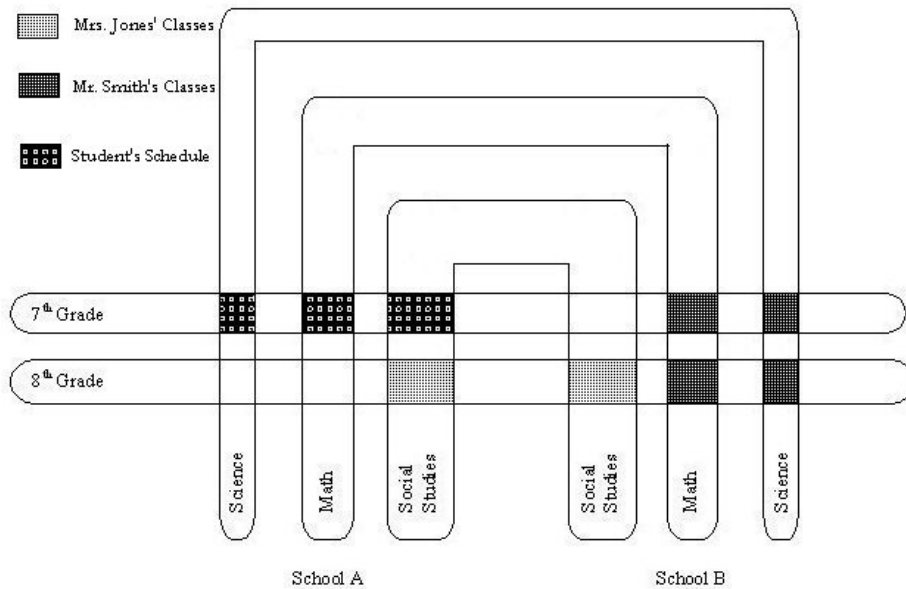



Figure 3: Overlapping Domains

administrative boundaries. Some sample domains are shown in figure 3. Mrs. Jones travels between schools A and B and teaches Social Studies to the 7<sup>th</sup> grade in both schools. A search on the paths authored by her yields paths from both the schools. Mr. Smith teaches Math and Science to students in the 7<sup>th</sup> and 8<sup>th</sup> grades in school B. A search for paths authored by him returns paths for both the grades and subjects that he teaches. A student schedule for a 7<sup>th</sup> grader in school A is also shown. In the scenario illustrated by figure 3, no fixed hierarchy of domains (and directories corresponding to these) will yield an arrangement of paths that is suitable for all the students or teachers. It is in this case of overlapping domains and unclear boundaries, that the virtual directories can be exploited to their full potential. The teachers as well as the students can create virtual directories over the physical path structure and easily access paths of their respective interests. Also, the teachers may create virtual



[Walden's Paths Main Page](#)

**Create a New Virtual Directory**

**Basic Search**

Type in the search terms:

Create a new virtual directory  
 Search existing virtual directories  
 Include a link to the main Directory

**Advanced Search**

Type the words to search for in each of the fields:

Keywords:

Topic/Subject:

Path Title:

Path Author(s):

Grades

Kindergarten  
 1    4    7    10  
 2    5    8    11  
 3    6    9    12

Create a new virtual directory  
 Search existing virtual directories  
 Include a link to the main Directory

This prototype path server is provided by Texas A&M University's Center for the Study of Digital Libraries (CSDL) and its continuing development is funded by the National Science Foundation.

Figure 4: Virtual Directory Creation Interface

directories and provide these to the students to protect them from a deluge of mostly irrelevant paths.

Possible search attributes for paths could be the grade levels, keywords, subjects, authors and date of creation. To support searches on these attributes, the existing path structure was augmented to include additional metadata, for example the grade levels that the path is suitable for, and lists of keywords associated with each of the pages, as well as with the path as a whole are also stored. The lists of keywords contain related words that do not appear in the annotations or anywhere else in the path.

The creation of virtual directories allows readers to view lists of paths that match certain criteria. It enables readers to search paths that have specific attribute values and view them as one list. These information structures are analogous to directories that contain related files in the traditional file systems. However, these structures do not exist on the server physically as files or directories, but are dynamically generated from their criteria every time they are accessed. We call these structures as virtual directories. The virtual directories are physically stored as the parameters provided by the readers of paths as attribute-value pairs in plain text files. A virtual directory returns a URL to the reader upon its creation. This URL acts as a handle for future

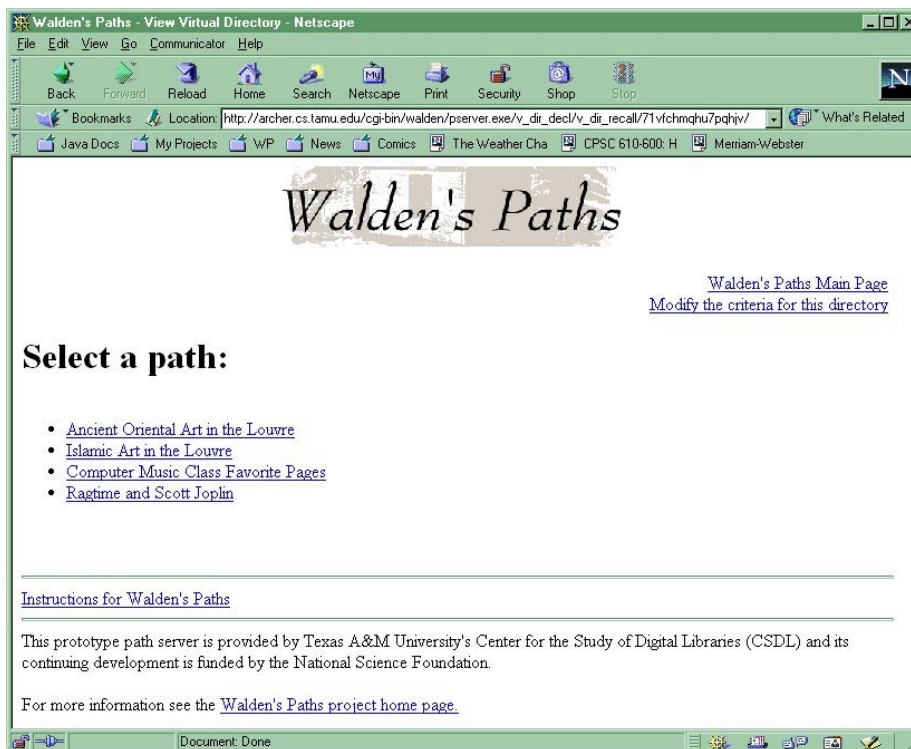


Figure 5: View of a Virtual Directory

access to the directory. Readers can modify virtual directories by changing the search criteria associated with them. The readers may save the modified criteria as a new directory, or in the current directory, overwriting the existing criteria. By their very nature, the virtual directories display only the paths available in the system at the time of access. When the reader accesses a virtual directory, the system searches for paths that match the criteria for the virtual directory and display these paths to the reader. Visually, they have an interface identical to the one shown in figure 1.

The virtual directory creation interface is shown in figure 4. The interface supports two search modes, basic and advanced. In the basic mode, the reader only needs to type in the list of keywords to search on. These keywords are matched against all attributes in the paths and a directory of all the paths that contain the keyword is generated and displayed. In the advanced mode, the reader has a finer control on the keywords and can specify the path attributes where these keywords must appear. In this case, the returned directory lists paths that contain the keywords only in the specified attributes. The interface also allows readers to create a link back to the main page of the Walden's Paths server (displayed in figure 1).

Figure 5 displays a virtual directory of all paths that contain either of the terms "Music" or "Art". It also contains a link to the Walden's Paths server main page. The access mechanism for paths in virtual directories is identical to that for the paths that are accessible from the main page of the Walden's Paths server. When displaying a virtual directory, the location bar contains the URL that acts as an access handle for

*Walden's Paths*

[Walden's Paths Main Page](#)

**Modify the Criteria for a Virtual Directory**

**Basic Search**

Type in the search terms:

Create a new virtual directory  
 Search existing virtual directories

Include a link to the main Directory

Overwrite the old search  
 Save as new search

Figure 6: Virtual Directory Modification Interface

future references. A reader may bookmark this URL and return to it at will, a standard functionality provided by most Web browsers. The invocation of this URL causes the contents of this directory to be rebuilt and presents the reader with all the paths that match the specified criteria.

Readers may modify the virtual directory by following the link that is displayed in the top right corner of the browser window. Editing the search criteria used in its creation results in a different list of paths, thus modifying the virtual directory. The interface for modifying a virtual directory, displayed in figure 6, is similar to that for creation of new virtual directories. When the reader follows a link to modify the virtual directory, the current parameters of the virtual directory are displayed to assist the modification process. The modification interface allows readers to either overwrite the existing criteria for the virtual directory or create a new virtual directory with the changed criteria. This option is added only to the current search mode. Thus, for the case shown in figure 5, the option to overwrite or create a new directory is not added to the advanced search interface. Conversely, if the reader chose to modify a search that was created as an advanced search, while modifying it, this option would be added only to the advanced search and the interface for the basic search would remain as shown in figure 3.

The directory creation interface in figure 4 also allows readers to search for existing virtual directories that match the specified criteria. When the existing virtual directories are searched, no new directories are created. In this case the basic search mechanism returns a list of all existing virtual directories that were generated using any of the terms specified by the reader, while the advanced search returns a listing of all virtual directories which contain the specified search terms in the corresponding fields.

## **6 Discussion and Comparison**

Yahoo! was the first Web portal that cataloged and categorized various Web sites. The Yahoo! repository spans over a vast range of subject areas. It is still a special case as it is one of the few portals that catalog all Web sites included in the repository manually [11]. Yahoo! provides search-based as well as navigable interfaces to the repository. The navigable interface is analogous to the hierarchical file systems, while the search interface is analogous to the virtual directory mechanism in Walden's Paths.

Yahoo! returns category matches and site matches in response to user queries. If we consider the list of returned results as a virtual directory, the category matches in these results can be considered to be subdirectories. Against this, the Walden's Paths virtual directories currently only support only single level structures. All the paths that match a reader-specified query are treated at par and returned as a list of paths. We do not make any attempt to further categorize the paths that match the search criteria.

The reuse of queries to recall an old search returns new results, based on the current contents of the repository. This is definitely an advantage while using the Walden's

Paths virtual directories. Readers can recall old criteria to retrieve the current contents of the directories, thus eliminating the need to remember successful searches. Mostly, users are unable to reconstruct the exact queries that yielded the best results and hence are unable to retrieve any information that was found earlier. With the use of virtual directories, path readers can be certain that all the results that matched their query will be returned, as long as they exist (the author has not removed the path from the repository) and are relevant (the author has not modified the path).

Cataloging of data for personal use only need fit the mental model of a user and can easily be perceived by others as idiosyncratic. However, when data must be cataloged for general usage by millions of people, there needs to be a standardized categorization of concepts and a set of general decisions must be followed. However, even when rules are followed, some data fits into more than one slot in a hierarchy. Yahoo! implements this via the use of virtual links in the category hierarchy. Thus, the Yahoo! hierarchy is more controllable, and tailorable to various scenarios than most machine-generated hierarchies that depend on keywords for classification. Walden's Paths currently has no centralized scheme for categorization of paths. The onus for providing the keywords and subjects for the paths is entirely on the authors. In the absence of clear guidelines, or a designated person to catalog paths, there arises a possibility of conflict between the thoughts of different authors. This can result in the usage of different terms, leading to confusion regarding location of paths in a hierarchy.

## **7 Future Work**

The current implementation of virtual directories is a satisfactory baseline. It must be further enhanced to provide a good utility value to the path readers and to provide better control to the teachers, path authors and path readers.

Currently, only registered authors can add paths to the Walden's Paths system. However, there is no control over the number of readers. Thus the number of virtual directories is expected to rise rapidly. The system must provide the creators of the virtual directories more control in order to protect their directories from access and from modification. Thus a reader can specify whether the directory should be returned as the result of a search over virtual directories. If the creator decides that the directory may not be listed in the results of searches over directories, it is effectively a private directory that can only be accessed via the handle returned upon its creation. The creator may allow others to view the directory but prevent anyone from modifying it. Thus, the directory created is immutable or, in more colloquial terms, a read-only directory. It must be noted that the creator cannot modify an immutable directory as the system is unaware of his identity. The Walden's Paths system must permit readers to create temporary directories. The readers may opt not to save a directory when it is created. This will enable readers to perform short-term searches and will help reduce load on the system. The directories that are saved, may also be subjected to certain timeouts, the limits of which can be set by the creator. Thus, directories can remain valid till certain dates or forever. This will further help clean up directories that are unlikely to be used in the long run.

The paths need to be cataloged using a well-defined scheme. The path authors may still provide keywords and subjects for their paths. However, these will be treated as suggestions and be subject to review by designated path catalogers to ensure consistency.

The search interface must also allow authors to control their searches better. Currently the search returns all paths that match at least one of the search terms, that is, they perform a disjunctive search. Readers of the paths will require a greater control over defining their searches. The system must handle conjunctive searches, as also phrase and boolean searches. The interface must permit the readers to search based on the dates of creation of the paths. As an example, this feature will enable readers to search for all paths created since their last visit

Readers also need a facility to exclude certain paths from a search. This will help readers achieve better precision when the search is performed in the future. Realization of this feature requires that each path in the system have a unique identifier. The readers may then choose the paths that they would like to exclude from the directory when the search is performed in the future. This also highlights the need to be able to include discarded paths back into the search.

The URLs returned by the search have an identical caption. This could result in confusion if readers save multiple searches (as they probably will). A facility to clearly identify each directory must be provided. This feature may require additional work on the readers' part, but it must be performed in order to remove any ambiguities that may result from saving multiple directories in the bookmark list.

In the future, the readers may also be required to log in if they wish to have their searches stored on the server, instead of on their desktops. Some readers may be reluctant to log in as it involves an additional step in viewing the paths. Currently users are dependent on their desktops or browser installation to access their virtual directories where the handles are stored. This feature will permit readers to access their virtual directories even when they do not have access to their desktops. The login feature will also enable the system to store and recall user preferences. This feature will be considerably useful in the educational setting, where exams may be presented in the form of a path and the users are expected to take them online.

## **8 Conclusion**

In this paper we have presented the Walden's Paths virtual directories. Readers no longer have to adapt to the single view of paths provided by the Walden's Paths path server. Virtual directories permit readers to reorganize the paths according to preferences, to suit their individual needs and mental models without requiring modification of the underlying file structure. This feature enables users to search for and mark the paths of their interest from a large repository of paths. Readers can dynamically build contexts over the available paths and be sure that future accesses to these contexts will always return all the paths that they match. The virtual directories

thus decouple the storage structure of the paths from the logical structures constructed by the readers. They also address privacy issues by allowing readers to retain control over sharing their contexts and authors over their paths. While the current implementation is a good baseline, it needs to be refined and enriched with more functionality before its true potential can be exploited.

## References

1. Bush, V., "As We May Think", *The Atlantic Monthly*, August 1945, pp. 101-108.
2. Furuta, R., Shipman, F., Marshall, C., Brenner, D. and Hsieh, H., "Hypertext Paths and the World-Wide Web: Experiences with Walden's Paths", *Hypertext '97: the Eighth ACM Conference on Hypertext*, Southampton, U.K., April 1997, pp. 167-176.
3. Furuta, R., Shipman, F., Francisco-Revilla, L., Hsieh, H., Karadkar, U. and Hu, S., "Ephemeral Paths on the WWW: The Walden's Paths Lightweight Path Mechanism", *WebNet (1) 1999*, pp. 409-414.
4. Gifford, D., Jouvelot, P., Sheldon, P. and O'Toole, J., "Semantic File Systems", *Proceedings of the thirteenth ACM Symposium on Operating Systems Principles*, Pacific Grove, CA, October 1991, pp. 16 - 25.
5. Gopal, B. and Manber, U., "Integrating Content-based Access Mechanisms with Hierarchical File Systems", *Proceedings of the third symposium on Operating systems design and implementation*, New Orleans, LA, February 22-25, 1999, pp. 265-278.
6. Halasz, F., Moran, T. and Trigg, R., "Notecards in a Nutshell", *Proceedings of the ACM CHI+GI Conference 1987*, Toronto, Ontario, April 1987, pp. 45-52.
7. Hardy, D., and Schwartz, M., "Customized Information Extraction as a Basis for Resource Discovery", *ACM Transactions on Computer Systems*, 14(2), May 1996, pp. 171 - 199.
8. Shipman, F., Marshall, C., Furuta, R., Brenner, D., Hsieh, H. and Kumar, V., "Creating Educational Guided Paths over the World-Wide Web", *Educational Telecommunications, 1996: Proceedings of ED-TELECOM 96*, June 1996, pp. 326-331.
9. Shipman, F., Furuta, R., Brenner, D., Chung, C. and Hsieh, H., "Using Paths in the Classroom: Experiences and Adaptations", *Proceedings of Hypertext '98*, ACM, Pittsburgh, PA, June 1998, pp. 267-276.
10. Trigg, R., "Guided Tours and Tabletops: Tools for Communicating in a Hypertext Environment", *ACM Transactions on Office Information Systems*, 6(4), October 1988, pp. 398-414.
11. Yahoo! - Frequently Asked Questions, <http://docs.yahoo.com/info/faq/faq.html> (accessed April 2000).
12. Zellweger, P., "Scripted Documents: A Hypertext Path Mechanism", *Proceedings of Hypertext '89*, ACM, New York, November 1989, pp. 1-26.