organizing digital documents

chapter ten

In this chapter we examine techniques for organizing files to enhance the user's ability to find and retrieve information—almost instantaneously!

The documents themselves are electronic files, and they physically reside on some storage device, whether it be magnetic disk, CD-ROM, optical media or some other form. As files on storage systems, all these documents have file names, such as FILENAME.TXT. For any collection of documents, the limitations of such short file names, or even long file names such as those permitted under UNIX, Mac and Win95, soon become apparent.

To achieve instant access, there is a critical need for a better way to organize files than using simple directory listings of file names.

The ability and tendency of HTML documents to provide dynamic links to other documents is the key difference between Web-centric documents and other document types. As stated earlier, the breakthrough of HTTP links between all of the servers on the Web offers unlimited access to data and knowledge resources. This universal linkage of resource defines the philosophy of Web-centric documents.

This limitless branching, which led to the description of a "Web" originally, presents the contents of libraries as an ongoing succession of lists of hot links. A "normal" Web page looks like a pyramid, or river delta, of previously organized folders of files. The branches of the streams, or lists of files, are the points where hypertext links are connecting separate documents or separate parts in one document.

The most important understanding of simple HTML linked collections is that they are the products and offerings of a specific author or group. For every link in an HTML document produced by traditional processes, some individual author or editor has made a decision to link two points on the Web. Even in the case of automatically bookmarked files, at some point a decision on the orderliness of the documents has been made. It is for this reason that text search is so important for collections of electronic documents.

Such singular decision presupposes a singular path and may one day leave a file or a collection of files out on an unconnected island. This dependence upon a single author or publisher is one of the weak links of the Web.

The long-term utility of individual collections can be assured through offline backup of the files. Future researchers will be able to use data-mining software to sift through large collections and retrieve valuable information. But the intent of this book is to

create digital collections that have orderliness built in and offer far greater access to the contents than has ever before been possible.

The goal of the Web, whether on Internets or Intranets, is instantly accessible information. On the communications level, this fantastic goal has been achieved. On the user level, the great advantage of HTML is that the linking of files and great collections can be organized through anchors and links.

tip

On the file level, it is very important to consider the contents of the file and what resources will be required on the user's end. Users coming into the Web via modem can't easily handle gigantic files.

Using HTML Documents To Organize Files

The definitive source of information for the structure of HTML is the IETF HTML 2.0 specification. The URL (Universal Resource Locator) of this document is: http://www.w3.org/hypertext/WWW/MarkUp/MarkUp.html. Any user on the Web can select this URL in his browser and view the documentation.

In an HTML document, the name can be marked up as the anchor, and the corresponding link will be the URL. Any user with a Web browser can now retrieve the document by clicking on the marked-up text.

For example, the first paragraph of this section could be modified with one such anchor and link to accomplish this click navigation:

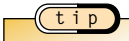
> "The definitive source of information for the structure of HTML is the *IETF HTML 2.0 specification.*"

The rest of the paragraph is assumed by the user because of the convention of *underlining and highlighting* text to represent a hyperlink. So, a familiar Web user immediately recognizes the "hot text" of the link as a pointer.

When the HTML source, which is the ASCII characters that make up the HTML document, is viewed, the linked statement appears as:

> "The definitive source of information for the structure of HTML is the < A HREF = "http://www.w3.org/hypertext/WWW/MarkUp/MarkUp.html" > IETF HTML 2.0 specification." < /A >

Most HTML editing packages provide a graphical user interface (GUI) that allows statements such as the above to be created with word-processor-like simplicity. For example, selecting a portion of text to serve as the anchor is as easy as selecting text to add bold or underlined attributes. The user is then prompted to enter the linked file or text. For the purposes of explaining the hypertext linking, the ASCII shows the nuts and bolts.



HTTP Anchors Turn Text Into Links

Technically, just like bold or underline, the anchor code is an "attribute" of the highlighted text that serves as the clickable link in the document.

In HTML, an anchor is defined as an element that points to:

- A specific location in the current document
- Another document
- A specific location in another document

In the previous example, the separate parts of the Anchor are:

Begin anchor (A) element: < A

Hypertext REFerence: HREF =

URL of the Link: "http://www.w3.org/hypertext/WWW/MarkUp/MarkUp.html" >

Highlighted text: IETF HTML 2.0 specification."

End anchor (/A) element: < /A >

The separate parts of the above URL are:

Access method: "http:" means to use HyperText Transfer Protocol

Server name: "//www.w3.org"

Top-level directory: "/hypertext"

Subdirectories: "/WWW/MarkUp"

Linked document: "/MarkUp.html"

tip

It is important to note that in addition to http, other Internet communications functions, or schemes such as ftp (File Transfer Protocol), gopher, and even e-mail through the "mail to:" function can be used to access files with a properly configured browser. Files on a local file server use the scheme called "file" and are accessed as "File:///" in the Browser Location window.

Popular Web home pages are more like magazines than any other publication. Everything is new every week or month. The Adobe site maintains an immediately understood layout, while changing the contents, just like magazine cover pages.

Running the mouse over the varying geography of the image maps (see details on image maps later in this chapter) displays the many geysers of information under the home page.

Anchor Link To A Specific Location

The author or publisher of an HTML document can provide the reader with preordained paths through the document. For example, in a traditional book, this is the equivalent of taking the table of contents and turning it into an instant page locator.

Or, in a word-processing document, every chapter or section heading could appear in a list, and the reader would be able to instantly jump to that area by clicking on the list item. This basic theory of ordering documents is used throughout many sites.

tip

The outline feature of popular word processors can be easily converted to HTML equivalents. With a little forethought, print and HTML design easily merge.

This is a core hypertext technique, and it is successfully employed not only in Web documents, but also in many of the online Help files that are included in popular software. In fact, a short review of one or more familiar Help files will provide tips on how professional electronic publishers take advantage of this feature. Most Help files offer two types of navigation, contents and search. In this case, contents is an example of hypertext organization.

However, without any advanced word-processing or electronic publishing tools, it is easy to create HTML files that offer this feature. For example, many FAQs (Frequently Asked Questions lists) follow this format.

There is FAQ for Acrobat 3 at

```
http://www.adobe.com/Acrobat3/acrobat3.html
```

Of course, this means that the document *acrobat3.html* is in the subdirectory (or server) called *Acrobat3*, and it is on the main Adobe Web site at *www.adobe.com*.

The first page of the document consists of a list of hot links, which are arranged in about 10 categories. To make it very simple for the reader to find what he is looking for, the text of each of the anchors is a simple phrase describing the content of that section. In fact, as we see when we read the rest of the document, the anchor text represents the section headings for the rest of the document.

Integrate PDF Files

When the user clicks on a link, he is instantly taken to the chosen point in the document. Clicking on the link below would present the relevant or related text.

What do I need to do to serve PDF files a page at a time?

There are four pieces to the Acrobat-on-the-Internet picture:

- The Acrobat 3 Reader for integrated viewing of the web
- Web servers that can “byteserve” optimized PDF files a page at a time to the Acrobat Reader
- Optimized PDF files that offer many common-sense advantages over megalithic mega-files: progressive display and maximum file compression
- Web links to connect your PDF files to other content on the Web

This example of a well-prepared HTML document provides many insights about advantages to the user, the most important of which is the ability to quickly find and retrieve the information of interest. Other advantages for both the author and user include ease of construction and efficient transmission.

In Adobe Acrobat, the Bookmarks feature offers the equivalent function to this intra-document organization and provides the same quick navigation features.

Anchor Link To Other Documents

The keystone of the Web is this ability to create source links in a document that point to target links in another document, either as a whole or to a specific location. It was this hyperlinking capability combined with HTTP network communications that created the World Wide Web.

By using the same structure as described above for links between sections of the same document, a page of links can refer to an entire collection of documents. These pages can be organized in several different ways according to the type of documents and the needs of the users.

Simple Lists

The needs of digital document collections may be met by simple lists. This is particularly true in technical and scientific fields, where the information is pre-organized by the subject of the documents.

For example, a manufacturing company may have a series of documents for each of its products. No matter how many manuals are involved in the entire collection, they can be presorted by the subject. All manuals that go to Satellite X are in one list, all that pertain to Rocket Y are in another, and so on.

For example, many times the "patch" software offerings provided by vendors are presented this way on the Web. The reason for this is that users who want these files are highly motivated to find them and don't mind scrolling to do so.

These types of applications lend themselves to simple list organization, and this requires the least amount of effort to organize the files. However, significant value can often be added to the collection if these simple lists are combined with other forms of navigation.

Most professionally produced home pages of big and popular Web sites contain extensive identifying information about the site, some hints on how to use the site, and only a few links. These few links immediately take the user down one of a small number of easily recognized paths.

For example, a software vendor's page will often have the following elements, or categories, each of which offers particular information or services:

What's New!

Software

Upgrades

Support

Hot Stuff!

Another example of this is found on home pages where there may be entire discrete categories of documents. For example, on the Adobe Home Page, it is necessary to present the user with a choice of platforms, so an early branch in the list might be:

Macintosh

Windows 95 and Windows NT

Windows 3.1

OS/2

By choosing one of the early branches, the user is then directed to only the body of files relevant to his needs. This greatly reduces the number of files that the user would otherwise have to review, thereby speeding up the process of locating the desired information.

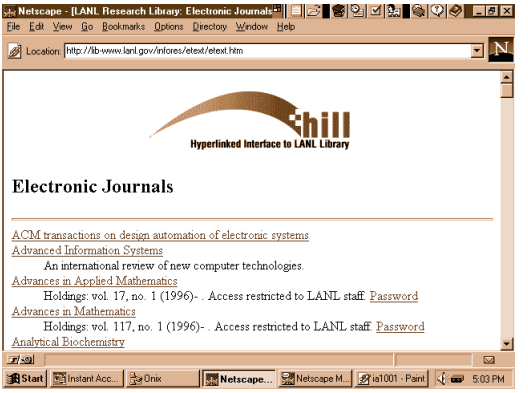
Mingled Lists

There are many applications where there are many references among the documents that contain links to the same pages. The simplest examples are the marketing pages, where many of the original branches eventually lead back to an order form.

In the above example, "What's New," "Hot Stuff," "Software" and "Support" could all ultimately refer back to the page for "Upgrades," where the user can purchase any or all of the above.

A mingled-list approach can produce an extensively cross-referenced database of HTML documents. The depth and functionality of such mingled lists are limited only by the author's creativity and labors.

The following example shows construction, but also more important, *contents*. This annotated list shows how useful short comments can be in finding interesting information. It also shows the early state-of-the-art of this list with the relatively recent dates of some of the oldest electronic text sources on the list.



Annotated lists allow rapid browsing through large collections. Note that both public and password-protected files appear.

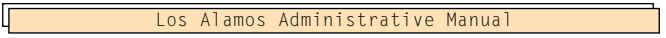
tip

When reading an online document, you can simply copy the text of a URL address and paste it into the location field on your browser. With a single carriage return, you go right to the copied link.

The Los Alamos Research Library provides an annotated list of electronic text sources at this site:

```
http://lib-www.lanl.gov/infores/etext/etext.html
```

To access the Administrative Manual, you need to have Adobe Acrobat installed on your system or local server. For more information, contact the LANL Index Project by sending e-mail to index@lanl.gov. The following are hyperlinks.



The Los Alamos Research Library contains these references:

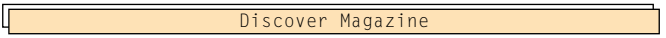
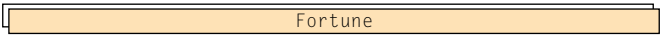


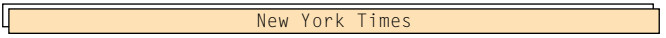
Table of contents to current and back issues with some full-text articles.



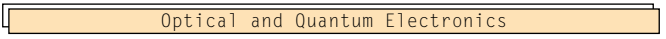
Tables of contents and some full text since Sept. 1996. (sic)

A monthly newsletter of links to new Internet resources.

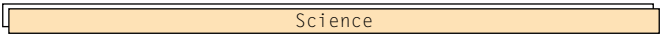
Full text of articles beginning with January 1995. Some of the articles may contain hypertext links.



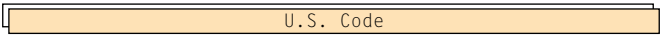
Today's New York Times. Registration is required for free access.



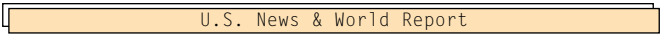
A trial subscription.



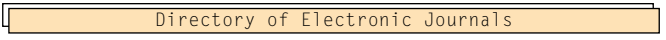
Summaries/abstracts of items in Science since October 1995, with full text of This Week in Science: Research Highlights. Contents only for June 30-September 19, 1995.



Experimental searchable version of the entire U.S. Code (with amendments through Jan. 4, 1993).



Online version with top stories, etc. Includes the annual College Fair with rankings.



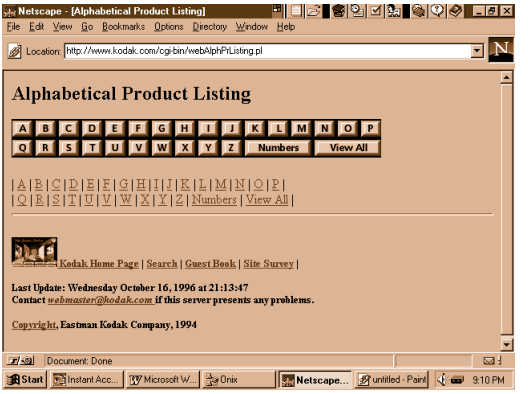
Gopher and URL addresses for electronic journals and newsletters; list is maintained by Association of Research Libraries.

“Speed Dialer” Links

To speed up access to very long lists, “speed dialer” links can be built into the document. For example, in a list organized alphabetically, users can be prompted to click on a letter of the alphabet to move directly to that section of the long list.

By providing this simple toolbar, rapid navigation is available:

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z



Eliminate scrolling through long lists with Pushbutton Dash. The same techniques can be used for other long, orderly lists such as date order lists or numerical lists.

Special Uses Of Links

Hot links can be used for many other functions to make the document more accessible to the reader. For example, a series of buttons can be used to perform the handy functions of Page Forward and Page Back. In this usage, the hot link takes the user to a specific page or to the previous or following page.

The well-rendered HTML version of “As We May Think” demonstrates this type of functionality. The buttons are unobtrusive but still allow smooth full-page turning as an alternative to scrolling vertically through the pages.

An authorized copy of this spectacular, future-predicting article from Atlantic Monthly 50 years ago is available at:

http://info.cs.vt.edu/AWMT/

This approach can also be used to allow instant access to any special pages in a document. For example, every page can offer a button to go to the table of contents, or to an index, or to any other frequently visited page or function set.

Building Organization Into New Documents

It is always a good idea to capture as much structural information from a file as possible. The simplest example is capturing the Document Information fields from a word-processing file, including everything from the author's name to the source application and system information. Ideally, as much of this baseline information as possible should be accessible via search techniques.

tip

Title, Subject, Author and many other standard fields might serve as great hyperlinks for future users, and they are easy to include.

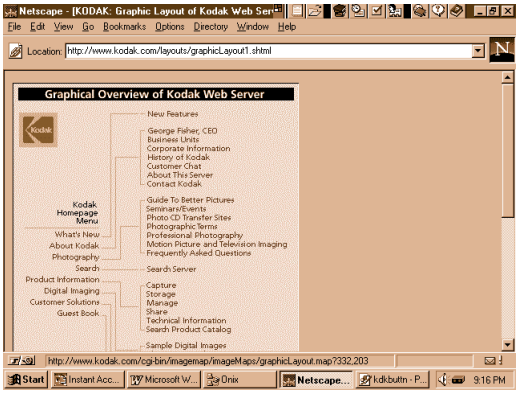
The Meta field in HTML files may be adapted to carry this type of information. In "encapsulated formats" such as PDF, much of this information can be readily captured from the source applications and fully exploited in a Web database.

Using Images to Organize Files

During the first week of May 1996, Adobe Systems introduced a new Web page that takes advantage of all the latest extensions currently available on the Web to present at once an attractive visual page and a highly interactive hyperlink menu.

Any graphics program, such as Illustrator or Photoshop, can be used to create graphically rich and appealing presentation pages. These attractive pages can then be used as image maps, which allow the user to navigate by clicking on a certain area on the page. Image maps function by watching the position of the cursor and assigning a specific URL to specific areas of the page.

In this way the most gorgeous traditional graphic pages can be used as a precisely defined grid of "pushbutton" hyperlinks in HTML. From the most mundane usage, where a schematic drawing may be hot linked to each component in the drawing, allowing a user to click on a certain point in the graphic and be instantly linked to relevant files...to the very attractive and pleasant-to-use file folder metaphors common to Windows applications ... image maps offer a very friendly navigation style. To view the HTML code for an image map, simply chose View Source Code from your Web browser.



This page offers a graphic layout to a large, complex Web site through a simple image map of a conventional organization table.

It is completely feasible to mimic the current organization and present the user with a series of ordered lists of the contents. Just as in a card catalog, the user can search by author, title and subject.

Of course, every digital collection will face a unique set of challenges, and many will only have file names to start from. Assuming that the files contain the common word-processing fields of Title, Author and Subject, all of this document-management information can be reused when moving to HTML pages.

Separate lists for each of the three categories would provide instant links to the source files. A simple automatic process could link the same source file, via its unique URL, to each of the Index fields. This allows the user to retrieve files through all of the traditional Title, Author and Subject fields.

tip

By viewing a list of files, the user can simply click to read a chosen file and then click Back to view the list again.

Convert your directory listing (DIR) to an HTML structure with hot-linked file names by copying names to a home page. Most home pages are stored under a specific server and top-level directory. This means that the base directory appears after the primary URL.

The key is to convert each directory to hyperlinks under the following convention:

```
http://www.ISP.com/HomePage/Index.html.
```

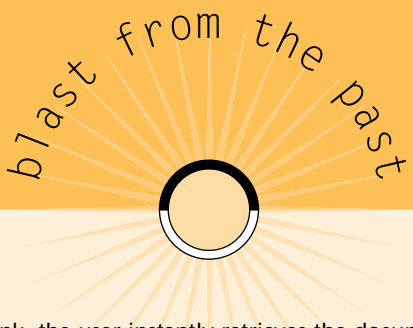
By not renaming your home page, you can shorten your URL to the first term after the home server.

Directory list		**Universal Resource Locator List**
filenam1.htm	=	http:\\www.ISP.com\HomePage\filenam1.htm
filenam2.htm	=	http:\\www.ISP.com\HomePage\filenam2.htm
filenam3.htm	=	http:\\www.ISP.com\HomePage\filenam3.htm

In the above example, filenam1.htm and so forth are directory names. By converting directory and file names to hypertext links, hierarchical sets of files can be automatically generated. For example, filenam1.htm might refer to all Acrobat files, filenam2.htm might be all Photoshop files, filenam3.htm all PageMill files and so on. The point is that entire directories can be converted to branches on hypertext trees.

Everything under each limb is linked in the same way, like the veins on a leaf, like the tributaries of a river, like neurons and axons. HTML is a software model of endless complexity, via simple branches.

The above routine converts a directory list to a series of URLs, or Universal Resource Locators. A URL connects a file on any attached server to the World Wide Web or to an Intranet Web. It's a message that never gets old: The Internet is a real network, with an infinite number of shared servers.



By clicking on a link, the user instantly retrieves the document. Compared to the traditional method of using the card catalog and then walking the aisles and searching the shelves of a physical library, the online catalog in a digital library can deliver the document directly to the user's desktop. Don't underestimate the fact that the library can actually be any spot in the world that is linked to the global Internet.

tip

Since the Web Internet is a real network, many times the hard-wired lists we are describing here are actually generated "on the fly" when the results of a directory listing are served up to the user.

All the provider has to do is create an attachment to the URL link by including the full path to the document from the Web. The Web server software takes care of all the technical details, such as the real IP address of the server as registered with the Inter-NIC and so on. Authors and publishers can build upon current network directory structures through this conversion of actual file names to Universal Resource Locators by connecting LAN and WAN systems to the Internet and Intranets.

Utilizing TOC, Index, Glossary, Appendices

The internal structure and built-in organization of many documents provide users with powerful, albeit manual, methods for the nimble handling of large quantities of information. Once again reaffirming the theme of this book, digital documents should work better than the traditional equivalents.

- Table of contents should offer elevator access → rapid level changes
- Index should offer escalator, level-by-level access → sequential changes
- Glossary, appendix should be value-added tools → universal access

Based on the principle that any hot link in a hypertext document can lead to anyplace else, whether it be a page or a library, HTML is built to offer author-published collections of files with a pre-defined series of links and logical paths through the documents.

Summary

There are a few principles of electronic document design that are just common sense:

- Make navigation tools always available; don't strand the reader in a sea of info.
- Make the reading as "thoughtless" as possible, as natural as possible, "just like a book."
- Deliver the desired information first, and presentation later.