Grass-Roots Cataloging and Classification: Food for Thought from World Wide Web Subject-Oriented Hierarchical Lists

David G. Dodd

The explosion of the use of the Internet by the general public, particularly via the World Wide Web, has given rise to an interesting phenomenon: the proliferation of semiprofessional attempts to give some subject-based access to Internet resources via hierarchical guides (hotlists) such as Yahoo. In this paper, the author examines the structure and principles of various hierarchical lists, which were examined during a period between September and October 1995. The lists are compared, when possible, to broad Library of Congress and Dewey classification schemes and to Library of Congress subject heading structures. The author also explores the approaches taken by nonlibrarians in their efforts to organize and provide access to materials on the Internet. In particular, the author focuses on the dichotomy between the hierarchical "browse" and the analytical "search" approaches to finding materials, as exemplified by these various attempts to organize the Internet.

Librarians are only beginning to make significant strides in providing access to materials on the Internet. The advance guard in this endeavor includes computer nerds, scientists, or just plain folks who want to impose some kind of order on the exponentially expanding mass of information distributed by various means over the Internet. Specifically, the introduction of the World Wide Web (hereafter the Web) has given this avant-garde a powerful tool—the use of hypertext linking—to allow for rapid movement around the Internet.

This is not to say that the profession has been inactive in this regard. Evidence that the library profession is moving to secure a role in the organization of the Internet can be found in the joint OCLC Online Computer Library Center, Inc./Department of Education project, "Building a Catalog of Internet Resource"; OCLC's work on providing Persistent Uniform Resource Locators (PURLs); the introduction of the 856 tag into the MARC record to allow for the recording of Uniform Resource Locators (URLs); and the increasing appearance of advertisements by libraries seeking to hire librarians specifically to evaluate, select, and catalog electronic resources. Taylor (1995) includes a brief section regarding subject access to the Internet, and the various projects being undertaken

DAVID G. DODD is Assistant Professor, Cataloger and Archivist at the Kraemer Family Library, University of Colorado at Colorado Springs (e-mail: ddodd@brain.uccs.edu). Manuscript received December 15, 1995; revised March 20, 1996; accepted for publication March 30, 1996.

by librarians to provide controlled vocabulary access to the chaos of the Internet.

In our rush to apply standards of librarianship to this exploding medium, I believe it would, however, be unwise to ignore the efforts of those who have gone ahead and attempted, on their own, to provide access to the medium. The hypertext link is easy to construct. The Web is graphically attractive. Perhaps that is why web users have shown so much interest in organizing this information.

Please note that the observations and conclusions drawn in this paper are based on a snapshot of a quickly moving target taken in late September to early October of 1995. I selected a small subset of available subject-oriented hotlists (Yahoo, Magellan, and the Whole Internet Catalog) and search engines (Lycos, Intercat) for examination, and scrutinized both their use of principles of categorization and the language they used. The instability and flux inherent in the Web mean that by the time this paper sees print, much will have changed and evolved. For instance, the Yahoo list has evolved significantly since October 1995: It now provides access to multiple search engines and also conducts searches of its own subject category words, a feature not available at the time of the original snapshot.

An earlier snapshot approach to the question of web access to information resources was conducted by Kambitsch (1994), who presented an anecdotal approach to evaluating various Internet search engines. These included Archie, Veronica, and the Clearinghouse for Subject Oriented Internet Resources at the University of Michigan, based on a search for a known item.

General Principles of Web Hotlists

The subject-oriented hierarchical classification system used by many web indexes represents one of the two major streams of thought on how best to provide access to resources on the Internet, and, in particular, on the Web. Hotlists of this type (called "hotlists" because they are hotlinked to the resources they list) function via a browsing mentality. A person looking for a resource works down into the hierarchies beginning at the top with the broadest category, and proceeds through the subcategories until arriving, it is hoped, at the goal—the resource that will deliver the sought-after information.

The other major trend in accessing resources on the Internet involves search engines that exhaustively scan the Web for matches on keywords. Examples of these include WebCrawler and Lycos. The most sophisticated indexes provide both search and browse capability. Later, I will provide some comparison between this type of searching and the results obtained by browsing hotlists.

Yahoo, a project begun in April 1994 by computer science Ph.D. students Jerry Yang and David Filo at Stanford University, began as a straightforward alphabetical listing of subjects. Click on a subject, and you would be sent to an alphabetical listing of resources on that subject. As the database evolved, often with the addition of as many as 800–900 new resources in a twenty-four-hour period, the hierarchies became gradually more complex. Yahoo's designers developed their current look by adhering to the general hypertext markup language (HTML) design principle of menu-driven hierarchies (Lemay 1995, 30–34). Thus, the opening screen (figure 1) contains bold pointers to fourteen broad categories, with some of the most frequently sought subcategories noted in a smaller font beneath them.

According to an interview conducted via e-mail with Srinija Srinivasan, who is in charge of Yahoo's "ontological and hierarchical" matters, there are several major factors driving Yahoo's structure (Srinivasan 1995):

1. Headings should be as concise and precise as possible. That means taking into account the context, or full path, of the heading in question so that, e.g., the same category is just called "Therapy" under Entertainment/Music/ but is called "Music Therapy" under Health/Alternative Medicine/.
2. Headings should be commonly used words and phrases that are likely to be utilized in a search query.

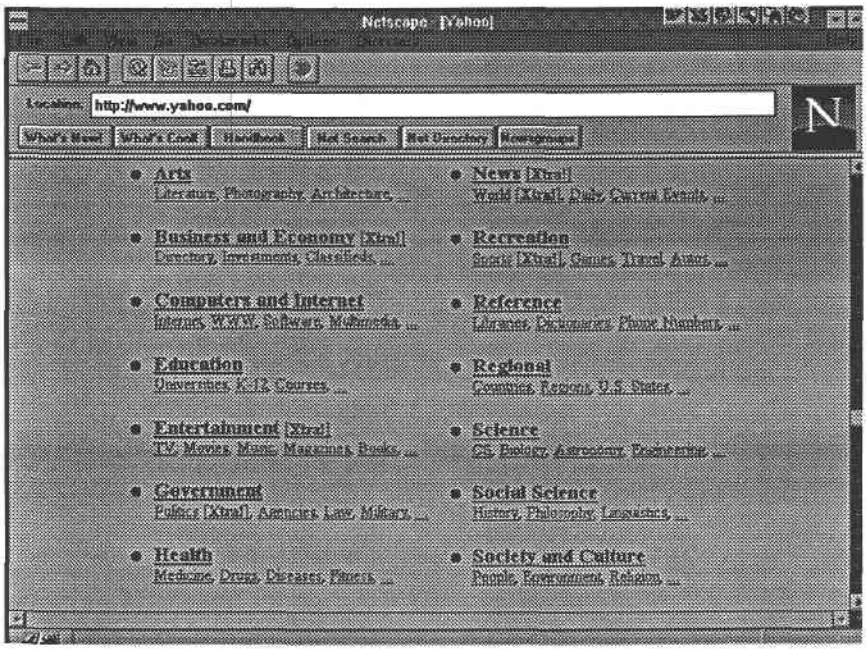

Figure 1. Yahoo's Opening Screen Categories.

3. Headings should be consistent across Yahoo so that, e.g., "Indices" means roughly the same thing in any Yahoo category, specific to that category.
4. Reference texts (dictionaries, almanacs, encyclopediae, etc.) and human specialists, when necessary, are consulted to create useful, familiar categorization schemes.

Asked at that time about future plans, Srinivasan stated that Yahoo intended to implement keyword searching on its own subject categories in the near future (which has since become operational) and hoped to use transaction-based analysis of queries to evaluate its choice of categories. She stated that they are also planning to investigate the use of thesauri to improve search results (Srinivasan 1995).

Magellan, another sophisticated site (figure 2), has fifteen broad categories. Although it lists no subcategories on its opening screen, activating the link on any of the main categories brings up a second level of subcategories. These range in number from a low of eight (under "Religion & Philosophy") to a high of twenty-nine (under "Business & Economics").

The Whole Internet Catalog (WIC) goes into greater detail, listing all of its subcategories on its opening screen, but adheres to the keynote of web organization: grouping the subtopics under thirteen broad categories.

Also worth examining are three good examples of the "labor of love" (i.e., classification systems not under the auspices of any institution or business): Hackstadt's Hierarchical Hotlist (H3), Hay's Ways, and Joel's Hierarchical Subject Index (JHSI). H3 has twenty main categories, three of which are directly related to computing. Hay's Ways has a compact graphical user interface that allows selection from a menu of ten broad subject categories. JHSI has six main categories, with thirty-two subcategories.

JHSI also includes a section on the theoretical background of hierarchical subject indexes (Jones 1995). I will quote from it at length because it indicates the sophistication of the thinking behind

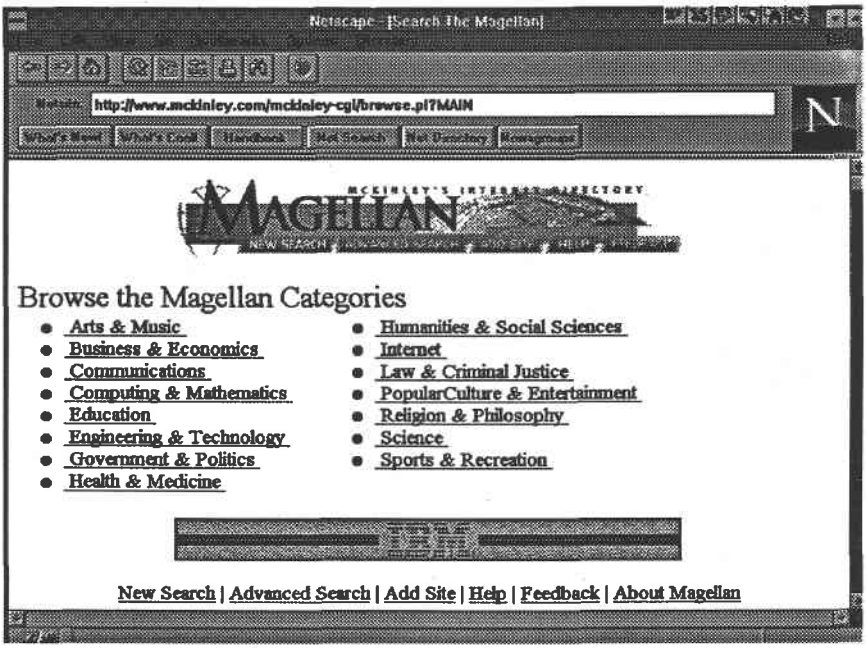

Figure 2. Magellan's Opening Screen Categories.

these hotlists. Note that the author, a graduate student in Computer Science at the University of Illinois, Urbana-Champaign, is a little fuzzy on, but by no means completely ignorant of, the principles of librarianship, and seems intent on finding the best way to provide access, regardless of prior conceptions.

Why a hierarchical index? Because there isn't a good one available yet. Existing attempts at providing an index based on a list of subject headings thus far have been very shallow, having only one or two levels.

Wouldn't another kind of index be better? For some kinds of searches, yes. There are already many different keyword and subject-heading based indexes to sources on the Internet. The problem with these indexes is that knowledge of keywords are needed by the person doing the search. In a traditional library setting, reference librarians can provide help to patrons in choosing the appropriate terminology. This is less true on the Internet.

So how does a hierarchical index solve the keyword ignorance problem? By having a hierarchical structure, browsing the subject headings becomes possible. It is assumed that the person doing the search will have some knowledge of how their search targets fits into the over wealth of human knowledge. For example, if someone is looking for a C++ compiler, they will try to find a reference to it in the areas of knowledge related to technology and then computers, rather than religion.

Why not pick Dewey or some other well-established classification system used in libraries? The problem with Dewey and Library of Congress subject headings is that they are all "a mile wide and an inch deep." Also, they don't closely match the sort of divisions that a domain expert would use, except in certain circumstances.

If some system isn't chosen, won't this lead to chaos? The indexers on [JHSI] are strongly encouraged to use pre-existing classification schemes, such as the ACM Computing Review Classification Scheme, or the Encyclopedia of Social Sciences scheme.

> *Shouldn't the index have classification codes that are recognizable by someone from library science so that experts from the library field can quickly find their way around?* This would be a great addition. However, since [JHSI] is to have a distributed management model, we have assumed that it would be easier if domain experts who are responsible for indexing a topic area chose a classification scheme that works best for their field. If this happens to match LC or UDC, fine. If someone wants to use a traditional subject heading classification to classify the sources on the Internet, they are welcome to do so.

The Correspondence of Broad Subject Terms to Traditional Classification Schemes

Although there is quite a bit of variation in the way specific topics are placed under these broad umbrellas (table 1), there is nevertheless wide consensus that people can use this hierarchical approach quite readily to narrow a search.

The column on the left in table 1 shows the main classification divisions for both Library of Congress Classification (LCC) and the Dewey Decimal Classification (DDC). The columns under each hotlist show the first-level category that contains the corresponding LCC or DDC category. Depending upon the wealth of material available, browsing via hierarchical lists may require as many as five levels of hierarchy, as in Yahoo's typical chain: Broad subject/Subcategory/Level 2 subcategory/Level 3 subcategory/topic. For example, the resources on rock musician Jerry Garcia are under "Entertainment/Music/Artists/Grateful Dead/Jerry Garcia." This arrangement sometimes leads users down a fruitless path if they guess incorrectly at the first two levels of categories. For instance, note that Magellan puts "General Reference" under "Popular Culture and Entertainment."

Terminology

A sampling of terms from each of three hierarchies was checked against the OCLC subject authority file. Fifty terms were selected from the 197 terms used by WIC; 17 were valid subject headings, 4 were valid see references, and the remaining 29—over half—were not found in the authority file. Fifty terms were selected from the 63 terms used by Yahoo; 30 were valid subject headings, 8 were valid see references, and the remaining 12 were not in the authority file. Sixty terms were selected from the 228 terms used by Magellan; 25 were valid subject headings, 12 were valid see references, and the remaining 23 were not in the authority file. Of the resources evaluated, Magellan uses an approach in its assignment of categories to indexed resources that is most similar to the *Library of Congress Subject Headings (LCSH).*

Subject vs. Form

In general, there is a tendency among the compilers of these sites to mix together words that describe the aboutness of the resources being pointed to, and words that describe the form of the resources. For example, "Dictionaries & Reference Guides" is used by WIC, as is "Economics." This is always an issue, and presents a trap into which this avant-garde seems to have fallen. It is disconcerting to see what appears to be a separate heading "Careers & Employment" or "Indices" repeating itself again and again under various major categories, until one realizes that the heading is only meant to apply to that particular category.

Both Magellan and Yahoo, in particular, use this strategy, and it makes good sense after a little while. Srinisivan's comment above makes it clear that this is a conscious decision on the part of the hierarchy's designers, corresponding to the cataloging practice of free-floating subdivisions.

Search vs. Browse

Some work has been described in the literature that attempts to define and differentiate between "browsing" and "searching" for information. The definition of

TABLE 1A

BROAD TOPIC NAMING IN SELECTED HIERARCHICAL LISTS COMPARED TO BROAD LC CLASSIFICATION OUTLINE

	WWW Hierarchical Lists		
LC Class. Term	Yahoo	WIC	Magellan
General Works	Reference	Education	Popular Culture & Entertainment
Philosophy	Social Science	Humanities & Social Sciences	Religion & Philosophy
Psychology	Social Science	H & SS	None
Religion	Society & Culture	Life & Culture	R & P
History & Social Science	Social Science	H & SS	Humanities
Geography	Science	H & SS	H & SS
Anthropology	Social Science	H & SS	None
Recreation	Recreation	Recreation, Sports, & Hobbies	Sports & Recreation
Social Sciences	Social Science	H & SS	H & SS
Political Science	Social Science	Government & Politics	Government & Politics
Law	Government	G & P	Law & Criminal Justice
Education	Education	Education	Education
Music	Entertainment also: Arts	Arts & Entertainment	Arts & Music
Fine Arts	Arts	A & E	A & M
Languages & Literature	Arts	H & SS	H & SS
Science	Science	Science & Technology	Science
Medicine	Health	Health & Medicine	Health & Medicine
Agriculture	Science	Business & Finance	Science
Technology	None	S & T	Engineering & Technology
Military Science	Government	Government	Government & Politics
Naval Science	None	None	None
Bibliography & Library Science	None	None	H & SS

"browsing" used in this paper was given in Chang and Rice (1993, 258):

> Browsing is the process of exposing oneself to a resource space by scanning its content (objects or representations) and/or structure, possibly resulting in awareness of unexpected or new content or paths in that resource space.

A series of six searches, three for known items, and three for general sub-

TABLE 1B

BROAD TOPIC NAMING IN SELECTED HIERARCHICAL LISTS COMPARED TO BROAD DEWEY CLASSIFICATION OUTLINE

	WWW Hierarchical Lists		
Dewey Topics	Yahoo	WIC	Magellan
General Knowledge	Reference	Education	Popular Culture & Entertainment
Psychology & Philosophy	Social Science	H & SS	R & P
Religion	S & C	L & C	R & P
Social Sciences	Social Science	H & SS	H & SS
Language	Social Science	H & SS	H & SS
Science	Science	S & T	E & T
Applied Science	Science	S & T	E & T
Art	Arts	A & E	A & M
Literature	Arts	H & SS	H & SS
History & Biography	Social Science	H & SS	H & SS

jects, were conducted in each of five databases: Yahoo, WIC, Magellan, Intercat (the catalog of the Building a Catalog of Internet Resources Project), and Lycos. The first three of these were subject-oriented lists. Of the three, one (WIC) had no keyword search capability. Yahoo and Magellan were searched using both the keyword search capability and the browse via subject hierarchy techniques. Intercat is based on the MARC format, and represents the closest thing to a traditional library approach to the Internet. When the search tests for this paper were conducted, the Intercat database was too small to test adequately its browse capability, although it does allow one to browse in a purely alphabetical list of words. The remaining database, Lycos, is purely a string-matching searchable index (table 2).

In table 2 it is demonstrated that for known-items, search capability is optimum. For subject-type queries where specific items are not known, searching is often not as effective as browsing, especially in Yahoo, which did not at that time allow for searching of its own categories.

Figures 4 through 7 give the entry of a search for the National Museum of American Art conducted in each database. Magellan (figure 7) and Intercat (figure 4) provide the most disciplined approach to their records. Intercat uses MARC, and Magellan assigns key words, language, publication information, and other details. Yahoo (figure 5) contains only a brief summary. WIC (figure 6) gives a fairly subjective review. And figure 8 shows the opening screen for the sought-after source itself. Here the value of descriptive cataloging becomes apparent, as the content of the Web page is well-reflected in the MARC record that describes it.

CONCLUSIONS

It is possible that we have a microcosm in the Web, under development, of two distinct patterns of human behavior vis-á-vis how we find things. Subject-oriented hierarchies are preferred by "browsers," while search-oriented indexes are for the more analytical-minded among us. Perhaps this is a left-brain vs. right-brain distinction; this may be an area for further research. One might make a case for a correspondence to subject-heading access versus shelf organization. It seems

TABLE 2

COMPARISON OF *SEARCH* AND *BROWSE* IN FOUR RESOURCES

Known-item searches

I. The Nine Planets
- A. Yahoo:
  1. *Search*: Delivered site and a mirror site.
  2. *Browse*: Science/Astronomy/Planets/Nine Planets.
- B. Magellan:
  1. *Search*: Delivered site as first record.
  2. *Browse*: Science/Astronomy/ (then prompted for additional words—added "nine" which delivered the site).
- C. WIC: *Browse only*: Science and Technology/Astronomy/Nine Planets.
- D. Lycos: *Search only*: Delivered site as #2 hit.
- E. Intercat: *Search only*: Delivered site.

II. National Museum of American Art (NMAA) (see figures 5–8):
- A. Yahoo:
  1. *Search*: Found five matches, including NMAA.
  2. *Browse*: Arts/Museums/NMAA.
- B. Magellan:
  1. *Search*: Delivered NMAA at top of list of more than sixty results found.
  2. *Browse*: Arts & Music/Art Museums: (then prompted for additional words—added "American Art" which delivered a list: NMAA at #4).
- C. WIC: *Browse only*: Art and Entertainment/Museums and Art Historical Resources/NMAA.
- D. Lycos: *Search only*: Found seventeen documents, all related to NMAA.
- E. Intercat: *Search only*: Delivered site.

III. Internet Public Library (IPL):
- A. Yahoo:
  1. *Search*: Eleven matches, including the site.
  2. *Browse*: Reference/Libraries/IPL.
- B. Magellan:
  1. *Search*: Timed out twice in succession.
  2. *Browse*: Educational/Libraries/ (then prompted for additional words—added "internet"—IPL not included in results).
- C. WIC: *Browse only*: Education/Libraries/IPL.
- D. Lycos: *Search only*: Delivered 418 documents; top five were for IPL.
- E. Intercat: *Search only*: Delivered site.

Topic searches

I. "Solar system":
- A. Yahoo:
  1. *Search*: Thirty-four matches.
  2. *Browse*: Science/Astronomy/nothing under "solar system."
- B. Magellan:
  1. *Search*: Found more than sixty records containing either or both words.
  2. *Browse*: Science/Astronomy/ (then prompted for additional words—added "solar system" which delivered twenty-nine records).
- C. WIC: *Browse only*: Science & Technology/Astronomy/ found three pertinent records.
- D. Lycos: *Search only*: Retrieved 350 documents.

II. "Impressionist painting":
- A. Yahoo:
  1. *Search*: Three matches.
  2. *Browse*: Arts/Art History/Genres/Impressionism (nothing under impressionist painting, per se).

B. Magellan:
  1. *Search*: More than sixty records found.
  2. *Browse*: Arts & Music/Art History/ (then prompted for additional words—added "impressionist" which delivered no records).

CONTINUED ON NEXT PAGE

TABLE 2 (continued)

C. WIC: *Browse only*: Arts & Entertainment/Museums and Arts Historical Resources: no further results.
D. Lycos: *Search only*: Delivered 283 documents.

III. "Quotations":
A. Yahoo:
1. *Search*: Twenty-eight matches.
2. *Browse*: Reference/Quotations.
B. Magellan:
1. *Search*: Twenty-eight matches.
2. *Browse*: Popular Culture and Entertainment/General Reference: (then prompted for additional words—added "quotations" which delivered four sites).
C. WIC: *Browse only*: Education/Dictionaries & Reference Guides/ (Bartlett's is listed)
D. Lycos: *Search only*: Delivered twenty-three documents.

that there are simply not enough ways to group things together for improved access; new methods of conglomeration will always emerge.

Some of the recent developments in library automation are reflected in, or possibly reflect, the proliferation of subject-oriented hierarchical hotlists. CARL Corporation's popular product, the Kid's Catalog, and its corollary for adults, Everybody's Catalog, certainly have elements of this browsability, as evidenced by the "explore" capability in Kid's Catalog (under development for Everybody's Catalog).

Works Cited

Chang, Shan-Ju, and Donald E. Rice. 1993. Browsing: A multidimensional framework. *Annual review of information science and technology* 28: 231–76.

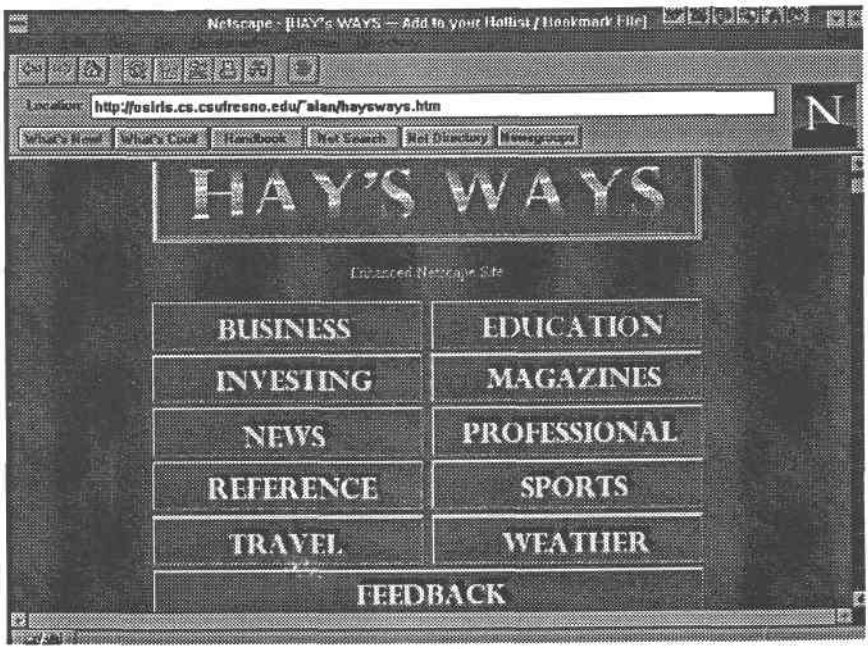

Figure 3. Hay's Ways Opening Screen.

000 nmm Ia
001 33004857
003 OCoLC
005 19950615000908.0
008 950818m19959999dcun u eng d
245 00 National Museum of American Art $h [computer file].
246 3 Smithsonian Institution National Museum of American Art
256 Computer data.
260 Washington, D.C. : $b National Museum of American Art, $c 1994-
500 Title from title screen.
505 0 Director's welcome -- General information -- Research resources -- Artworks -- Education -- What's new? -- Museum departments -- Publications -- Feedback -- Special exhibitions -- Renwick gallery -- Search.
520 "Welcome to the National Museum of American Art's new World Wide Web site. We will be adding materials daily, so come back often. In the meantime, enjoy viewing and reading about over 500 works of art; reports of upcoming and recent events; and interacting with staff and some of the artists in the collection."--Welcome screen
538 Mode of access: Internet. Address: http://www.nmaa.si.edu
610 20 National Museum of American Art (U.S.)
650 0 Art museums $z United States $z Washington, D.C. $x Databases.
856 7 $2 http $u http://www.nmaa.si.edu

Figure 4. MARC Record for the National Museum of American Art (from OCLC's Intercat Database).

Government:Agencies:Smithsonian Institution:National Museum of American Art

National Museum of American Art
Publications - Back issues of the scholarly journal, American Art; complete text excerpts of museum catalogs.

Figure 5. Text of Yahoo's Entry for the NMAA.

National Museum of American Art

The online home of the Smithsonian Institution's National Museum of American Art. The inaugural Internet exhibition is "The White House Collection of American Crafts," a rich multimedia presentation of 72 works by contemporary American craft artists. The exhibit features scores of images, videos, and sound files. I particularly enjoyed the virtual tour, which presents pictures of the works as they were exhibited in the White House. The interviews with the artists are also a nice touch. Other resources here include a gallery of GIF images of famous paintings in the museum collection ("Highlights of the Permanent Collection") and a catalog of museum publications, including excerpts and images.

Figure 6. Text of WIC's Entry for NMAA.

National Museum of American Art

Arts & Music , Art Museums

Keywords: American Art, The Smithsonian Institute, National Museum of American Art

Audience: Art Enthusiasts, Art Historians, Art Students

Description: The National Museum of American Art Web site offers a wide array of information and images. Users will find a complete history of the museum, information on events, exhibits, visiting, and more. There are also several art galleries that users can visit on the Web, with images and information on the artists. Information on traveling exhibits and membership is also available here.

Language: English

Producer: National Museum of American Art

Contact E-mail: NMAA.NMAAInfo@IC.SI.EDU

No Cost

Non Commercial

Not Moderated

Figure 7. Text of Magellan's Entry for NMAA.

15) National Museum of American Art home page [0.9053, 4 of 4 terms, adj 1.0]

Abstract: Click anywhere on the image. (Click for text mostly version.) Welcome to the National Museum of American Art's new World Wide Web site. We will be adding material daily, so come back often. In the me...
http://www.nmaa.si.edu/ (6k)

Figure 8. Text Retrieved for NMAA Search on Lycos.

Filo, David, and Jerry Yang. 1995. Yahoo. (http://www.yahoo.com).

Hackstadt, Steven. 1995. Hackstadt's hierarchical hotlist. Version 1.5. (http://www.cs.uoregon.edu/~hacks/hotlist).

Hay, Alan. 1995. Hay's ways. (osiris.cs.csufresno.edu/~alan/hayways.htm).

Intercat: A catalog of Internet resources. 1995. (http://www.oclc.org:6990).

Jones, Joel. 1995. Joel's hierarchical subject index. (http://www.cen.uiuc.edu/~jj9544/index.html).

Lemay, Laura. 1995. *Teach yourself Web publishing with HTML in a week*. Indianapolis: SAMS Publishing.

Liebscher, Peter, and Gary Marchionini. 1988. Browse and analytical search strategies in a full-text CD-ROM encyclopedia. *School library media quarterly* 16: 223–33.

Lycos. 1995 (http://www.lycos.com).

Magellan. 1995. (http://www.mckinley.com).

Noerr, Peter L., and Kathleen T. Bivins Noerr. 1985. Browse and navigate: An advance in database access methods. *Information*

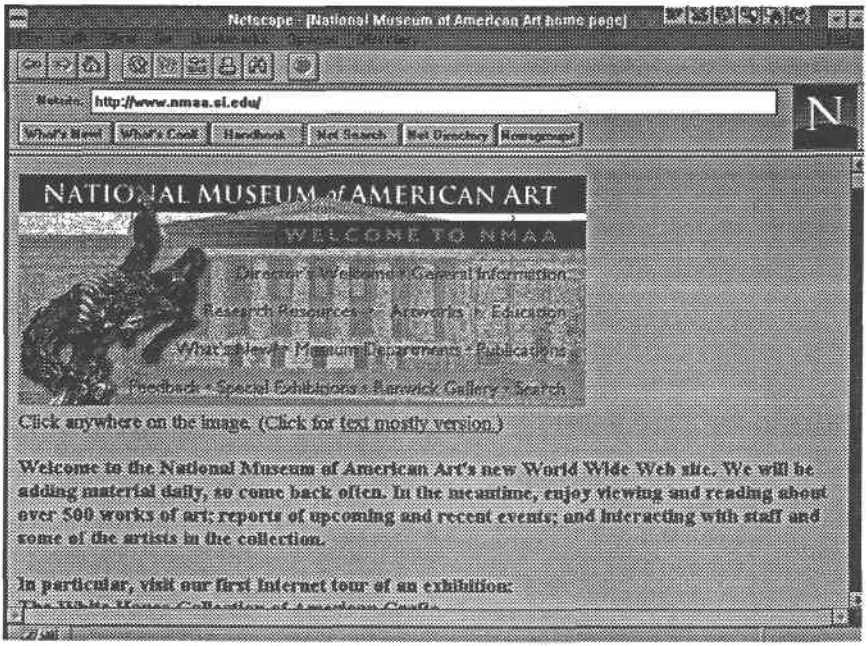

Figure 9. Home Page for the National Museum of American Art.

processing and management 21: 205–13.

Qiu, Liwen. 1993. Analytical searching vs. browsing in hypertext information retrieval systems. *The Canadian journal of information and library science* 18, no. 4: 1–13.

Srinivasan, Srinija. 1995. E-mail interview with the author. October.

Subject Cataloging Division. Processing Department. 1978. *LC classification outline*. 4th ed. Washington, D.C.: Library of Congress.

The whole internet catalog. 1995 (http://gnn.com/gnn/wic).

The WWW virtual library. 1995. (http://www.w3.org/pub/DataSources/bySubject/Overview.html).