DOCUMENT RESUME

ED 403 886 IR 056 258

AUTHOR Callery, Anne
TITLE Yahoo! Cataloging the Web.
PUB DATE 26 Apr 96
NOTE 10p.; A downloaded version of one of 29 conference papers (for those in ERIC, see IR 056 258-274) placed on the Internet under the following overall title: "Untangling the Web. Proceedings of the Conference Sponsored by the Librarians Association of the University of California, Santa Barbara, and Friends of the UCSB Library" (Santa Barbara, CA, April 26, 1996). A list of all 29 papers is appended to this paper.
AVAILABLE FROM Electronic version: http://www.library.ucsb.edu/untangle/
PUB TYPE Reports - Descriptive (141) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cataloging; Classification; *Information Retrieval; Information Sources; Internet; Keywords; *Online Searching; Subject Index Terms; *World Wide Web
IDENTIFIERS Browsing; *Search Engines

ABSTRACT

The Internet has the potential to be the ultimate information resource, but it needs to be organized in order to be useful. This paper discusses how the subject guide, "Yahoo!" is different from most web search engines, and how best to search for information on Yahoo! The strength in Yahoo! lies in the subject hierarchy. Advantages to searching a hierarchical subject index include the following: a higher relevancy rate of items retrieved; the user doesn't need to know all the synonyms of a search term to bring up a topic; and the serendipitous discovery of related items. As opposed to using standard library classification systems, Yahoo! creates its own classification system. Yahoo! currently receives thousands of submissions each day. Every site added is examined by a human being. The suggested category (that which the submitter selects) is used as a guide. Subject lists are organized on a dedicated server and distributed among the catalogers. The cataloger selects an item from the list and a display is brought up. There are fields for title, URL, contact person, geographic location, descriptive comment, and indicators for the presence of Java and VRML. Users can search for information in Yahoo! in two ways. One is by browsing the subject tree and the other is by keyword search. The Yahoo! search can also be incorporated into browsing. (AEF)

Yahoo! Cataloging the Web

Anne Callery
Cataloger
Yahoo! Inc.

Abstract

The Internet has the potential to be the ultimate information resource, but it needs to be organized in order to be useful. I will discuss how Yahoo! is different from most web search engines, and how best to search for information on Yahoo! Libraries are forging ahead and beginning to catalog the Internet, but Yahoo! catalogs differently, not following traditional library procedures. I will explain why this is so, and demonstrate Yahoo!'s entire cataloging process. This presentation should be of interest to general users as well as catalogers.

The Internet is full of information that needs to be organized and made accessible in order for it to be useful. Yahoo! organizes information on the Internet, particularly on the World Wide Web.

Yahoo! is Not Just a Search Engine

There is often confusion about the functions of subject guides, such as Yahoo!, as opposed to search engines, such as Lycos, Alta Vista, WebCrawler, et al. Yahoo! can perform as a search engine (through its Open Text searches), but its strength lies primarily in the subject hierarchy.

There are several advantages to searching a hierarchical subject index, for example:

Higher relevancy rate of items retrieved; less false hits. For example, try running a search for information about surfing. In order to find the sites about riding a board on the waves, you'll have to wade through an awful lot of sites using the popular Internet metaphor.

The user doesn't need to know all the synonyms of a search term to bring up a topic. For example, if a user wants to find sites for organizations in the field of physics, she doesn't have to search for physics plus organizations or societies or associations, etc. She looks under the category Physics, browses a short list of subcategories, selects a subcategory called Organizations, and there are the sites. It's not necessary for the user to bring these entities together herself; they are already arranged that way.

Another benefit of browsing is the serendipitous discovery of related items. In cases in which the user may be looking for a specific site and doesn't see it in its subject area, chances are that other sites grouped in the same area may have something useful.

Contrary to the claims of many enthusiastic users, Yahoo! doesn't presume to catalog everything on the web. Rather, Yahoo! is a filter and organizer of useful information, and we plan to continue in that capacity. While a search engine may have many individual URLs in its database, and searches on all of them, not all of these URLs may be truly useful individually. For example, someone might publish a 50-page thesis on the web. A search engine would index the URL of each of the 50 pages (being individual HTML documents, each with a different URL), so a search could pull up a few random pages in the middle of the thesis, such as a single illustration or reference. (And, unfortunately, not every web publisher thinks to add handy links back to the main page on these little side pages.) Yahoo! would index only the top page of the document (or its significant sections) and bring it together as a whole.

Yahoo! Creates its Own Classification System

Librarians are among those cataloging the Internet. One major project is OCLC's Internet Cataloging Project. Although this project is great for library cataloging, there are a few reasons why this kind of cataloging isn't practical for Yahoo!.

First, libraries catalog Internet sites to integrate them with other, long-existing, materials. This is why they apply traditional cataloging and make MARC records for Internet sites. There is even a new MARC field -- 856 -- which contains information particular to this kind of material (e.g., URL). Yahoo! is cataloging only items on the Internet, so we don't have an existing collection to which we need to conform. There aren't any pre-existing standard rules to follow; we make the rules ourselves.

Second, the information in web documents is not as static as a print publication (or sound recording, film, map, etc.). Full description can be made of these items, and if changes are made to the work, supplements or new editions are released, which are then cataloged separately. On the web, a page could go up about a particular topic, but the actual information therein or the authorship of the works could change from time to time. Although the "object" being cataloged -- the site at that URL -- is consistent, the material within is easily changeable. So a record with full description could become inaccurate within a short time, defeating the purpose of such a record.

Third, a full description is also not as necessary on the web because it is so easy for the user to access a site and make her own decision about its usefulness. On the web, a site is just a click away, not a scribbling of a call number and a trip to the stacks.

The decision to depart from standard library classification systems was a carefully considered one. With so much to do already, we would have been happy to adopt an existing system and save a lot of time and energy. However, for various reasons, no one system could meet all our needs. We do look to other systems P e.g., Library of Congress Classification (LCC) P for ideas and guidelines for the organization of certain areas. I like to compare Yahoo!'s subject hierarchy with the early Dewey Decimal Classification, except in our case, it's much easier to expand and grow! Yahoo! may have started out a little heavier in some areas than others, but some of those initially smaller areas have really taken off.

How Yahoo! Catalogs Web Sites

Yahoo! currently receives thousands of submissions each day. Although our cataloging staff is continually growing and we have made many improvements to the add process, we're still a little short of meeting this demand. Every site added to Yahoo! is examined by a human being. The suggested category (that which the submitter selects) is used as , and we reserve final editorial judgment. Having the user suggest a category helps us organize the

submissions, which are grouped each week by subject. The subject lists are organized on a dedicated server and distributed among the catalogers. Most of us specialize in certain areas, which ensures that each category has a small group of people who know it fairly well. The cataloger selects an item from the list, and a display is brought up.

There are fields for title, URL, contact person, geographic location, descriptive comment, and indicators for the presence of Java and VRML. We're not using all of these fields in the actual Yahoo! display at this time, but they could be implemented later. Below the fields is a snapshot of the submitted page. Occasionally, just looking at this top page will tell us enough about the site to place it in a category (e.g., an X-Files fan page), but more often we will explore the site a little to get a feel for its content. We select categories (using another application which is an interface to the Yahoo! database) and add the site. Then we send off an e-mail to let the submitter know the site has been added or, in some cases, explaining why the site was not added.

We do have some standardization in the form of add guidelines. The most important of these is just to use common sense. We look at the site carefully to determine the best subject area, sometimes consulting reference material and each other. The category the user submits it under may not be the best category for that site. (For example, the Texas Beef Council once submitted their site under Health/Fitness and Exercise.) Occasionally, we'll e-mail the submitter and ask for more information to help us place the site correctly. We also look for content. Often people put up a page and submit it to Yahoo! before any substantial content is added. We don't want to list a site containing nothing but "under construction" signs, or a company's site with nothing but an address and phone number with the instructions to call them for more information. Users are quickly turned off by underdeveloped sites, and it reflects poorly not only on the site itself, but on Yahoo! as well.

Because our subject hierarchy is dictated by whatever we find, we often create new subcategories and develop the hierarchy as we go. This is a "bottom up" approach, as opposed to more traditional "top down" systems. In some cases we try to use the most common terms a person might look for. For example, we have a category Recreation/Hobbies/Model Airplanes. When we first received a site about model helicopters, we wondered whether to change the name of the category to Model Aircraft. However, we decided to include the helicopter site and retain the first name because Model Airplanes is the more idiomatic phrase.

We also try to maintain a consistent vocabulary in the naming of common subcategories. A good example of this is Universities. Within the directory for a particular university, we have chosen to divide the institution's individual sites thus.

Some of these subcategories and many of the sites are linked back to their appropriate subject areas. For example, the Athletics directory is linked to Recreation/Sports/College/ and named for the University. Individual departments are linked to their academic disciplines.

Such a detailed structure became necessary because universities are such a large presence on the web. As a category grows, sites may become more specific and logical subdivisions emerge, so we make them. We are currently creating similar structures for the Regional category. In the early days of the web (which, of course, is fairly recent!) the majority of sites originated from large institutions or organizations. As more people and businesses get on the web now, more of them are smaller, regional operations. A good example is Internet service providers. It doesn't make sense for us to list them all in one big alphabetic list under Business/Companies/Internet Service Providers. The typical user is only going to be interested in providers in her area, and wouldn't be too happy to have to sift through hundreds of entries. By listing these services regionally, we make sure that someone interested in Internet services in, for example, Chicago could simply go to that area in the Regional category and find local providers listed there.

Specificity to a region is one of the main distinctions we look for when placing a site. Another important distinction is whether the site is commercial. All commercial sites are added under Business and Economy, in either Companies or Products & Services. This has to do with the nature of the material, which is usually just advertising the company or selling its products, as well as with the general attitude of the Internet community towards commercialism on the net.

Searching Yahoo!

Users can search for information in Yahoo! in two ways. One is by browsing the subject tree. The other is by keyword search. A keyword search looks in the Yahoo! database for words in the title and comment fields of individual entries and in the names of categories. The search results display has three parts. First, whole categories containing the keyword(s) are shown, then individual sites. The third section is the result of an Open Text search . A great feature of Yahoo! searching is the links to other search engines displayed at the bottom of the search results screen. Selecting one of these will not only take the user to the search engine, but will also automatically execute the same search.

The Yahoo! search can also be incorporated into browsing. When users select a Yahoo! category, they have the option of searching only within that category . This is a significant aid in categories which contain a large number of sites, or when searching a subject that could be listed under more than one subcategory. The results of this search show where the keywords occur within the category. As with a regular search, Open Text results are displayed, along with links to other search engines.

Yahoo! is continually evolving, and we welcome feedback and suggestions on how we can make Yahoo! better. If any of you has ideas or suggestions, please feel free to use the forms in the "Write Yahoo!" section under the Info button on the Yahoo! banner, or send me an e-mail, to anne@yahoo.com.

UNTANGLING THE WEB

Proceedings of the Conference Sponsored by the Librarians Association of the University of California, Santa Barbara and Friends of the UCSB Library

April 26, 1996
University Center, University of California, Santa Barbara

Editor: Andrea L. Duda
Editorial Assistant: Madeleine N. Kempton
Conference Chair: Rosemary Meszaros

	Session 1: 9:15-10:15 am
A	**Yahoo! Cataloging the Web** Anne Callery *Yahoo! Inc.* | **Biography** | **Abstract** | **Paper** |
	Technical Services and the World Wide Web Virginia M. Scheschy *University of Nevada, Reno* | **Biography** | **Abstract** | **Paper** |
B	**Spinning a Web Search** Mark Lager *California Lutheran University* | **Biography** | **Abstract** | **Paper** |
C	**The Alexandria Digital Library on the World Wide Web** James Frew *University of California, Santa Barbara* | **Biography** | **Abstract** | **Paper** | (Paper is by Terence R. Smith)
D	**Wow! This Librarian and These Librarians are Magic! The World Wide Web and Its Use in a Medium-Sized Public Library** Michael Charton *Parsippany Public Library, Parsippany, NJ* | **Biography** | **Abstract** | **Paper** |
	Journals Online News: Dispersing Collection Management Information on the World Wide

E	**Web** Anne Langley and Sandra Leach *University of Tennessee, Knoxville* | **Biography** | **Abstract** | **Paper** |
	Session 2: **10:30-11:30 am**
F.	**Weaving the Web into Course Integrated Instruction** Ruth Wallach, Linda McCann *University of Southern California* | **Biography** | **Abstract** | **Paper** |
G.	**Untangling the Web: Using the World Wide Web for Art and Humanities Reference Services** Mary N. Hernandez, Karen Dalziel Tallman *University of Arizona* | **Biography** | **Abstract** | **Paper** |
H.	**The MELVYL Web Interface** Laine Farley *University of California Division of Library Automation* Mary Horres *University of California, San Diego* Mike Berger *University of California Division of Library Automation* | **Abstract** | **No paper available** |
I.	**The Art and Science of Web Server Management** Roy Tennant *University of California, Berkeley* | **Biography** | **Abstract** | **Paper** |
J.	**The Librarian as Publisher: A Case Study of a World Wide Web Publishing Project** Mark Stover *Phillips Graduate Institute* | **Biography** | **Abstract** | **Paper** |
	Scholarly Communication and Electronic Publication: Implications for Research, Advancement, and Promotion Lizbeth Langston *University of California, Riverside* | **Biography** | **Abstract** | **Paper** |
	Session 3: **1:00-3:00 pm**

K.	**UCSD's Infocat Cataloging Project** Linda Barnhart, Crystal Graham, Becky Ringler, Bradley Westbrook *University of California, San Diego* **| Abstract | No paper available |**
L.	**Cooperative Web Development and Maintenance** Michael Adams, Julius Bianchi, Lynda Fulford, Mark Lager, Carole Thompson, Nathan Tierney *California Lutheran University* **| Biography | Abstract | Paper |**
M.	**Untangling the Tangled Webs We Weave: A Team Approach to Cyberspace** Ellen Broidy, Kathryn Kjaer, Christina Woo *University of California, Irvine* **| Biography | Abstract | Paper |**
N.	**Collection Development Tools/Methods for Virtual Libraries and Subject Lists in Selected Major Subject Areas** Charlene Baldwin *University of California, Riverside* Romeilia Salinas *University of California, Los Angeles* Richard Chabran *University of California, Los Angeles* Lynne Reasoner *University of California, Riverside* Lucia Snowhill *University of California, Santa Barbara* Garrett Bowles *University of California, San Diego* Fred Yuengling *University of California, Santa Cruz* Carlos Rodriguez *University of California, Riverside* Steve Mitchell *University of California, Riverside* **| Biography | Abstract | Paper |**
O.	**First Amendment Issues and the Web: The Internet Porn Panic and Restricting Indecency in Cyberspace** Dorothy Imrich Mullin *University of California, Santa Barbara* **| Biography | Abstract | Paper |**

Session 4:
3:15-4:15 pm

P.	**The Jumpstart Program: Fostering Partnerships to Integrate the Web into the Curriculum** Lucy Wegner, Karen Howell, Rick Lacy, Carl Sutter *University of Southern California* **| Biography | Abstract | No paper available |**
Q.	**Spiders and Worms and Crawlers, Oh My: Searching on the World Wide Web** Ann Eagan, Laura Bender *University of Arizona* **| Biography | Abstract | Paper |**
R.	**Casting the Net: The Development of a Resource Collection for an Internet Database** Gerry McKiernan *Iowa State University* **| Biography | Abstract | Paper |**
	Instruction on the Web: Authoring Tutorials in HTML Katherine M. Whitley *University of California, San Diego* **| Biography | Abstract | Paper |**
S	**From Lab to Library: The Web's Effect on Teaching the Internet** Lisa Lehman *University of Alaska, Fairbanks* **| Biography | Abstract | Paper |**
	Critical Thinking in an Online World Debra Jones *Cabrillo College* **| Biography | Abstract | Paper |**
T.	**Bringing Order to the Web: Future Approaches to Organizing the Digital Universe** Cecily Johns *University of California, Santa Barbara* Brian Schottlaender *University of California, Los Angeles* Bruce Miller *University of California, San Diego* Daniel Pitti *University of California, Berkeley* **| Biography | Abstract | No paper available |**

Poster Sessions:
4:30-5:30 pm

Weaving the Web into the Job Seeking Process
Carol Doyle, Don Lubach, and Janet Martorana
University of California, Santa Barbara

| **Abstract** |

Involving Reluctant Patrons in Web Use and Development
Leroy D. Smith, Robin Satterwhite, and Susan Brown
The Colorado College

| **Abstract** |

Web Access for Library Users: Planning and Implementing a Multimedia Internet User Workstation
Daniel Taylor
California Institute of Technology

| **Abstract** |

Library of Congress Subject Headings as Subject Terminology in a Virtual Library: The INFOMINE Example
Steve Mitchell
University of California, Riverside

| **Abstract** | **Paper** |

Web Weavers: Integrating Web Instruction into Library Skills Classes
Sylvia Curtis, Sherry DeDecker, Nerea Llamas, Lorna Lueck, Janet Martorana
University of California, Santa Barbara

| **Abstract** |

Using Research and Reference Skills to Find Agricultural Information on the Internet
May Ying Chau
Oregon State University

| **Abstract** |

Intellectual Property and Copyright in the Academic Environment: So What?
Denise A.D. Bedford
University of Southern California

| **Abstract** |

Cooking with Prepared Foods: Using Research Assistant and RfGuide to Provide Library Service and Instruction to Distance Learning Students
Marjorie F. MacKenzie
Lewis-Clark State College

| **Abstract** |

Last modified: May 16, 1996

Direct questions or comments about this page to webmaster@library.ucsb.edu.

Search "Untangling" Papers | "Untangling the Web" Home | InfoSurf Home

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research & Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE:

I. DOCUMENT IDENTIFICATION:

Title: Yahoo! Cataloging the Web

Author: Anne Callery

Corporate Source: Yahoo! Inc.

Publication Date: April 26, 1996

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthy abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

_X__ Permission is granted to the Educational Resources Information Center (ERIC) to reproduce this material in microfiche, paper copy, electronic, and other optical media (Level 1).

or

____ Permission is granted to the Educational Resources Information Center (ERIC) to reproduce this material in other than paper copy (Level 2).

Sign Here, Please A E Callery

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from teh ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission fro the copyright holder. Exception is made for non-profit reproeudction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature: A E Callery Position: Cataloger

ed Name: Anne Callery Organization: Yahoo! Inc.

Address: 3400 Central Expwy, Suite 201, Santa Clara CA 95051

Telephone Number: 408-731-3321

Date: 7/31/96

III. DOCUMENT AVAILIBILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of this document from another source, please provide the floowing information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more strigent for documents which cannot be made available through EDRS).