Maintaining a Catalog of Manually-Indexed, Clinically-Oriented World Wide Web Content

**William Hersh, M.D., Andrea Ball, M.L.S., Bikram Day, M.S.,
Mary Masterson, M.P.H., Li Zhang, M.S., Lynetta Sacherek, M.L.S.
Division of Medical Informatics & Outcomes Research
Oregon Health Sciences University
Portland, Oregon, USA**

With no quality controls and a highly distributed means of posting information, finding high-quality, clinically-oriented content on the World Wide Web can be difficult. Maintaining a catalog of such information can be equally challenging. CliniWeb is a catalog of quality-filtered and clinically-oriented content on the Web designed to enhance access to such information. This paper describes a group of semi-automated tools have been developed to maintain the CliniWeb database. One allows easier identification of content by utilizing Web crawling techniques from high-level pages. Another allows easier selection of content for inclusion and its indexing. A final one checks links to help keep the database current. These are augmented by general plans to adopt more detailed metadata and linkages into the medical literature.

INTRODUCTION

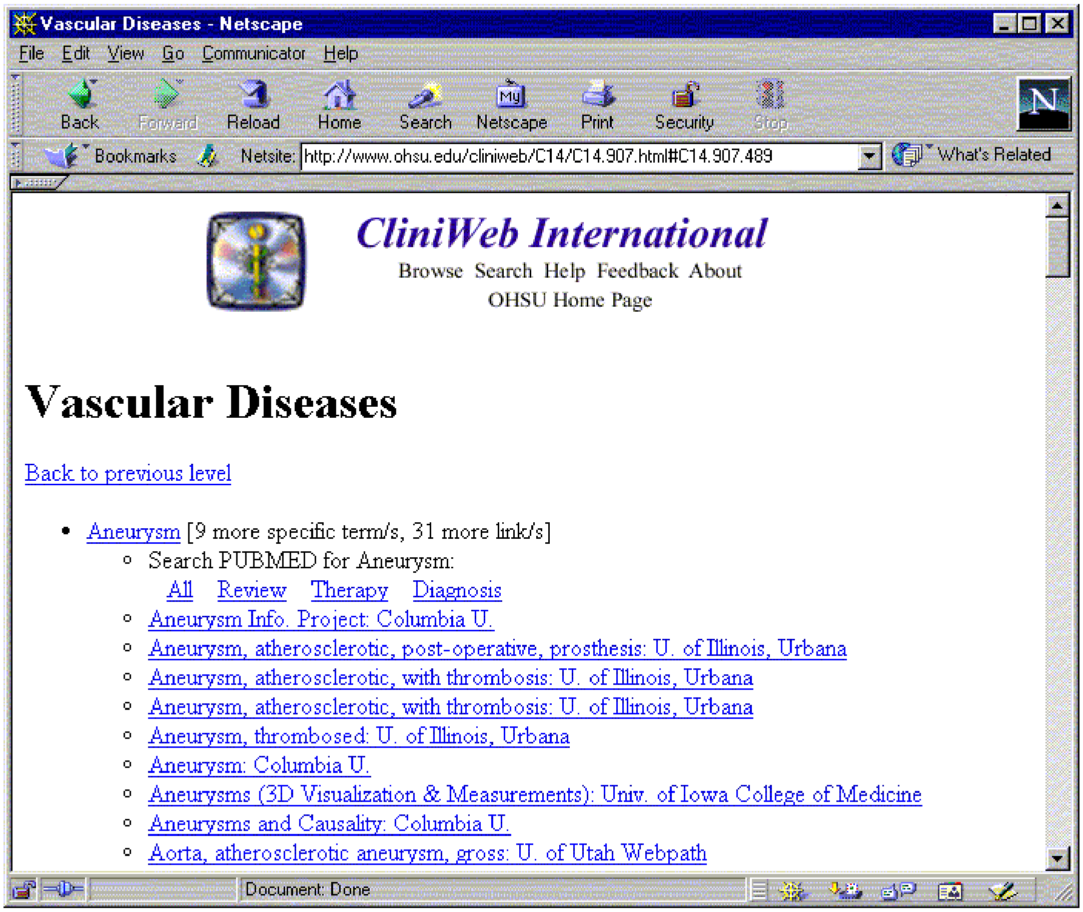
The World Wide Web has changed the way that information is produced, disseminated, indexed, and retrieved, and it challenges conventional notions of information retrieval (IR). Prior to the advent of the Web, users searched for on-line information in discrete databases, which were usually bibliographic in nature. Although large, these databases had circumscribed boundaries and the user could make assumptions about their content and quality.

The Web, however, has changed all of this. Content is easy to produce and disseminate by virtually anyone with an Internet Service Provider (ISP) account. The information is likely to be indexed by one or more of the many *Web crawlers* that find it and make it retrievable from their search engines. Web crawlers are computer programs that identify pages on the Web. They work by starting at an individual page and processing it to identify links to other pages. They next follow the links and repeat the process. The main use of Web crawlers is to quickly identify all pages for inclusion into general search engines (e.g., Excite, AltaVista, HotBot). While discrete databases can still be searched on the Web (e.g., accessing MEDLINE through PubMed or the sites of commercial database providers), much searching is done with these search engines, which index unfiltered Web content. They are limited by the undiscriminating nature by which they add information to their databases, including a great deal of which is of poor quality [1].

Another approach to IR on the Web is the development of catalogs (also called portals or meta-sites), which filter content based on pre-defined criteria. In the clinical realm, for example, there are many sites (e.g., Medical Matrix, Yahoo Health, HealthFinder) that catalog health-oriented sites on the Web which meet certain quality criteria. Another example of a filtered site is CliniWeb (http://www.ohsu.edu/cliniweb/), which has additional unique features based on its goals of providing access to quality-filtered, clinically-oriented information [2], including:

1. Cataloging content on a per-page as opposed to per-site basis
2. Including only pages that have clinical content, i.e., excluding individual and institutional home pages, advertisements, and lists of links
3. Indexing with a higher level of specificity, i.e., using the National Library of Medicine (NLM) Medical Subject Headings (MeSH) vocabulary [3] as opposed to broad subject categories such as *Orthopedics* or *Cancer*

CliniWeb provides access to Web pages manually indexed by a large subset (trees A-G) of MeSH, including the major trees, *Diseases*, *Anatomy*, and *Chemicals and Drugs*. Each page in the site is one level of the MeSH tree, linked to the parent term above it and children terms below it in the hierarchy. In a list under each term are the Web pages from the database that have been indexed by that term. Figure 1 shows a sample CliniWeb page.


Vascular Diseases - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks
Netsite: http://www.ohsu.edu/cliniweb/C14/C14.907.html#C14.907.489
What's Related
Document: Done

CliniWeb International

Browse Search Help Feedback About

OHSU Home Page

Vascular Diseases

Back to previous level

- Aneurysm [9 more specific term/s, 31 more link/s]
  - Search PUBMED for Aneurysm:
    All Review Therapy Diagnosis
  - Aneurysm Info. Project: Columbia U.
  - Aneurysm, atherosclerotic, post-operative, prosthesis: U. of Illinois, Urbana
  - Aneurysm, atherosclerotic, with thrombosis: U. of Illinois, Urbana
  - Aneurysm, atherosclerotic, with thrombosis: U. of Illinois, Urbana
  - Aneurysm, thrombosed: U. of Illinois, Urbana
  - Aneurysm: Columbia U.
  - Aneurysms (3D Visualization & Measurements): Univ. of Iowa College of Medicine
  - Aneurysms and Causality: Columbia U.
  - Aorta, atherosclerotic aneurysm, gross: U. of Utah Webpath

Figure 1 – CliniWeb page for Vascular Diseases, showing children terms (e.g., Aneurysm) with links to PubMed searches and specific pages.

Users can find information in CliniWeb by searching or browsing. When searching, the user enters a textual query and a list of possible matching MeSH terms is returned. The user then clicks on the term of interest and is taken to the page containing the portion of the MeSH tree where that term resides. From this point the user can browse up and down the hierarchy. The user can also start browsing from the top level of MeSH. Aside each term at each level is an indication of how many more specific terms and links are present at the next level below the term in the hierarchy.

As with many sites on the Web, maintaining and enhancing the site was much more difficult than initially building it. As new content appeared and old sites changed, we found it necessary to devise tools to manage maintenance and growth of the site. This paper describes the tools that maintain CliniWeb, providing insight into the requirements for effectively maintaining a catalog of manually-indexed, clinically-oriented content. It concludes with a description of some additional new features and future plans.

TOOLS TO MAINTAIN CLINIWEB

The major aspects of CliniWeb to benefit from automation have been the identification of clinical content, its selection and indexing, and its maintenance.

Identifying clinical content

As the goal of CliniWeb is to maintain a portal of quality-filtered, clinically-oriented content, tools must be devised to find such content efficiently. Manual review of all individual pages, the approach taken in the first version of CliniWeb, was not scalable [2], so we sought an approach that was as automated as possible to identify qualified content. We therefore devised an approach of using a Web crawler with modifications to identify pages which were likely to be of sufficient quality and clinically relevant [4]. This approach has been used by another medical Web catalog [5], but this site does not use human review and filtering of its content.

In order to use a Web crawler to improve the efficiency of identifying pages likely to qualify for inclusion, we had to modify the basic crawler approach. First, we had to point the crawler at a page likely to have or link to qualified content. Second, we had to keep the crawler from linking to pages having less likelihood of being qualified. This was done by developing a list of *sentinel* pages that represented the top level or table of contents of content likely to be clinically oriented and of good quality. As many Web sites are organized hierarchically, identifying such pages was not difficult. The next step was to keep the crawler from including as many non-qualified pages as possible. This was done by placing restrictions on links the crawler could follow. The following rules were found to be most effective:

1. Do not link across domains.
2. Do not link to other directories within a site.
3. Do not link when the URL shortens.
4. Do not link to non-HTML files.

Following these rules, each sentinel page yielded a list of *candidate* pages for inclusion in the CliniWeb database. Our initial review of medical Web sites identified 531 sentinel pages. The above mentioned identification process yielded 52,173 candidate pages for possible inclusion in the CliniWeb site.

The code base for the crawler is based on the lwp library of Perl (http://www.linpro.no/lwp/). The code was modified to add the above rules and process an input file of sentinel pages. As candidate pages are identified, they are stored in tables using the Oracle Relational Database Management System (Oracle Corp., Redwood Shores, CA) for later indexing.

Selecting and indexing content

Once candidate pages are identified, they must be judged as to whether they meet the standard of quality and clinical orientation. These attributes are judged by indexers, who tend to be inclusive if the page appears to come from a reputable source (e.g., health-oriented government agency or academic medical center). A variety of individuals have indexed CliniWeb, all of whom either had some type of health care or medical librarian background.

Another tool deemed necessary was one to assist in the selection and indexing of content. Indexing in the initial version of CliniWeb was problematic, due to the lack of tools to assist this process. The only automation of indexing in the first version was to use the SAPHIRE concept-matching system [6] as an aid to identifying indexing terms. To improve this process, we developed an interactive indexing tool that allows the page being indexed to be viewed and to use SAPHIRE interactively. The tool consists of two HTML frames, one which allows viewing of the Web page being indexed and the other which provides an interactive version of SAPHIRE and a description of the indexing status.

The indexing tool begins by displaying the page in a frame and prompting the indexer whether they wish to index the page, omit it from the database (because it is not qualified), or revisit it (index it later). If the indexer chooses to index the page, the title is passed to SAPHIRE and an initial set of one or more indexing terms are suggested. The indexer can select one or more of these terms or interact further with SAPHIRE. Options include:

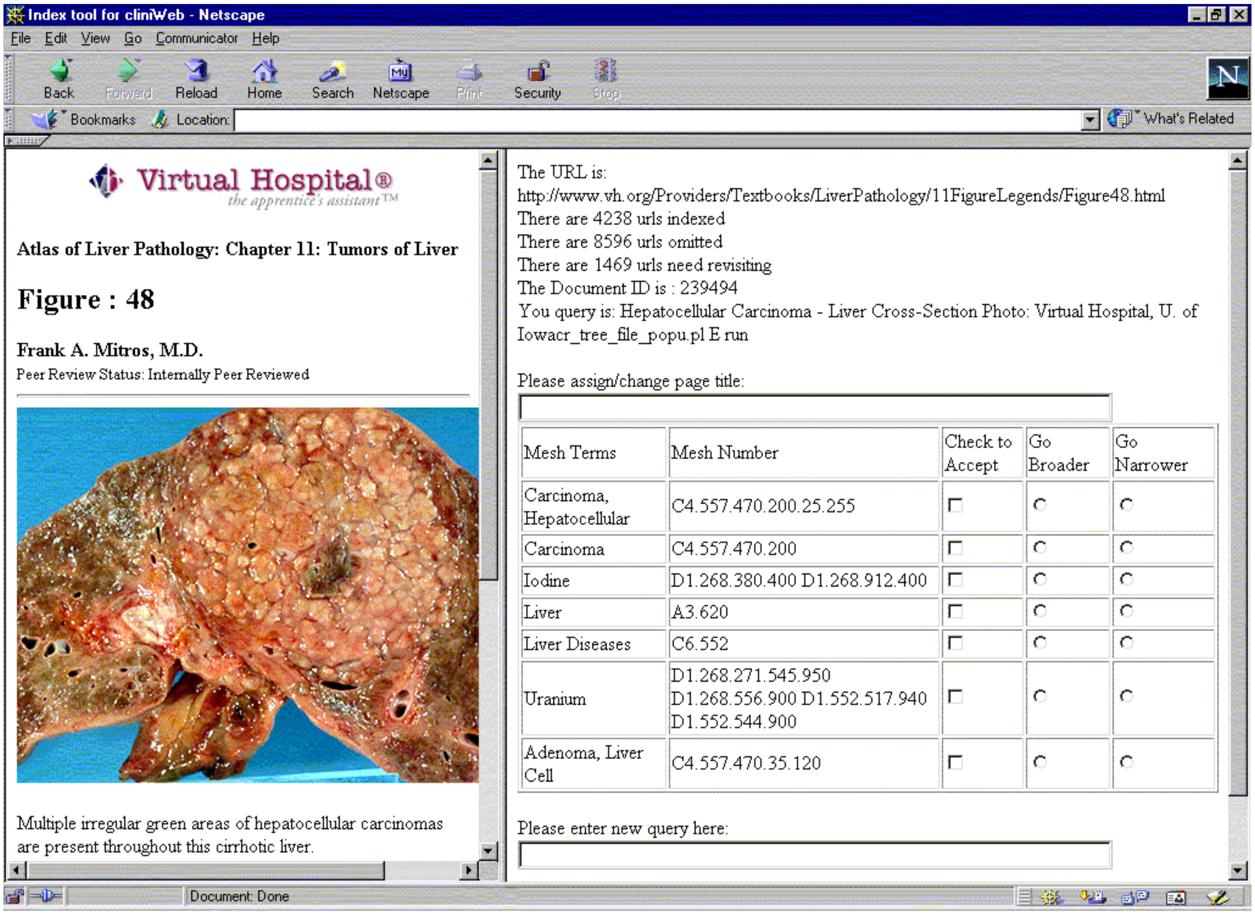
1. Sending more or different text to SAPHIRE for suggestion of additional terms.
2. Browsing up or down the MeSH hierarchy to identify additional terms.

When the user is done indexing, the selected terms are stored to the Oracle database. Figure 2 shows a screen display of the indexing tool.

The indexing tool is based on HTML with use of JavaScript to manage user input, navigation, and interaction with SAPHIRE. The Oracle database is used to store all indexing data. Interaction with Oracle is provided by the DBI-DBD Perl-Oracle API (http://www.hermetica.com/technologia/DBI/). The pages for the CliniWeb site are generated by a program that creates the hierarchical page structure and then populates it with individual links.

Maintaining content

Another problem with Web catalogs is the constant change (usually reorganization but sometimes deletion) of Web sites. To identify links that are no longer valid, a program called checklinks has been developed which goes through the database of indexed pages and uses the HTTP protocol to determine if they are accessible. Those which are not accessible are flagged as such and not included in the next output of the database. It also signals an indexer to visit the site to see whether the content has been moved to a different URL or is deleted. If a new URL can be identified, then the URL is changed in the Oracle database.


Index tool for cliniWeb - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks Location:
What's Related
Virtual Hospital®
the apprentice's assistant™
Document: Done

Atlas of Liver Pathology: Chapter 11: Tumors of Liver

Figure : 48

Frank A. Mitros, M.D.

Peer Review Status: Internally Peer Reviewed

Multiple irregular green areas of hepatocellular carcinomas are present throughout this cirrhotic liver.

The URL is:
http://www.vh.org/Providers/Textbooks/LiverPathology/11FigureLegends/Figure48.html
There are 4238 urls indexed
There are 8596 urls omitted
There are 1469 urls need revisiting
The Document ID is : 239494
You query is: Hepatocellular Carcinoma - Liver Cross-Section Photo: Virtual Hospital, U. of Iowacr_tree_file_popu.pl E run

Please assign/change page title:

Mesh Terms	Mesh Number	Check to Accept	Go Broader	Go Narrower
Carcinoma, Hepatocellular	C4.557.470.200.25.255	□	○	○
Carcinoma	C4.557.470.200	□	○	○
Iodine	D1.268.380.400 D1.268.912.400	□	○	○
Liver	A3.620	□	○	○
Liver Diseases	C6.552	□	○	○
Uranium	D1.268.271.545.950 D1.268.556.900 D1.552.517.940 D1.552.544.900	□	○	○
Adenoma, Liver Cell	C4.557.470.35.120	□	○	○

Please enter new query here:

Figure 2 – CliniWeb indexing tool.

RECENT ENHANCEMENTS TO CLINIWEB

Two additional enhancements have been made to CliniWeb since the original version. The first is the addition of links to MEDLINE through PubMed (http://www.ncbi.nlm.nih.gov/PubMed/). This enables the user to find more detailed references available in the medical literature. The basic approach passes the MeSH term to PubMed, along with some limits to reduce the output of the search. All terms are provided a general search and another limited to review articles. Disease terms have the option of further limiting either to therapy and diagnosis, using the clinical filters developed at McMaster University [7]. Once the search has been passed to PubMed, further interaction with PubMed can continue. Figure 3 shows a list of the search strategies currently used. The NLM has recently adopted a similar approach in their MEDLINEPlus system (http://www.nlm.nih.gov/medlineplus/).

Another enhancement is the ability to enter query terms in languages other than English. This capability is based on SAPHIRE International, which uses the non-English terms in the UMLS Metathesaurus to allow selection of terms in different languages [8]. This gives an entry into CliniWeb and medical information in general via non-English languages, although all content currently indexed in CliniWeb is in English.

FUTURE PLANS

Continued work on CliniWeb will build on the foundation that currently exists. One area currently being investigated is the development a deeper model of metadata to represent page content beyond simple MeSH terms. Based on the Medical Core Metadata (MCM) Project [9], this will include the use of other MeSH features (e.g., subheadings) as well as resource types to identify type of content (e.g., case report, topic review, etc.).

We also plan to improve the generic PubMed searches. For most MeSH terms, a search on the term itself, even when qualified to obtain review or therapy articles, yields an excessive number of references. We will look at means for reducing this output, such as:

1. Limiting the number of journals
2. Allowing the user to add qualifying terms, such as treatment terms to a disease term

The major limitation of CliniWeb and other manually indexed and filtered Web sites is the time and effort required to maintain them in the highly dynamic milieu of the Web. The only certainty about the Web is its uncertain evolution. Continued work will not only require developing better approaches to managing access to its content, but also the ability to keep up with its evolution.

ACKNOWLEDGEMENTS

This work has been supported by Department of Energy Grant DE-FG03-94ER61918 and NLM Fellowship Training Grant LM07088.

REFERENCES

1. Hersh WR, Gorman PN, and Sacherek LS, Applicability and quality of information for answering clinical questions on the Web. *Journal of the American Medical Association*, 1998. 280: 1307-1308.
2. Hersh WR, et al., CliniWeb: managing clinical information on the World Wide Web. *Journal of the American Medical Informatics Association*, 1996. 3: 273-280.
3. Lowe HJ and Barnett GO, Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *Journal of the American Medical Association*, 1994. 271: 1103-1108.
4. Day B, *The Design and Implementation of a Hypertext Document Management System*, 1998, Unpublished Master's Thesis, Oregon Health Sciences University: Portland, OR.
5. Suarez HH, Hao X, and Chang IF. Searching for information on the Internet using the UMLS and Medical World Search. in *Proceedings of the 1997 Annual AMIA Fall Symposium*. 1997. Nashville, TN: Hanley & Belfus 824-828.
6. Hersh WR and Leone TJ. The SAPHIRE server: a new algorithm and implementation. in *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care*. 1995. New Orleans, LA: Hanley-Belfus 858-862.
7. Haynes RB, et al., Developing optimal search strategies for detecting clinically sound studies in MEDLINE. *Journal of the American Medical Informatics Association*, 1994. 1: 447-458.
8. Hersh WR and Donohoe LC. SAPHIRE International: a tool for cross-language information retrieval. in *Proceedings of the Annual AMIA Fall Symposium*. 1998. Orlando, FL: Hanley-Belfus 673-677.
9. Malet G, et al., A model for enhancing Internet medical document retrieval with "medical core metadata". *Journal of the American Medical Informatics Association*, 1999. 6: 183-208.

```
<A HREF="http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=4&term=Hypertension[MAJR]
+AND+Human[MESH]+AND+English[LANG]&dopt=d&relpubdate=1+Year&dispmax=20">All</A>

<A HREF="http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=4&term=Hypertension[MAJR]
+AND+Human[MESH]+AND+English[LANG]+AND+review[PTYP]&dopt=d&relpubdate=1+Year&dispmax=20
">Review</A>

<A HREF="http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=4&term=Hypertension[MAJR]
+AND+Human[MESH]+AND+English[LANG]+AND+(randomized+controlled+trial[PTYP]+OR+drug+therapy[
MESH]+OR+therapeutic+use[MESH]+OR+random*[WORD])&dopt=d&relpubdate=1+Year&dispmax=20">Thera
py</A>

<A HREF="http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=4&term=Hypertension[MAJR]
+AND+Human[MESH]+AND+English[LANG]+AND+(sensitivity+and+specificity[MESH]+OR+sensitivity[WOR
D]OR+diagnosis[MESH]+OR+pathology[MESH]+OR+radiography[MESH]+OR+radionuclide+imaging[MESH]+
OR+ultrasonography[MESH]+OR+diagnostic+use[MESH]+OR+specificity[WORD])&dopt=d&relpubdate=1+Year
&dispmax=20">Diagnosis</A>
```

Figure 3 – HTML code for ClinWeb PubMed links.