USING DATA FROM WEB-SERVER LOGS TO DEVELOP A CONCEPT-BASED BROWSE SET FOR WATER QUALITY AND AGRICULTURE

Stuart R. Gagnon, M.S.L.S.
University of Maryland Libraries
College Park, Maryland

Joseph R. Makuch, Ph.D.
Water Quality Information Center, National Agricultural Library
Beltsville, Maryland

ABSTRACT: Search engine and OPAC transaction logs have been evaluated for a variety of purposes, such as the analysis of demographic and behavioral data; and search frequency, length and success. Transaction logs contain statistics on many facets of a user's interaction with web servers. Web-user communication statistics, in the form of search engine term-frequency logs, were collected for a web site covering water quality and agriculture at the National Agricultural Library's Water Quality Information Center. In this paper, we look at enhancing user success in locating desired documents in the site's online database by creating a hyperlinked browse list. This browse list appears on the database search page and represents concepts drawn from analyzing several months of web logs. Due to government policies, user-identifying information is rarely available from Federal government servers. However, anonymous communication feedback sources, such as logs of search terms entered in database queries, may provide useful information to information managers. An evaluation of frequently-used terms provides a mechanism to identify and meet the needs of people seeking information on water and agriculture.

Introduction

Controlled access to meta-information (databases of bibliographic records, etc.) through web search-engine programs can provide useful value-added benefits to users searching record sets designed to cover specific subject areas. However, with a small amount of consistent metadata representing a narrow subject area, e.g., water and agriculture, the ability to search this information, especially for first-time or novice users, can be hampered by lack of search skills and subject understanding. We sought to assist users unfamiliar with searching a small database aimed at a specific subject area.

In this paper, we describe a process through which information stewards may attempt to gage the needs of an anonymous group of web users by evaluating search strings created for searching a database covering a specific subject. The Water Quality Information

Center's Database of Online Publications on Water and Agriculture, our studied resource, is part of an effort to provide information by the National Agricultural Library.

Background on the Water Quality Information Center at the National Agricultural Library

Since 1862, the National Agricultural Library (NAL) has been helping people who need information about agriculture and related topics. Located in the Abraham Lincoln Building in Beltsville, Maryland, NAL is one of the United States' national libraries, which include the Library of Congress, the National Library of Medicine, the National Library of Education and the National Transportation Library (a virtual library). NAL is part of the U. S. Department of Agriculture's (USDA) Agricultural Research Service. For a history of NAL, see Fusonie (1988).

NAL's mission is to ensure and enhance access to agricultural information for a better quality of life. To do this, NAL

- acquires, organizes, manages, preserves and provides access to information and provides quality stewardship of its unique collection.
- assists, trains and educates people based on assessment of their information needs.
- provides leadership in information management.
- maximizes access to information through collaborative efforts and utilization of technology.
- enhances global cooperation through international exchange of information and the provision of services and technical assistance (United States Department of Agriculture 1994).

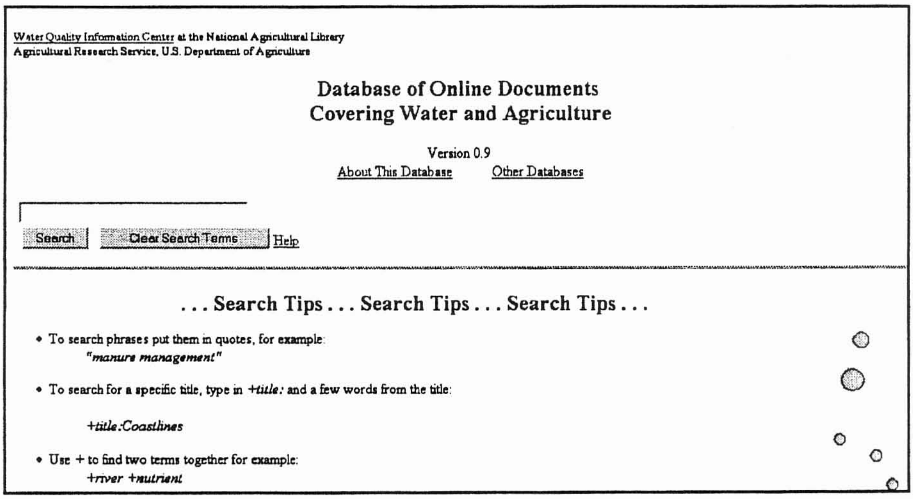
Additional information on NAL is available at the library's web site, www.nal.usda.gov, and in the fact sheet *The National Agricultural Library: A Public Resource for the Public Good* (2003).

As the world's largest agricultural library, NAL is an important information source for people working on agricultural issues, including issues related to water and agriculture. To strengthen its capacity in the water resources area, NAL established the Water Quality Information Center in 1990. The center allows the library to provide special emphasis on meeting the needs of people seeking water-related information.

As the focal point of NAL's water quality efforts, the center collects, organizes and communicates the scientific findings, educational methodologies and public policy issues related to water resources and agriculture. Agricultural nonpoint-source pollution is the center's major focus.

The center primarily serves water resources professionals—including scientists, policy makers, economists, engineers and many others—who are working on solutions to water quality problems associated with agriculture. These individuals are usually affiliated with federal and state governments, academic institutions, or agricultural or environmental organizations. Members of the public with an interest in water issues are also served by the center.

To serve the needs efficiently of the many people seeking water information, the center strives to maximize the use of appropriate information technologies and develop integrated water information systems that allow users themselves to locate and access the information they need directly.



Water Quality Information Center at the National Agricultural Library
Agricultural Research Service, U.S. Department of Agriculture

Database of Online Documents
Covering Water and Agriculture

Version 0.9
About This Database Other Databases

. . . Search Tips . . . Search Tips . . . Search Tips . . .

- To search phrases put them in quotes, for example:
  "manure management"
- To search for a specific title, type in *+title:* and a few words from the title:
  +title:Coastlines
- Use + to find two terms together for example:
  +river +nutrient

Figure 1: Search Page for WQIC Database

Water Quality Information Center's Database of Online Documents

The Water Quality Information Center has been providing people with web access to a database of freely available, online publications related to water and agriculture since May 2000. The database contains bibliographic records, including URLs, of publications of interest and provides a single-point of access to publications that are produced by many organizations. A history and more details regarding the database are available in Makuch and Gagnon (2000, 2002). Figure 1 shows the database search page which is located at www.nal.usda.gov/wqic/wqdb/esearch.html.

There is also an internal version of the database. The internal version (the "working" database) uses ProCite software and is where all additions, deletions and changes are made.

To make the database available on the web, records from the ProCite database are saved as an ASCII text file, uploaded to the NAL server and then Perl scripts are run that convert the file to consecutively numbered HTML files. These HTML files represent individual records in the database and can be searched by the search engine (Inktomi) used on the NAL web site. The Inktomi search engine software <http://www.inktomi.com/> creates logs of search terms entered by users. Later we will examine this facet in more depth.

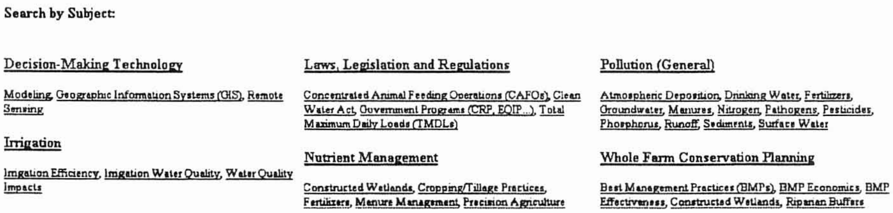
The number of records in the database has grown from an initial 415 to 1,442 (as of September 12, 2003). In addition, the completeness and quality of the metadata in the records has also been enhanced. This improves searching and retrieval of desired documents. Water Quality Information Center staff are also working with NAL information technology staff to convert the current database to MySQL and enhance the user interface.

The database is one of the most utilized sections of the Water Quality Information Center web site. Nine times during the twelve-month period September 2002 - August 2003, the database search page has been among the top ten center web pages accessed (with regard to the number of page "hits"). During this time the page never was less than the twelfth most accessed page. We interpret these hit levels as user interest in our valuation of the database.

Searching Online Water and Agriculture Documents at NAL

The consistent popularity of the database and its content links led us to seek ways to improve access to the descriptive metadata. Collection development strategies—important to our efforts to improve the resource—are outside the scope of this paper. Implementation of metadata elements consistent with the Dublin Core Metadata Initiative < http://dublincore.org/ > has fitted our next generation database (in MySQL version) with standards compliance and provided the power to search fields more easily and with more accuracy (Gagnon & Makuch 2001). Our current redesign work involves improving the search interface, including search page structure, search box placement and search power variables, such as the use of relevance.

As a first step, we are developing a subject list for browsing on the search page, which should inform first-time users of the general subjects covered by the database and help to access those subjects. We based our first instance of this list upon WQIC staff knowledge of the subject area, the subject content of the database and possible difficulties encountered with subject-area terminology, such as components of laws. Figure 2 shows the first version of our browse topics list as a simple hierarchy. Since our goal is to help people find the information they need easily in the database, we sought a source of feedback from users to guide changes to the search interface.



Search by Subject:

Decision-Making Technology

Modeling, Geographic Information Systems (GIS), Remote Sensing

Irrigation

Irrigation Efficiency, Irrigation Water Quality, Water Quality Impacts

Laws, Legislation and Regulations

Concentrated Animal Feeding Operations (CAFOs), Clean Water Act, Government Programs (CRP, EQIP...), Total Maximum Daily Loads (TMDLs)

Nutrient Management

Constructed Wetlands, Cropping/Tillage Practices, Fertilizers, Manure Management, Precision Agriculture

Pollution (General)

Atmospheric Deposition, Drinking Water, Fertilizers, Groundwater, Manures, Nitrogen, Pathogens, Pesticides, Phosphorus, Runoff, Sediments, Surface Water

Whole Farm Conservation Planning

Best Management Practices (BMPs), BMP Economics, BMP Effectiveness, Constructed Wetlands, Riparian Buffers

Figure 2: Browse Topics List, First Version

Transaction Log Analysis

Bates (2003) has long promoted the use of "lead-in" terms which can be employed as "user access vocabulary" aids. Since search engine transaction "logs can be studied and analyzed to determine what kinds of resources people are using" (Janes 2003), WQIC looked into a transaction log analysis approach.

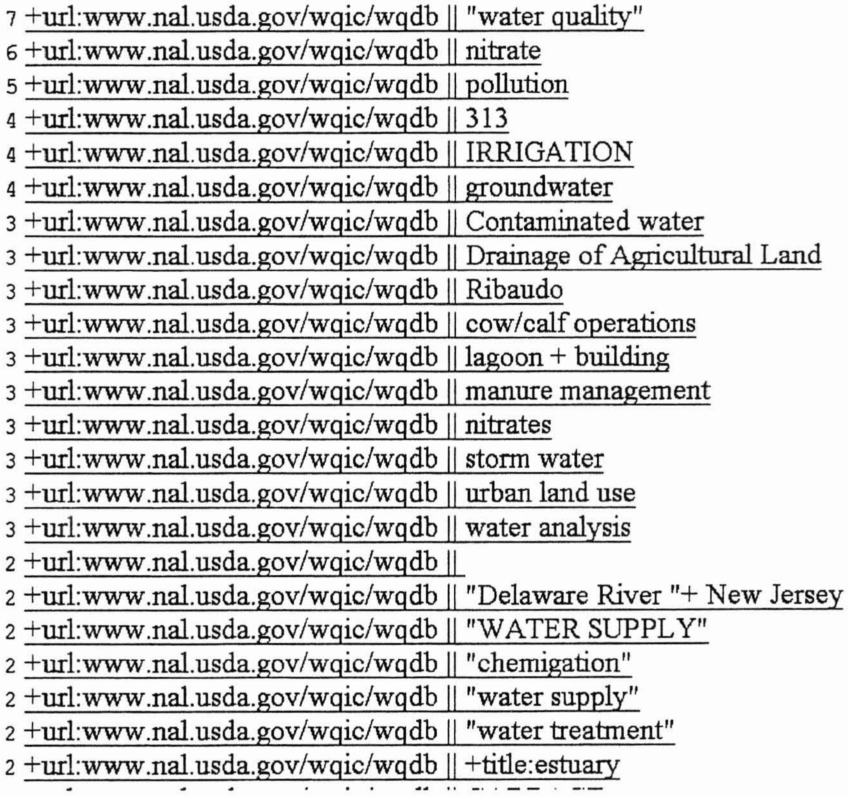
The science of analyzing large sets of data from search engine transaction logs continues to move forward. Some researchers have used computer and human evaluation of very large sets of search engine logs (over 5 million queries) to, among other things, enhance access terminology. Experimental work in "[s]ubject content analysis of queries, such as exploring the search topics and observing changes in their frequency distributions over time," yielded an auto-categorization approach which has shown good performance (Pu, Chuang & Yang 2002). Another attempt to leverage user logs employs a probabilistic model for query term expansion which works to provide related concepts for the routinely short web query (Cui, Wen, Nie & Ma 2003). WQIC staff considered that Inktomi's logs of search queries run against our local subject set, while small in scope as compared with auto-categorization approaches, might assist us in a strategy to enhance our subject-based interface, the search-page browse list.

Appearance and Manipulation of Logs

Some locally-loaded search engine software products offer "query by frequency," or "frequent search queries," reports. Search term frequency logs are easily accessible, but may not be utilized to their full potential at NAL. We sought to make use of the tool, if possible. NAL information technology staff has set-up Inktomi to provide these web transaction logs on a monthly, or bi-monthly, basis. Query by frequency reports are available to all web management staff—linked from the NAL web statistics page. Analysis of these logs can provide useful information to content providers.

NAL logs reside as HTML pages on a local server and include each query transaction sent to all NAL search pages. In other words, reports are not provided for each search page—just as one single combined file. Approximately 20 search-page instances from various units of the library feed into each monthly log. So, individual search page owners must extract their own query transaction logs.

Once one is signed onto the web server, manipulation of the NAL-wide log is performed with a simple UNIX *grep* command to remove all lines with the URL of the WQIC database search page. This leaves a much smaller and more manageable log including only WQIC database transactions. Because all of these files are very large and the process to create the parsed files leaves many copies, any process file used in the transfer is deleted to preserve storage space. Once the WQIC monthly searches are isolated into separate files, simple analysis can performed with the aid of common desktop, or freely-downloadable, software.

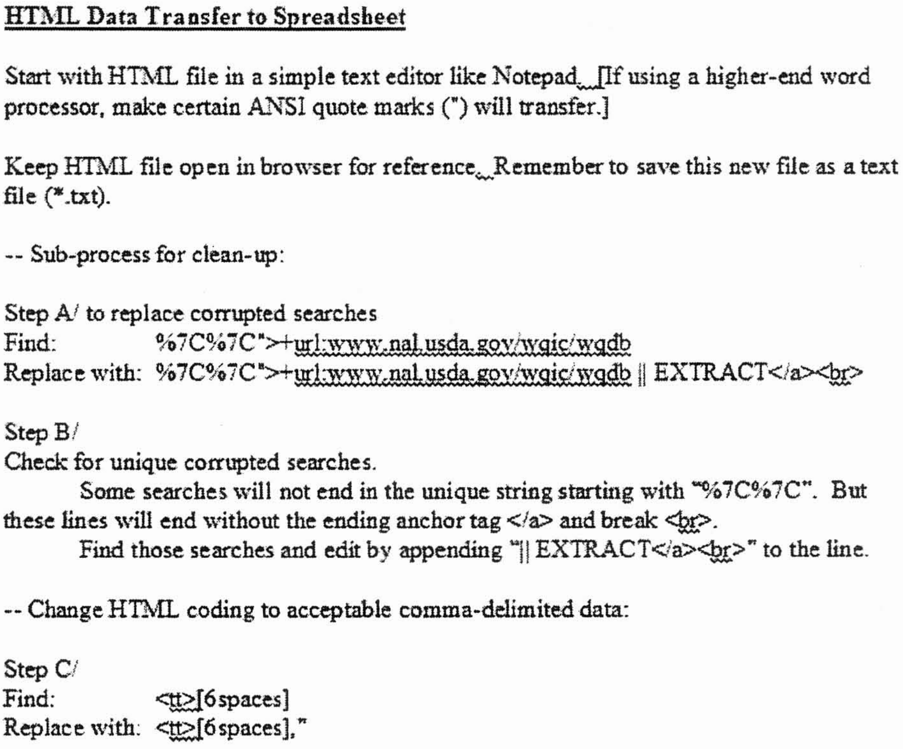


```
7 +url:www.nal.usda.gov/wqic/wqdb || "water quality"
6 +url:www.nal.usda.gov/wqic/wqdb || nitrate
5 +url:www.nal.usda.gov/wqic/wqdb || pollution
4 +url:www.nal.usda.gov/wqic/wqdb || 313
4 +url:www.nal.usda.gov/wqic/wqdb || IRRIGATION
4 +url:www.nal.usda.gov/wqic/wqdb || groundwater
3 +url:www.nal.usda.gov/wqic/wqdb || Contaminated water
3 +url:www.nal.usda.gov/wqic/wqdb || Drainage of Agricultural Land
3 +url:www.nal.usda.gov/wqic/wqdb || Ribaudo
3 +url:www.nal.usda.gov/wqic/wqdb || cow/calf operations
3 +url:www.nal.usda.gov/wqic/wqdb || lagoon + building
3 +url:www.nal.usda.gov/wqic/wqdb || manure management
3 +url:www.nal.usda.gov/wqic/wqdb || nitrates
3 +url:www.nal.usda.gov/wqic/wqdb || storm water
3 +url:www.nal.usda.gov/wqic/wqdb || urban land use
3 +url:www.nal.usda.gov/wqic/wqdb || water analysis
2 +url:www.nal.usda.gov/wqic/wqdb ||
2 +url:www.nal.usda.gov/wqic/wqdb || "Delaware River "+ New Jersey
2 +url:www.nal.usda.gov/wqic/wqdb || "WATER SUPPLY"
2 +url:www.nal.usda.gov/wqic/wqdb || "chemigation"
2 +url:www.nal.usda.gov/wqic/wqdb || "water supply"
2 +url:www.nal.usda.gov/wqic/wqdb || "water treatment"
2 +url:www.nal.usda.gov/wqic/wqdb || +title:estuary
```

Figure 3: WQIC-only Search Frequency Log

Figure 3 is a web log for the WQIC's database search page. Search syntax is mostly intact in the Inktomi logs. It could prove of some value to analyze user search strategies in this form alone. However, the sort is on "frequency" which means that any syntax differences (plus signs, capitalization, etc.) will place the terms out of sequence.

Studying the logs as HTML pages (or worse, as code) does not lend itself to quick analysis. So we devised a clean-up procedure through which they might be imported to a spreadsheet program for more control over column order, row sorting, stray or useless characters, and other anomalies. The logs are imported into Microsoft Excel after replacing HTML code with commas and quotes so that the text resembles a comma delimited file, a much easier format to import. Inside Excel, the search queries can be cleaned-up, resorted, examined and exported to text analysis software.



HTML Data Transfer to Spreadsheet

Start with HTML file in a simple text editor like Notepad. [If using a higher-end word processor, make certain ANSI quote marks (") will transfer.]

Keep HTML file open in browser for reference. Remember to save this new file as a text file (*.txt).

-- Sub-process for clean-up:

Step A/ to replace corrupted searches
Find: %7C%7C">+url:www.nal.usda.gov/wqic/wqdb
Replace with: %7C%7C">+url:www.nal.usda.gov/wqic/wqdb || EXTRACT</a>

Step B/
Check for unique corrupted searches.
Some searches will not end in the unique string starting with "%7C%7C". But these lines will end without the ending anchor tag </a> and break <br>.
Find those searches and edit by appending "|| EXTRACT</a>
" to the line.

-- Change HTML coding to acceptable comma-delimited data:

Step C/
Find: <tt>[6spaces]
Replace with: <tt>[6spaces],"

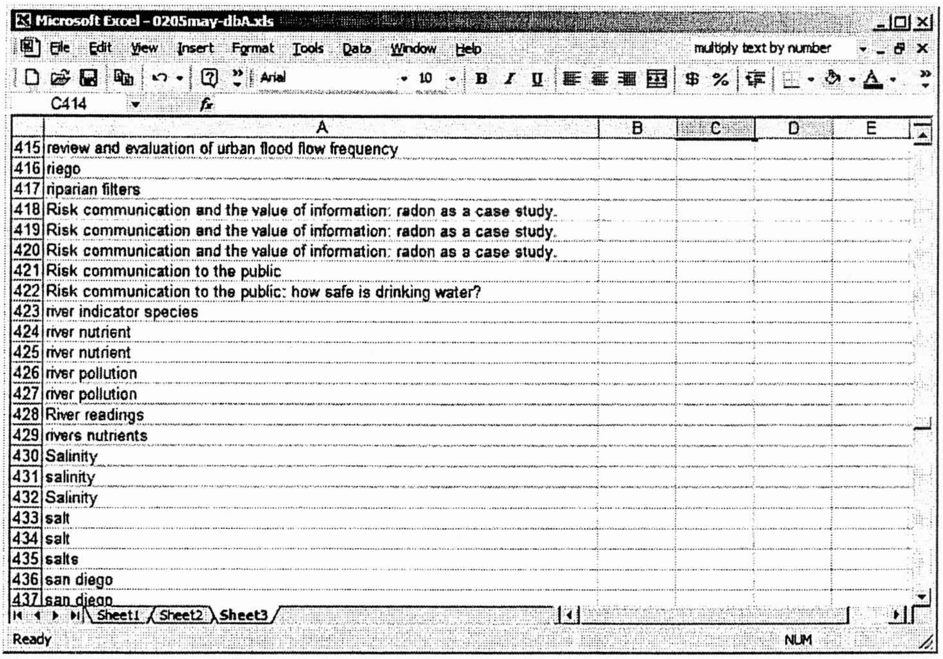
Figure 4: Procedure Section for File Manipulation

Figure 4 is an example of a procedure for HTML file manipulation prior to spreadsheet importing. The code must be made consistent. For example, very long searches had corrupted the original Inktomi file and needed to be fixed in our smaller set in order to preserve any data loss. HTML code was mostly replaced with comma delimited syntax to inform the spreadsheet software where to create columns. Many HTML files can be manipulated in this way. Of course, there are other ways to go about editing for import. However, editing the HTML code guarantees better control over what information gets moved.

Text Analysis

The utility of having log data in HTML files, text and spreadsheets is to provide more power to view and sort the data while preserving access to the actual search which, through these Inktomi-provided pages, can be run automatically by selecting the accompanying link. For example, Initial query term comparisons can be made in a spreadsheet by re-sorting alphabetically on the term column. Results the users may have obtained (depending of course upon the accessible records on the search date) can be checked against the database by performing the search through the link. However, these searches will be re-introduced to the currently active transaction log and should be isolated out whenever possible so as not to skew any results.

In our logs, through quick reviews of the HTML pages, we found short, highly-generic queries were used, such as "water quality" or "water pollution." Many ill-formed searches (poor syntax, typographical or spelling errors, etc.) were sent, as well. And, what we found in rudimentary searches seemed to match earlier evaluations. For example, some researchers examined known-title searches and found them prevalent (Sullenger 1997). Quick review of our logs did show title queries, but not in large numbers.



	A	B	C	D	E
415	review and evaluation of urban flood flow frequency				
416	riego				
417	riparian filters				
418	Risk communication and the value of information: radon as a case study.				
419	Risk communication and the value of information: radon as a case study.				
420	Risk communication and the value of information: radon as a case study.				
421	Risk communication to the public				
422	Risk communication to the public: how safe is drinking water?				
423	river indicator species				
424	river nutrient				
425	river nutrient				
426	river pollution				
427	river pollution				
428	River readings				
429	rivers nutrients				
430	Salinity				
431	salinity				
432	Salinity				
433	salt				
434	salt				
435	salts				
436	san diego				
437	san diego				

Figure 5: Alphabetical Sort with All Instances in Spreadsheet

Figure 5 shows an alphabetical list with potential concepts mappings. The concepts for "rivers" and "nutrients" occurred in several months' worth of logs. Coupled with our understanding of related terms, a new term for our browse topics list might be "nutrients in rivers."

Commonly Occurring Search Phrases and Subject Tools

Term sources are available to confirm how terminology relates or to provide insights into user vocabulary at a conceptual level. The NAL Thesaurus (NALT) is a library priority for indexing AGRICOLA, NAL's database of agricultural literature, and the local online catalog. In some examples we first consulted the NAL Thesaurus for associated concepts. Not finding a term within the thesaurus, we can turn to a variety of databases. Locally, the WQIC Subject File Term List, which forms the organization of the center's vertical files, can also be utilized for additional terms and integration of concepts.

Problems with Approach

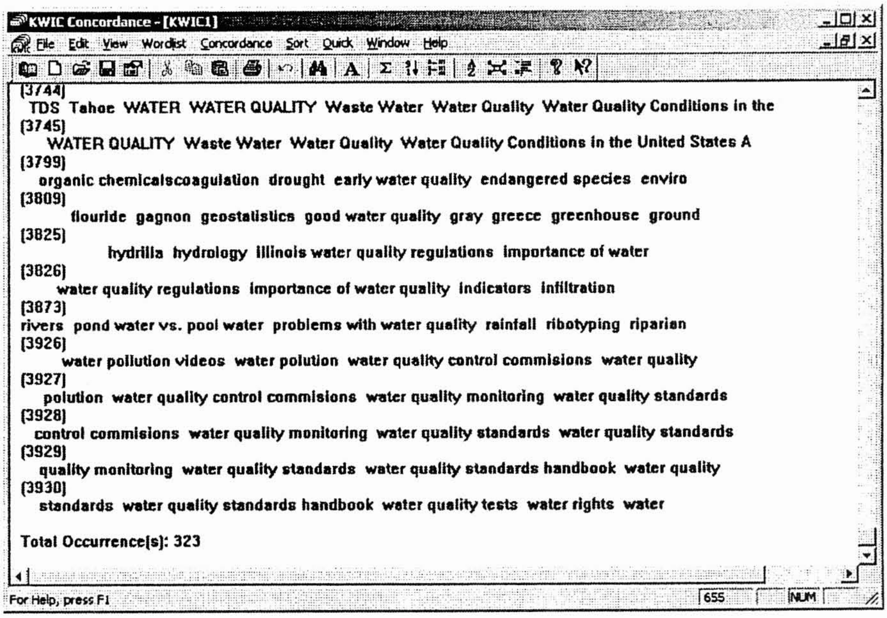
Manipulation and analysis of files is time consuming. Since our initial stages also included procedure writing, manipulation time may be reduced to a manageable level. Also, if an analysis process becomes part of a monthly routine, the time should be reduced. We can also import text files from the text analysis output into the spreadsheet program. This may provide better ways to handle statistics on the frequency of similar terms and other elusive data.

As we learn more about the data and the more common phrases become documented, we will become more familiar with working with these logs. Time expended upon the project should lessen, but still show benefits. But, one of the biggest challenges is to develop a process for tracking changes in user success. At this point, the best we can expect is to document an increase in traffic levels for the search page once we implement the browse topics list.

Future Directions

We sought to go further in our analysis by using text analysis software. Corpora and concordance tools—parts of a computational linguistics tool set—offer novel ways to examine web transaction logs. See Ball (1996), Kita (2000) and Manning (2003) for introductory information on linguistics and Morris (2000) for its application to web search-term analysis.

While individual web search statements are short, transaction logs can contain larger amounts of data. For example, each month the WQIC database search page generated logs with over 600 search queries. We used a free program developed by Satoru Tsukamoto at Nihon University in Japan called KWIC Concordance which is "a corpus analytical tool for making word frequency lists, concordances and collocation tables." With such tools, we sought to utilize these tools in order to more closely examine query term frequency. For example, based upon preliminary term analysis through the spreadsheet, the frequency of a common phrase in context can be studied. See Figure 6 for a keyword-in-context concordance on "water quality." KWIC Concordance can be used to more easily find a phrase if it occurs beyond the first term in a query. Other free software allows for stop word lists. In our analyses, we might use these lists to remove common terms that we have already addressed.


KWIC Concordance - [KWIC1]
File Edit View Wordlist Concordance Sort Quick Window Help
For Help, press F1
655
NUM

```
[3744]
 TDS  Tahoe  WATER  WATER QUALITY  Waste Water  Water Quality  Water Quality Conditions in the
[3745]
    WATER QUALITY  Waste Water  Water Quality  Water Quality Conditions in the United States A
[3799]
  organic chemicalscoagulation  drought  early water quality  endangered species  enviro
[3809]
       flouride  gagnon  geostatistics  good water quality  gray  greece  greenhouse  ground
[3825]
             hydrilla  hydrology  illinois water quality regulations  importance of water
[3826]
     water quality regulations  importance of water quality  indicators  infiltration
[3873]
rivers  pond water vs. pool water  problems with water quality  rainfall  ribotyping  riparian
[3926]
      water pollution videos  water polution  water quality control commisions  water quality
[3927]
    polution  water quality control commisions  water quality monitoring  water quality standards
[3928]
   control commisions  water quality monitoring  water quality standards  water quality standards
[3929]
   quality monitoring  water quality standards  water quality standards handbook  water quality
[3930]
   standards  water quality standards handbook  water quality tests  water rights  water

Total Occurrence(s): 323
```

Figure 6: KWIC concordance run against "water quality"

As stated in our title, we are poised to utilize results transaction log examination to enhance the browse topics list for the new MySQL interface.

Conclusions and Summary

We were still evaluating the log data at press time. However, some generalizations can be made in relation to the analysis of search queries run against our database:

1.) Terms may fall into general subject categories and therefore help WQIC better portray subject material in a browse topics list;
2.) Some special queries may trigger collection development;
3.) Terms used in queries may show levels of subject area understanding characterizing users; and
4.) User syntax errors may show a need to improve online help.

Our research will continue to seek ways to integrate feedback from our users and improve user access. We hope to uncover a variety of forms of feedback and to develop procedures to process the information. Our ultimate goal is to improve WQIC information products.

References

Ball, C.N. 1996. Concordances and Corpora: Tutorial. Available: http://www.georgetown.edu/faculty/ballc/corpora/tutorial.html [Accessed: September 25, 2003]

Bates, M.J. 2003. Improving user access to library catalog and portal information: Task Force Recommendations 2.3, Research and Design Review, Final Report (Version 3). Library of Congress Bicentennial Conference on Bibliographic Control for the New Millennium. [Online.] Available: http://www.loc.gov/catdir/bibcontrol/2.3BatesReport6-03.doc.pdf [Accessed: 2003-09-25].

Cui, H., Wen, J.-R., Nie, J.-Y. & Ma, W.-Y. 2003. Query expansion by mining user logs. *IEEE Transactions on Knowledge and Data Engineering* 15 (4): 1-11.

Fusonie, A. E. 1988. The History of the National Agricultural Library. *Agricultural History* 62(2): 189-207.

Gagnon, S. R. and Makuch, J.R. 2001. Lessons Learned Associating and Developing Specialized Terminology Schema for a Web Database at the National Agricultural Library. Collected Presentations, 16th Annual Computers in Libraries Conference (March 14-16, 2001): 64-71.

Janes, J. 2003. Counting what counts (Internet Librarian column). *American Libraries* 34 (5): 72.

Kita, K. [2000.] Software tools for NLP. Available: http://www-a2k.is.tokushima-u.ac.jp/~kita/NLP/nlp_tools.html [Accessed: October 20, 2003].

Makuch, J. R. and Gagnon, S. R. 2000. The National Agricultural Library's database of online documents covering water and agriculture. Presented at the Third Water Information Summit, Supporting Sustainable Water Resources Management Through Information Distribution, November 3-5, 2000, Miami, Florida. Available at: www.waterweb.org/wis/wis3/presentations/18_Makuch_paper.pdf [Accessed September 12, 2003].

Makuch, J. R. and Gagnon, S. R. 2002. Information access initiatives of the Water Quality Information Center at the National Agricultural Library. Paper and poster presentation at the 5th Water Information Summit, October 23-25, 2002. Ft. Lauderdale, Florida. Available at: www.waterweb.org/wis/wis5/presentations/Makuch.pdf [Accessed September 12, 2003].

Manning, C. 2003. Statistical natural language processing and corpus-based computational linguistics: An annotated list of resources. Available at: http://nlp.stanford.edu/links/statnlp.html [Accessed September 29, 2003].

Morris, R.W. 2000. Knowledge management in the professional services: Lessons from functional linguistics. Proceedings of the 2000 IEEE Engineering Management Society: EMS 2000, Albuquerque, NM (13-15 Aug. 2000): 637-641.

National Agricultural Library. January 2003. The National Agricultural Library: A public resource for the public good. Beltsville, MD: National Agricultural Library.

Pu, H.-T., Chuang, S.-L. & Yang, C. 2002. Subject categorization of query terms for exploring Web users' search interests. *Journal of the American Society for Information Science and Technology* 53 (8): 617-630.

Sullenger, P. 1997. A serials transaction log analysis. *Serials Review* 23 (3): 21-26.

United States Department of Agriculture. September 1994. National Agricultural Library...ensuring and enhancing access to agricultural information for a better quality of life. Washington, D. C.: United States Department of Agriculture.

Disclaimer
The use of trade, firm, or corporation names in this publication is for the information and convenience of the reader. Such use does not constitute an official endorsement or approval by the United States Department of Agriculture, the Agricultural Research Service or the National Agricultural Library of any product or service to the exclusion of others that may be suitable.