The End of Zero-Hit Queries:
Query Previews for NASA's Global Change Master Directory

Stephan Greene, Egemen Tanin, Catherine Plaisant, Ben Shneiderman#
Human-Computer Interaction Laboratory
University of Maryland Institute for Advanced Computer Studies
#Department of Computer Science
{plaisant, greene, egemen, ben}@cs.umd.edu
http://www.cs.umd.edu/projects/hcil

Lola Olsen#, Gene Major*, Steve Johns*
Global Change Master Directory,
#also NASA Goddard Space Flight Center,
*also Hughes STX Corporation
{olsen, major}@gcmd.gsfc.nasa.gov

Abstract
The Human-Computer Interaction Laboratory (HCIL) of the University of Maryland and NASA have collaborated over the last three years to refine and apply user interface research concepts developed at HCIL in order to improve the usability of NASA data services. The research focused on dynamic query user interfaces, visualization, and overview +preview designs. An operational prototype, using query previews, was implemented with NASA's Global Change Master Directory (GCMD), a directory service for earth science data sets. Users can see the histogram of the data distribution over several attributes and choose among attribute values. A result bar shows the cardinality of the result set, thereby preventing users from submitting queries that would have zero hits. Our experience confirmed the importance of metadata accuracy and completeness. The query preview interfaces make visible problems or holes in the metadata that are unnoticeable with classic form fill-in interfaces. This could be seen as a problem, but we think that it will have a long-term beneficial effect on the quality of the metadata as data providers will be compelled to produce more complete and accurate metadata. The adaptation of the research prototype to the NASA data required revised data structures and algorithms.

1- Introduction

Soon, users (scientists, teachers, students, etc.) will be able to access NASA's Earth Observing System Data Information System (EOSDIS) to retrieve earth science data from hundreds of thousands of datasets containing pictures, measurements, or processed data, from centers around the country. Data about the datasets (called metadata) is available and is searched as a preliminary step before retrieving the huge datasets. Standard EOSDIS metadata includes spatial coverage, time coverage, type of data, sensor type, campaign name, level of processing, etc. Existing EOSDIS prototypes use classic form fill-in (or form-completion) interfaces. They allow users to search the already large holdings but zero-hit queries are a problem, and it is difficult or impossible to estimate how much data are available on a given topic and what to do to increase or reduce the result set.

More robust querying interfaces are needed for a system that will support intermittent users with extensive or limited computer experience. Users will come to the EOSDIS with problems that vary according to several factors, including: how well-defined the solution to the problem is (e.g., ranging from simple facts to interpretations for complex phenomena) and how well-defined the problem is in the information seekers mind. Such information systems must accommodate different levels of experience with the content area, with the information system itself, and with information seeking in general.

Our approach is to employ overview + preview representations of metadata (Greene et al., 1997) that allow users to rapidly and dynamically eliminate undesired data sets, while at the same time previewing result set sizes to avoid zero-hit queries (Doan, Plaisant and Shneiderman, 1996; Doan et al., 1997; Plaisant, Bruns et al.97). The reduced volume of the metadata allows queries to be previewed and refined locally before they are submitted over the network. This two-step approach of query preview and query refinement extends the technique of dynamic queries [Ahlberg, Williamson and Shneiderman, 92;

Shneiderman, 1994] to large distributed information systems.

In this paper we describe our original design and prototype, then discuss the technology transfer process from our research prototype to an operational prototype by reviewing the issues and problems encountered, the options examined, and the solutions adopted.

Related work

The growth of the internet and the formidable push to make information available online or digitize existing materials has been only partially matched by the development of tools to assist users in their tasks. Generic improvements to the browsing of web pages have been proposed using a book metaphor [Card et al., 1996], "decks of cards" metaphor [Brown and Shiller, 1996], or tiled elastic windows [Kandogan and Shneiderman, 1996].

Search interfaces are being refined. Research is being conducted to improve the search engines which rely on the analysis of machine readable textual materials, while powerful tools are created for specialized formats (e.g. Informedia for videos [Watclar et al, 1997]). Other projects combine the text search with visualization methods to display the result of the searches (e.g. tilebars [Hearst et al., 1996] or Envision [Heath et al., 1995]). When no metadata or text is available to be searched, browsing becomes essential (Plaisant, Marchionini et al, 97; Marchionini et al, 97, Nation et al., 97)

[ACM special issue 1995] and [IEEE special issue 1997] provides several examples of large digital libraries projects, and shows that the focus of many projects remains on digitization and infrastructure. Comprehensive evaluations will be the key to understanding the benefits of new designs [Hill et al,97; Marchionini and Crane, 94].

2-From research prototype to operational prototype: issues and solutions adopted

2.1 Original research prototype

After a rudimentary interface mockup in Visual Basic [Doan, 96] we implemented a prototype in Tcl-Tk and then in Java when it became available. The research prototype used simulated data, randomly generated to illustrate our concept for a possible EOSDIS interface.

Query Preview

Our original query preview prototype features a single-screen overview of all earth science data sets available from a designated information service provider (Figure 1). The data are characterized with three high-level attributes: location, temporal coverage, and topic content. For each attribute value in the user interface, the number of available data sets bearing that attribute value is shown (Figure 1a). This gives users an overview of the distribution of the available data sets. By clicking on combinations of attribute values, users can prune to reflect only those data items that satisfy selected attributes (Figure 1b). When the preview indicates a reasonable result set size, users submit the query and move the query to a refinement phase.

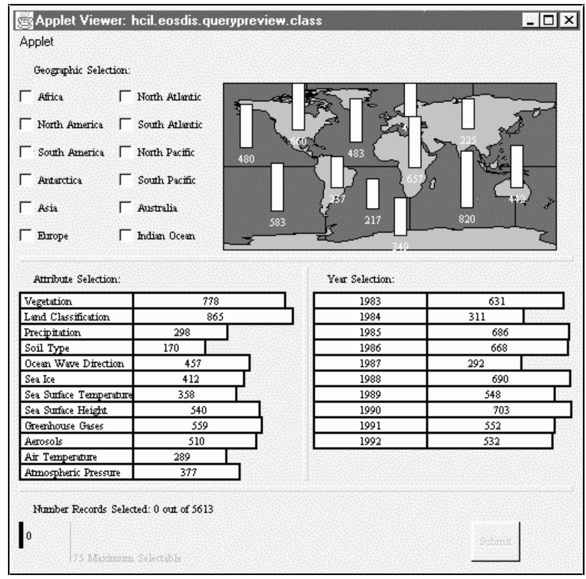

Figure 1a: Research prototype: The query preview screen displays summary data on preview bars. Users learn about the holdings of the collection and can make selections over a few parameters.

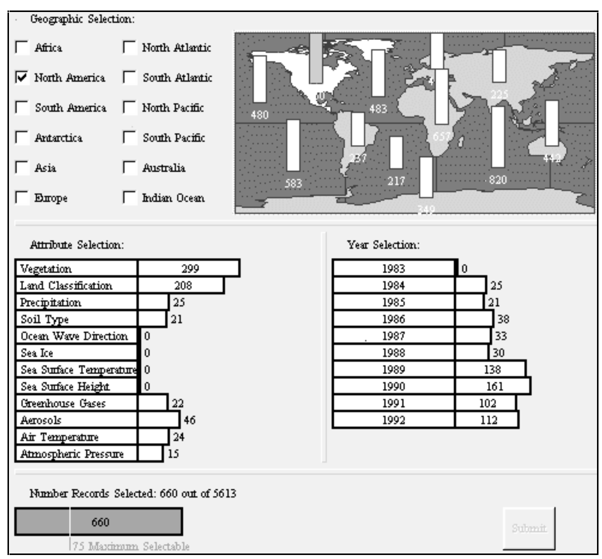

Figure 1b: Research prototype: Following the principles of dynamic queries, the preview bars are updated immediately (in less than 100msec.) when users select an attribute value (here North America). The result bar at the bottom shows the total number of selected data sets.

Query Refinement

During the refinement phase, the user interface operates on the result set of the preview phase and shows all the metadata. In our research prototype, each data set is represented as a line in an overview display, whose axes are the size (vertical axis) and the time period (horizontal axis) of the data sets respectively. Users further refine the query by selecting more precise values for the parameter, sensor, platform, project, data archive centers, processing data level, etc. (Figures 2a and 2b). When they think the query is sufficiently refined, users submit the query and retrieve full data set records and can be directed to the web location where the data are available.

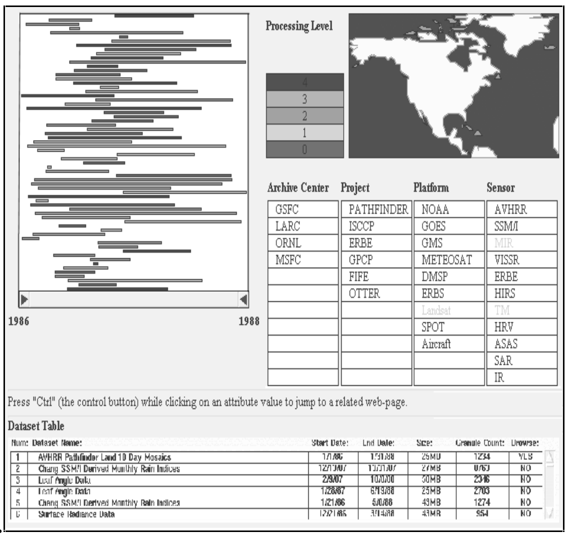

Figure 2a: In the research prototype of query refinement, users can browse more information about individual data sets. The result set can be narrowed again by making more precise selections on more attributes.

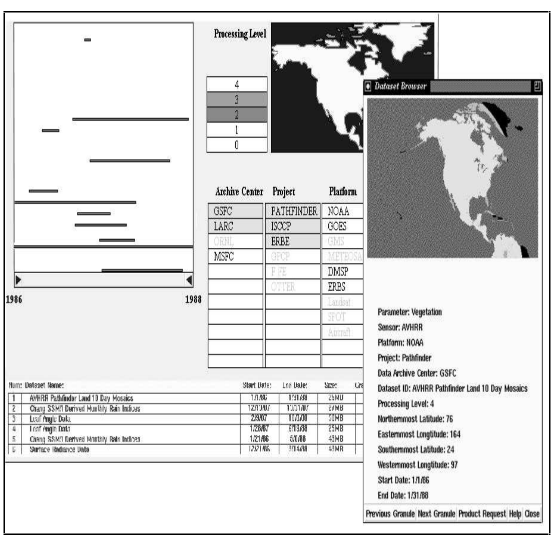

Figure 2b: Here the query has been refined by selecting 2 archive centers, 3 projects and 2 processing levels. More filtering could be done by zooming on the timeline or on the map. The timeline overview and the data set table reflect the remaining data sets. Details and sample images can be downloaded from the network (window on the right) before the long process of ordering the large data sets.

Evaluation of the prototypes

The original research prototype was informally evaluated during one of the prototype workshops organized by NASA and Hughes. It received positive feedback and showed that users understand the interface without training. A controlled experiment evaluated the benefits of query previews[Tanin et al, 1997]. Twelve computer science students searched a database of films either with a form fill-in interface or with a query preview followed by a form fill-in. Results indicated that the query preview led to 50% faster task completion times when query preview attributes were related to the tasks, and higher subjective satisfaction.

2.2 Development of the Operational Prototype for the GCMD

We now describe the operational prototype and the issues involved in its development for the query preview and then for the query refinement. The prototype is implemented in Java, based on code from the final version of the research prototype, and uses a series of Perl scripts to generate the query preview tables and the result html pages. In appearance the query preview operational prototype is very similar to the research prototype, but several underlying features had to be redesigned. The query refinement is very different from the research prototype as it was replaced by a second phase of previewing with a second set of attributes followed by a more traditional result screen using frames.

Query Preview Issues

ATTRIBUTES AND ATTRIBUTE VALUE CHOICES

An early task was to capture the GCMD database in a manner appropriate for representation in the interface. The three attributes of location, temporal coverage, and topic content, as used in the prototypes, were confirmed by GCMD earth science specialists and database administrators as the most salient and universal from among all the attributes typically associated with earth science data resources. These were kept as the main attributes for the interface. For each of the three attributes, the large set of attribute values was aggregated up to around ten high-level attribute designations. A domain expert selected date ranges, high level topics and regions (Figure 3). The granularity of the attribute values chosen for use in the interface is deliberately crude, in order to be able to represent vast amounts of diverse scientific data in a single overview. The careful choice of appropriate attribute values and attribute value aggregations represents a significant portion of the design effort.

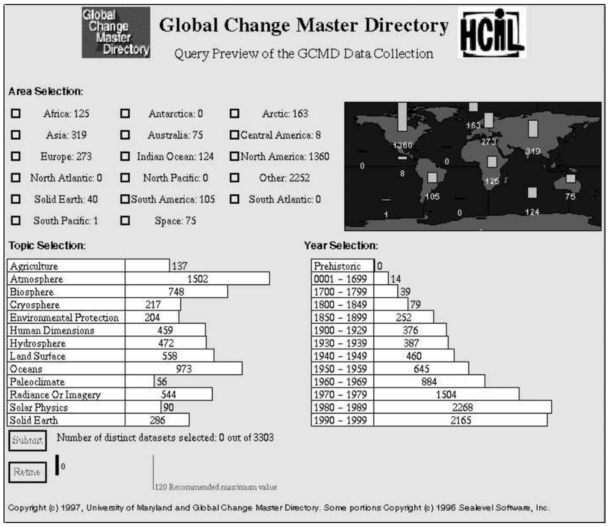

Figure 3: First operational query preview interface for GCMD

QUERY PREVIEW ALGORITHM IMPROVEMENTS

Our original plan was to duplicate data sets as needed to account for multiple attribute values (e.g. a data set covering North and South America would be counted twice within the attribute totals for area). Our assumption was that there would be a maximum of 20 to 30 per cent duplication. At this rate, duplication would be barely noticeable by users reviewing the values on the bars. However, we discovered that on average, each data set had to be duplicated seven times, which made the problem overly apparent in the interface (e.g. the total result bar could be 150 data sets, while individual bars showed up to 1000 data sets). Because GCMD has a reasonable number of data sets (around 5000) we were able to correct this problem by listing the dataset IDs in the preview table and calculating the exact number of distinct data sets after each user selection. The additional calculations of this hybrid technique first introduced unacceptable time delays, and required improvements to the data structures and algorithms used to determine exact data set counts. The bars are now adjusted in less than 100ms on a PC when users make a selection, easily meeting the requirements of dynamic query interfaces.

QUERY PREVIEW TABLE GENERATION AND DATABASE ADJUSTMENTS

An input data generation module was written in Perl. It sends SQL queries to the GCMD data server and generates a series of query preview tables and a master file describing those files. These files in turn generate the interface. One relation was added to the database to reflect the mapping between data attributes and preview attributes.

CORRECTION OF VALUES IN DATA

Some inconsistencies and missing values were discovered during the process of both generating the data tables and visualizing them in the new user interface. The query preview design crucially provides a complete overview of a database. Implementing it can thus be expected to reveal problems that would likely remain unnoticed in a traditional text interface. This brings benefit to database administrators as well as users attempting to understand the data for the first time. Until the metadata is repaired, a temporary method to deal with those problems is to add ad-hoc attribute values such as "not specified".

INTERFACE DESIGN CHANGES

Numerous adjustments were needed in the interface to accommodate the structure and distribution of GCMD data. Overall, the data sets were more numerous, more intricate, and less uniformly distributed than the simulated data used in the early prototypes. First, the preview bars were changed to use logarithmic scales when necessary, and the map and other display widgets were generalized to accommodate dynamic size requirements and layout changes.

Once it became clear that this interface was feasible for GCMD and was likely to be made available to users, we conducted a final detailed review of the user interface and found improvements in the area of layout, consistency, feedback and window coordination.

- Headers were revised and made consistent in the preview, refinement and results screens.
- “Help” and “About…” buttons were added in each screen in a consistent location. They lead to normal HTML pages, which can be edited easily. A short instruction sentence was added the preview and refinement screens.
- The result bar was modified to also use a non-linear scale and was labeled appropriately. The “recommended number “ was removed since there was no basis for it in this implementation. The result bar was changed to yellow to match the attribute preview bars.
- The labels of the buttons were made more descriptive: e.g. “refine with 3 other attributes”
- The “area” checks boxes (see Figure 3) were replaced with a 3rd set of bars for the location (Figure 5). This was found to be more consistent. Even though some of the areas cannot be shown on the map we decided to keep the map for its visual properties giving users an immediate way to deal with most area selections.
- The attribute labels were changed to mixed case (instead of upper case, which is known to be hard to read). But this should really be done at the data level, not the interface level.
- The generation date of the preview table is mentioned next to the total number of datasets in the result.
- To facilitate the browsing of the results a frame version was prepared.
- A reset button was added.

Most of those changes can be seen when comparing:

- Figure 1 (research prototype)
- Figure 3 (first operational prototype)
- Figure 5 (final version as of November 1997)

RESULTS GENERATION

The query preview returns the list of records corresponding to the query. Because of the relatively small number of records in GCMD, we facilitated rapid result generation by including the list of record identifiers in each cell of the query preview table. This technique is not appropriate for holdings with constant updates and a real search will have to be performed. The HTML pages showing results sets are generated by another Perl script that sends an SQL query to the database to retrieve full record names.

Query Refinement Issues

Returning the list of data sets immediately after the query preview phase is only effective if the number returned is small. Otherwise, a refinement phase is needed. Implementing an elaborate query refinement interface as designed in the original research prototype was found to be a significant implementation effort. Since our goal was to have an operational prototype by the end of 1997 we examined four alternatives for the refinement phase:

- provide a second preview step with 3 additional attributes
- include two or three more attributes in the query preview (to increase its discriminating power)
- connect to the existing form-based interface
- explore the use of off-the-shelf systems (such as Spotfire, a visualization and data mining product based on HCIL research—see www.ivee.com)

The first option was selected and will be described in details below. The second option (i.e. including more attributes) was rejected because it would generate a large query preview table that would lead to long loading time and slow preview updates. The third option (connecting to existing form-based interface) was judge to have a high implementation cost for little return. The use of Spotfire was rejected because no Java version was then available.

The first option was implemented rapidly and provides a second preview step with 3 other attributes (Figure 4). This progressive refinement, where overviews lead to previews of subsets and in turn serve as overviews for more fine-grained objects illustrates how interactive overviews and previews can scale to large, complex digital information spaces. Moreover, it can be implemented with significant reuse of existing software (for GCMD a single applet is used for query and refinement, and the data in the table triggers changes in the interface).

A domain expert chose the attributes. They are: data center, source and sensor.

There were no major problems with the Java interface development but once again the development of this interface made visible underlying problems of metadata incompleteness.

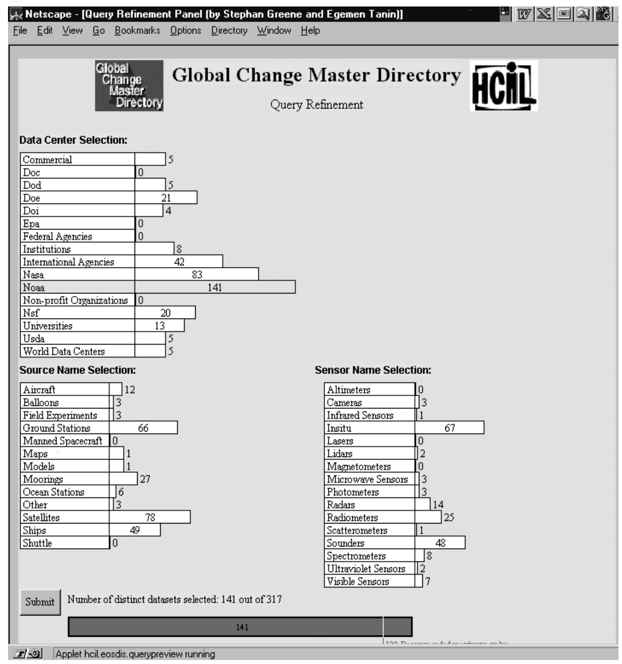

Figure 4: First query refinement design, with three additional attributes.

Maintenance issues

The query preview approach relies on the availability of up-to-date query preview tables.

Those tables are generated in less than an hour. They can be generated at regular intervals or whenever new data sets are added. The date of last update must be available to users.

Providing alternative preview tables

In the early phase of the GCMD project we designed an interface offering alternative query preview tables (e.g. for different science categories such as ocean science or atmospheric science, or data origin such as GCMD vs. CIESIN). To simplify the GCMD interface we chose to provide only one table but the Java applet is still capable of handling several tables, providing buttons for them at the top of the screen to select alternative tables.

2.3 Final version

The GCMD staff has taken over responsibility for completing the final version and making it publicly available. Figures 5-8 show the latest version of the screens and Figure 9 summarizes the architecture of the system implementation.

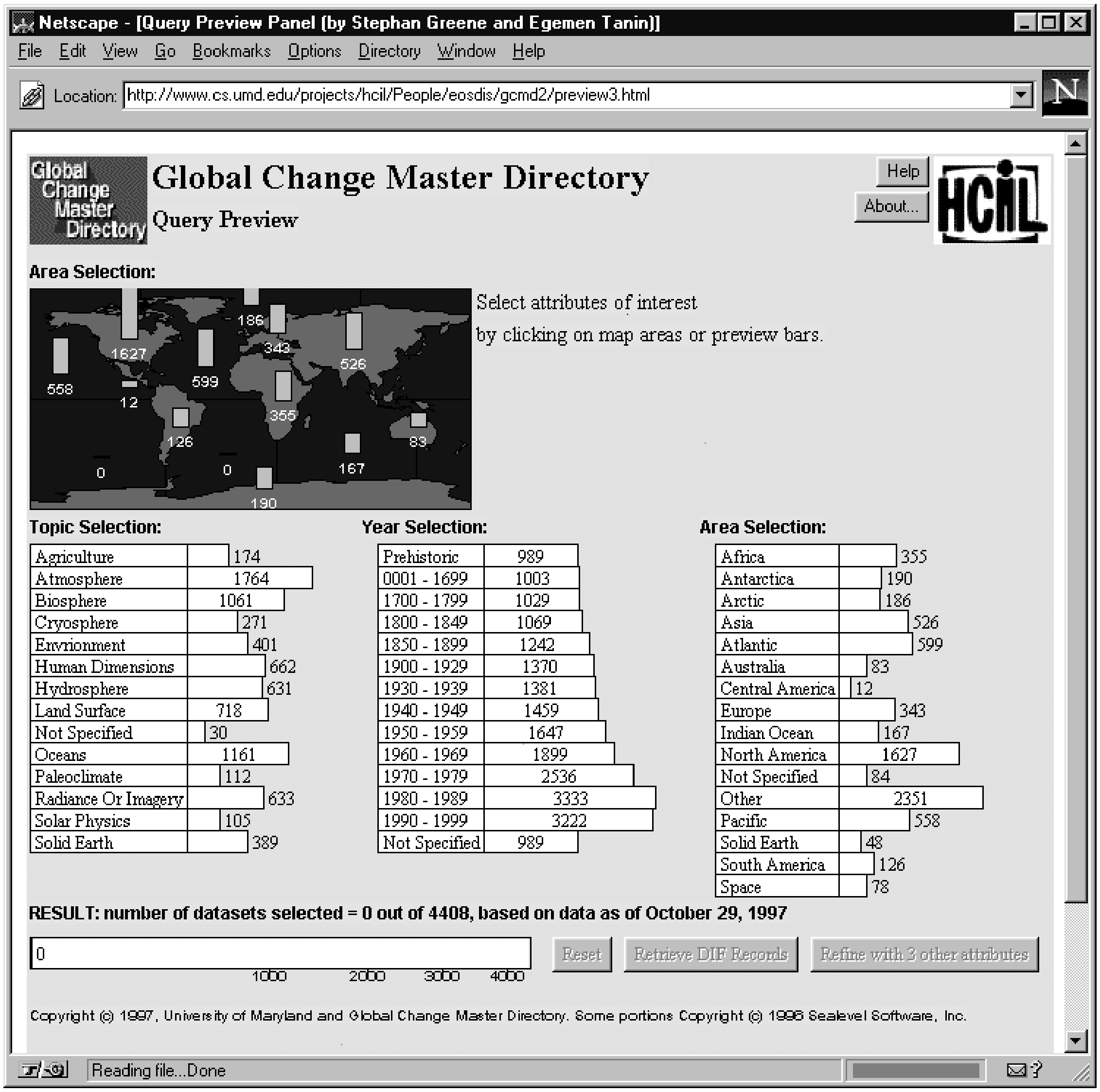

Figure 5a: The query preview screen displays summary data on preview bars. Users learn about the holdings of the collection and can make selections over three parameters.

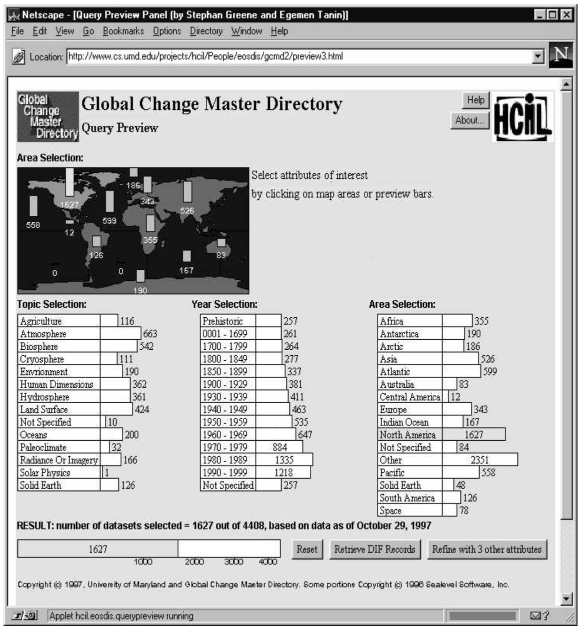

Figure 5b: Following the principles of dynamic queries, the preview bars are updated immediately (in less than 100msec.) when users select attribute values (here North America). The result bar at the bottom shows the total number of selected data sets.

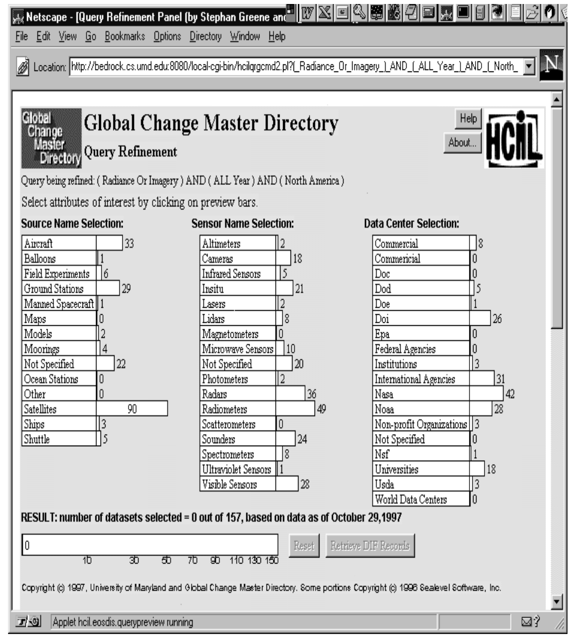

Figure 6a: The query can be refined with three other attributes.

Figure 5c: An additional attribute value is selected.

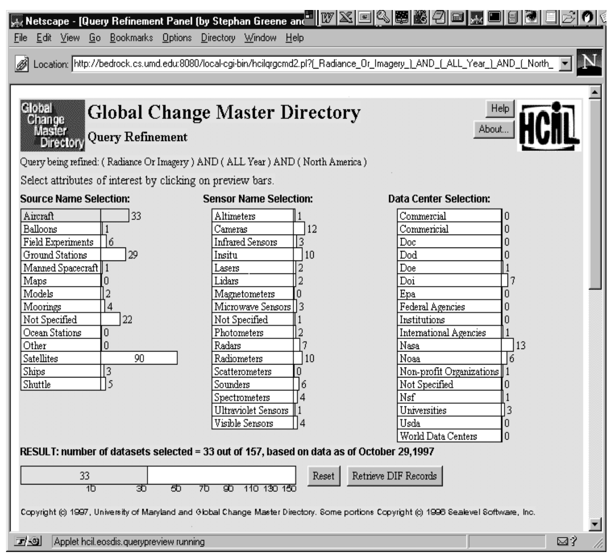

Figure 6b: Refinement by selecting a Source, reducing the results to 33 datasets.

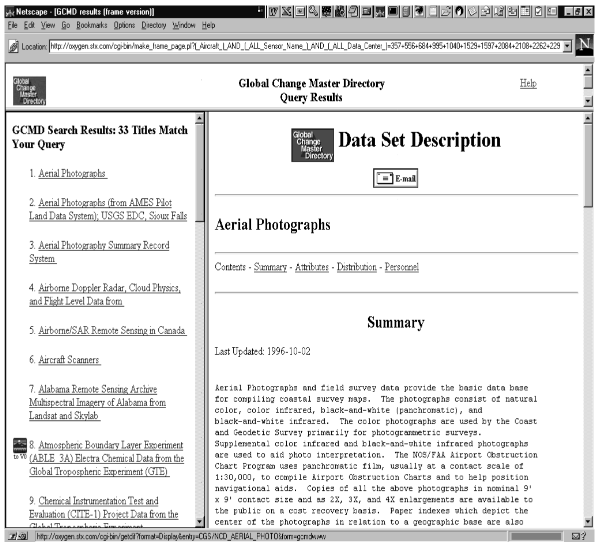

Figure 7: Results are displayed in a frame window to facilitate the browsing of the records. Links to the NASAV0 interface are available when an inventory search is possible for the dataset (i.e. the search can be continued within the dataset).

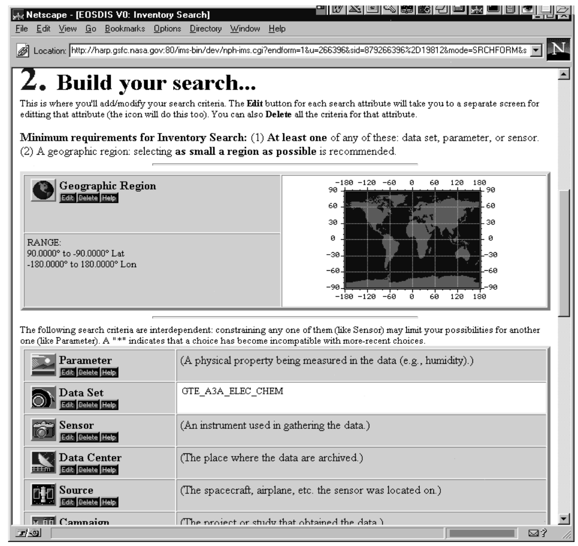

Figure 8: When clicking on the NASA V0 logo next to the dataset name, users can jump directly to a V0 inventory search. Here we see that the name of the dataset was passed to the V0 search. Other values such as location or time could be passed as well.

3- Lessons Learned and Future Research Possibilities

We believe that query preview and query refinement interfaces have substantial benefits to users. This project demonstrated that the concepts are feasible in a large operational system, such as the EOSDIS directory environment. Consensus was reached on attributes and values selection and performance is satisfactory (speed issues were all solved). Our experience confirmed the importance of metadata accuracy and completeness. The query preview interfaces make visible any problems or holes in the metadata that are unnoticeable with classic form fill-in interfaces. This could be seen as a problem but we think that it will have a long term beneficial effect on the quality of the metadata as data providers will be compelled to produce more complete metadata.

In going from a research prototype to an operational system, our data management algorithms and graphical user interface mechanisms all required modification to accommodate unanticipated data characteristics. Our next challenge is to scale up the software architecture to accommodate much larger and more varied data collections.

Our work illustrates that the number of queries processed may not be the most appropriate measure of the success of a query system. The query preview interface will most likely reduce the number of queries sent to the database server as users will be able to refine their query formulation locally before submitting it. A high number of queries might very well indicate that users are not finding what they need. Measuring the average size of returned result sets for queries might be a more reasonable benchmark. And measuring the proportion of zero-hit queries and mega-hit queries to the total number of queries submitted would also begin to more effectively measure the success of user interface search systems.

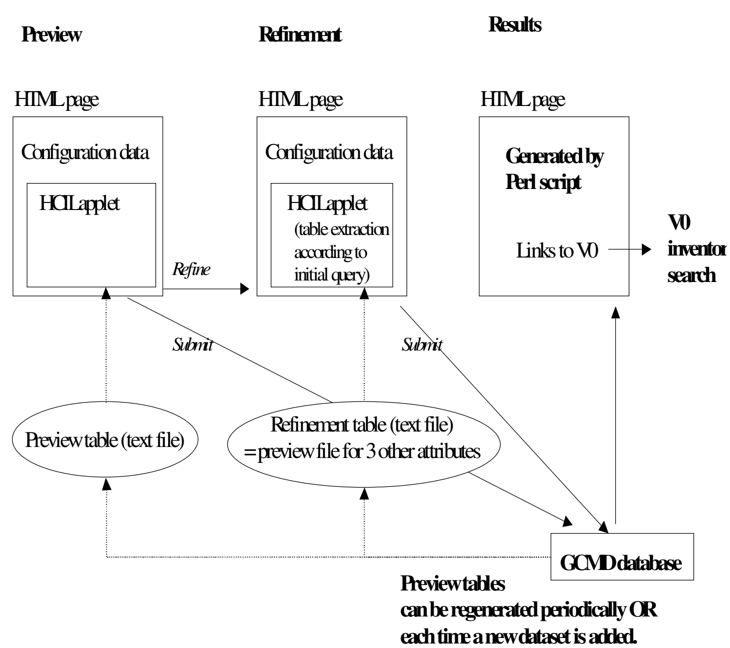

Figure 9: Summary of the system architecture

Acknowledgements

This work is supported in part by NASA (NAG 52895 and NAGW 2777) and by the NSF grants NSF EEC 94-02384 and NSF IRI 96-15534.

Project website:

Project URL at HCIL (early prototypes, local GCMD prototype, papers and reports)
http://www.cs.umd.edu/projects/hcil/Research/1995/dq-for-eosdis.html

Related reports and publications

Ahlberg, C., Shneiderman, B., Visual Information Seeking: Tight coupling of dynamic query filters with starfield displays, *ACM CHI '94 Conference Proc.* (Boston, MA, April 24-28, 1994) 313-317.

Brown, M, and Shillner, R., The deckscape Web browser, *ACM CHI '96 video*, Summary in *CHI' 96 Companion*, 418-419, ACM New York.

CACM 95, Special issue on Digital Libraries, *Communications of the ACM*, April 1995, **4**, 29-39, ACM New York.
Card, S., Robertson, G., York, W., The WebBook and the Web Forager: An Information Workspace for the World-Wide Web, in *CHI'96 Proceedings*, ACM, New York, 1996.

Computer 97, Special issue on Digital Libraries, *Computer,* May 1997, IEEE.

Doan, K., Plaisant, C., Shneiderman, B., Query previews in networked information systems, *Proc. of the Third Forum on Research and Technology Advances in Digital Libraries, ADL '96* (Washington, DC, May 13-15, 1996) IEEE CS Press

Doan, K., Plaisant, C., Shneiderman, B. and Bruns, T., Interface and Data Architecture for Query Preview in Networked Information Systems, *University of Maryland technical report*, CS-TR-379, UMIACS-TR-97-42, ISR-TR-97-57 (May 1997)

Greene, S., Marchionini, G., Plaisant, C., Shneiderman, B.
Previews and Overviews in Digital Libraries: Designing Surrogates to Support Visual Information-Seeking, *University of Maryland Technical Report* CS-TR-3838

Hearst, M., Tilebars: visualization of term distribution information in full text information access. *Proceedings of CHI'95,* ACM New York, 1995.

Heath, L., et al., Envision: a user centered database of the computer science literature. *Comm. Of the ACM,* April 1995, pp. 52-53

Hill, L., Dolin, R., Frew, J., Kemp, R., Larsgaard, M., Montello, D., Rae, M., and Jason S., (1997). User evaluation: Summary of the methodologies and results for the Alexandria Digital Library, University of California at Santa Barbara. *Proceedings of the 60th Annual Meeting of the American Society for Information Science* (November 1997) Information Today (Medford, NJ) 225-243.
http://www.asis.org/annual-97/alexia.htm

Kandogan, E., and Shneiderman, B., Elastic Windows: A Hierarchical Multi-Window World-Wide Web Browser, *Proc of UIST' 97 conference*, ACM, New York

Marchionini, G., Crane, H., Evaluating hypermedia and learning: methods and results from the Perseus project. *ACM Transactions on Information Systems, 12,* 1 (January 1994) 5-34.

Visualizing websites using a hierarchical table of contents browser: WebTOC
in *online Proceedings of 3rd Conference on Human Factors and the Web*, Denver, CO, June 1997.
http://www.uswest.com/web-conference/proceedings/nation.html

Plaisant, C., Bruns, T., Doan, K, Shneiderman, B., Query Previews in networked Information Systems. *CHI 97 Technical Video program (March 22-27, 1997 - Atlanta GA)* ACM, New York. Also includes a 2 page summary in *CHI 97 Companion (March 22-27, 1997 - Atlanta GA)*, ACM New York.

Plaisant, C., Marchionini, G., Bruns, T., Komlodi, A., Campbell, L. (1997) Bringing Treasures to the Surface: Iterative design for the Library of Congress National Digital Library Program. *Proceedings of CHI '97,* March 1997, ACM New York. 518-525.

Marchionini, G., Plaisant, C., Komlodi, ., Interfaces and Tools for the Library of Congress National Digital Library Program. *HCIL Technical report (November 1997*)

Rao, R. Pedersen, J, Hearst, M, Mackinlay, J., Card, S., Masinter, L., Halvorsen, P., Robertson, G., Rich interaction in the digital library, *Communications of the ACM*, April 1995, **4**, 29-39.

Shneiderman, B., Byrd, D., and Croft, B. (1997). Clarifying Search: A User-Interface Framework for Text Searches. *D-Lib Magazine, January 1997.*

http://www.dlib.org/dlib/january97/retrieval/01shneiderman.html

Smith, T. A digital Library for geographically referenced materials, *Computer,* May 1997, 54-60, IEEE.

Tanin, E., Beigel, R. and Shneiderman, B., Design and Evaluation of Incremental Data Structures and Algorithms for Dynamic Query Interfaces, Proceedings *of the 1997 IEEE Information Visualization workshop,* pp. (Oct. 1997)

Tanin, E., Lotem, A., & Haddadin, I. 1997. Evaluation of Query Previews: User Preference and Performance.

Manuscript: www.otal.umd.edu/SHORE

Wactlar, H., Kanade, T. Smith, M., Stevens, S., Intelligent access to digital video: Informedia project. *IEEE Computer,* May 1997, 46-52..

Williamson, C., Shneiderman, B., The dynamic HomeFinder: Evaluating dynamic queries in a real-estate information exploration system, *Proc. ACM SIGIR `92* (Copenhagen, June 21-24, 1992) 338-346.