10/10-15-91 JSO (1)

LBL-30953
UC-605

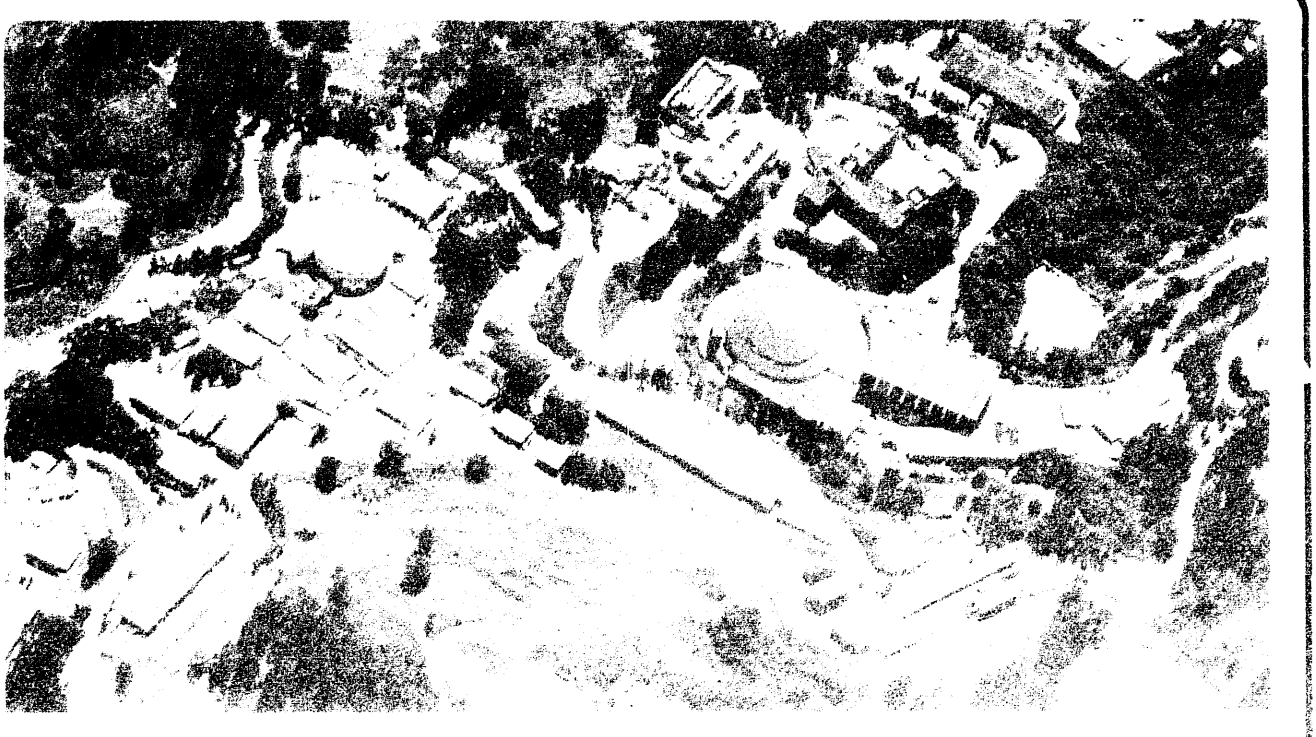
Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

Information and Computing Sciences Division

MetaBrowser: A Combined Browsing, Query, and Analysis Tool

A. Shoshani and E. Szeto

April 1991



Prepared for the U.S. Department of Energy under Contract Number DE-AC03-76SF00098

DISCLAIMER

LBL--30953

DE92 000714

MetaBrowser: A Combined Browsing, Query, and Analysis Tool

Arie Shoshani and Ernie Szeto
Information & Computing Sciences Division
Lawrence Berkeley Laboratory
University of California
Berkeley, CA 94720

April 1991

This work was supported by the Director, Office of Epidemiology and Health Surveillance, Office of Health, Office of Environment, Safety and Health of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

MetaBrowser: A Combined Browsing, Query,
and analysis tool

1. Basic requirements

The MetaBrowser design is based on the premise that scientists should not be forced to learn new languages or commands for finding the data they are interested in and for selecting subsets of the data for further analysis. Furthermore, there should be a single system that permits browsing, query, and analysis of the data, so that the scientist does not have to switch between systems. The current version of the MetaBrowser was designed for the DoE CEDR (Comprehensive Epidemiological Data Resource) project, but the same principles can apply to other scientific disciplines.

Browsing and query should be combined. It is quite natural for a user to explore the information in the database before deciding what subset of the data to select for further analysis. In general, if there is a large number of datasets (i.e. databases) in the system, then the user would want to find out information about the various datasets (called *metadata*), before choosing one or more datasets for further exploration. Thus, a *metadatabase* that holds information about datasets in the systems must exist.

2. The federated metadata

Since each dataset has its own metadata, we refer to the combined metadata for all datasets as the *federated metadata*. This is illustrated in Figure 1. Each dataset has *structural* and *descriptive* metadata associated with it. The structural metadata is often referred to as the database schema definition. This includes the files (or relations) names, and the name, type, and length of each column (attribute). The descriptive metadata consists of information about the dataset, such as who created it, the history of modifications made to it, relevant citations, and what kind of data it contains. The descriptive metadata is domain specific. Thus, in CEDR, it has information about the cohort it contains, such as "white males over 40", and the type of radiation exposure figures it contains, such as "internal exposure to tritium".

As shown in Figure 1, the Federated Metadatabase consists of the structural and descriptive metadata information of *all* the datasets.

3. User interaction model

The MetaBrowser was designed as a four step process, discussed below. In a typical scenario, the scientist will proceed as follows:

a) **Find a dataset of interest.** The scientist explores the federated metadata in order to find datasets of interest. This can be done by browsing the information about the datasets and/or by issuing a query for searching the datasets with certain properties.

b) **select a subset.** After selecting a dataset, the scientist may explore further the information about that selected dataset and/or specify a query for selecting a subset. At this point, he/she

may wish to go back to step a) and look for other datasets, or proceed to the next step.

c) **Inspect the selected subset.** Once a subset has been selected, the selected data (or part thereof) are displayed for inspection. Various visualization modules can be used depending on the application. At a minimum, the user should be able to see instances of the selected subsets in a tabular form, and browse over them. At this point, the user should be able to go back and modify the query (for example, add/remove columns), or proceed to create a file containing the subset selected for further analysis.

d) **Invoke the analysis software.** The analysis software is usually a software package or (special purpose) application that the scientist is familiar with. Thus, in order to permits a smooth transition between the data selection subsystem and the analysis subsystem, data conversion needs to be supported. The data conversion may be as simple as reformatting, or may require restructuring of the data. Typically, only a single file is passed on to the analysis software. The analysis software is application dependent, although some general purpose packages can be used in various applications. In CEDR, statistical packages, such as SAS or S, are the preferred analysis tools for Epidemiologists.

Note that the datasets, the metadata for the datasets, as well as the federated metadata are managed by the underlying (relational) data management system.

4. The user interface

There are two guiding principles to the user interface.

a) **Object presentation.** The information should be presented to the user independently of any particular physical or system organization. While interacting with the system, the user should only be aware of objects, such as people, departments, cities, etc., rather than relations, attributes. Furthermore, the selection and manipulation of data items should be in terms of these object, rather than system dependent manipulation operators, such as "join" or "project". Thus, in our context this means that users are not required to learn a new language, such as SQL. Consequently, it is necessary to provide a translation between to object level view presented to the user and the data management system used to manage the data.

b) **Self-guiding.** The user should be completely guided by the interface, without any preknowledge of the system. Ideally, the user should be able to interact with the system without any instructions, and quickly find the datasets of interest, create subset, inspect the results, and generate a file for further analysis. A "guided tour" or an on-line tutorial may also be provided for first time users.

c) **Simplicity.** Simplicity is achieved if the user is exposed to the minimum number of concepts or construct in order to understand what's presented, or express desired operations. In our context, we do not expects users to know a query language, such as SQL. Rather, the user is guided by the information on the screens on how to browse and express queries. We have chosen to limit the interaction to windows with very few concepts (scroll lists and buttons) which are introduced in the initial screen (see Appendix 1). At this time we are not using

other graphical techniques, although we are not ruling out their usefulness in the future. Also, customization of terms that captures the application specific terminology is helpful in increasing clarity and understandability. Thus, in the examples below, we use "variable" rather than "attributes" or "columns", since such terms are more meaningful to epidemiologists.

5. Implementation details and examples

This prototype system was developed especially for the CEDR application. Thus, the screens shown have been tailored for a limited set of goals.

This prototype uses a window display methodology, with scroll lists to display and select data items. Buttons are used to invoke desired functions. Only a single mouse button is required for manipulating the screens.

Appendix 1 shows the main introductory screen. This single page guides the user as to the purpose of the system and how to proceed. The system has gone through several iterations of refinement in order to eliminate areas that seemed confusing to users.

The last pages of this document contain a series of screens that illustrate the system as the user sees it. The first screen, the "Startup Screen", introduces the information from Appendix 1. Clicking on the "Search for Datasets" button brings the next screen, "Datasets Selection". In this screen, we show that certain "subject terms" and "sites" were selected. The next two screens show details about the selected "subject terms" and "sites" respectively, as a result of clicking on the corresponding buttons. Note that the term "show details" is used here rather than "show attributes", as this was more clear to users.

At this point the user is interested in finding the datasets that qualify under the conditions specified (i.e. "subject terms", "sites", and "years"). The user proceeds by clicking on "Select Datasets", and gets the next screen, labeled "Dataset Selected". Note that this amounts to a query with three selection conditions. The user is still browsing, so a particular dataset is selected, and "show details" is clicked. The next screen "Dataset Information" is generated, with details about this dataset, its origin, a brief description, etc. More detailed information is available, by clicking on the "Sites", "Variables", and "Bibliography" buttons. The "variables" and "bibliography" details are shown in the next two screens. Note that when "variables" are shown more details about them can be displayed as shown. A similar display is available for "bibliography".

Assuming that the user is satisfied with this dataset, he/she proceeds to a subset selection phase, by going back to the "dataset selected" screen, and clicking on the "Query Selection" button. This produces the "Query Selection Output Variables" screen, where the user can select several variables of interest. The order in which the items are selected is the order of the output that will be produced. This is indicated under the "Output Order" label on the screen. Clicking on the "Select Sort Variables" button produces the next screen "Sort Selection". Here the user can specify the sort order of the output that will be produced. It is sometimes desirable to have multiple sort elements: primary, secondary, etc. We show four such variables selected. The notation "A" or "D" next to the sort order stand for "ascending" and "descending" which were specified by the user in a pop-up window, not shown here.

Now the user wants to retrieve and inspect the results. Clicking on the "proceed to

Retrieve" button generates a query that produces the next screen "retrieved Data". Note that the user may specify the number of rows to be displayed (not shown here), so as to reduce access time for large datasets in the data inspection phase. Also, if the user changed his/her mind, the Abort Data retrieval" button can be clicked to abort. During inspection of the data (scrolling rows) the user can display each row individually (see next screen "row selected"). This is handy in case that the number of attributes is large. It is also possible to display the information on a single variable in case that the user wants to review its content (see the next screen "Column selected"). Finally, the user can click on the "Output to File" button to save the file for further analysis. In this version, a limited formatting is provided for delimiting the data fields, as shown in the "Output to File" screen.

At any point during the above process, the user can go back and change selections or conditions. The system remembers the last state, so that changes can be made with minimum repetition.

Appendix 1: The text of the introductory screen of the MetaBrowser

Welcome to the CEDR browsing and query interface.

This interface will help you browse information about datasets of interest. For each dataset of interest to you, you can find descriptive information, such as variables, sites involved, and bibliography. In addition, you can extract a desired subset of the dataset, and save it in a file.

To see the next pages of this introduction, use the scroll bar on the right, by clicking on the up or down arrows. On a SUN workstation, click the leftmost button of the mouse only. You can also click on the middle of the scrollbar, drag it down, and release the click button, to move quickly through the pages. Please experiment with the scroll bar as it will be used in subsequent windows as well.

The introductory pages below are the only instructions given, and are sufficient for working with this interface, so please read them carefully. Next, there is a brief explanation of the function of the buttons shown below on this screen.

LIST DATASETS button:

Clicking on this button (below) will display an alphabetic list of the datasets known to CEDR. You can then select a particular dataset from the list, and find additional information about it, such as bibliography and variables it has. For each dataset selected you can proceed to extract a desired subset of the dataset and save it in a file.

SEARCH FOR DATASETS button:

As an alternative to displaying an alphabetic list of the datasets, you can click on this button (below) to search for desired datasets by specifying the subject terms, sites, and years you are interested in. Accordingly, one or more datasets will be selected. You can then select one of these datasets and continue as above to find additional information about it and to extract a subset from it.

RESTART button:

The system will remember selection conditions of your query. If you want to start over again, click on this button.

This application uses an X window system and window manager. For this application it is sufficient to know that lists are shown in display windows that have scroll bar on their right, similar to this introductory window. In subsequent windows, items can be selected from lists by clicking on an item once. To cancel a selection click on the item again.

This prototype browsing and query system was developed at the Lawrence Berkeley

Laboratory by the Data Management Group. The goal is to have an interface that users can operate without any prior knowledge of data management systems or query languages. Suggestions as to the kind of interface that would be useful for the cedr information system are welcome. For suggestions, please send electronic mail to E_Szeto@lbl.gov or A_Shoshani@lbl.gov.

FEDERATED METADATABASE: A DATABASE ABOUT DATABASES

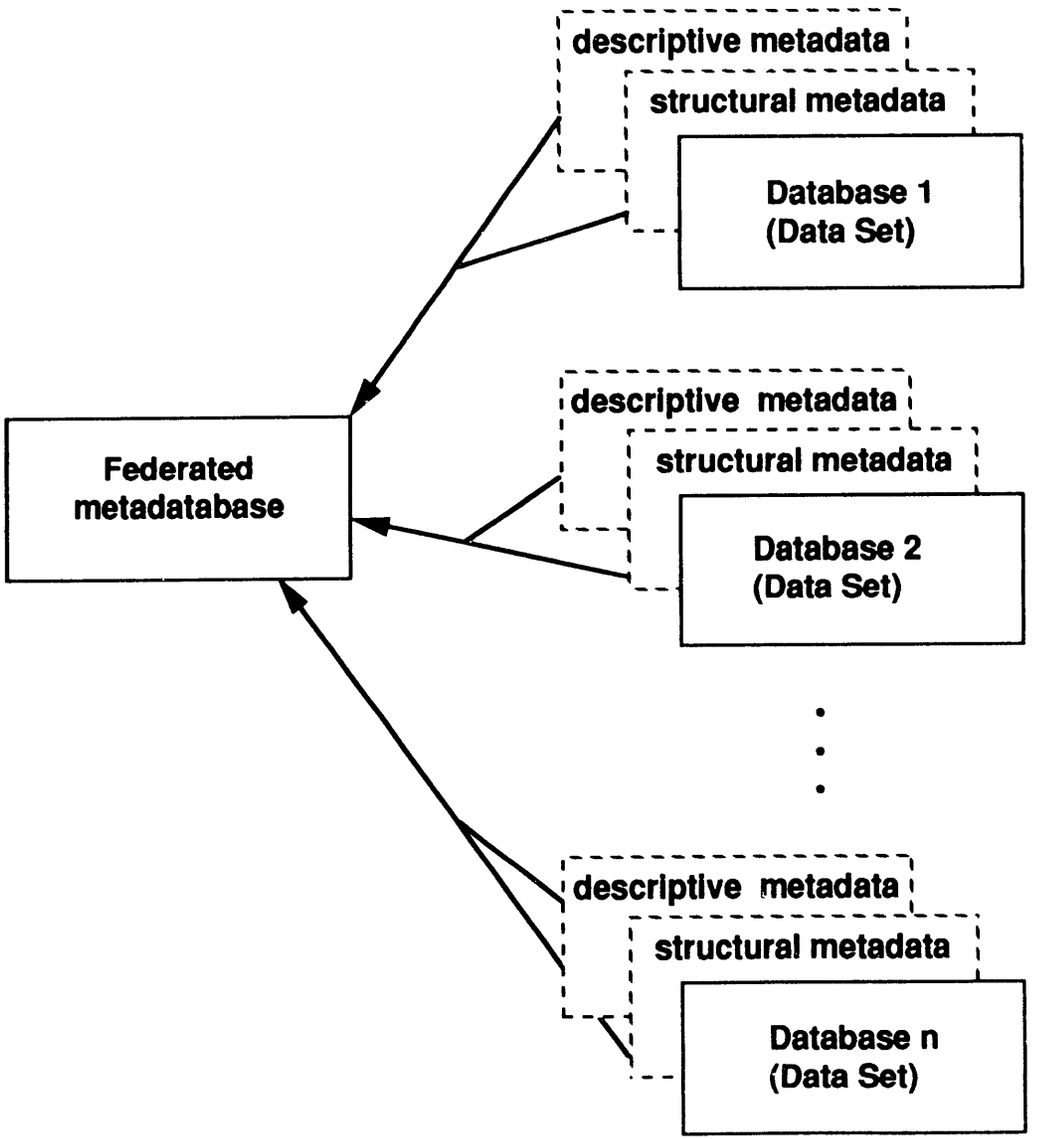
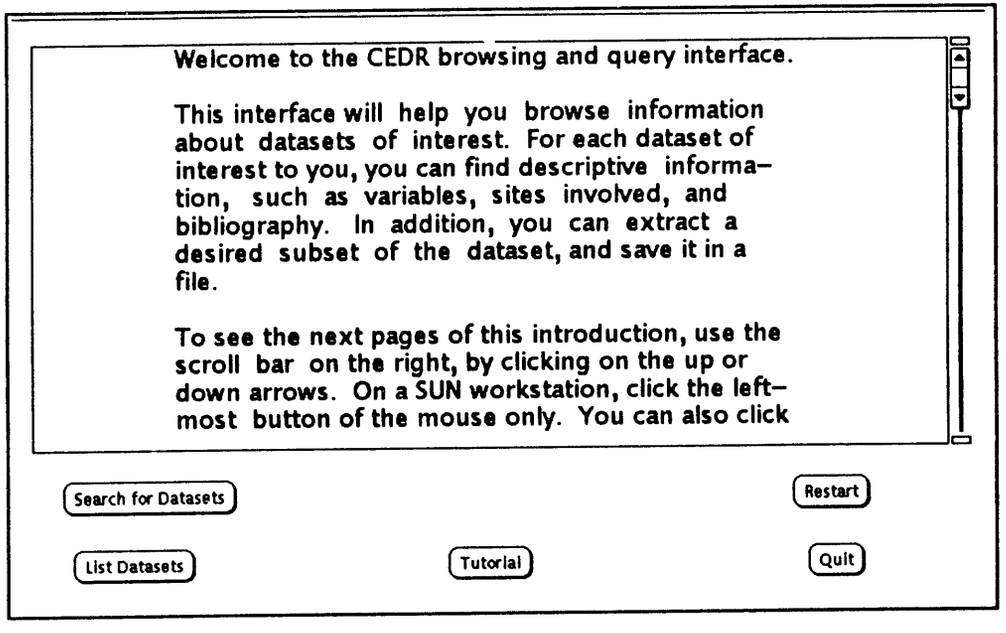
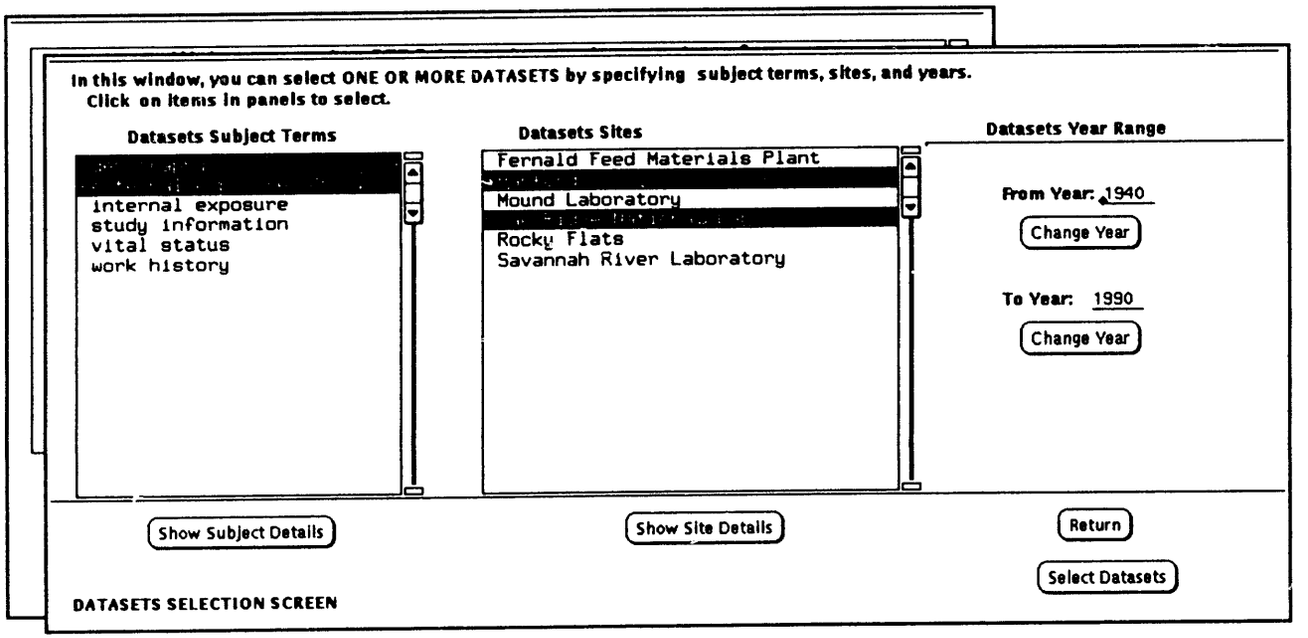


Figure 1

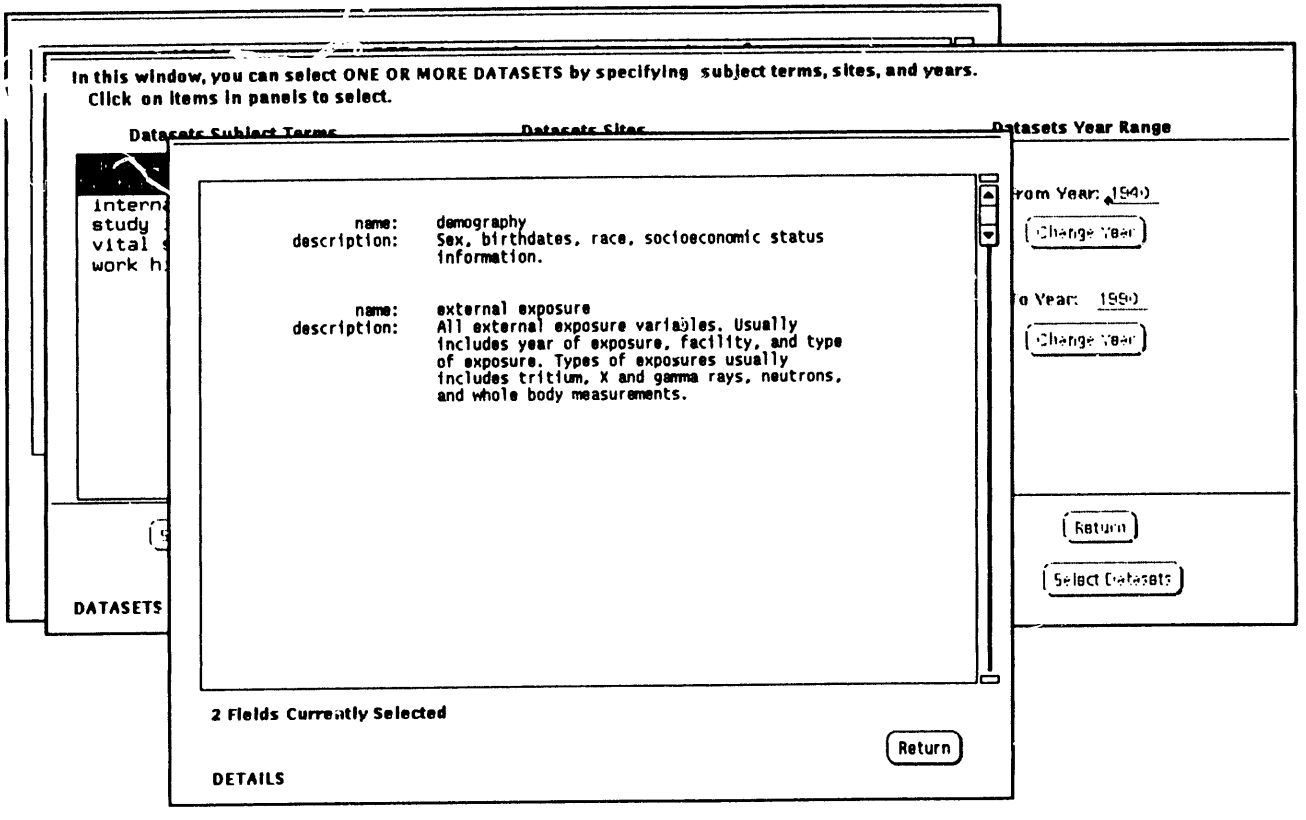
Startup Screen



Datasets Selection

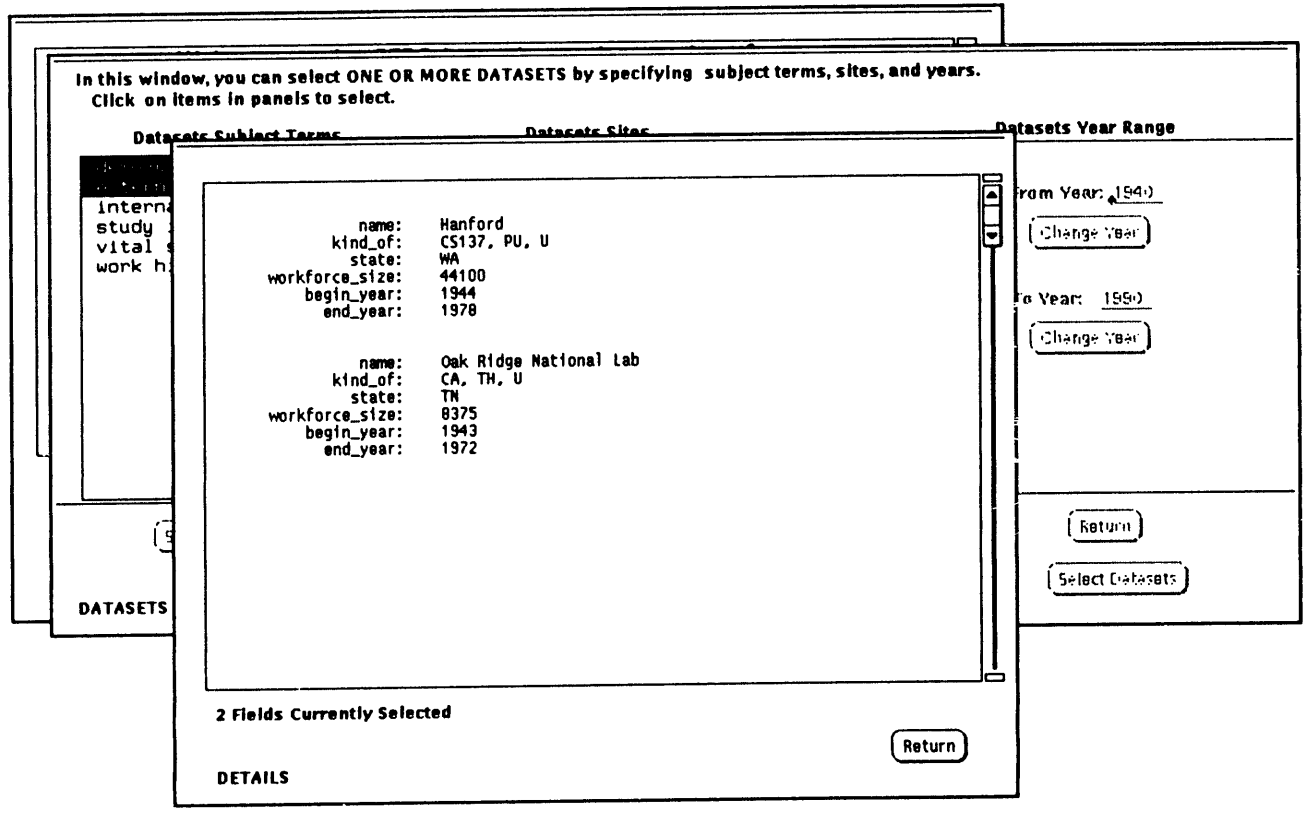


Subject Terms Details



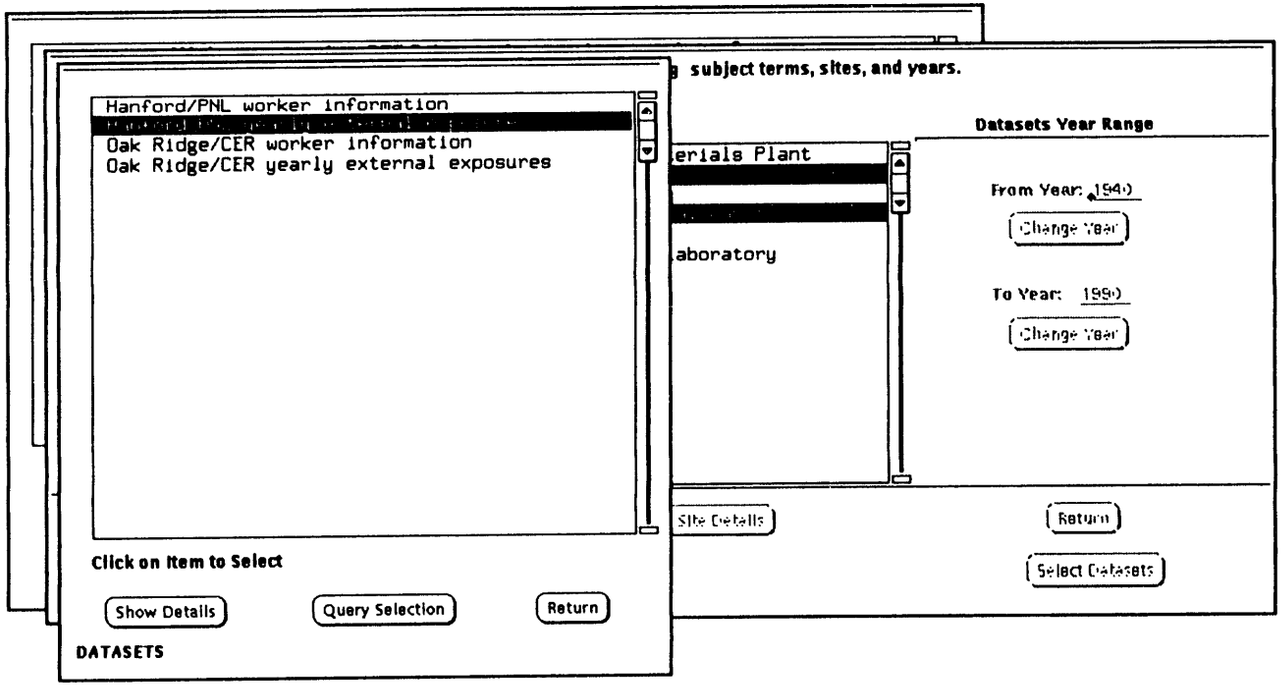
xclock browse

Site Details



xclock browse

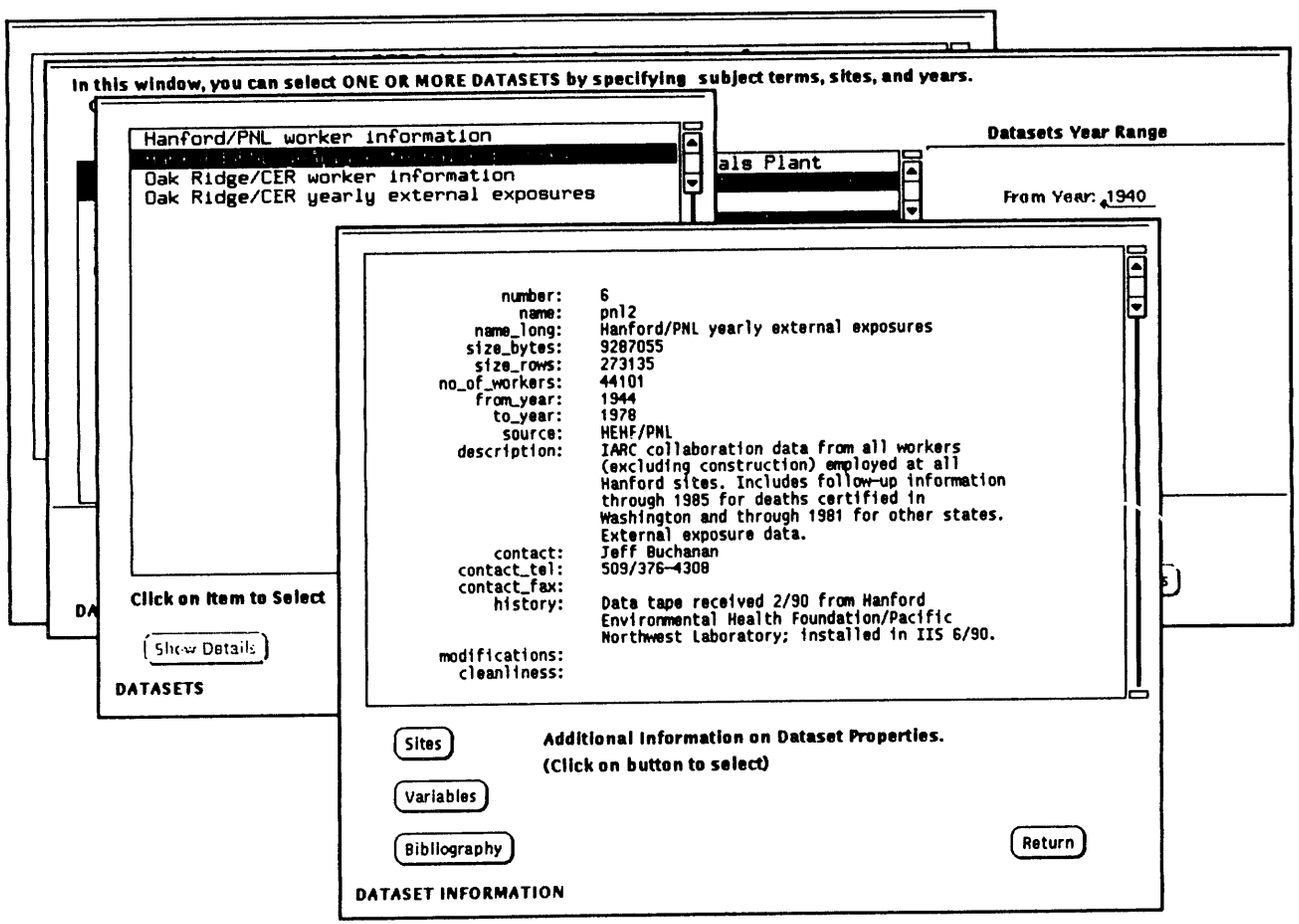
Dataset Selected

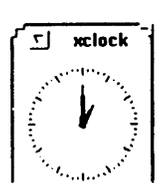


browse

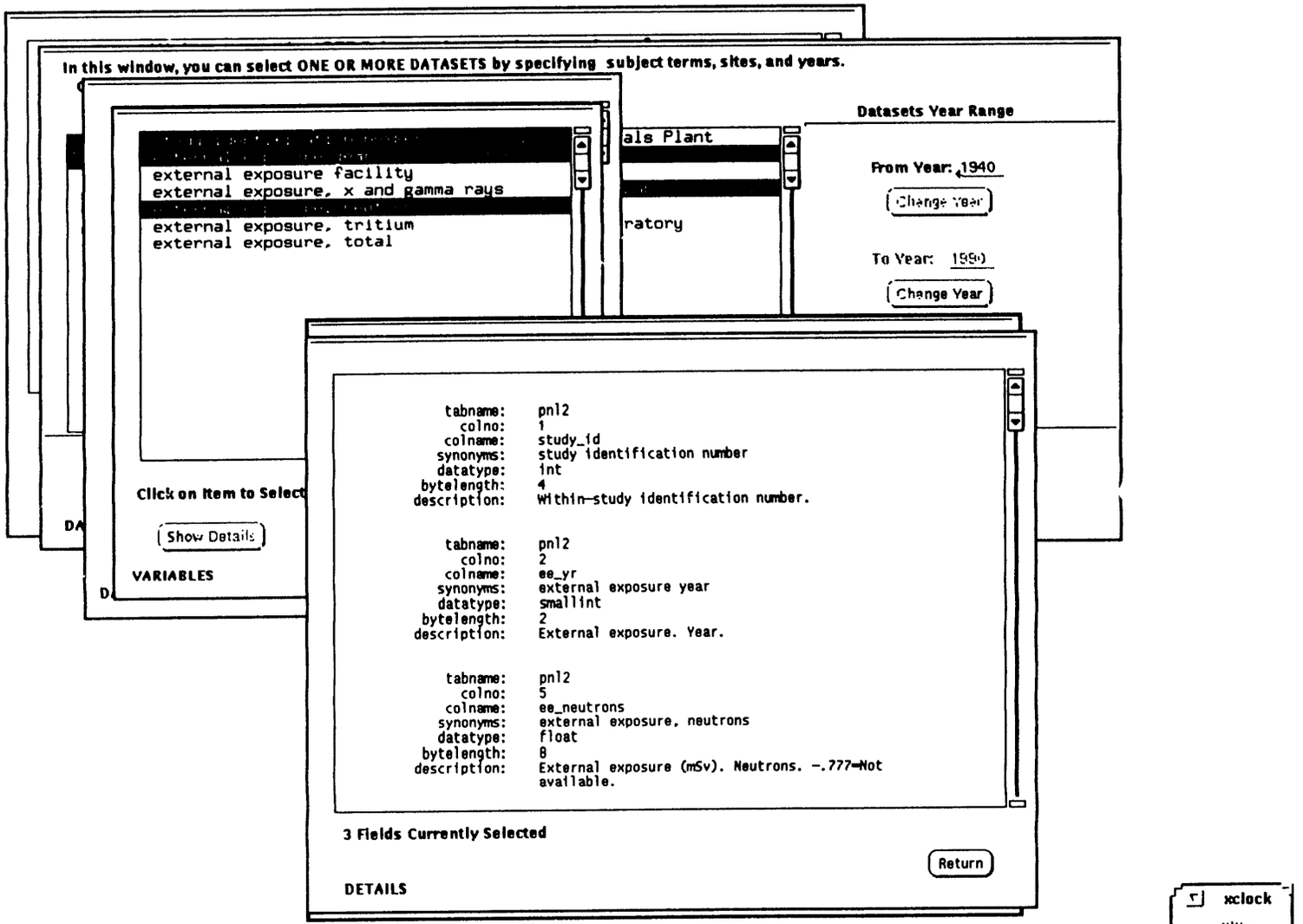
xclock

Dataset Information

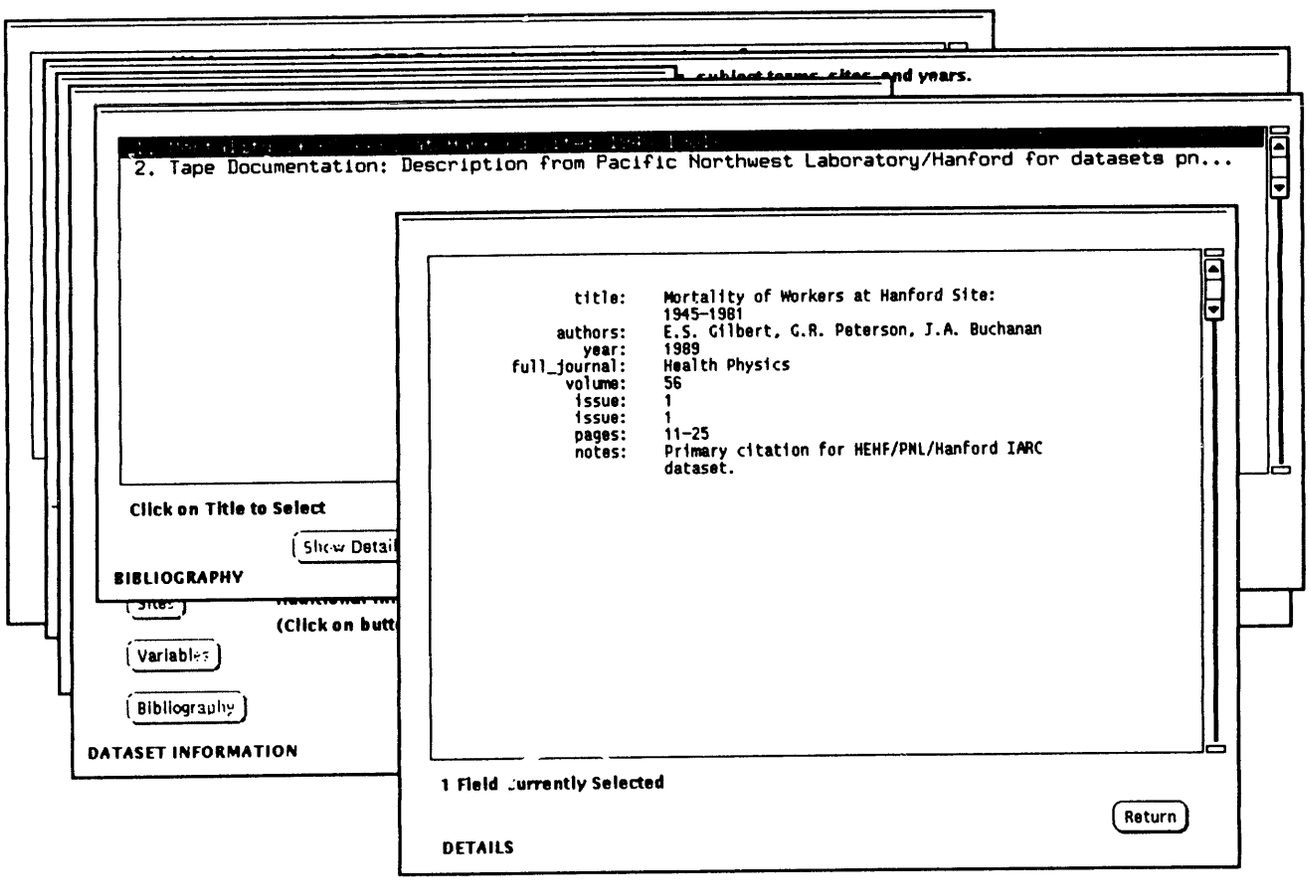




Details on Variables

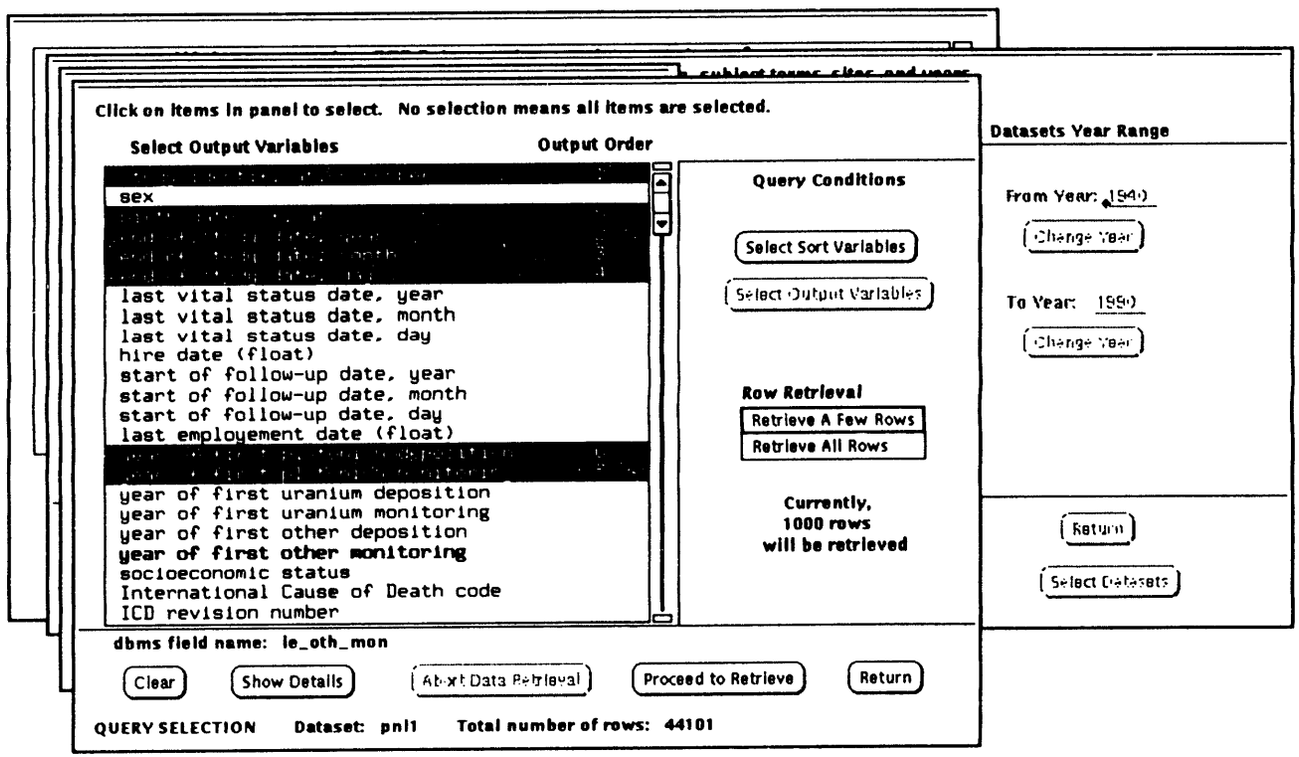


Bibliography



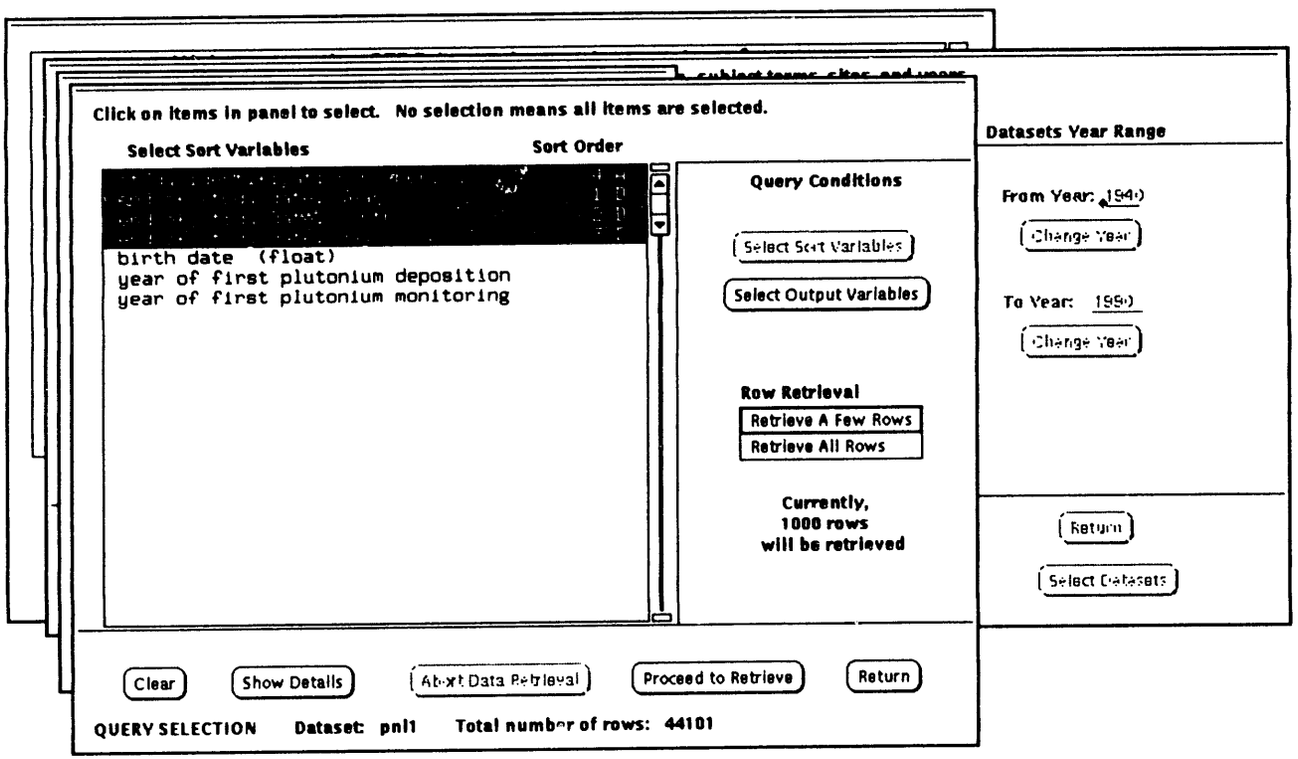
xclock browse

Query Selection Output Variables



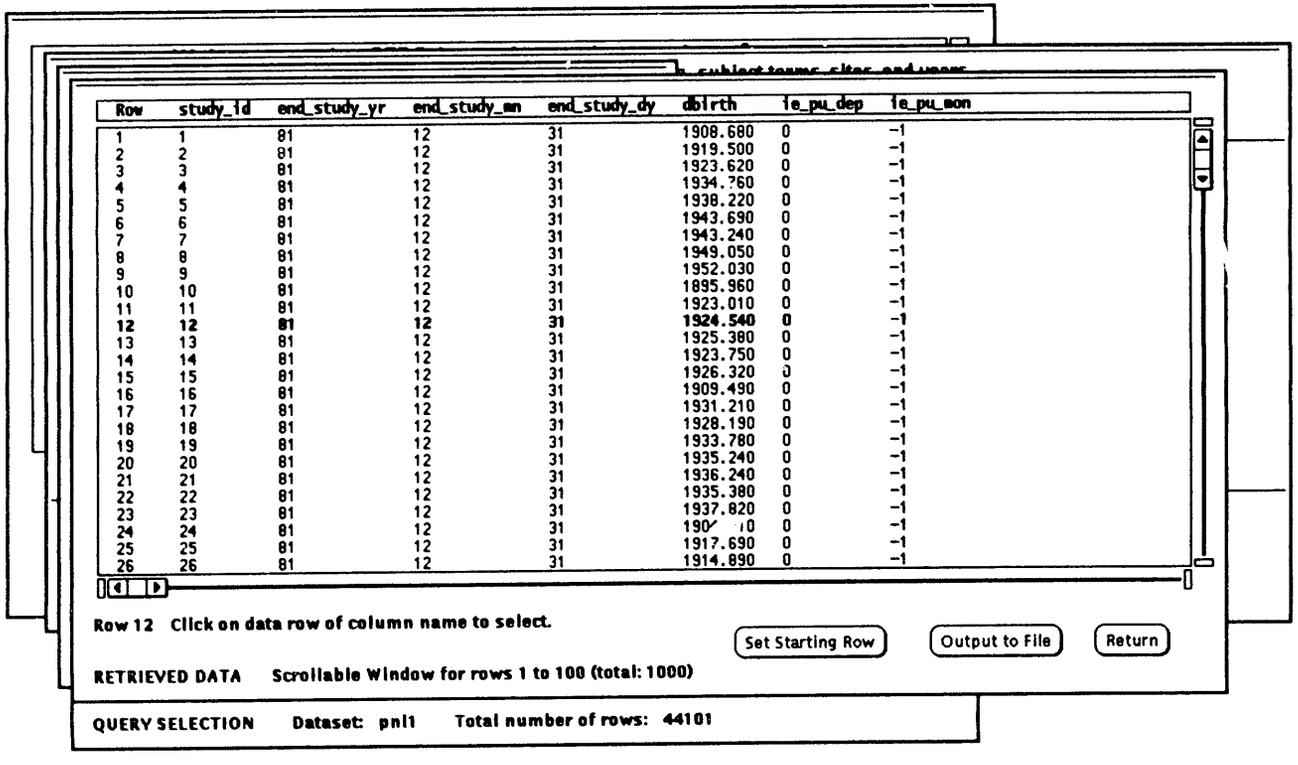
xclock browse

Sort Selection



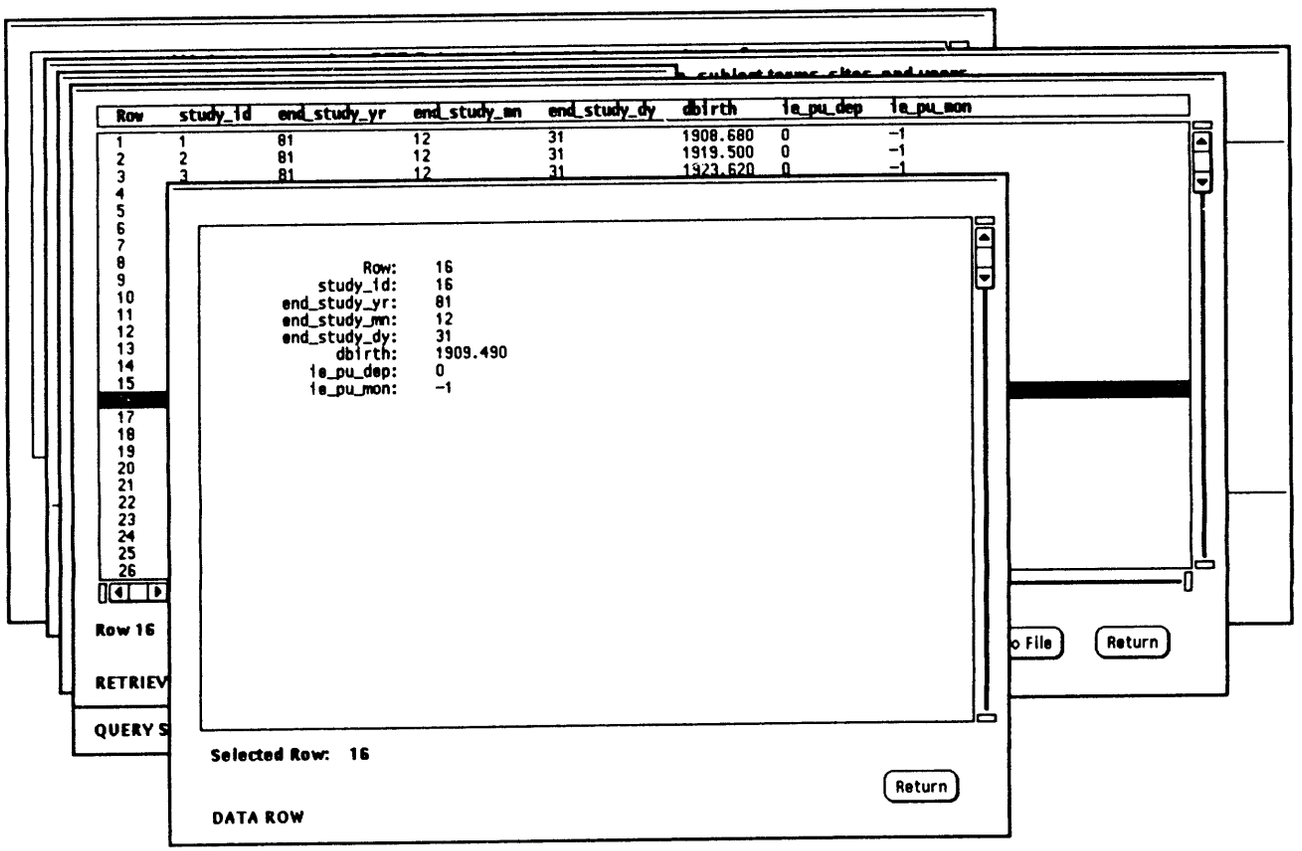
xclock browse

Retrieved Data



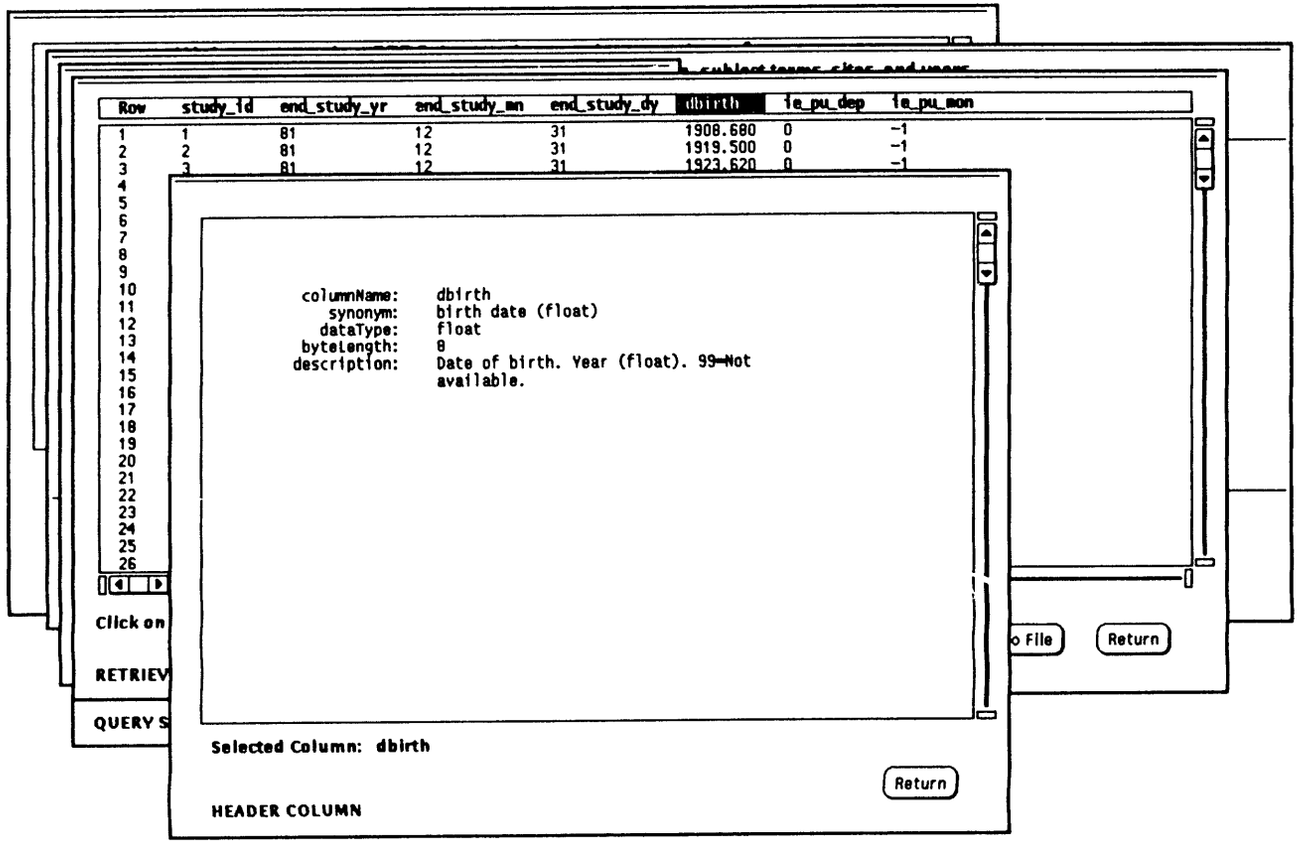
Row	study_id	end_study_yr	end_study_mn	end_study_dy	dbirth	fe_pu_dep	fe_pu_son
1	1	81	12	31	1908.680	0	-1
2	2	81	12	31	1919.500	0	-1
3	3	81	12	31	1923.620	0	-1
4	4	81	12	31	1934.760	0	-1
5	5	81	12	31	1938.220	0	-1
6	6	81	12	31	1943.690	0	-1
7	7	81	12	31	1943.240	0	-1
8	8	81	12	31	1949.050	0	-1
9	9	81	12	31	1952.030	0	-1
10	10	81	12	31	1895.960	0	-1
11	11	81	12	31	1923.010	0	-1
12	12	81	12	31	1924.540	0	-1
13	13	81	12	31	1925.380	0	-1
14	14	81	12	31	1923.750	0	-1
15	15	81	12	31	1926.320	0	-1
16	16	81	12	31	1909.490	0	-1
17	17	81	12	31	1931.210	0	-1
18	18	81	12	31	1928.190	0	-1
19	19	81	12	31	1933.780	0	-1
20	20	81	12	31	1935.240	0	-1
21	21	81	12	31	1936.240	0	-1
22	22	81	12	31	1935.380	0	-1
23	23	81	12	31	1937.820	0	-1
24	24	81	12	31	190[illegible]0	0	-1
25	25	81	12	31	1917.690	0	-1
26	26	81	12	31	1914.890	0	-1

Row Selected



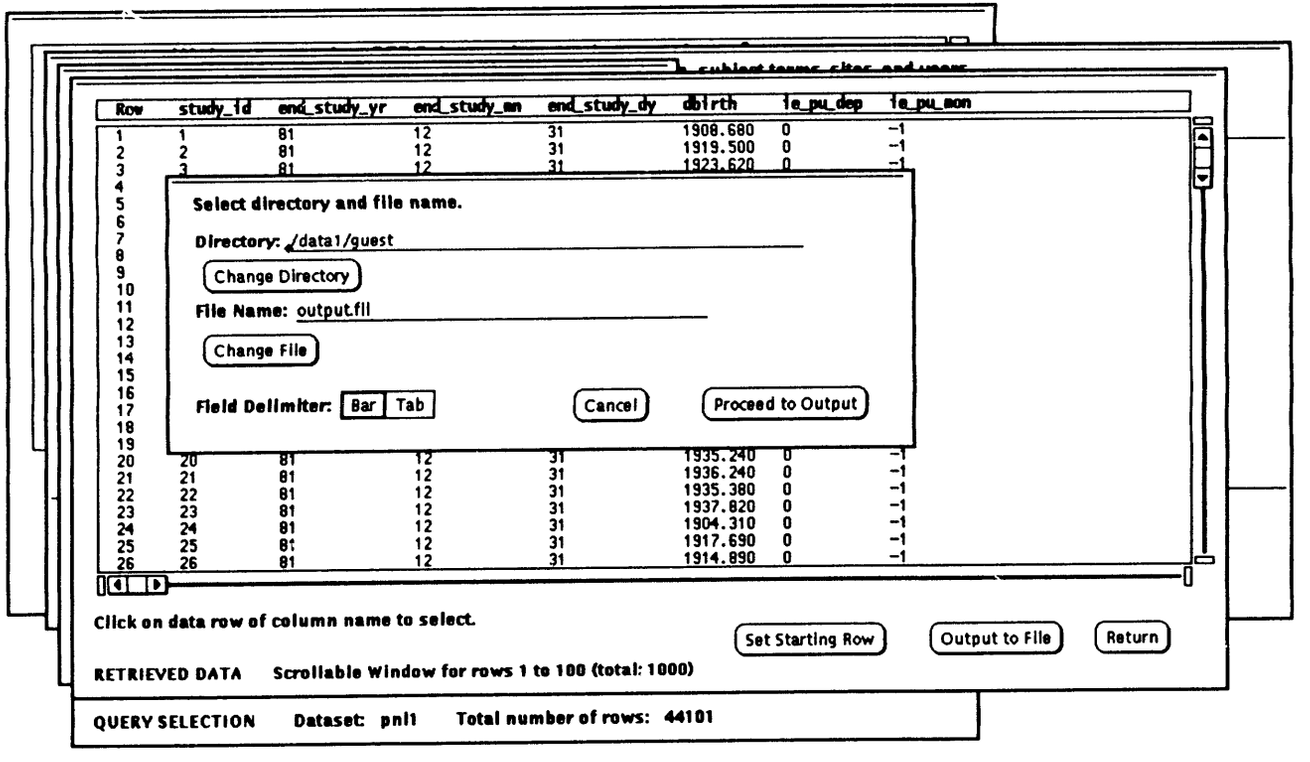
xclock browse

Column Selected



xclock browse

Output to File



xclock browse

END

DATE
FILMED
11/06/91