

10
10-15-91 JSD

LBL-30953
UC-605



Lawrence Berkeley Laboratory

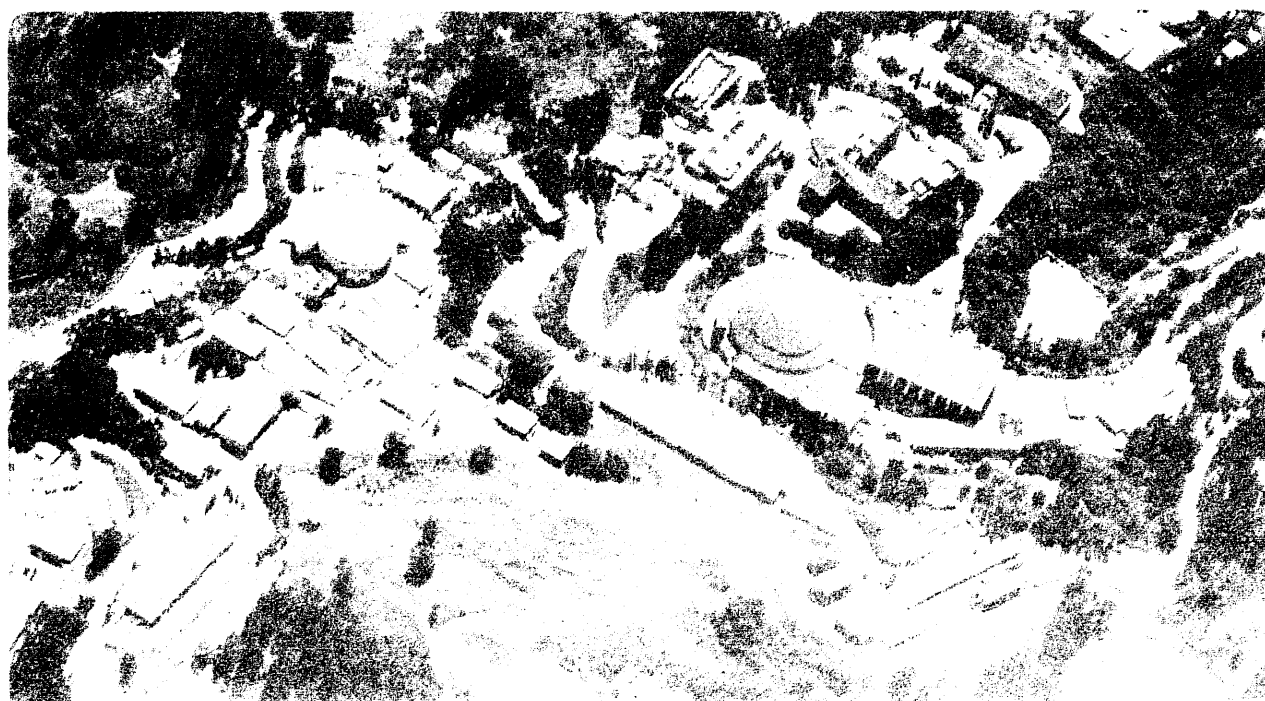
UNIVERSITY OF CALIFORNIA

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

This document was prepared as an account of work sponsored by the United States Government. Neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California and shall not be used for advertising or product endorsement purposes.

This report has been reproduced directly
from the best available copy.

Available to DOE and DOE Contractors
from the Office of Scientific and Technical Information
P.O. Box 62, Oak Ridge, TN 37831
Prices available from (615) 576-8401, FTS 626-8401

Available to the public from the
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Road, Springfield, VA 22161

Lawrence Berkeley Laboratory is an equal opportunity employer.

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.

MASTER 
DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

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 permit 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 objects, 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 the 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 pre-knowledge 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 constructs in order to understand what's presented, or express desired operations. In our context, we do not expect 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

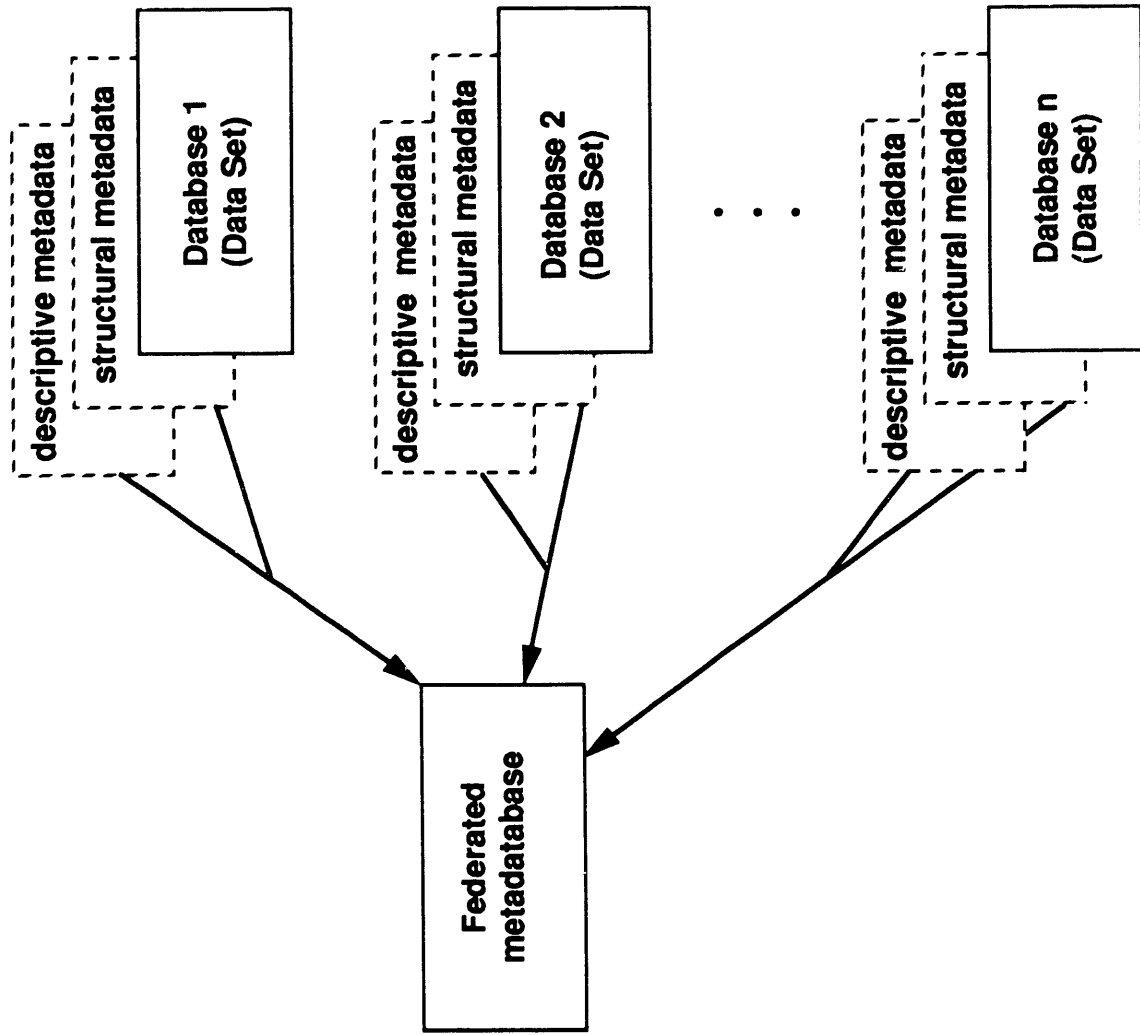
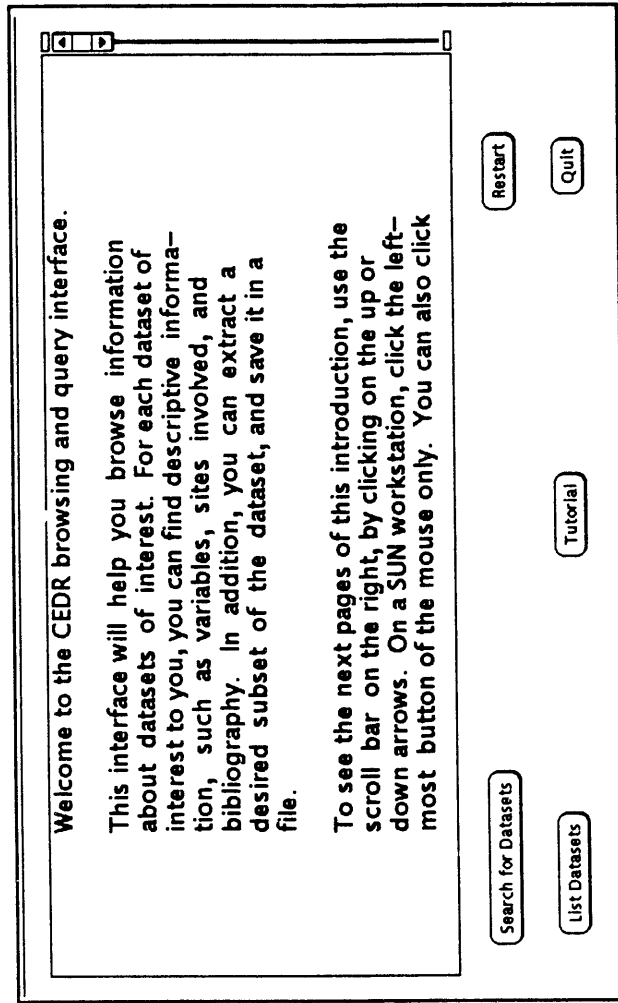


Figure 1

Startup Screen



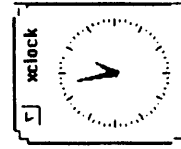
Datasets Selection

In this window, you can select ONE OR MORE DATASETS by specifying subject terms, sites, and years.
Click on items in panels to select.

Datasets Subject Terms	Datasets Sites	Datasets Year Range
<ul style="list-style-type: none">internal exposurestudy informationvital statuswork history	<ul style="list-style-type: none">Fernald Feed Materials PlantMound LaboratoryRocky FlatsSavannah River Laboratory	<p>From Year: 1940 Change Year</p> <p>To Year: 1990 Change Year</p>

Buttons: Show Subject Details, Show Site Details, Return, Select Datasets

DATASETS SELECTION SCREEN



Subject Terms Details

In this window, you can select ONE OR MORE DATASETS by specifying subject terms, sites, and years.
Click on items in panels to select.

Datasets Subject Terms

intern
study
vital
work h

Datasets Year Range

From Year: 1990
Change Year

To Year: 1990
Change Year

Return

Select Datasets

Datasets Sites

name: demography
description: Sex, birthdates, race, socioeconomic status information.

name: external exposure
description: All external exposure variables. Usually includes year of exposure, facility, and type of exposure. Types of exposures usually includes tritium, X and gamma rays, neutrons, and whole body measurements.

Return

DATASETS

2 Fields Currently Selected

DETAILS



Site Details

In this window, you can select ONE OR MORE DATASETS by specifying subject terms, sites, and years.
Click on items in panels to select.

Datasets Subject Terms

Intern
study
vital s
work h

Datasets Sites

name:	Hanford
kind_of:	CS137, PU, U
state:	WA
workforce_size:	44100
begin_year:	1944
end_year:	1978
name:	Oak Ridge National Lab
kind_of:	CA, TH, U
state:	TN
workforce_size:	8375
begin_year:	1943
end_year:	1972

Datasets Year Range

From Year: 1940

To Year: 1990

2 Fields Currently Selected

DETAILS



Dataset Selected

subject terms, sites, and years.

Serials Plant

laboratory

Site Details

Return

Select Object:

Datasets Year Range

From Year: 1940

Change Year

To Year: 1990

Change Year

Return

Select Object:

Hanford/PNL worker information

Oak Ridge/ER worker information

Oak Ridge/ER yearly external exposures

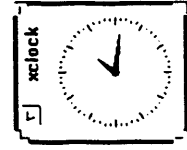
Click on item to Select

Show Details

Query Selection

Return

DATASETS



Dataset Information

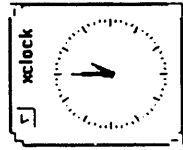
In this window, you can select ONE OR MORE DATASETS by specifying subject terms, sites, and years.

6
 number: Hanford/PNL yearly external exposures
 name: Hanford/PNL yearly external exposures
 name_long: 9287055
 size_bytes: 273135
 size_rows: 44101
 no_of_workers: 1944
 from_year: 1978
 to_year: 1978
 source: HENF/PNL
 description: TAPC collaboration data from all workers (excluding construction) employed at all Hanford sites. Includes follow-up information through 1985 for deaths certified in Washington and through 1981 for other states. External exposure data.
 contact: Jeff Buchanan
 contact_tel: 509/376-4308
 contact_fax:
 history:
 modifications:
 cleanliness:

DATASETS

Additional Information on Dataset Properties.
(Click on button to select)

DATASET INFORMATION



Details on Variables

In this window, you can select ONE OR MORE DATASETS by specifying subject terms, sites, and years.

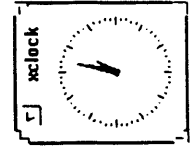
Datasets Year Range
 From Year: 1940
 To Year: 1990

VARIABLES

tabname:	pn12
colno:	1
colname:	study_id
synonyms:	study identification number
datatype:	int
length:	4
description:	Within-study identification number.
tabname:	pn12
colno:	2
colname:	ee_yr
synonyms:	external exposure year
datatype:	smallint
length:	2
description:	External exposure. Year.
tabname:	pn12
colno:	5
colname:	ee_neutrons
synonyms:	external exposure, neutrons
datatype:	float
length:	8
description:	External exposure (mSv). Neutrons. -.777=Not available.

3 Fields Currently Selected

DETAILS



Bibliography

2. Tape Documentation: Description from Pacific Northwest Laboratory/Hanford for datasets pr...

Click on Title to Select

BIBLIOGRAPHY

Variable: Bibliography

DATASET INFORMATION

1 Field Currently Selected

DETAILS

Return

title: Mortality of Workers at Hanford Site:
1945-1981
authors: E.S. Gilbert, G.R. Peterson, J. A. Buchanan
year: 1989
journal: Health Physics
full_journal: Health Physics
volume: 56
issue: 1
pages: 11-25
notes: Primary citation for HEHF/PNL/Hanford IARC dataset.



Query Selection Output Variables

Click on items in panel to select. No selection means all items are selected.

Select Output Variables	Output Order
sex	
last vital status date, year	
last vital status date, month	
last vital status date, day	
hire date (float)	
start of follow-up date, year	
start of follow-up date, month	
start of follow-up date, day	
last employment date (float)	
year of first uranium deposition	
year of first uranium monitoring	
year of first other deposition	
year of first other monitoring	
socioeconomic status	
International Cause of Death code	
ICD revision number	

dbms field name: le_oth_mon

QUERY SELECTION Dataset: pn11 Total number of rows: 44101

Query Conditions

Row Retrieval

Currently, 1000 rows will be retrieved

Datasets Year Range

From Year: 1940

To Year: 1990

Sort Selection

Click on items in panel to select. No selection means all items are selected.

Select Sort Variables

birth date (float)

year of first plutonium deposition

year of first plutonium monitoring

Sort Order

[Select Sort Variables]

[Select Output Variables]

Query Conditions

[Select Sort Variables]

[Select Output Variables]

Datasets Year Range

From Year: 1940 [Change Year]

To Year: 1990 [Change Year]

Row Retrieval

Retrieve A Few Rows

Retrieve All Rows

Currently, 1000 rows will be retrieved

[Return]

[Select Datasets]

[Clear] [Show Details] [Abort Data Retrieval] [Proceed to Retrieve] [Return]

QUERY SELECTION Dataset: pnli Total number of rows: 44101



Retrieved Data

Row	study_id	end_study_yr	end_study_mon	end_study_dy	dbirth	le_pu_dep	le_pu_mon
1	81	12	12	31	1908.680	0	--
2	81	12	12	31	1919.500	0	--
3	81	12	12	31	1923.620	0	--
4	81	12	12	31	1934.760	0	--
5	81	12	12	31	1938.220	0	--
6	81	12	12	31	1943.690	0	--
7	81	12	12	31	1943.240	0	--
8	81	12	12	31	1949.050	0	--
9	81	12	12	31	1952.030	0	--
10	81	12	12	31	1895.960	0	--
11	81	12	12	31	1923.010	0	--
12	81	12	12	31	1924.540	0	--
13	81	12	12	31	1925.380	0	--
14	81	12	12	31	1923.750	0	--
15	81	12	12	31	1926.320	0	--
16	81	12	12	31	1909.490	0	--
17	81	12	12	31	1931.210	0	--
18	81	12	12	31	1929.190	0	--
19	81	12	12	31	1933.780	0	--
20	81	12	12	31	1935.240	0	--
21	81	12	12	31	1936.240	0	--
22	81	12	12	31	1935.380	0	--
23	81	12	12	31	1937.820	0	--
24	81	12	12	31	1907.10	0	--
25	81	12	12	31	1917.690	0	--
26	81	12	12	31	1914.890	0	--

Row 12 Click on data row of column name to select.

RETRIEVED DATA Scrollable Window for rows 1 to 100 (total: 1000)

QUERY-SELECTION Dataset: pnt1 Total number of rows: 44101

Buttons: Set Starting Row, Output to File, Return



Row Selected

Row	study_id	end_study_yr	end_study_mn	end_study_dy	dbirth	ie_pu_dep	ie_pu_mon
1	81	12	31	1908.680	0	-1	-1
2	81	12	31	1919.500	0	-1	-1
3	81	12	31	1923.620	0	-1	-1
4							
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16	81	12	31	1909.490	0	-1	-1
17							
18							
19							
20							
21							
22							
23							
24							
25							
26							

Row: 16
 study_id: 81
 end_study_yr: 12
 end_study_mn: 31
 end_study_dy: 1909.490
 dbirth: 0
 ie_pu_dep: 0
 ie_pu_mon: -1

Selected Row: 16
 DATA ROW



Column Selected

Row	study_id	end_study_yr	end_study_mn	end_study_dy	dbirth	le_pl_dep	le_pl_mon
1	81	12	31	1908.680	0	-1	-1
2	81	12	31	1919.500	0	-1	-1
3	81	12	31	1923.620	0	-1	-1

ColumnName: dbirth
Synonym: birth date (float)
datatype: float
bytelength: 8
description: Date of birth. Year (float). 99=Not available.

Selected Column: dbirth
HEADER COLUMN

Click on RETRIEVE QUERY S

Return



Output to File

Row	study_id	end_study_yr	end_study_mn	end_study_dy	dobirth	le_pu_dep	le_pu_mon
1	81	12	31	1908.680	0	--	--
2	81	12	31	1919.500	0	--	--
3	81	12	31	1923.620	0	--	--
4	81	12	31	1923.620	0	--	--
5	81	12	31	1923.620	0	--	--
6	81	12	31	1923.620	0	--	--
7	81	12	31	1923.620	0	--	--
8	81	12	31	1923.620	0	--	--
9	81	12	31	1923.620	0	--	--
10	81	12	31	1923.620	0	--	--
11	81	12	31	1923.620	0	--	--
12	81	12	31	1923.620	0	--	--
13	81	12	31	1923.620	0	--	--
14	81	12	31	1923.620	0	--	--
15	81	12	31	1923.620	0	--	--
16	81	12	31	1923.620	0	--	--
17	81	12	31	1923.620	0	--	--
18	81	12	31	1923.620	0	--	--
19	81	12	31	1923.620	0	--	--
20	81	12	31	1923.620	0	--	--
21	81	12	31	1923.620	0	--	--
22	81	12	31	1923.620	0	--	--
23	81	12	31	1923.620	0	--	--
24	81	12	31	1923.620	0	--	--
25	81	12	31	1923.620	0	--	--
26	81	12	31	1923.620	0	--	--

Select directory and file name.

Directory:

Change Directory

File Name:

Field Delimiter:

Click on data row of column name to select.

RETRIEVED DATA Scrollable Window for rows 1 to 100 (total: 1000)

QUERY SELECTION Dataset: pnt1 Total number of rows: 44101



END

**DATE
FILMED**

11 106191

I

