Overview of the Technical Architecture for FERRET

A Federal Electronic Research and Review Extraction Tool for Data Access and Dissemination of Micro and Macro Data

Cavan Capps: Chief, Survey Modernization Programming Branch
Demographic Surveys Division
Bureau of the Census

The objective of the tool is to provide access to data for both public and internal users. Enhancing the access to data should increase it's use and value at all levels of policy development. Potential identified users include:

- Internal Government Users
- Government Policy Analysts
- Academic and Federal researchers
- Public Users (including Federal Depository Libraries and academic students)

The tool will provide internal analysts and the public the same view of publicly accessible data and documentation. Internal government analysts will also have access to all data covered by the privacy act. The data will be protected by multiple layers of security including the state of the art in Internet firewalls and proxy servers, C2 databases etc. Such security should allow internal users to do research on data sets previously inaccessible (theoretically potentially permitting the merging of data from different data sets based on encrypted SSN and establishment address.)

The tool will promote data access by users of different levels of expertise. Both novice and expert users will be supported. Long term research and quick turn around policy oriented descriptive research will also be supported.

Access to data is defined as:

- Ability to quickly locate the right data from a variety of data sets. (Provided by a navigation strategy.)
- Ability to quickly understand the data and make comparisons between data elements from different data sets. (Provided by a comprehensive on-line documentation strategy)
- Ability to easily manipulate and extract the data to desktop and larger computers for further processing. (Provided by a collection of mathematical and graphical summarization and extraction tools)

The critical components of the FERRET architecture are:

- Navigation
- On-line Documentation
- Mathematical and Graphical Summarization
- Data Object Model
- Layered Client Server Data Base Access

Navigation

Initial research indicates that users can be largely separated into two general groups. Frequent users of the data often access the data by subsets of variable related to particular concepts. Infrequent users often find these hierarchical groupings confusing. Since FERRET must support both expert and novice users, the tool will allow variables to be subsetted by concept or topic, and to be searched for by a key word search across a long descriptive name and attributes of all the variables. Surveys can be considered the highest level in the hierarchy of concepts and keyword searches can be limited to any single or set of concepts. Context sensitive help will provide access to variable definitions and permit the user to make appropriate comparisons of similar data elements from different survey data sets. Since users may search for a variable name used by any survey data set, a thesaurus is planned.

The tool will provide access to Micro data and Macro data in the forms of tabular reports and time-series. In order to make the tool easy to learn and use, the same navigation strategy will be used to access each of these data types.

On-Line Documentation

The ability to understand the data will be determined by the quality of the on-line documentation. As a result, it is critical that subject matter experts own the data and have tools to update on-line documentation easily. On-line documentation will be provided in three major ways:

- Context Sensitive Help
- Comprehensive Table of Contents of Survey Documentation (Meta Data)
- Key word and phrase searches through all the documentation

All surveys, concepts and variables on the screen will provide 'context sensitive help' that will explain the concept or variable. Help for micro and macro variables will include but not be limited to:

- Variable definition
- Weighting information
- Variance, Standard Error, and other quality measures
- Related Variables
- Breaks in Series or changes in questionnaire wording
- Warnings concerning appropriate data usage

If the documentation is to be accurate, the on-line public documentation must be the same documentation used by internal users. Where ever possible, documentation changes (such as new wording for a survey question) should be automatically propagated from the instrument or other sources of working electronic documentation.

Since several organizations work together to produce any large survey, the tool acknowledges that different organizations contribute to different pieces of the documentation. Since users need the most current documentation, tools are provided for each suborganization to update the documentation that it develops in an electronic distributed fashion, with coordination and review provided at the survey level.

A centralized Internet electronic discussion about the survey, is planned to provide a feedback mechanism to allow subject matter owners to continually improve the documentation. Such discussion might allow serious users to share needs, advice and research with other researchers and data providers.

Mathematical and Graphical Summarization

Internal users will have access to statistical packages like SAS. Both public and internal users should have access to an increasing set of tools that aggregate, chart and map the data using the desktop computer. The goal is to provide such tools in a modular fashion and to integrate tightly with the current generation of commercial desktop software including spreadsheets and word processors. Users should be able to easily produce reports using the desktop software that they are familiar.

Data Objects

Organizing data elements as objects will permit the tool to integrate variables from different surveys into one access tool. Each survey will be considered an object with an associated set of attributes or characteristics. Each variable, time-series, or tabular report will also be considered an object with associated characteristics. These attributes will allow the FERRET tool to display the required information graphically, provide paths to the necessary context sensitive help and the detailed information found the comprehensive table of contents. Although the table of contents associated with each survey will provide the primary set of meta data, the attributes associated with each variable could also be considered meta data. However, the primary purpose of the attributes associated the data objects are to provide FERRET the means to construct the graphical user interface. The objects would be managed in a relational data base that would allow different surveys to provide access and to administer their surveys relatively independently without impacting surveys or work from other organizations. Coordination could be provided as required at different levels.

Survey Attributes would include:

- Survey Name
- Link to Survey Home Page
- Data Base Machine Location
- Data Base Interface Type

Micro Data Attributes would include:

- Variable Name
- Survey Name
- Mnemonic
- Long Description
- Questionnaire question or Recode Definition
- Weighting information
- Variance information
- Breaks in series information
- Link to survey table of contents
- Related Variables
- Security Code
- Concept Mapping
- Geographic Use Code
- Link to rules for appropriate use
- Link to Source and Method of Collection

Macro Time-Series Attributes might include:

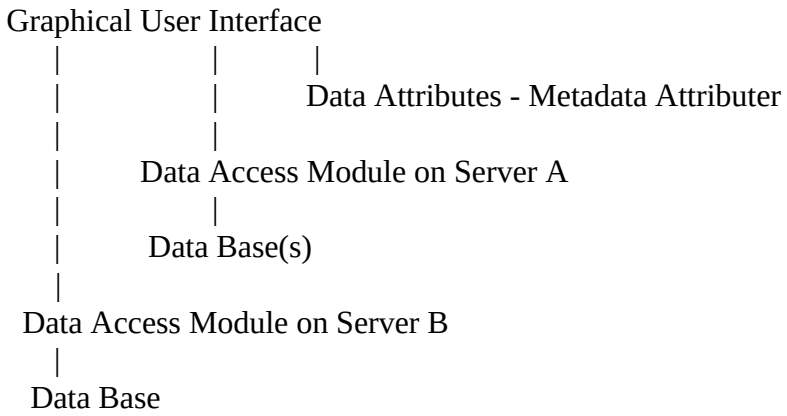
- Variable Name
- Survey Name
- Mnemonic
- Long Description
- Construction of the Data Item
- Questionnaire Cross-Reference
- Standard Error
- Security Code
- Seasonal Adjustment information
- Breaks in Series
- Link to survey table of contents
- Related Variables
- Concept Mapping
- Geographic Use Code
- Link to rules for appropriate use
- Link to Source and Method of Collection

Macro Tabular report attributes might include:

- Report Title
- Table Title
- Survey Name
- Long Description
- Construction of the Data Items in Table
- Questionnaire Cross-Reference
- Standard Error information
- Security Code (internal work tables and published tables)
- Link to survey table of contents
- Related Tables
- Geographic Use Code
- Link to rules for appropriate use
- Link to Source and Method of Collection

Layered Client Server Data Base Access

The tool will use a module on the server that will provide to different data bases located on distributed machines. The data will be staged on the data base machines in a standard way, allowing users to extract large data sets that would normally overwhelm desktop computer in a standard client-server environment. This configuration will also permit data aggregation and other data usage rules to be applied at the server.



Summary
The goal of the FERRET tool is to provide a single location to shop for data from several surveys, providing a repository for each survey's micro data, macro data and associated documentation. The tool assumes a modular system that supports a variety of data base backends located on different computer hardware platforms. This will allow different surveys to be added to the tool in an incremental progression and allow for new data base technologies to be added as they become available. The tool acknowledges the collaborative nature of effective survey data production and uses a distributed approach to provide access to continually evolving documentation. Most importantly, the tool requires that subject matter experts own the data and documentation. It is critical that they have the ability to update documentation in a labor saving and timely basis, where possible tools to support such efforts will be pursued.