

Dr. Cavan Capps
Director, FERRET Project
FERRET: A Joint Census Bureau/Center for Disease Control Project

Dr. Capps' presentation centered on the Federal Electronic Research and Retrieval Tabulation Tool (FERRET) program, a joint project of the Census Bureau and the Center for Disease Control.

What is FERRET? FERRET is a website-based project meant to help users find data. Specifically, it provides access on demand for the current population survey and all demographic surveys done inside the census. The Internet is the primary source of data dissemination because it is the cheapest method available.

The Internet provides challenges such as information overload, difficulty finding data (because you cannot search a database through a search engine, and data is stored by organization), stovepipe solutions (each provider of data wants to solve all of your problems, but there is no interfacing among them), and confusing tool sets (everyone has done it differently).

Data democracy would mean open access for data suppliers, providing the ability to support all databases and find and authenticate data sources. This would also mean open access for data users, with free software for all end users and free or near free access to data. This would be beneficial to all because as the number of users increases, the value of the data increases as well.

The goal should be a virtual data library with links to data from large government sites, data archives, state and local sources. This would maintain data producer accountability. The objective is data access that would allow the user to find the right dataset, understand the data through useful documentation that is tightly linked to the data, and manipulate the data as needed.

Needs include tools to find the data and multiple kinds of available data, such as individual respondent data, time series, maps and tables. There should be the same kinds of interfaces across all datasets, so users need not be retrained to use each dataset. Users should be able to find data through documentation and metadata searches. Definitions and attributes of the data should be included and linked to the data, along with information on data relationships describing how the data can be used. Rules for use also need to be included, describing rules for weighting, graphical presentation, reliability of the data, and linking to geography. There should be security and access controls so agencies can link confidential data they have access to with other data. Other needs include ad hoc tabulations, file extraction and the ability to do simple descriptive statistics online before downloading large datasets.

Challenges include providing core software that will provide searching and access across geographically separated data sources, suggest how to use the data appropriately (by recommending weights, for example), and provide the ability to combine data appropriately. Other challenges include the enormous size of the datasets, the number and platform variety of datasets, rapid obsolescence of both hardware and software, and the maintenance of software, data and documentation. The people who understand the data and the software are becoming the most expensive component—use their expertise, don't try to train them in something new.

Data usage rules need to be stored with the data in a metadata repository. Data directory services include weighting rules, matching rules over time and between component datasets, and rules for geography, graphing and reliability.

The FERRET project has partnerships with agencies such as FedStats, ICPSR's data archives, the National Science Foundation, State Data Centers, BLS, BEA, the National Center for Health Statistics, CDC, NCHS, HUD, the Department of the Interior, the EPA and the Department of Agriculture.

Currently available data includes surveys from the Census Bureau, BLS, CPS, SIPP and SPD (income & poverty data), NCHS (health data), and HANS (Health and Nutrition Survey). Data that will be available soon includes the 2000 census, the economic census, county business patterns, population estimates, mortality rates, hospital discharge rates, crime statistics, and trade statistics, among others. Dr. Capps' agency is trying to develop a data network and a tool base. They will also provide software and advice on how to set up and maintain a system for an individual agency.

In response to a question from a conference participant, Dr. Capps said that decentralization of the data results in cost efficiencies and accountability for the data itself. He noted that a centralized system would be prohibitively expensive and that no one could be held accountable for the data.