The First Annual State Data Conference:
Assessing the Availability of Connecticut State and Local Data
May 10, 1999

Summary

This conference was sponsored by the Consortium for Public Policy Research at the University of Connecticut.

The goal of this conference is to begin a conversation about what conference participants recognize to be a serious issue in the state–the lack of complete, readily available and easily accessible data for the state of Connecticut–and to discuss methods of dealing with the problem, including the establishment of a state data center in Connecticut. Dr. Fred Carstensen noted that good data is the bedrock for intelligent policy discussion and complete understanding of many issues including economic development, demographic trends and social health statistics.

Connecticut had a state data center in the past, but it was a casualty of state budget cuts. Some of the current efforts undertaken by individual agencies to organize data and make it readily accessible are recognized as quite good, but in terms of establishing a state data center, we must begin again nearly at ground zero. [The State's Office of Policy and Management retains a "State Data Center," but its staff, once eight, is now but one, William Kraynak, and its function is now as the liaison with the U.S. Census; it is not responsible for state or local level data.]

The desired outcomes of this conference are an understanding of the data issues in the State of Connecticut, and the laying of a broad foundation for a new state data center in Connecticut. Conference participants and speakers discussed what data is currently available and what additional data is needed. In addition, conference participants left with an understanding of the available opportunities for creating a state data center and for founding a new, vastly strengthened state data system in Connecticut that can serve both the public and private sector.

The conference focused on four main areas:
What do we have to build on?
What do we need?
What have others accomplished?
Where do we go from here?

What do we have to build on? There are many individual sources of useful data in Connecticut. The University of Connecticut's Center for Geographic Information and Analysis and Homer Babbidge Library are both making great strides in making geographic and geo-spatial data available to the university and state communities. The FERRET Project of the Census Bureau and the Center for Disease Control have also been working hard to make all population and demographic data obtained from the census available on demand online. Individual datasets available from various other sources include information on businesses and public health information on a large variety of topics. However, as noted by conference speakers as well as the conference participants surveyed, many of the individual datasets are not always complete, compatible or fully accessible. Jeffrey Blodgett of the Connecticut Economic Resource Center (CERC) noted the unavailability of concrete economic and business information, while Dr. David

Gregorio of the University of Connecticut Health Center discussed confidentiality issues surrounding access to individual health information.

The Map and Geographic Information Center (MAGIC) at the University of Connecticut has been working in recent years to create an online collection of spatial data for the State of Connecticut. The Center collects, describes and provides access to all kinds of spatial data for the State of Connecticut from the state level to the town level, taken from multiple federal and state sources. Pat McGlamery notes that while the primary intended users are UCONN faculty, students and staff, two-thirds of the users of the MAGIC website are from outside the University. In total, approximately 9000 gigabytes of data are currently downloaded per month, in compressed form, mostly at the state level. Currently, MAGIC is working to amass a complete set of maps of Connecticut before 1800.

The University of Connecticut Center for Geographic Information and Analysis is another UCONN agency dealing with geographic information. The main functions of UCCGIA are to train and educate individuals in the use of geographic data, facilitate archiving of geographic information and facilitate geographic research. As Dr. Robert Cromley explains, UCCGIA is currently working on a website called CTDATA whose goal is to provide geo-referenced attribute data for Connecticut that will be compatible with the geography data files that MAGIC provides. The CTDATA website will eventually contain information on topics such as the economic characteristics, demographics and population of an area, all geo-referenced to be compatible with MAGIC's spatial data. The data will be available at many levels, including user-defined districts.

What do we need? There are a large number of individual agencies at the state and federal level that produce and publish datasets, in printed and/or website form. These datasets would benefit from being linked together and/or integrated into a whole. In addition, there is a lot of data that is recorded but not reported due to budget issues or perceived lack of interest in the data. More of the data that is recorded needs to be reported. Data sources should be linked to one another, be easily searchable, and easily accessible. Noting the unavailability of reliable and complete business information, Jeff Blodgett recommends creating a master business registry and undertaking a longitudinal analysis of businesses in Connecticut to provide data for both public and private concerns.

Connecticut business data is out there, but hard to come by. Business demographics are important tools for economic analysis and forecasting, yet much of this useful data that *is* collected is inaccessible. In addition, many different definitions exist for what a business is, further complicating any analysis. Jeff Blodgett therefore recommends the creation of a master business registry in Connecticut. The master business registry, once created, would integrate existing state and federal databases into a common file, linking agencies and records, with confidential data stored at the appropriate agencies. He also recommends implementation of a longitudinal panel study of Connecticut businesses. The study would include an analysis of the growth, decline and transformation of Connecticut businesses over a ten to thirty year period.

Individual health information and public health information are available in varying degrees in Connecticut. As noted by Dr. David Gregorio, individual health information is harder to access because of consent and confidentiality issues. Public health information is more easily accessible on these fronts, but still faces analytical issues such as generalizability and continuity. Still, a lot of health information is available in the State of Connecticut. For example, Connecticut's Annual

Registration Report of births, deaths, marriages and divorces has been kept for one hundred and fifty years, and the Connecticut Tumor Registry is the oldest and most complete registry of its kind in the United States. In addition, data is available in areas such as immunizations and infectious diseases, other chronic diseases, health behaviors and mental health and addictions, among many others.

What have others accomplished? The Indiana State Data is a shining example of what can and should be done in the State of Connecticut in terms of a state data center. The Statistics Canada project for complete digital individual health data records shows us what good, complete data can deliver. Finally, the FERRET project of the Census Bureau and the Center for Disease Control illustrates the benefits to both data suppliers and users of creating an online, decentralized data center.

The Indiana State Data Center is an example of what a state data center can and should be. It receives approximately $500,000 per year through the Indiana Business Research Center at the Indiana University Kelley School of Business. As noted by Dr. Morton Marcus, a state data center need not be funded by a single source, however. Sometimes it is better to get individual pieces of funding from different sources. The main purpose of a state data center should be to organize available data and make it available to users. It should be a broad attempt to organize and disseminate data that is already being collected. In some cases, data collected by individual federal, state, and local agencies is either never published or is discarded after a period of time. A state data center can organize such data and provide continuous access to it, sometimes even providing data to the agency that originally collected it. A state data center can make data available through websites, publications, and presentations. All three of these methods are important. Linking multiple websites together is a valid way to provide access to data, rather than trying to contain it all at one site. Publications are also important because they provide visibility and an opportunity for agencies that may not otherwise publish a chance to do so. Presentations also provide visibility. Finally, it is important that the data center understand the underlying meaning of the data to which it provides access. In that respect, it is important to build relationships with the recorders of the data. This takes time, but is well worth it.

Compressed Digital Health Records are an example of what can be done with complete data and new technologies. Traditional health care records have many problems including incompleteness, bulk and inaccessibility. Using new technology, a patient's entire health history from birth can be stored on a card and archived, accessible to any doctor or hospital. Patients benefit from the increased availability of their health information, as do HMO's, insurance companies, home care organizations, and the public. Such records can reduce storage costs, prevent adverse drug interactions, and allow medical care providers to give better care. Peter Gunther, President of Smith Gunther Associates, Ltd., suggests that Connecticut begin the new millennium by getting newborns on this system, and then building coverage from there.

FERRET, a joint project of the Census Bureau and the Center for Disease Control, provides access to the current population survey and all of the demographic surveys done inside the Census through its website. The Internet provides some difficulties, but is still the cheapest and fastest method for disseminating data. Dr. Cavan Capps, the Director of the FERRET Project, stresses that the goal of a state data center should be a virtual data library, with links to multiple sources of information and tightly linked documentation for the data that is available. Users should be able to find and manipulate the data as needed, and should be able to easily access documentation and

data attributes. Rules and suggestions for using the data should also be made available, as well as simple descriptive statistics for determining the usefulness of the data to the user before downloading large files. Finally, decentralization is advisable, as it leads to cost efficiencies in providing the data, as well as accountability for the data itself.

Where do we go from here? Now that we understand what data is currently available, what is needed, and what can be done, the next step is to begin the process of establishing a state data center for the State of Connecticut. This center would be responsible for making data on the State of Connecticut available and accessible to individuals and public and private establishments in the State of Connecticut, as well as elsewhere.

Most conference speakers noted the importance of focusing on the whole. That is, recognizing that providing access to all of the available data, regardless of its type, should be the main focus of a state data center in Connecticut. In addition, linking and providing access to existing data sources should be the first undertaking of such a data center, while increasing the stock of available data should be the second priority.

There is a lot of data available in the state of Connecticut–the goal of a state data center should be to make that data more readily available and accessible. Many individuals have access to data that is more current than what is published, and would be quite willing to enter the data themselves and take responsibility for keeping it current.

The data center does not need to be a single independent organization–it can and should be a collaborative effort of a variety of agencies. It is important to consider the individual incentives of the various people and agencies involved in the process–giving credit to the individuals and agencies that provide the data is necessary and important. In the long run, this is the only way to sustain the data center. People must have ownership, a vested interest in preserving the data center, in order for it to survive in the long run. In this sense, broad participation is vital.

The hope is that the legislature will approve a study bill to analyze the data issues in Connecticut, and to determine the amount of investment needed to create a Connecticut State Data Center. In addition, an institutional framework, in the form of an advisory panel, would be useful to deal with issues of confidentiality of data, uses and sources of data, and new data development.

Pat McGlamery
Assistant University Librarian & Director, MAGIC, Homer Babbidge Library, UCONN

MAGIC: The Connecticut Spatial Data Center

Mr. McGlamery's discussion centered on the University of Connecticut's Map and Geographic Information Center (MAGIC) website. The University's map library was renamed MAGIC approximately 4 years ago.

Since 1992, MAGIC has been working at the federal and state level to bring together an online collection of spatial data for Connecticut. There are currently 20,000 data sets for Connecticut on the website. These data sets include attribute data, digital cartography and photographic data.

What is special about spatial? Spatial data is not a map–it is data that is to be used in a Geographic Information Software (GIS) program. All of the data is integrated into a geo-referenced system, so that there are tables of data associated with each map. The user must have a GIS program to be able to use it. Map libraries then collect spatial information regardless of its format. They collect maps, tables (such as Excel files) and data.

What role does MAGIC play in the delivery of information as a library? MAGIC collects, describes, and provides access to data. The website contains data, reference aids, and web links. All of the data on the MAGIC website is spatial, and is integrated into a geo-referenced system. The data is in the public domain (free). The primary users are UConn faculty, students, and staff (1/3 of users). Secondary users are other citizens of Connecticut. However, people outside of the state access and use the data as well. 9000 gigabytes of data are downloaded per month from the MAGIC site, mostly state-level data, with spikes in March and November, probably corresponding to the academic calendar.

The data has been collected from many sources such as the Department of Environmental Protection, Census Bureau, U.S. Geological Society, Department of Transportation, and the Department of Economic and Community Development. Available data includes town boundaries and profiles, roads, hydrography, railroads, land use and cover, power lines, census tracts, arial photographs, etc. MAGIC is currently working with the Library of Congress, Harvard University, Yale University, Brown University and the Royal Library of the Netherlands to amass a comprehensive collection of maps of Connecticut before 1800 that will soon be available on the website.

What does the future hold? Future prospects for MAGIC include realistic 3-D visualization, interactive modeling and query with local and regional data, and discovery and extraction of Connecticut spatial data.

Dr. Robert Cromley

Director, University of Connecticut Center for Geographic Information and Analysis

UCCGIA: The University of Connecticut Center for Geographic Information and Analysis

Dr. Cromley's discussion centered on the University of Connecticut Center for Geographic Information and Analysis. The UCCGIA is a joint venture between the College of Liberal Arts and Sciences and the Homer Babbidge Library at the University of Connecticut. UCCGIA works closely with the Map and Geographic Information Center and is part of the Consortium for Public Policy and Research at the University.

The primary functions of the UCCGIA include training and education in the use of geographic data, facilitating archival processes related to geographic information, and facilitating geographic research.

UCCGIA's current projects include creating and maintaining metadata records, system administration for an Environmental Systems Research Institute (ESRI) software site license for all public institutions of higher education in Connecticut and establishment of the CTDATA website. The UCCGIA also coordinated UConn's application to be a member of the University Consortium for Geographic Information Sciences (UCGIS), a national lobbying effort to promote geographic information sciences.

CTDATA is a work-in-progress website for the dissemination of geo-referenced data within the State of Connecticut. The long-term goal for this website is to extract and serve attribute data files that are compatible with the geography data files served by the MAGIC website. Each file will be in the form of a table. The CTDATA website will eventually have economic databases and educational, political and population data by state, town, county, LMA, regional planning districts and user-defined districts. By clicking on a series of menus, the user will be able to "order" the data and then be presented with it.

Jeffrey Blodgett

Vice President for Information Resources, Connecticut Economic Resource Center, Inc.

Sources and Availability of Business Data: The Good, The Bad, and The Suppressed

Mr. Blodgett's presentation centered on the sources and availability of business data in the State of Connecticut, and what would be important elements in a study of business demographics. Mr. Blodgett advocates such an undertaking in the State of Connecticut.

What is business demographics? It is the science of dealing with the distribution, density, size, mix and decline of businesses over time. The three main components of a study of business demographics would be the absolute size, components of change, and population characteristics of businesses. The absolute size includes the number of businesses, their growth, decline, and mix. Components of change include the migration of businesses, mergers and acquisitions, changes in product line, start-ups and closings. Population characteristics include industry affiliation, employment size, geographic distribution, legal form of organization, inputs and outputs.

Why should we care about business demographics? Business establishments provide the grist for our economic statistical mills in the form of jobs, wages, profits, output and exports. We can get measures of job growth and change, indicators of entrepreneurial activity, insights for regional local planning and benchmarks for workforce training. This would permit the study of gross changes in business. For example, according to the New England Economic Indicators published by the FED, Connecticut new business incorporations have been experiencing a downward trend in recent years–from March of 1996 to March of 1998 the number of new business incorporations dropped from 375 to less than 225–but we don't know why.

How many businesses are there in CT? It depends on whom you ask and how you measure it.
Census bureau: 91,925 (all non-farm, non-government establishments with paid employees, 3/12/96)
Dept of Labor: 94,507 (current number of registered employers, doesn't include multiple locations)
Dun & Bradstreet: 160,000 (number of registered businesses)
Revenue Services: 178,408 (registered sales and use tax permits, 88,000 of which are corporate filers)
Secretary of the State: 184,442 (number of registered corporations and partnerships in CT)
IRS: 205,083 (number of filers of schedule C in CT)

A conference participant suggested one might also look at the number of business tax returns in Connecticut.

Key Public Sources of Business Data:
CT Dept of Labor: ES 202 theoretically provides the greatest level of industry and geographic detail. Contains data on establishments, wages, payroll and average wage, covers all non-farm establishments with paid employees, and provides monthly, quarterly and annual data. However, access to the data is limited.
Dept of Revenue Services: Provides data on corporate income taxes and sales and use tax permits. However, the type of economic activity is coded only at the time of initial registration with no updates. It is still, however, the best measure of sales activity at a 2-digit SIC level.
Secretary of the State: A constitutional agency empowered to register corporations and partnerships. The records are public, but there is no data on the type of economic activity.

Bureau of the Census: "Standard Statistical Establishment List" forms the basis for economic censuses and surveys. However, access is *extremely* limited.
Bureau of Labor Statistics (BLS): "Business Establishment List" for the United States. It is similar to the census, but again, access is limited.

Current Conditions: There are large datasets at both the federal and state level. However, they suffer from redundancy, a lack of interagency coordination, and no federal-state cooperation.

Mr. Blodgett recommends creating a master business registry and undertaking a longitudinal analysis of businesses in Connecticut. For the master business registry, he recommends first undertaking a feasibility study of a master business file. The first issue in that case is how to define "business." The master business registry would then integrate existing databases into a common, shared file, with agencies linked to master records and confidential data housed at the appropriate agencies. This would create cost-savings and efficiencies across state agencies. For the longitudinal analysis, he recommends planning and implementing a longitudinal panel study of business growth, decline, and transformation in the state of Connecticut over a ten to thirty year period.

Mr. Blodgett closed with the following quote from Business Week: "Without good statistics, we don't know whether what we are doing is working. Better economic data will mean better economic policymaking by government, better decisions by investors and corporations and, ultimately, a higher standard of living for everyone–and we'll even be able to measure it."

Dr. David Gregorio
Department of Community Medicine, University of Connecticut School of Medicine
The Health of State Public Health Statistics

Dr. Gregorio discussed the sources and availability of individual and public health data in the State of Connecticut and the issues surrounding access to such data. He began by noting that public health was "decidedly underrepresented" at the conference.

Health Information is separated into two categories: individual health information and public health information. Individual health information includes biographical information about individual patients such as the patient's health history, illness episodes, medical care provider, insurance coverage and treatment options. Public health information includes historical information about populations of people such as their general health status, incidence and prevalence rates, medical care delivery systems, health care finance and disease control strategies. Both types of data incur data issues of confidentiality and consent, as well as analytical issues of generalizability, continuity and hidden arguments.

Current Health Data Sources:
Annual Registration Report of Vital Statistics includes births, deaths, marriages and divorces. Connecticut has 150 years of these reports.
Immunization and Infectious Disease Reporting by sources such as the Connecticut Immunization Registry and Tracking System (CIRTS), Infectious Disease Registry, STD Registry, Lyme Disease Registry and Food-Borne Illness Registry.
Chronic Disease Reporting by the Connecticut Tumor Registry (CTR) which is the oldest and most complete registry in the United States and second in the world. Other areas of reporting include lead poisoning surveillance, occupational diseases, disorders and birth defects surveillance.
Health Behavior Surveillance Systems such as the Behavioral Risk Factor Surveillance System, Mother and Child Health Indicators (collects birth weights, prenatal care, birth outcomes), Family Health Indicators (pregnancy prevention activities), and Connecticut Health Check (a periodic survey of students in grades 6-12).
Other Areas of Data include mental health and addictions, lab services, professional licensure and regulation, health care facilities and medical care utilization.

Spatial Locators for the data vary according to the type of data. Some public health data, such as the tumor registry, is linked to street addresses. For personal data, issues of consent are major barriers–rules about accessing the data are evolving and becoming "much more stringent" than they have been in the past.

A conference participant noted that there is a need for socioeconomic statistics to link to public health profiles. In response, Fred Carstensen noted that income inequality, not income level, is linked to the condition of the public health profile.

In response to a question from a conference participant, Dr. Gregorio noted that overall, Connecticut is "not at the top or bottom" of any health indicator. From a national perspective, the State of Connecticut does quite well, however, Dr. Gregorio noted that there are "very dramatic pockets of disadvantage which are reflected in compromised health status." Public health researchers are most interested not in the medical causes of death, but in the underlying behavioral ones, and in that respect, there are "less than satisfactory levels of behavior than we would like to

see," and "some serious pockets of concern" in Connecticut. Also, there are newly growing morbidities, such as neglect and interpersonal abuse.

Dr. Morton Marcus

Director, The Indiana Business Research Center, Kelley School of Business at Indiana University

ISDC: The Making of a State Data Center

Dr. Marcus discussed the origin and operation of the state data center in Indiana.

The Indiana State Data Center Does Not Exist. However, the Indiana State Library does exist. It is the lead agency for the nonexistent Indiana State Data Center (ISDC). The library covers the administrative details and phones. The ISDC is a "wholesaler" of information, while the library covers the "retail" end.

History: The Indiana Department of Commerce originally commissioned the data center project as the Indiana Business Research Center at the Indiana University Kelley School of Business in 1970. Dr. Marcus was in charge, and realizing that there was no verification of the data being entered, he subsequently closed it down and got set up in the State Library system using Library Services and Construction Act funds. Later, the State Data Center Program was formed. The money is allocated to the Indiana State Library and the director of the library directs the funds to the State Data Center Program. Currently, there is an appropriation of approximately $500,000 per year to the Indiana Business Research Center at the Kelley School of Business, which supports the ISDC's activities.

Purpose: The ISDC provides information to the Lieutenant Governor, economic development organizations within Indiana, and to anyone else who needs data. The problem that arises with any data center is that you can only work with the data that is available, and you must decide how to organize it so that people can use it.

Important Players:

The **unit of analysis** (the person or business being monitored or measured)
The **recorder (**the person or agency that records the data)
The **storage agent**, (the person or agency that stores the data)
The **reporter**, if any (the person or agency that reports the data)
The **analyst** (a person who wants to get into the details)
The **user** (the person who makes policy or business decisions)

The storage agent is very often the same as the recorder, but can be a separate entity such as a library. The storage agent is very important because some recorders discard information after a certain period of time. In that case, the storage agent will have data that the recorder no longer does. The reporter is very often the same as the recorder, though a great deal of information is never reported, either due to budget constraints or lack of willingness to report. "A state data center has to be concerned about the legitimate interests of each of these individuals."

The State Data Center need not know everything or take a position on anything. It is important, however, to understand where the numbers are coming from and what they mean. Dr. Marcus stressed that you must talk to the people who put the numbers together to understand what the numbers mean at the recording level. The State Data Center also does not need to exist as a formally funded entity. For that, you would need an appropriation as such. It is sometimes easier to get different pieces of funding from different sources and then put it all together.

Outputs/Products:
Websites can contain static components (such as reports), dynamic components (such as unemployment figures), and interactive components (wherein the user has more control over the specific data output they get). Dr. Marcus noted that multiple websites that are linked together make sense. "Every agency wants its own website…Let them have it. Don't fight duplication of effort." The marginal cost of this duplication is very small.
Publications are also very important. It is necessary for every agency involved in supplying the data to have ownership of it. Relationships with the agencies are vital to maintaining the data. Printed publications, press releases and profiles can give the people who supply the data visibility by allowing them to publish where their own agency may not have the resources to do so.
Presentations are also important–talk to anyone who will listen.

In response to a question from a conference participant, Dr. Marcus explained that over time, you can build relationships with agencies and play a role in the generation of data. The ISDC currently works with the Census Bureau and the BEA in Washington on how to improve the entering and presentation of data, both off the record and without charge. But these kinds of roles and relationships happen over an extended period of time.

Perhaps most important, Dr. Marcus stressed that a state data center in Connecticut should be a broad attempt to bring information collected by state and federal agencies to the State of Connecticut. The specific kinds of data are not important–the agency should be concerned with the organization and dissemination of what is available.

Peter Gunther
President, Smith Gunther Associates, Ltd., Ontario, Canada
Digital Compression and Patient Care Records

Mr. Gunther's discussion centered on data issues and new technologies in patient care records.

Problems with traditional health records in both the United States and Canada include that they may be illegible, incomplete, not fully portable or inaccessible. For example, many hospital records are stored off campus 31 days after one has left the hospital. In some cases, it becomes easier to redo tests than to access old records.

What are the advantages of a digital health records system? They will be accurate, legible, complete, compressed (and therefore easily and efficiently stored on a single file archived on a card), can utilize artificial intelligence (to check prescriptions for conflicts), and data will be accessible on a need to know basis.

Who wins in establishing these systems? Patients are insured against receiving incorrect or conflicting drugs, do not need to unnecessarily repeat tests, and have a complete health record on file which is accessible to all of the patient's medical providers. The public wins in terms of earlier identification of plagues. The medical profession can handle more patients. HMO's can send people home to better care sooner, and don't need to incur super storage costs for items such as X-rays. Home Care Providers can service more patients with the same level of personnel and improve their services. Insurance companies also win because extended longevity means greater profits.

What are the necessary components of the system? Communication systems (secure notepads for medical professionals), archives (local, state and perhaps national), digital compression and decompression facilities (at all hospitals and in all medical professional offices), set-top boxes (hooked into the communication system to connect to the archive), patient care cards (which contain all of the patient's information) and software (the artificial intelligence to run the system). In Ontario, the system is about six months off from implementation.

Mr. Gunther challenges Connecticut to begin the new millennium by getting babies born now on such a system, and then build from there. The records could be done on a voluntary basis.

A conference participant inquired about the cost to alternative medical practitioners such as massage therapists. Mr. Gunther responded that the set-top box would cost less than $2,000.

Finally, Mr. Gunther indicated that we will need information filters within the system so that people are not harassed to donate blood, bone marrow, etc.

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.