

Expanding Researcher Access to EIA Microdata
Jacob Bournazian
SMG, EIA
jacob.bournazian@eia.doe.gov

Abstract

EIA has provided restricted access to approved researchers through the ASA/EIA Fellowship Program that required the researcher to work on-site. In the future, EIA will provide similar on-site access to approved researchers under the NISS/EIA student research program. EIA is evaluating alternative methods for expanding researcher access to energy microdata. One option is to enter into an arrangement with another federal agency to host EIA data at a Research Data Center. Another option is to grant researcher access through some type of secure remote access facility. Some agencies manage such facilities themselves. However, there is also a capability that is now being offered by the National Opinion Research Center that is being used or considered for use by smaller statistical agencies.

Nine major statistical agencies provide researchers with some form of restricted access to confidential data. Four agencies (Census, NCES, NASS, and NCHS) also provide on-line data base query systems that allow researcher access to suitably protected files from personal computers at remote locations. EIA's research access program is much more limited than those offered by other statistical agencies. Expanding researcher access to EIA microdata would increase the agency's rate of return on its investment for data collection. This paper discusses options for expanding researcher access to EIA microdata while maintaining the data protections that it pledged to survey respondents.

Questions for the ASA Energy Committee

1. Which mode(s) of researcher access should EIA provide?
2. Which mode(s) of access works best for the ability to do research?
3. What mode of access would researchers utilize the most?
4. What issues arise for researchers when accessing microdata from federal statistical agencies?

Introduction.

For the majority of information that the Energy Information Administration (EIA) collects through its surveys, it pledges to protect the company identifiable information according to certain laws or policies. For twelve (12) surveys, this protection is based on the Confidentiality Information Protection and Statistical Efficiency Act of 2002

* This is a working document prepared by the Energy Information Administration (EIA) in order to solicit advice and comment on statistical matters from the American Statistical Association Committee on Energy Statistics. This topic will be discussed at EIA's Spring, 2007 meeting with the Committee to be held April 19 and 20, 2007.

(CIPSEA). For the over 50 remaining surveys, this protection is based on policy and a determination that the information has commercial value that is exempt from disclosure under a listed exemption of the Freedom of Information Act.

Section 12(f) of the Federal Energy Administration Act of 1974 requires EIA to share information, in a manner designed to preserve its confidentiality, with other federal agencies that is consistent with their official use and purpose.¹ In performing its data stewardship role as a statistical agency, EIA needs to maintain a balance between protecting the survey information it collects and providing information to make informed decisions, develop policy, and evaluating and analyzing energy information. EIA has two main options for protecting the confidentiality of the information it releases.

EIA's current approach is to only release aggregate statistics and protect any sensitive table cells from revealing company identifiable values by withholding those cell values from release. This approach is called "restricted data." Another approach is to impose conditions on who may have access, for what purposes, at what locations and on what terms and conditions. This second approach is called "restricted access." This paper discusses current options EIA is considering for expanding researcher access to survey data.

Alternative Programs That EIA Is Considering.

In the past, EIA provided restricted access to approved researchers through a joint fellowship program between the American Statistical Association and EIA. A total of six (6) researchers participated in this research program from 2002 – 2006. EIA is considering providing similar on-site access to approved researchers under a new arrangement with the National Institute of Statistical Sciences. A second option EIA is considering is to enter in to an arrangement with the U.S. Census Bureau to host EIA data at a Research Data Center. A third option is to grant researcher access through some type of secure remote access facility currently offered by the National Opinion Research Center. EIA is considering placing survey data from the Residential Energy Consumption Survey and Commercial Buildings Energy Consumption Survey at either the U.S. Census Bureau's RDCs or at NORC's data enclave. Survey data from the Manufacturers Energy Consumption Survey is already available at the Census Bureau's RDCs.

Licensing is another alternative where a researcher signs an agreement that allows them to install the restricted data on their computer in return for meeting the agency's conditions relating to maintaining confidentiality of the data. Licensing is the least common mode for permitting researcher access to microlevel data.² EIA is not considering a licensing program and so that option is not discussed in this paper.

Expanding researcher access to EIA company level data would increase the agency's rate of return on its investment for data collection. Finding new ways to use existing data may also lessen the need for new collection efforts and reduce burden on respondents. Researcher access to more survey data also allows an agency to learn more about its data,

statistical procedures, and validation of its survey responses.³ EIA needs to develop new policies that protect the reported survey responses from unauthorized disclosure while maximizing data quality and utility for users. Effective analysis and research many times requires that researchers be able to access the actual company level data.⁴ This may be necessary even though de-identified public use files are available. One example is utilizing outliers in a model. In determining the variance of the price level for a product in a certain geographic area, it is critical to know the highest or lowest prices in the market to accurately assess demand and consumer preferences.

In a recent report by an external study team, the group recommended that EIA increase its level of interaction with the research community to a level that is on par with EIA's level of interaction with the energy industry."⁵ Later the report states "Ideas to promote this interaction included providing a program of visiting scholars to bring academic researchers to EIA, and expanding research access to micro-data made under appropriate safeguards to protect the confidentiality of the reporting entities."⁶ EIA currently has no formal process in place for academic researchers to access its survey data.

EIA has always applied data safeguards to prevent the unauthorized disclosure of reported survey values. The reason that EIA pledges to protect information provided by respondents is that these pledges along with the data safeguard procedures maintain and improve the quality of the reported data. A breach of confidentiality or pledge to protect may or may not cause a particular EIA respondent direct harm or subject them to criminal or civil penalties. However, at a minimum, it will cause a loss of respect from those participating in EIA surveys and those concerns have been shown to affect respondent behavior and survey response rates.⁷ EIA's effort to establish and develop restricted access programs should consider adopting the necessary policies and procedures to avoid increasing the actual and perceived risks of a breach of a pledge to protect data. Also, selecting the proper mode for granting researcher access is important because public or industry perception that company level data are being misused may have an equal negative impact on response rates as an actual breach of a confidentiality pledge.⁸

Discussion of Options for Expanding Restricted Access.

Research fellowship programs are usually funded so that the researcher works at the agency's location. These programs allow researchers to work with company level data that otherwise would only be available to agency staff. The researcher becomes an agent of the agency and is subject to the agencies' laws and adheres to the same confidentiality obligations as regular agency employees.⁹ The advantage is that the research is done on-site, communication between staff and the researcher is maximized, and the ease of monitoring the research activity by the agency. The disadvantages of these programs are that the available program slots are limited, the research proposals are limited to certain areas, and the researcher must travel to the agency's office. The NISS/EIA student research program is one example of these type of fellowship programs. The NISS/EIA research projects would be performed on site at EIA's headquarters and limited to specific research topics.

Research Data Centers (RDCs) allow researchers to use company level data at the offices of the data holder, or at a controlled site designated by the data holder, under highly restricted conditions. Researchers travel to the designated site to access data where they are monitored by staff and/or a RDC administrator. The National Center for Health Statistics (NCHS), National Agricultural Statistics Service (NASS), and U.S. Census Bureau operate controlled access research center sites. Researchers do not have access to any network within the RDC or through the Internet. Staff is usually available for questions and the researchers' work is subject to an extensive review.¹⁰ The duration of a researcher's stay at an RDC may be as short as a week or extend several months. User fees are assessed for using an RDC. The greatest advantage of using RDCs is that they currently offer the largest universe of business databases for performing research.

RDCs have been an effective mode of access to control the risk of identification of company reported values while allowing researcher access to the underlying microdata. The main criticism of using RDCs is that it requires travel and lodging if a researcher is not located near an RDC and many times traveling to an RDC is not convenient. Another limitation for researchers is the possible requirements to use unfamiliar data analysis software.¹¹ The project approval process for researcher access to an RDC may also be lengthy.

Remote access facilities are a third mode of access that allows researcher access to company level data from a remote computer using the Internet. The clear benefit is that researchers work at their office and are not physically required to travel to a physical site in order to do research. Access from a remote computer creates several issues for maintaining data protections because users may submit multiple queries to the database and may potentially infer company level attributes by comparing results for some table cells against those generated from prior runs.¹² Additional data safeguards are required in terms of user authentication and other protections to prevent the theft or corruption to the data bases or unauthorized disclosures of survey data.

There are two types of remote access. The first type is "monitored remote access." This approach allows a researcher to submit a query to a database from a remote site but the researcher is not allowed to view the microlevel data. This may be accomplished by limiting the type of analysis that can be done with the data or allowing the output to pass through a series of automated or manual filters that block certain types of queries and results. The second type is "unmonitored remote access" where the researcher has complete access to analyze and view the microdata subject to certain technical and legal protections against disclosure.¹³

The development of computer science technology has enhanced the potential for providing online remote access systems and resolving data security concerns. Various protocols have been developed such as secure personal computers that prevent downloads and other features, monitoring PC activity through audit logs and trails, and restricting user access from specific pre-screened Internet Protocol addresses. Statistics Netherlands is currently experimenting with using a researcher's fingerprint as a form of biometric identification to verify the identity of the researcher trying to connect and

access data from a remote location.¹⁴ The databases that are available through remote access on the websites of federal statistical agencies are generally de-identified using similar disclosure avoidance techniques as those used for preparing public-use files.¹⁵

A novel mode of access that was developed by the National Organization for Research and Computing (NORC) creates a data enclave that hosts data from several agencies that researchers may access from either a remote computer or from a secure room at NORC offices in Washington, DC and Chicago, IL.¹⁶ NORC plans to open use of the data enclave beginning May, 2007. For using remote access, the researcher accesses data inside the enclave via an encrypted connection through Virtual Private Network (VPN) software controls. Citrix technology is also applied to prevent the user from downloading data or using “cut and paste” functions.¹⁷ The NORC approach is a portfolio approach that offers a range of services, depending upon the needs of the researcher and agency, such as review of proposals, creating new data sets by combining data from multiple data sets placed in the enclave, and researcher training.¹⁸ There is a similar project approval process that does not require the stringent backgrounds checks for approved access at an RDC. Research proposals would be submitted to NORC for initial review and then forwarded to EIA for approval. The researcher would need to be approved by EIA. There are also user fees charged to the researcher for using the data enclave but these are lower than those charged by an RDC.

Summary.

EIA is considering three options for expanding researcher access to company level survey data. The first is student/researcher projects that are approved under the new NISS/EIA student research program. A second option is to place survey data at the Census Bureau’s Research Data Centers. A third option is making the data available through the NORC data enclave project. Each option has its advantages and limitations. The options must balance the need for expanding research access while maintaining the confidentiality of the reported survey data. The optimum choice would be to expand researcher access to EIA survey data in the least burdensome manner to researchers, at a reasonable cost to EIA, so that it does not increase the risks of an unauthorized disclosure or breach of confidentiality among survey respondents.

¹ 15 USC 771(f), as amended.

² National Research Council, “Expanding Access to Research Data.” P. 33 (2005).

³ Id. at p. 39.

⁴ Lane, Julia. “Optimizing the Use of Micro-data.” P.10. Journal of Official Statistics. Statistics Sweden. Vol _ No. _ (2007).

⁵ “Challenges, Choices, Changes: An External Study of the Energy Information Administration” p. 37 (2006).

⁶ Id at p. 38.

⁷ National Research Council, “Expanding Access to Research Data.” P. 51 (2005).

⁸ Id.

⁹ Jabine, Thomas. “Procedures for Restricted Data Access.” Journal of Official Statistics. Vol. 9(2). P. 549 (1993).

¹⁰ Confidentiality and Data Access Committee, “Restricted Access Procedures,” Federal Committee on Statistical Methodology. P.2.

¹¹ Lane, Julia. “Optimizing the Use of Micro-data.” P.10. Journal of Official Statistics. Statistics Sweden. Vol _ No. _ (2007).

¹² National Research Council, “Expanding Access to Research Data.” P. 79 (2005).

¹³ Lane, Julia. “Optimizing the Use of Micro-data.” P.10. Journal of Official Statistics. Statistics Sweden. Vol _ No. _ (2007).

¹⁴ Id. at p.11.

¹⁵ National Research Council, “Expanding Access to Research Data.” P. 33. (2005).

¹⁶ Lane, Julia. “Creating a Data Enclave For Sensitive Microdata.” Presentation to EIA. April 10, 2007.

¹⁷ Id.

¹⁸ Id.