# Lessons learned from Internet dissemination of confidential farm survey results: USDA's Agricultural Resource Management Survey

Mitch Morehart and Charles Towe,*
Economic Research Service, USDA

**Abstract**

ERS and NASS annually collect a wealth of data that describe farming in America through the Agricultural Resource Management Survey (ARMS). While ARMS provides a rich set of economic data, customers outside the government, especially researchers, have gone largely under-served. ARMS data had been available only within USDA except when summary information was released publicly, when users requested ERS do special tabulations, or when academic researchers entered into special research agreements with ERS/NASS to access the data in USDA offices. In response, ERS developed easy to use web-based data delivery tools that expand access to farm survey data as a public good, while maintaining the security of the confidential data. Experience in the construct and maintenance of the tool since its inception, stakeholders use and satisfaction, and recent changes and enhancements are discussed.

**Keywords**: confidential survey data, ARMS, Internet, farm surveys

## 1 Introduction

Economic research is highly data dependent. Often, the most interesting research problems are stimulated by policy questions on distributional issues that simply cannot be addressed without microdata on establishments and the firms that own them. The importance of microdata to economic research was the theme of Heckman's Nobel Lecture where he suggested that: "The availability of new forms of data has raised challenges and opportunities that have stimulated all of the important developments in the field and have changed the way economists think about economic reality (Heckman, 2001)." Establishment data are vital in understanding individual or firm behavior and are necessary to determine the marginal impacts of changes in policy or from other internal or external events. Microdata enable analysts to do multivariate regressions, whereby the marginal impact of key variables, controlling for other factors, can be isolated (Lane, 2003). Widely accessible microdata also have the additional benefit of allowing replication and verification of research results.

USDA's National Agricultural Statistics Service (NASS) and Economic Research Service (ERS) have been collecting annual, farm-level economic data for more than twenty years in what is now known as the Agricultural Resource Management Survey (ARMS.)[1] The ARMS is critical to the research and analysis mission of the Economic Research Service, and is a key input to estimates across the Department of Agriculture and in other agencies. It is a valued and unique resource, since it is the only national survey from which observations of field-level farm practices, the economics of the farm business and the characteristics of the household operating the farm, are all collected annually in a representative sample.

While ARMS provides a rich set of information, customers outside the government, especially researchers, have gone largely under-served (U.S. General Accounting Office, 1992). Data from ARMS had been available only within USDA except when summary information was released publicly, when users requested ERS do special tabulations, or when academic researchers entered into special research agreements with ERS and NASS to access the data in USDA offices to test specific hypotheses. The limited access to ARMS data frustrated those who saw the broad benefit of the information it contained. This friction between data access and data confidentiality is a common occurrence, particularly as it pertains to Government data. The Committee on National Statistics (CNSTAT) recently conducted a comprehensive review of the risks and opportunities of expanding access to confidential data which highlights the main issues and documents current procedures and solutions employed by government agencies. (Panel on Data Access for Research Purposes, 2005). In essence, the tradeoffs involve the desire to get the highest return possible for substantial data collection costs and respondent burden to gather information necessary to produce official statistics and support economic research on one hand and the requirement to uphold the pledge of confidentiality and ensure the future participation of respondents.

To expand external researcher access, ERS and NASS developed dynamic, technologically advanced, and easy to use web-based data delivery tools that are readily available through the ERS website (www.ers.usda.gov). Internet services have become part of a comprehensive suite of on-line services offered by the agency for its external customers. Traditional means of disseminating this vital but sensitive information did not meet the needs of users who we found wanted better access, transparent processes, and the ability to work with the data on demand; not lim-

*The views expressed here are those of the author(s), and may not be attributed to the Economic Research Service or the U.S. Department of Agriculture.

[1] For a detailed perspective on the origin and use of ARMS as a principal USDA survey, see Johnson and Morehart (2006).

ited to sets of pre-programmed tables. The wealth of data was dispersed across the website and hard to find. The new suite of tools that provide selective access to ARMS data not only expand access to farm survey data as a public good, but maintain the security of the confidential data. Researchers now have instant access to tailored information about agricultural production technology, farm business viability, and the structure of U.S. agriculture. This paper describes the framework that guided development of dissemination tools and describes the procedures used to manage data access with adherence to confidentiality. We will also provide a perspective about some of the lessons learned in the construct of the tools and their acceptance by various stakeholders. The final section of the paper presents some recent improvements and future direction of the project.

## 2 Developing a two-tiered data dissemination system

One outcome of the initial planning stages for improving data access and dissemination was the recognition of two distinct audiences for ARMS information.[2] A large portion of the customer base was interested in having on-line access to summary tabulations of the data. Another group, primarily researchers, wanted the ability to access the raw data from their desktop and perform statistical analysis. With this in mind, two separate development tracks were initiated. One involved building a user friendly web tool that enabled users to select among survey data sets to build custom reports, refine queries with specific populations, group summary statistics for comparisons, and choose among several output options for results. The second track envisioned development of an experimental remote access for registered users (via a secure Extranet) to perform statistical analysis and economic modeling. Common to both of these products was the need to develop delivery methods and security protocols that ensured data confidentiality.

Considerable effort has gone into developing disclosure limitation methods for tabular data that effectively lower disclosure risk and provide useful products to data users. These techniques include cell suppression, local suppression, global recoding, rounding, and various forms of perturbation. Cell suppression techniques have been around for some time and were part of our internal data masking procedures prior to publishing survey results. A variety of automated and complex routines have been offered with varying degrees of success (see, for example, Kirkendall and Sande (1998), Fischetti and Salazar (2000), and Giessing (1999)). With guidance from these studies, we wrote a series of SAS IML macros to implement a sophisticated cell-suppression algorithm. Although table specific, it did recognize data relationships within rows of the predefined tables and across columns of output. The primary disclosure rules being implemented in the algorithm are know as the $(n, k)$ rule. The $n$ part of the rule identifies a threshold (3 observations in this case) for which sample size is small enough that the possibility of re-identification is too high. The $k$ part of the rule establishes a threshold for dominance (60 percent in this case) where identity and attribute disclosure risk become too high when a single observation accounts for the threshold amount or more of the total estimate for the cell item considered.

The final, and perhaps most challenging, aspect of the cell-suppression routine, was defining heuristics for complementary cell suppression. The secondary cell suppression problem is to apply these complementary suppressions to the set of sensitive cells in such a way as to ensure that the complementary suppressions create the required uncertainty about the true values of the sensitive cells, while still preserving as much information in the table as possible. Using known equation relationships the equation checking functionality (ECF) was designed to recommend to the requested application an appropriate list of variables to obscure in order to prevent solving across row variables to determine the cell value that failed legal disclosure. The process is as follows. First, a cell check is completed to determine the row variable names that fail legal disclosure. Second the row variable name is passed to the ECF. The ECF routine utilizes the existing row variable list and the lookup table defining the implicit functional relationships to determine the optimal answer, consisting of variable names to obscure. The ECF procedure is designed to select the optimal variables, which is the minimum number of variables, to complete the obfuscation task. Dimensionality is a curse with this type of brute force program in terms of efficiency. However, with the functional structure of the ARMS data tables this was not of significant consequence.

An additional controlling feature of the tabular summaries was that they followed standard accounting guidelines, which prescribed the row content, but allowed some flexibility in the level of aggregation. So, for example, categories of expenses for which there were limited responses (having greater potential for disclosure risk) such as livestock leasing were combined into a more general category of other livestock-related expenses. Micro-aggregation is a data perturbation method, characterized by the publication of only small aggregates instead of the original data. Other aggregation approaches involve combining geographic areas or other attributes to minimize disclosure risk. One example of this approach is the Web-based query systems developed by the National Institute of Statistical Sciences (NISS) that disseminate NASS data on usage of agricultural chemicals (fertilizers, fungicides, herbicides and pesticides) on farms. This principal of aggregation was also applied to classification variables used in the tables created from the ARMS.

Once the disclosure routines were in place, we were able to develop prototype delivery systems for evaluation on internal servers. The first working system ar-

[2]Several focus groups and usability studies were conducted during 2003.

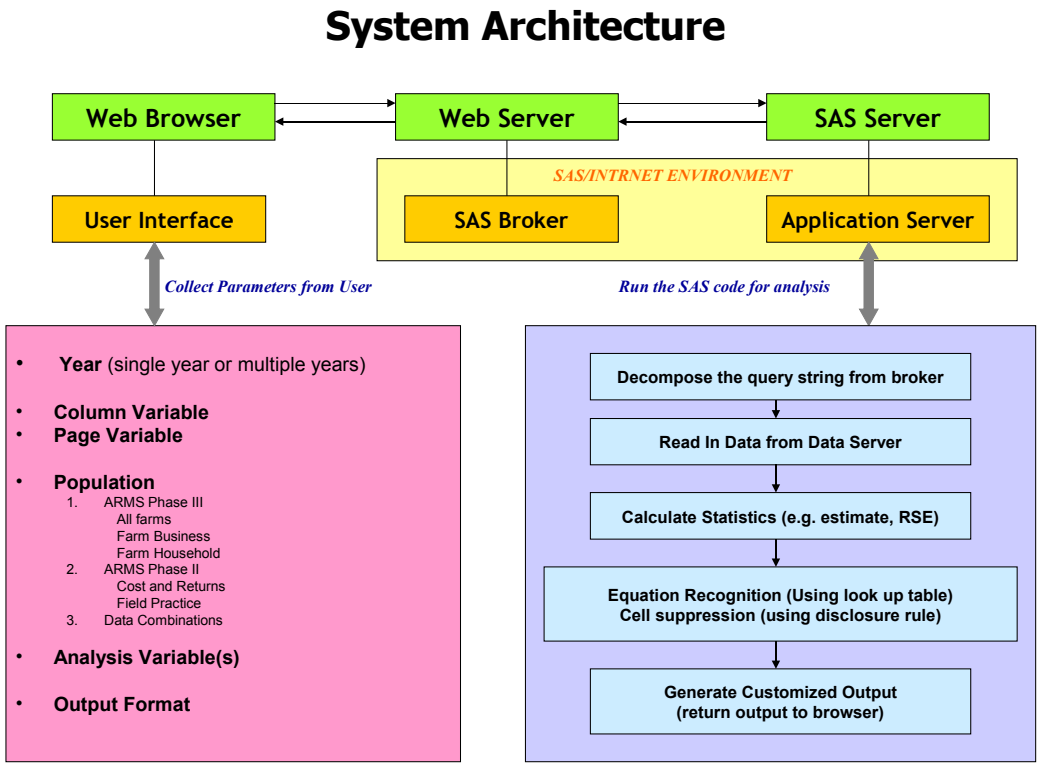


Figure 1: Initial data delivery system structure

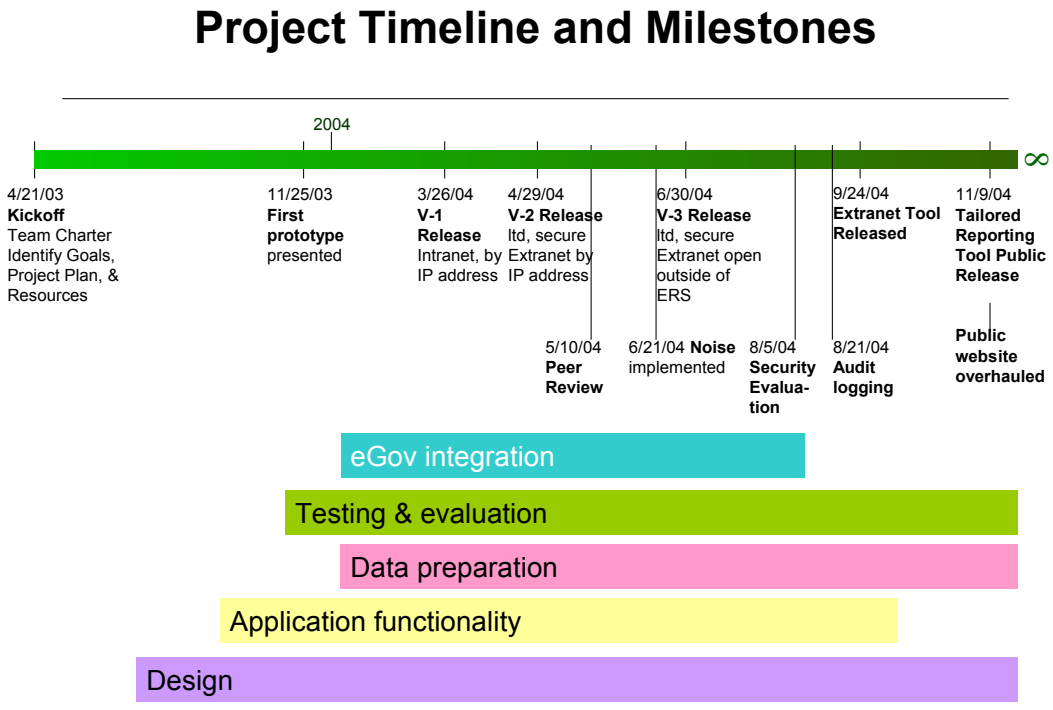


Figure 2: Initial data delivery system structure

chitecture placed a web interface in front of a dynamic SAS Intrnet data query structure. The user interface and menu utilized html coding and Microsoft .Net capabilities. The backbone of the system involved three servers, a web server, SAS server, and separate server to house the encrypted data (figure 1). User input was fed to the SAS broker, which decomposed the query string into SAS parameters. Data were then read into SAS based on the request. Requested statistics were then calculated along with measures of statistical reliability (relative stand error). The next step imposed the primary and complementary cell suppression algorithm. The final step generated a masked table for return to the web browser. This same backbone was used to service the advanced research user interface.

A peer review meeting that consisted of 32 attendees across two USDA Agencies (ERS, NASS), as well as individuals from other areas of the Federal Government was held on May 10, 2004. Its purpose was to provide a forum for open discussion of the ERS/NASS initiative to improve ARMS data dissemination via a controlled access web delivery tool, with specific regard to security and data confidentiality issues. The format consisted of a series of presentations by ERS and NASS staff that provided background on the Agricultural Resources Management Survey, current data dissemination methods, and the proposed solution to more broadly deliver ARMS data to the user community. The reviewers from five external agencies asked probing questions to which ERS and NASS staff responded. At the conclusion of the presentations, the reviewers were asked for their reactions to and concerns about any and all of the issues raised during any part of the presentations, to make suggestions for improvements, and to assist the project team by guiding its members to better solutions.

## 3 From prototype to final product

Peer review made an invaluable contribution that identified strengths and weakness of the working prototypes and established the major improvements necessary prior to implementation. There were three primary areas of concern 1) strengthening data security and protecting confidentiality, 2) delivery speed and system load capabilities, and 3) access tracking capabilities. External contractors were consulted in addressing some of these issues. Advances in computer technologies and software also facilitated some improvements to the system architecture. Figure 2 show the project time line from first prototype to a public release of the tools and highlight the major activities during this period.

An inherent layer of confidentiality protection for the ARMS is that it is a sample survey rather than a census. For sample surveys, estimates are made by multiplying an individual respondent's data by a sampling weight before they are aggregated. Since sampling weights are not published, this weighting helps to make an individual respondent's data less identifiable from published totals. Only providing weighted summary statistics also provided an opportunity to implement data masking directed toward the weights themselves. This involved adding zero mean noise to the determination of weights for each respondent such that the effect of the noise on items that were not at risk for disclosure was minimized Evans et al. (1998), Duncan, George T. and Mukherjee, Sumitra (2000), and Camden et al. (2003). In sample surveys, each respondent's data is generally weighted inversely proportionally to the probability of being selected in the sample. Individuals with the lowest valued weights are those most at risk for disclosure. With noise added to the weights maintained in the survey data base, we retained the $(n, k)$ cell primary suppression rules and added additional rules that flagged cells that contained a large percentage of noise.[3] As noise was imposed in the raw data at NASS, these protections were one component of the security applied to the advanced statistical analysis Extranet. Additional precautions were added to the ARMS Extranet Online Tool to limit statistical analysis to no less than 30 samples, trim the upper and lower distribution tails so that minimum and maximum queries did not disclose individual values, and exclude any potential identifier variables including the weights (Allen (1992)).

The original prototype which used real-time SAS running in the background against web queries was too slow for Internet delivery, was problematic with high use loads, and had significant system maintenance costs. To address this we reformulated the system architecture for the dissemination of tailored reports to include two new additional steps. First, all table cells would be calculated and stored in an SQL database with the appropriate suppression algorithms applied. The web user interface was then reconfigured to query against this SQL data base. As a result, processing time was dramatically improved, and system maintenance reduced to semi-annual processing to create the SQL database. The new system architecture for tailored reports also provided greater flexibility in the type of tabular presentations (for example showing years as columns rather than just classification variables) and more easily accommodated adding graphing capabilities. The advanced statistical Extranet application continued to rely on a broker and SAS running, real-time, in the background of requests. There were some enhancements to this process that improved response speeds, but the majority of time was spent reconfiguring the interface based on usability test results.

The ability to track access and store the results of user submissions was advocated as a necessary component of the advanced statistical Extranet application. By signing the Confidentiality Agreement users agree that any data file provided to them "...will be used only for statistical reporting and analysis and will not be published or released in identifiable form." In this context, the term

[3]The specific procedure used to add noise is not known to data users and therefore not provided in this paper. This non-disclosure of the specific parameters is a necessary layer of confidentiality protection.

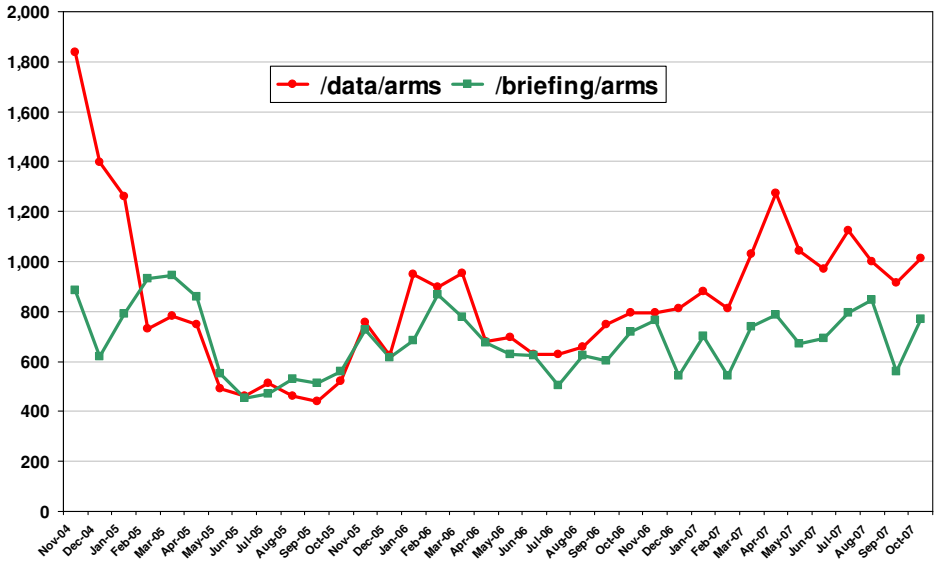


Figure 3: Initial data delivery system structure

"statistical summary information" means the result(s) of statistical analysis in any of the following forms: record listings, frequency tabulations, magnitude tabulations, means, variances, regression coefficients, correlation coefficients, graphical displays and any other result of an analytic process. While the Extranet application has several built-in mechanisms to reduce the potential for disclosure, the ultimate measure of disclosure risk is visual inspection of output. To minimize the burden of this activity the web delivery system must have the capability to track access and store results in order to allow for output review.

## 4 System performance

We are well into the third year of operating and maintaining the ARMS data delivery tools. Since its launch in November of 2004, the customized data summary tool has averaged 784 unique visits per month (figure 3). It has become a main feature of the ERS website data page and has been widely accepted and used by our customer base. Online access makes analyzing natural resource, technology adoption, farm business, and farm household issues less costly and more efficient. One-stop shopping improves value and provides a usable, standardized method of obtaining ARMS data. Separately produced outputs, some with product specific formats and programming, dispersed throughout the website were replaced with a single robust product. The improved access to data has reduced demand on staff to help find information and in requests for special tabulations. Data consistency is better managed in the update process, the data is easier to find for users, and programming is centralized for improved sustainability. The centralized system also facilitates better access to survey procedures and documentation so that users know what they are getting with ARMS data.

The capability to interact with data users is an important feature of the customized data summary tool. We have received an average of 6 inquiries a month. Topics range from specific data questions to general comments about the web interface and usability of the tool. Several changes have been made as a result of user feedback. For example, the farm production specialty classification variable was modified in two of the featured states to accommodate separate analysis of fruits, vegetable, and nursery and greenhouse operations that are normally collapsed into one category called specialty crops.

The advanced statistical analysis component of the web delivery tools that is provided via secure Extranet has had limited use and is much more resource demanding to update and maintain. Since its launch, there have been fewer than 30 users that have accessed the system. Beyond the initial requirements of having a memorandum

of understating that defines research goals and uses of the data and signing the confidentiality agreement, users are required to obtained customer USDA Level 2 eAuthentication ID. This involves some additional paper work and appearing in person at a local USDA designated Service Center. The advanced statistical tools that are currently available on the ARMS Extranet Online Tool are limited to exploratory variable summaries and linear regression analysis. This system has made access more convenient, but only provides data users the ability to do a preliminary evaluation of their research application.

## 5 Future Enhancements

The high level of user satisfaction and relatively low maintenance costs of the web tool that provides summary tabulations suggests that the primary focus for future development is better meeting the needs of researchers that want a convenient way to conduct statistical analysis. The capability provided by continued improvement of computer technologies has widened the scope of possibilities for access to restricted data (Wolf (2002) and King (2007)). Considerations for system costs and maintenance and delivering a high level of user services also are important, particularly for relatively small agencies such as the Economic Research Service. The wide acceptance and use of the tabular summaries provided on the ERS website had stimulated interest among data users, so there is no anticipation of a reduction in the demand for access. The other consideration, that we initially underestimated, is the sophistication of researchers in terms of computer software and data analysis. Their feedback confirms that most researchers prefer to have a more direct capability to write and submit code against the data rather than a complex web menu system that works as a front end for code processing.

With this in mind, we conducted a comprehensive review of existing systems that allow remote access to restricted data. Some of the systems reviewed included 1) the Luxembourg Income Study System (LIS), 2) Remote Data Access (RDA) of Statistics Canada, 3) Remote Access Data Laboratory (RADL) of Australian Bureau of Statistics (ABS), and 4) Research Data Center (RDC) of National Center for Health Statistics (NCHS). There were many common features across these systems such as an email or web interface and allowance for a variety of statistical processing code (SAS, SPSS, STATA, etc.). In each of these systems, security protocols were put in place such that researchers did not have direct access to microdata, results were subjected to confidentiality review before being sent back to the user, and there were usage logs kept. While construction of a similar system at ERS is feasible, costs are prohibitive, particularly as the number of users increase.

ERS and NASS recently initiated a 2-year pilot project to study the feasibility of remote access to ARMS data using

Of recent interest is the possibility of participating in a data enclave, with the capability to accommodate secure remote access. This type of system has been considered for other sensitive data such as the Health and Retirement Study (Nolte and Keller (2004)). The most promising prospect for small agencies, such as ERS, is the data enclave being proposed by the National Opinion Research Center (NORC). [4] It is designed to provide a secure mechanism for producers of sensitive data to enable more convenient access to approved researchers. Researchers will be able to access ARMS data from their office desktop computers using a secure Citrix environment. In addition to data warehousing and confidentiality protections, NORC will set up and manage an active outreach program to inform the national research community of the data and to foster the use of the data in research leading to conference presentations and journal publications. NORC also will establish an extensive education program to ensure appropriate use and disclosure of the data, including confidential aspects of the data. The enclave is setup as a collaborative environment so that reserachers can shre documentation and code.

## References

Allen, R. (1992), "Statistical Defensibility as Used by U.S. Department of Agriculture, National Agricultural Statistics Service," *Journal of Official Statistics*, 8, 481–498.

Camden, M., Daish, K., and Krsinich, F. (2003), "The Noise Method for Tables - Research and Applications at Statistics New Zealand," in *Joint ECE/Eurostat work session on statistical data confidentiality*, Luxembourg: United National Statistical Commissionand Economic Commission for Europe Conference of European Statisticians, no. Working Paper 28.

Duncan, George T. and Mukherjee, Sumitra (2000), "Optimal Disclosure Limitation Strategy in Statistical Databases: Deterring Tracker Attacks through Additive Noise," *Journal of the American Statistical Association*, 95, 720–729.

Evans, T., Zayatz, L., and Slanta, J. (1998), "Using Noise for Disclosure Limitation of Establishment Tabular Data," *Journal of Official Statistics*, 14, 537–551.

Fischetti, M. and Salazar, J. J. (2000), "Complementary Cell Suppression for Statistical Disclosure Control in Tabular Data with Linear Constraints," .

Giessing, S. (1999), "A Survey on Software Packages for Automated Secondary Cell Suppression," .

Heckman, J. J. (2001), "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *Journal of Political Economy*, 109, 673–748, available at http://ideas.repec.org/a/ucp/jpolec/v109y2001i4p673-748.html.

[4] For more information, see: http://dataenclave.norc.org/.

Johnson, J. and Morehart, M. (2006), *The Wye Group Handbook: Rural Households' Livelihood and Well-Being*, UNECE, Eurostat, FAO, OECD, World Bank, chap. Income and Wealth Statistics for Selected Countries: The Agricultural Resource Management Survey (ARMS), pp. 1–30, Chapter 14.1.1.

Karr, A., Lee, J., Sani, A., Hernandez, J., Karimiand, S., and Litwin., K. (2000), "Web-Based Systems that Disseminate Information from Data but Protect Confidentiality," Tech. rep., National Institute of Statistical Sciences.

King, G. (2007), "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing," Tech. rep., Harvard University.

Kirkendall, N. and Sande, G. (1998), "Comparison of Systems Implementing Automated Cell Suppression for Economic Statistics," *Journal of Official Statistics*, 14, 513–535.

Lane, J. (2003), "Uses of Microdata: Keynote Speech," in *Statistical Confidentiality and Access to Microdata: Proceedings of the Seminar Session of the 2003 Conference of European Statisticians*, Geneva, pp. 11–20.

Nolte, M. A. and Keller, J. J. (2004), "Research Use of Restricted Data: The HRS Experience," Joint Statistical Meetings, Toronto.

Panel on Data Access for Research Purposes, N. R. C. (2005), *Expanding Access to Research Data: Reconciling Risks and Opportunities (2005)*, The National Academic Press.

U.S. General Accounting Office (1992), "Data Collection: Opportunities to Improve USDA's Farm Costs and Returns Survey," Publication GAO/RCED-92-175, Washington, DC.

Wolf, V. D. (2002), "Issues in accessing and sharing confidential survey and social science data," *Data Science Journal*, 2, 66–74.