

Session 9

SHARING DATA FOR STATISTICAL PURPOSES

REALIZING THE PROMISE OF DATA SHARING

THOMAS L. MESENBURG
ASSISTANT DIRECTOR FOR ECONOMIC PROGRAMS
BUREAU OF THE CENSUS
WASHINGTON, DC

COUNCIL OF PROFESSIONAL ASSOCIATIONS
ON FEDERAL STATISTICS
BETHESDA, MARYLAND
NOVEMBER 13, 1996

REALIZING THE PROMISE OF DATA SHARING

I have three purposes today -- to (1) outline my perspectives on new data sharing authority, (2) engage you in the challenges of implementing new authority, and (3) suggest an available enhancement opportunity.

I. PERSPECTIVES ON DATA SHARING

Both the promise of new Federal data sharing authority, and the barriers to obtaining it, are real and worth some reflection.

Data Sharing Has Promise

The promise of data sharing is to improve official statistics, and strengthen supporting agency operations, without having to change our decentralized statistical "system."

- It offers a "win-win" opportunity for users of statistics, providers of business and household information, and participating Federal agencies.
- It won't eliminate Federal debt, or assure world order, but it does present an exceptionally favorable balance of likely benefits and modest risks.

Improved Statistics. Improving official statistics is a goal that has widespread and long-standing support. Data sharing would help achieve this goal for many economic, and demographic data series by:

- Making single reports more complete or reliable where issuing agencies can use similar or related content information shared by other agencies.
- Making related reports more comparable where separate issuing agencies can receive shared technical detail and applied methods.
- Encouraging wider interagency use of "best series" content and "best methods" statistical practices.

Strengthened Operations. Strengthening statistical operations across agencies is the means by which data completeness, reliability and comparability can be increased. It is also the means by which additional benefits can be realized. For example:

- Completeness of Census information on foreign-ownership of U.S. companies would be increased if it could include confidential BEA data.
- Reliability of BEA national income and product estimates would be higher if BEA's sources could include confidential Census and BLS data.
- Comparability of Census and BLS industry statistics would be greater if they could use the same 4-digit SIC classifications for each establishment.
- Statistical agency costs would be lower if separate data collections could be avoided, reduced in size, coordinated, or streamlined in process.

Despite the real promise of increased statistical data sharing, it cannot begin and will not be possible unless significant barriers are recognized and overcome.

- Data sharing authority is not on "cruise control;" enacting it will require navigation and direction by the interested parties!

Barriers To New Authority

The barriers to new Federal statistical agency data sharing authority are substantial, and include both legal considerations and practical concerns.

Legal Considerations. Legally, data sharing cannot greatly expand without changes in existing law. And to be fully effective, data sharing amendments must be adequate in scope and tailored to statistical agency differences.

Amendments proposed in 1996 would be adequate to remove existing barriers. Taken together, H.R. 3924 (introduced by Cong. Horn) and Internal Revenue Code amendments (forwarded by Treasury Assistant Secretary Samuels) would:

- Authorize eight major statistical agencies to receive shared data.
- Allow sharing only by written agreement of all participating agencies.
- Assure effective actual use of shared data and full data confidentiality.
- Provide conforming amendments for six statistical agencies.

It is critical that data sharing amendments allow tax information to be shared with the designated statistical agencies. Without including the Treasury-proposed Internal Revenue Code amendments, data sharing's promise cannot be realized. To illustrate:

- Since the early 1950's, the Census Bureau has received and used extensive tax information from the Internal Revenue Service (IRS).
- We use it primarily to help maintain a complete and current business register, and augment the coverage of five-year census statistics.
- The 1996 register includes tax information for some 5 million smaller businesses that is essential for efficient and reliable survey samples.
- For other data sharing agencies, use of this complete register is indispensable to permit smaller survey samples, reduced burden, and lower costs.

Practical Concerns. We expect the 1996 data sharing and companion bills to be repropose in 1997. However, I don't need to remind you that proposing and enacting legislation can be two different things. Timely action on data sharing amendments will require :

- Congressional leadership, and broad committee support.
- Interagency cooperation on action strategies and plans
- Demonstration of significant public and program benefits.
- Sensitivity to individual and information confidentiality concerns.

One likely legislative issue will be -- what should be the priority for different kinds of statistical changes? That is, can organizational and data sharing changes both be considered; if not, which should come first, and why? I:

- Assume there may be diverse views around the room on this issue.
- See data-sharing as clearly worth pursuing, with or without other changes.
- Urge that agencies and the Congress make it an early action priority.

A different kind of practical concern is our deep-seated tradition of decentralized statistical operations. It has strengths that we are familiar with and value. However, one of them is not natural and sustained cooperation on specific initiatives.

Obtaining and using adequate data sharing legislation is an initiative that will require our highly decentralized statistical system to get together and stay together until the “deed is done.”

- It will involve eight major agencies plus the IRS, several Congressional committees, and numerous data users and policy participants.
- It means that to enact legislation, and achieve “hard” benefits, affected agencies must institute and sustain new levels of specific cooperation.
- Instituting and sustaining the needed levels of interagency cooperation will be easier said than done, and require arrangements yet to be developed.

But let’s move on. Let’s assume adequate data sharing authority becomes law, and consider some of the implementation challenges we would face.

II. CHALLENGES TO SHARING DATA

Data sharing legislation would not be self-executing. It would “authorize” but not “require” implementing actions. Affected agencies would include the Office of Management and Budget (OMB), eight designated statistical agencies, the Internal Revenue Service, and perhaps other agencies. What actions might be needed, what might those actions require, and what risks might they present?

What should be the priorities for implementation?

As proposed in 1996, data sharing authority would have broad statistical program coverage. The eight statistical data centers are responsible for most general purpose statistics; potentially any Federal executive agency could participate in a data sharing project; and data sharing will require some resource commitments.

Whether by choice or by consequence, priorities will become part of the data sharing implementation process. I would prefer that we prioritize by choice, and seek projects that (a) have the biggest impact in terms of our statutory purposes, and (b) present the lowest risks in terms of threats to data security and confidentiality. For example, we might initially emphasize:

- Projects that seek to greatly improve sample frames, and can reduce burden and costs without compromising information quality.
- Data collections that involve business respondents, who seem to be less concerned than individuals about data confidentiality and security risks.

Should “third parties” have access to shared data?

The 1996 proposal required data centers to use shared data “exclusively for statistical purposes,” but did not require them to limit data access to their own employees. Centers could allow access by “agents” who work under the agency’s supervision, and accept the legal conditions under which data were shared.

But statistical agency operating practices vary widely; “agents” can hold quite different degrees of discretion; and our “information age” demands that agencies be sensitive to real and perceived concerns about individual information security. Specific issues regarding “third party” access may include:

- Whether use of an agent or contractor is new; should long-standing or existing uses be different from proposed new uses?
- Whether the basis for accountability is important; should agency-sworn individuals be different than contractor employees?
- Whether the functions purchased are important; should hired central computer services be different than local field interviewers?
- Whether other contractor activities are important; should small or specialized firms be different from major information companies?

Because data sharing will be new, carefully scrutinized, and easily misinterpreted, I would urge initial caution in allowing third parties access to shared data. For example, I would think that third party access to confidential data files:

- For production purposes as part of an ongoing contracted-out data processing service would be warranted, as clearly cost-effective.
- For sample selection purposes by a national household survey company would be unwise, as subject to real and perceived misuse.

Should historical data files be shared?

Substantial benefits can come from sharing data that have been collected and are now held by Federal statistical agencies. For example, Census Bureau business register information could help social program agencies that survey for-profit service providers, such as hospitals, medical clinics, and professional schools. However, virtually no existing data were collected with notice to respondents that their information might be shared among Federal agencies.

- At best, this presents a fairness issue to be resolved by data sharing participants. If historical data files are to be shared, how can this be done without unacceptably violated past assurances of limited use?
- At worst, this presents a potential legal barrier to sharing historical data. If data were collected for use by one agency, how can it be provided for the use of another agency?

The 1996 data sharing bill proposed a public notice-and-comment procedure to identify and resolve concerns regarding the use of historical data files. I think this offers a useful way to begin to address specific concerns and hopefully to avoid major objections. However:

- Not everyone reads the *Federal Register* from cover to cover each day, and a public notice that yields no objections should not be taken as sufficient.
- As data sharing arrangements become more widely used and known, newly interested parties can be expected to raise new policy and legal concerns.

In addition to a public notice process, I would urge that each data sharing arrangement include explicit (a) assessments of anticipated sharing benefits and fairness risks, and (b) determinations that the benefits clearly outweigh the risks.

What notices should future respondents receive?

Virtually all existing notices to census and survey respondents provide assurance that individual responses will be used only for statistical purposes; most provide this assurance only on behalf of the collecting agency; and none that I know of suggest that responses may be shared with another Federal agency. Future data collections that may be subject to data sharing will likely require changes in these existing respondent notices.

The issues we face in this area involve to what extent, and how notices should be changed. Here, I urge that we be inclusive, appropriate, and reasonably consistent.

- Inclusive in that a "sharing notice" should be included wherever future sharing might result (i.e., in virtually all substantive collections).
- Appropriate in that the notices should briefly outline the prospect of statistical sharing, without detailing expected recipients or arrangements.
- Consistent in that notices used for separate data collections should be identical if possible, or quite similar if variations are needed.

The reasons for consistency are that (a) the intended substantive message should be the same for all collections, (b) individual business and household respondents may receive notices from different programs and agencies, and (c) we should avoid creating real or suggesting false differences in sharing prospects.

III. ENHANCEMENT OPPORTUNITY -- NAICS

Let me conclude by commenting on data sharing and the implementation of NAICS -- the North American Industrial Classification System. Convergence of these two initiatives offers an important and immediate opportunity to enhance data sharing benefits.

- Data sharing will encourage more open and complete statistical use of information obtained and used by Federal agencies.
- NAICS will establish a greatly updated and more consistent system of industry classification for first-ever use throughout Canada, Mexico, and the U.S.

NAICS is scheduled for implementation beginning in 1998. NAICS will fundamentally redesign industry classifications for business establishments, and implementing NAICS will require new information to assure accurate reclassifications. For this purpose:

- In December 1997, the Census Bureau will request reclassification information from 5 million business locations as part of the five-year economic census.
- In fiscal years 1997 and 98, the Bureau of Labor Statistics will request similar information from the 3.5 million business locations most affected by NAICS.

If revised data sharing authority were available by mid-1997, (a) both of these overlapping data collections would not be necessary, and (b) a reduced collection effort could obtain better and more up-to-date business information. That is:

- The Census Bureau's collection of reclassification information could proceed for larger businesses operating in the census year.
- BLS' subsequent collections could be reduced in scope and focused on smaller businesses not covered in the census, and businesses formed after the census.
- Both agencies could receive and use the considerably expanded reclassifications resulting from the redesigned collection effort.

I noted that "revised" authority would be needed because the provisions of H.R. 3924 would need to be enhanced to permit this change. This is because under the 1996 proposal, data could be shared only for "exclusively statistical" uses, and BLS' information program uses are not exclusively statistical.

- BLS' industry reclassifications are intended for use by cooperating agencies in each state, and their uses are inherently regulatory as well as statistical.
- This limitation would prevent Census Bureau reclassifications obtained during the 1997 economic census from being made available for full use by the BLS.

Statistics Canada has followed a practice for 20 years that could resolve this problem. It is to legislate that industry classifications assigned to individual establishments by the statistical agency are not confidential, they are available as public information.

- This policy has two primary rationales -- classifications are (a) derived by statistical agencies, and (b) widely available in commercial business directories.
- Statistics Canada's practice is well-accepted, it has caused no public controversy, and it has facilitated both public statistics and private economic opportunities.

A provision that declares statistically derived information, such as industry and geographic classifications, as not confidential could be a very useful enhancement to 1997-proposed data sharing legislation.

In conclusion, let me confirm that I believe data sharing promises clear benefits for information providers, data users, and participating agencies. However, this promise will not be realized by relying on abstract policies, outside events, or the efforts of others. Realizing the promise of data sharing will require the "real time" attention, action, and cooperation of most of us in this room. And it will require an agenda that includes adequate legal authority, useful sharing projects, and new program relationships. Let us be aggressive in insuring that data sharing becomes a reality in 1998.

Data Sharing- A USDA/NASS Perspective

Douglas G. Kleweno
National Agricultural Statistics Service (NASS)

Council of Professional Associations on Federal Statistics
Bethesda, Maryland
November 13, 1996

Data Sharing- A USDA/NASS Perspective

I. Introduction

The principle mission of the National Agricultural Statistics Service (NASS) is "to serve the United States, its agriculture, and its rural communities by providing meaningful, accurate, and objective statistical information and services." This charge was established with the founding of the agency in 1863 within the Department of Agriculture (USDA) and continues to be the focus as described in NASS'S Strategic Plan for 1996 [1]. The task includes official USDA forecasts and estimates prepared by the NASS Agricultural Statistics Board (ASB). NASS conducts reimbursable survey work for other Federal, State, and producer organizations and provides technical assistance for other countries developing agricultural statistics programs. Most recently, the 1996 Congress appropriated funds for NASS to conduct the 1997 Census of Agriculture.

The statistical data series on U.S. agriculture are the result of data collection survey activities requiring cooperation from thousands of respondents each year surveyed on production and marketing decisions. Field observations and administrative data compliment this effort in the annual publication of approximately 350 reports prepared by the ASB. Additional survey data are compiled on agricultural resources and their management as related to economics, the environment, and farming practices.

NASS State Statistical Offices (SSO's) in 45 locations provide the infrastructure to efficiently service all 50 States with statistical activities from data gathering to data sharing. Each office operates under a cooperative funding arrangement with State Departments of Agriculture and/or land-grant universities. These arrangements facilitate collaboration of State and Federal agricultural survey data management resulting in efficiencies which reduce duplication of functions and increase services to data users. The challenge to continue improving the cooperative effort among State and Federal Agencies as it relates to this data process is vital to the role of a statistical agency.

As a federal statistical organization, NASS is charged with the dissemination of reliable, timely, and relevant information to data users. The agricultural data series are used for production and marketing decisions by farmers, ranchers, and other agribusinesses. These statistics also are critical to the agricultural industry and policy makers in responding to the dynamics of the sector and the Nations' economy, environment, and natural resources. The importance of data sharing must remain a priority; however, the preserving of data confidentiality cannot be compromised.

The following sections describe NASS's data sharing policy and

management, define current data sharing between NASS and the Census, illustrate other NASS external data sharing arrangements, and discuss future possibilities and challenges of data sharing by NASS.

II. NASS Data Sharing Policy and Its Management

Authority

As a Federal statistical organization, NASS has the authority and public financial support to collect and publish official U.S. agricultural statistics. This authority carries a major responsibility of maintaining data confidentiality. The authority is established by the United States Code, Title 18, Section 1905 and Title 7, Section 2276. Key provisions of this law that protect the confidentiality of reported data follow:

- Neither the Secretary of Agriculture, or any employee of USDA, or any other person may disclose such information to the public, unless such information has been transformed into a statistical or aggregate form that does not allow the identification of the person who supplied particular information.
- Such information shall be immune from mandatory disclosure of any type, including legal process without consent of such person.
- Any person who shall publish or publicly release such information collected shall be fined up to \$10,000 or imprisoned for not more than one year, or both.

A confidentiality pledge is required on all NASS questionnaires requiring data from a respondent and is a part of the introductory statement read to telephone respondents. Any request for an exception must be submitted in writing to the NASS Associate Administrator who is also referred to as the ASB Chairperson.

Other legislation also defines data sharing procedures as established for Federal statistical organizations including NASS. The Privacy Act of 1974 was enacted by Congress to protect individuals' interest in controlling identifiable information maintained by federal agencies. The act covers record systems which contain personal information and from which agencies retrieve information by personal identifiers such as name or social security number. Only a limited number of NASS systems of records qualify and then just a few items on those survey instruments are applicable to these provisions. Most NASS surveys request items of information on the farm business organization which are exempt from this act. Agency guidelines on the Privacy Act are documented in a Policy and Standards Memorandum Number 5-95 [2]. The Privacy Act does not halt research that can be accomplished with nonidentifiable data or dissemination of anonymous information.

The 1966 Freedom of Information Act regulates the disclosure for research and statistical records and applies to data for individuals and organizations. Public access is permitted to records from federal agencies except for specific exemptions. The exemption most frequently employed by NASS is disallowing access to identifiable records which would disclose sensitive information about individual units.

Management Plan

Data sharing is a significant and integral part of NASS's mission as a Federal statistical agency. The importance of this function is clearly stated in the 1996 NASS Strategic Plan which outlines specific goals relevant to the data sharing issue. Goal 1 states that NASS will be recognized as the Federal agency prepared to provide all essential statistics related to U.S. agriculture and the mission of the USDA. Goal 2 states NASS will be the first choice for customers seeking value and excellence in new agricultural and rural statistical services. Goal 3 states NASS will preserve its integrity as an impartial and highly credible source of statistics and as a protector of confidentiality for individual data sources. Associated with each goal are performance targets, strategies, and measures used to evaluate progress as required by the 1993 Government Performance and Results Act.

All NASS data sharing requests are reviewed and either approved or denied by the Associate Administrator. NASS organizational structure also includes a Standards Officer staff position whose responsibilities include coordination of Agency data sharing policy and other standards documented in NASS Policy and Standard Memoranda (PSM). These PSM's establish Agency program policies, standards, and guidelines relating to agricultural statistics and address in detail the topics of confidentiality of information, Privacy Act of 1974, access to lists and individual reports, release of unpublished summary data and estimates, and standards for suppressing data due to confidentiality.

Requests for and sharing of NASS data occurs in various ways. First, the most direct method of data sharing is official NASS data published by the Agricultural Statistics Board. Annually, estimates for about 120 crops and 45 livestock items in addition to environmental and economic data are published. These data must adhere to the NASS standard for suppressing data due to confidentiality [3]. This policy applies to all estimates and summary data published in either releases or research reports. To avoid disclosure of individual operations for a given item of interest, summary data and estimates must NOT be published or released if either: (1) the nonzero value for the item of interest is based on information from fewer than three respondents; or (2) the data for one respondent represents more than 60 percent of the value to be published. Any exception to this rule is granted only with written and signed permission of the respondent concerned.

Suppressed data may be aggregated to a higher level for publication. Care is required, however, to ensure that the suppressed data cannot be reconstructed within the published document or when the same data are republished at various time intervals.

USDA Regulation 1042-42, issued January 16, 1996 defines the functions and authority of the NASS Agricultural Statistics Board. Authority given the ASB Chairperson stipulates that each employee, before being given access to confidential reports and unpublished records, is required to read Departmental Regulation 1042-42; U.S. Code Title 18, Section 1905; and U.S. Code Title 7, Section 2276. A recertification of all employees occurs every 3 years. Individuals other than NASS employees must request access to unpublished data through the ASB Chairperson.

Second, unpublished NASS data sharing follows rigid guidelines outlined in an agency PSM titled, Release of Unpublished Summary Data and Estimates [4]. Unpublished summary data and estimates are made available only for appropriate research and statistical purposes which are considered beneficial to USDA, NASS, and public interest and not detrimental to the survey respondents. These statistics cannot be released while the estimates to which they apply are still current. Requests for release must be submitted on a specific form. Release of survey indications to external data users is permitted only when approved by the Associate Administrator. NASS research or staff reports meeting the above criteria may publish survey indications to support research conclusions and recommendations with the approval of the Associate Administrator.

When there is a perceived general demand and recurring requests are likely for certain unpublished information, the Associate Administrator will likely recommend publishing the summary statistics, given they are statistically reliable and confidentiality rules are maintained. This action ensures all interested data users are equitably served.

Third, the NASS policy in response to data requests for access to NASS lists and individual reports by external data users is described in PSM 6-90 [5] as follows. "List sampling frame names, area sampling frame names, and individual reports will not be made available outside of NASS if it is determined to be detrimental to NASS; the Department of Agriculture; or the farmers, ranchers, and agribusinesses included in the sampling frames. Such information can only be made available for approved statistical activities. In all activities, the identity of respondents or data included in individual responses must not be made available to unauthorized persons. All requests based on Freedom of Information Act or other legal aspects must be referred to the Associate Administrator."

Data Sharing Rules for NASS List Sampling Frames:

NASS list sampling frame rules ensure that respondents' status as

an agricultural enterprise and their reported survey data are kept confidential. The list sampling frames can only be used by external data users outside the Agency for limited uses following specific procedures. Notable examples follow:

1. The Bureau of the Census has historically had access to lists of farm operators on NASS list and area sampling frames for list building and coverage evaluation activities associated with the Census of Agriculture. The U.S. Code, Title 13, provided confidentiality protection.
2. General Accounting Office and Office of Inspector General auditors in program evaluation efforts have had access to reports from samples of NASS list frame names under tightly controlled arrangements. The requesting organization must provide sufficient information to allow NASS to evaluate how the sampled data will be used, and be assured that the review is of a statistical nature with no punitive or enforcement purposes directed at individual respondents. Reported data are released only with the written approval of selected respondents. Auditors must be certified with NASS confidentiality forms and any release of sampled data must be approved by the Associate Administrator.
3. NASS will conduct surveys for local cooperators using the NASS list sampling frame if a formal cooperative agreement or memoranda of understanding exists. Respondent burden and benefits to the total Federal/State statistical program are evaluated.
4. Other governmental officials are asked to review list frame names and addresses to assist NASS list building and maintenance efforts. Individual identifiers such as social security numbers and employer identification numbers are removed from the document before the review begins.
5. NASS will make mailings for State and federal government agencies of unbiased informational items or referendum ballots to portions of the list sampling frame if the purpose contributes significantly to the agricultural sector.
6. Other data user requests for NASS lists not established by the above procedures are addressed with a justification statement to the Associate Administrator.

Data Sharing Rules for NASS Individual Reports:

Information collected from individual or organizations is protected from disclosure by U.S. Code Title 18 and Title 7. However,

restricted on-site analytical use of microdata is permitted when certain procedures are followed. The site location is restricted to NASS offices or Economic Research Service offices with specific arrangements to ensure data security. Non-NASS users are limited to academia, private nonprofit organizations, and governmental analysts conducting research that serves the general public and contributes to the understanding of agriculture or statistical procedures used by NASS. These analysts must complete a data request form, subject to the approval of the Associate Administrator, which documents the data requirements, project schedule, statistical methodology, summary output, intended use and distribution of the report, and nondisclosure assurances.

Analysts must become familiar with concepts underlying the NASS data to ensure appropriate statistical interpretation and use of the data. Knowledge and understanding of the survey design, sampling procedures, response rates, data editing and imputation, outlier adjustments, and other factors impacting summarization and data usage are important. Most Agency data are obtained using complex sample survey designs and the Agricultural Statistics Board ensures that defensible statistical standards are met for all published data. Because sample surveys are designed to produce estimates for the NASS statistics program, it is important that the Board evaluate and provide guidance to ensure non-NASS analysts also adhere to appropriate statistical methods.

All direct identifiers such as name, address, telephone number, and social security numbers are deleted from the individual identifier data files before access is granted.

Another data sharing alternative used by NASS is special tabulations and data aggregations. The non-NASS analysts specifications are used to create the desired summary instead of direct access being granted to individual identifier NASS data files. The agency is reimbursed for expenses incurred.

NASS also conducts surveys and shares data with other federal, State, and producer organizations using cooperative agreements, memoranda of understanding, and contracts. NASS has cooperative arrangements with State departments of agriculture and/or land grant universities in each field office which service all 50 States. These reimbursable surveys offer the same protection to respondents as other NASS surveys with the same U.S. Code Title 18 and 7 rules for data access. The national level rules apply to the cooperative statistical programs in each State Statistical Office. All NASS employees and sworn agents are fully responsible for the various statutory confidentiality provisions and subject to the penalties for disclosure. Another benefit of this cooperative agreement to the public is the amortization of the cost of frame development and maintenance across more sample surveys.

Two methods of data sharing are acceptable to NASS for statistical research projects. First, special research can be conducted through the Statistical Survey Institute Program. Fellowships in

this program are solicited and supported by the American Statistical Association (ASA) and NASS. Fellows become agents of NASS and usually take a one year sabbatical. Research is conducted on-site at NASS with access to appropriate microdata files. For example, Bargmann [6] conducted research to develop a technique to generate artificial data sets with similar statistical characteristics for annually collected economic data.

Second, research agreements can be made for individual data analysis. A research proposal must be submitted for approval by the NASS Associate Administrator documenting individual data access requirements, research to be conducted, uses and benefit of the research, and data security procedures to be employed. The researcher(s) must read and sign the statement which documents the rules of confidentiality and disclosure compliance. Once the document is completed and approved, the cooperating institution researchers can have access to individual data at their site. To protect respondent identity, it may be necessary for NASS to use data masking techniques or remove unusual observations from the file. Microdata sharing is permitted outside NASS office facilities under these conditions but only if other alternatives such as the use of summary data or access within NASS State offices are not feasible for the study being conducted.

III. Current Data Sharing Between NASS and Census

The Agriculture Division in the Bureau of the Census and the National Agricultural Statistics Service in USDA have a long history of cooperation. NASS agricultural data collection began in 1842 as part of the U.S. Patent Office and the Census of Agriculture began in 1840 as part of the decennial Census. NASS collects and reports the current happenings in agriculture, i.e., acres planted, yields, effect of weather on crop production, livestock production, and prices, while the Census provides very detailed demographic and agricultural information by county every 5 years. Up through the 1920 decennial census, the agricultural information was collected with the population census. For the first quinquennial Census of Agriculture taken in 1925, NASS handled the field enumeration through its State offices. It is obvious that the Agencies managed to share data at that time.

However, later confidentiality legislation prohibited the sharing of data and a serious problem arose in 1948 between the two agencies due to the inability to share data collected from cotton gins. The Census was required by law to report the number of bales ginned by certain dates during the harvest season, and NASS was responsible for forecasting total cotton production each month and relied on cotton ginners as an excellent source of information to help forecast production. The Cotton Ginners Association successfully lobbied Congress to eliminate the duplicate monthly surveys and the Appropriations Committee threatened to reduce each agency's budget unless some means could be worked out to share data. Subsequently, an agreement was reached in which the Census agreed to obtain a "waiver" from each gin that provided them

authority to share the data with NASS.

The Reports Act of 1942 prohibits duplication of data collection by the Federal Government. However, over 50 years after its enactment, considerable duplication continues because NASS and the Census are unable to share data. Title 13 of the U.S. Code allows only Census Bureau employees (including special sworn employees) access to individually identifiable information. Also, the purpose of use must be within the scope of the Bureau's statutorily defined mission.

The National Agricultural Statistics Service has shared its individual farm list with the Census Bureau in preparation for each census of agriculture. NASS SSO staff, as sworn Census agents, also have participated in data edit processes at the Ag Census offices. Title 13 prohibits any reciprocal flow of information to NASS for use in its development and maintenance of a list of farm operations. The provisions of the Tax Reform Act of 1976 (P.L. 94-455) do not permit the Internal Revenue Service and Census, who uses farm tax return lists as a primary name source, to share the same lists with NASS. The result has been increased costs to NASS to construct and maintain its list sampling frame.

Funding for the 1997 Census of Agriculture was shifted by Congress from the Bureau of the Census to NASS starting with fiscal year 1997. However, the proposed legislation to transfer full authority of the Ag Census to NASS has not been approved yet. Because of that outcome, the Census Bureau has not permitted NASS offices to have access to previously reported data and names and addresses from the 1992 Ag Census. If NASS SSO's could access this data, it might be possible for those offices to identify further duplication in the mailing list and reduce expenses for the 1997 Census.

NASS also has not been permitted any access to the Internal Revenue Service (IRS) farm tax returns which will be used as input to the 1997 Census mailing list development. NASS hopes to work with the IRS to obtain access based on the shift of the Ag Census responsibility to NASS and the NASS confidentiality procedures. Legislation will be resubmitted to obtain the full Ag Census authority.

Proposed action for efficiencies and for improving the quality of agricultural and economic statistics is not new. For example, former Chairman Michael Boskin of the Council of Economic Advisers [7] recommended legislation to allow limited sharing of confidential statistical information for statistical purposes only among statistical agencies with rigorous safeguards. The Data Sharing Task Force, convened by the Office of Management and Budget (OMB) to assist in the formulation of legislation by OMB, stated in their December 1994 draft report [8], "...we believe that data sharing presents an opportunity for tangible benefits in our decentralized statistical system, and 1995 is the time to act on this opportunity." In the past year, however, the effort to pass interagency data sharing legislation was unsuccessful. Even with

projected data and cost efficiencies, less respondent burden, and data quality improvements, congressional approval will be difficult. The concern about loss of privacy with increased data sharing, regardless of any legislated protection, is a growing public concern.

IV. Other NASS External Data Sharing Arrangements

The NASS data sharing policies and their management, discussed earlier, establish a process to ensure respondent confidentiality while providing data access for legitimate statistical purposes. Described below are cases of data sharing which illustrate these agency procedures in use. The examples are not intended to be an exhaustive list but instead represent recent successful NASS data sharing programs.

1. Farm Injury Survey - The survey data were collected for all 50 States over a three year period under a cooperative agreement. The sponsor for the three year project was the National Institute of Occupational Safety & Health (NIOSH) of the Center for Disease Control (CDC). The client requested that NASS be as open with sharing of the data as possible within NASS policy. When requested, the agency has provided outside data users with not only the summary data, but also masked individual record data. Several steps were taken to avoid disclosure of the identity of individual records.
2. ARMS - From the initial development and design of the Agricultural Resource Management Study (ARMS), a cooperative data collection and sharing agreement between NASS, the Economic Research Service (ERS), and the Animal Plant Health Inspection Service was a critical decision. This new survey program integrates multiple survey activities and survey data requirements for an annual farm finance survey, cost of production surveys, cropping practice surveys, chemical use survey, and a national animal health management survey. Data sharing details are still being finalized. However, the basic premise is that by sharing in the survey design, data collection, and summary process, respondent burden will decrease, survey costs will decline, and each agency will be able to make better use of a richer data set. Each of the federal agencies will abide by the NASS confidential and data disclosure policies. Individual identifiers will not be included in any sharable files.
3. NAHMS/ERS - As part of the continuing effort to reduce burden and lower costs, the National Animal Health Monitoring System (NAHMS) and ERS are making more use of data already being collected as part of the regular NASS survey program. Several data sharing examples include the use of the Agricultural Survey death loss data for cattle, hogs, and sheep for NAHMS and the land use data from the June Area Survey for ERS. These data are provided to NAHMS and ERS at virtually no cost other than extra summary processing. In most cases, NASS is

providing these organizations with data aggregations. NAHMS and ERS then perform additional data analyses and set their needed estimates.

4. Sheep Herd Management Survey - During 1996, NAHMS and the American Sheep Industries (ASI) requested NASS to provide a list of producers to contact for a mail survey on sheep herd management practices. NASS could not give a list of names to NAHMS or ASI because of list confidentiality but was able to provide the sample. A sample was drawn from the NASS list sample frame and sent to the Colorado State Statistical Office. The questionnaires were provided to the State office by NAHMS in pre-stuffed and sealed envelopes. The envelopes were then labeled by the office and mailed out to the respondents. By not labeling the questionnaire, when the respondent returned the questionnaire to NAHMS, the individual or organization identity was not available unless the respondent provided it. Only the State and strata size were known by NAHMS for each response.
5. Farm Labor Survey - The California Employment Development Division (EDD) and NASS are finalizing a cooperative agreement to consolidate each organization's survey program into one joint activity to begin January 1997. NASS will maintain the sampling frames and select a California sample to meet both agencies data requirements. EDD will collect and edit the data and the NASS California State Statistical Office will process the survey data. NASS and EDD will coordinate data analysis, data summary review and estimation, and data publication. EDD plans to implement an estimation procedure similar to NASS. A joint review of survey data is intended to eliminate program differences. EDD will adhere to NASS data confidentiality and disclosure policies.
6. Cooperative Research - The ERS and NASS, through a cooperative agreement [9], share microdata from the jointly funded annual farm finance survey. NASS collects data with the survey instrument identifying the collaborative effort of the two federal agencies. NASS and ERS share in data analysis and summary review. National production expense estimates are set in an Agricultural Statistics Board session where both organizations are represented. ERS abides by NASS Title 18 and 7 policies for data publication and research activities. An extension of this agreement is the joint approval by both agencies to research requests, typically from land grant universities. A memorandum of understanding is required between the researcher(s) and ERS which follow NASS established data confidentiality and unpublished data access policies. The research is conducted at ERS or NASS State Statistical Offices where microdata are used with individual identifiers suppressed. The number of these arrangements continues to increase as the federal agencies attempt to be responsive to a 1992 General Accounting Office audit report [10] recommending an expansion of this data sharing program.

7. EMAP - The Environmental Monitoring Assessment Program (EMAP) was a joint survey project with a memorandum of understanding (MOU) between NASS, North Carolina State University, the Agricultural Research Service-USDA, and the Environmental Protection Agency (EPA). On-site access to microdata was granted to this agriculture group of researchers and analysts after signing appropriate security and disclosure statements. Also, NASS removed individual identifiers, such as the latitude/longitude coordinates, from the first microdata file and created a second public database for other EPA research groups associated with forestry, arid lands, and surface water areas. These later EPA groups were not part of the original MOU and were not bound by NASS confidentiality and disclosure rules. NASS verified the public data base met disclosure requirements and approved all data requests. This second data base had a hydrological unit as a first level record identifier to ensure confidentiality. The unit is a large area of land usually greater than a county in size. This complex data sharing arrangement with multiple data bases showed creative efforts were used to follow disclosure rules while providing data user access to microdata. A loss of EMAP funding by EPA, however, ended this project.
8. Nematode Study - A data sharing agreement between NASS and the Smithsonian for a Nematode Study, conducted as an extension of the EMAP project, is still pending. When funding of EMAP was lost, there was interest in preserving the research and analysis of nematodes removed from the soil samples. The Smithsonian initially asked for the latitude/longitude identifier. To ensure confidentiality of reported data, NASS agreed to accept this request with the condition the coordinates of the identifier on individual data be aggregated to a county level.

Numerous other requests for unpublished NASS data are recorded by the agency. For example, in 1995 the Associate Administrator approved 31 data requests. So far in 1996, 24 new requests for unpublished data have been filed. This does not include NASS responses to data sharing needs that extends over several years. For example, unpublished State level farm wage rate data and production input price data are provided annually to ERS for internal analytical uses, again subject to disclosure and publication constraints.

V. Future Data Sharing Possibilities and Challenges

For over 125 years, NASS data sharing procedures have been rigorously enforced by adhering to legislation and Agency policy that ensures individual respondent protection from disclosure, facilitates proper interpretation of data and sound statistical procedures, and responds to customer requests for access to survey data for statistical purposes beneficial to the public interest. The Agency's reputation and credibility have and will continue to

depend on keeping this commitment to excellence in service to both the survey respondent and data user.

Significant data sharing challenges face the future of NASS and other federal statistical organizations. Some broad issues include the changing U.S. public sentiment towards an individual's privacy, the downsizing and reinventing of the federal government for more efficient use of public resources, the rapid advancement of technology in the areas of data storage, data transmission, and data processing, and the increased domestic and international demand for information. Also, new legislation to permit restricted access to data by federal statistical agencies for statistical purposes must be a priority.

More specially, there are several issues which impact current and future NASS data sharing policy. First, the new role of conducting and being the caretaker of the Agricultural Census will impact how NASS collects survey data as well as disseminates that data. Current statutes require that data sharing will continue with safeguards to ensure data confidentially and individual disclosure protection. Still unknown is how legislative policy, new statistical procedures, and data processing and systems development will change how business is conducted.

Second, NASS has expanded the scope of survey activities to a more diverse subject matter and thus the type of data and customers requesting data are more diverse. For example, the NASS Associate Administrator in a 1995 memorandum to Agency management [11] discussed recommended guidelines for public use geographic data. Two of the principles addressed were:

NASS will work with others to create data files with maximum allowable geographic identifiers, and

NASS will consider computer classified remote sensing map products as publicly releasable.

Third, NASS must continue to explore innovative methods to increase data access within policy guidelines utilizing advancements in technology and statistical procedures for data capture, storage, masking, dissemination, and security. Off-site processing and research activities are becoming a greater likelihood with improved data systems security and high speed telecommunications capability. Data warehousing from centralized storage of microdata files will increase the efficiency and accessibility of longitudinal and time series data relationships but will require suppressing individual identifiers for any data sharing to occur within confidentiality rules.

The NASS research agenda, as presented by Bosecker [12] in the 1996 NASS Research Division Missions and Projects Executive Summary, describes projects related to data sharing efforts. For example, a cooperative agreement is being arranged with George Mason University to develop and evaluate data smoothing and display

techniques that have meaningful information content to data users but do not reveal any individual data. Also, a project is underway to evaluate the quality of California Environmental Protection Agency data used by NASS in lieu of collecting similar data from sampled growers during the Fruit and Vegetable Chemical Use Surveys.

To conclude, NASS considers data sharing an integral part of its mission as a federal statistical agency. Each data request is responded to in a manner that ensures data confidentiality is enforced; data integrity is achieved; and data accessibility as a service to the public interest is accomplished. As stated by Allen [13], "It will continue to be NASS policy that data sharing will occur on a case by case basis as needed to address an approved, specified USDA or public need." It is imperative that data users understand, respect, and protect the confidential use of data with the same zeal and commitment as NASS to maintain and potentially improve future data sharing policy.

References

- Allen, Rich, (1995), "Memorandum on Guidelines and Principles for Public Use Geographic Data." [11]
- _____, (1993), "Memorandum Summarizing the Office of General Counsel's Interpretation of Statutes Governing NASS Data Sharing." [13]
- Bosecker, Ron, and Research Division Staff (1996), "Research Division Missions and Projects Executive Summary." [12]
- Caudill, Charles E. (1990), "A Federal Agency Looks for Answers to Data Sharing/Confidentiality Issues." [6]
- Kibler, W. E., and John E. Lee (1986), "Memorandum Stating Conditions on Use of Farm Costs and Returns Data." [9]
- OMB Task Force (1995 draft), "Data Sharing Task Force Report." presented to Office of Management & Budget, December 1994. [8]
- Peach, J. Dexter (1992), "Data Collection, Opportunities to Improve USDA's Farm Costs and Returns Survey." A report by the U.S. General Accounting Office to the Chairman, Committee on Agriculture, Nutrition, and Forestry, U.S. Senate, July 1992. [10]
- Panel on Confidentiality and Data Access (1993), "Private Lives and Public Policies." National Academy Press, Washington, D.C. [7]
- USDA, National Agricultural Statistics Service (1996), "NASS Service to Agriculture Through Statistics, Strategic Plan 1996." [1]
- _____, (1995), "PSM No. 5-95, Privacy Act of 1974." [2]
- _____, (1990), "PSM No. 6-90, Access to Lists and Individual Reports." [5]
- _____, (1990), "PSM No. 7-90, Release of Unpublished Summary Data and Estimates." [4]
- _____, (1996), "PSM No. 12-96, Standard for Suppressing Data Due to Confidentiality." [3]

Comments by James R. Nunns¹
on Papers Presented at the session on
“Sharing Data for Statistical Purposes”
Seminar on Statistical Methodology in the Public Sector
Council of Professional Associations on Federal Statistics
November 13, 1996

Introduction

It may be useful to begin with some background information on the Office of Tax Analysis (OTA) and its role in the Federal statistical community before commenting on the papers. OTA is not a designated statistical agency, but relies heavily on many statistical sources in preparing the official Administration forecasts of budget receipts, revenue estimates for proposed legislation, distributional and other economic analyses of tax provisions, and related responsibilities. Our main source of statistical data is tax return information prepared by the Statistics of Income (SOI) Division of the Internal Revenue Service (IRS), but we also make extensive use of the Current Population Survey and Survey of Income and Program Participation from the Department of Commerce, the Consumer Expenditure Survey from the Department of Labor, the National Medical Expenditure Survey from the Department of Health and Human Services, the Survey of Consumer Finances from the Federal Reserve Board, and many others.

OTA's heavy reliance on non-Treasury statistical data gives us a stake in improving these data. One potential source of improvement could be through expanded sharing of IRS data. However, along with IRS we have very real concerns about the impact of sharing IRS data on the level of tax compliance, on the public's perception of IRS, on the IRS's ability to monitor the use of shared data, and on the right of taxpayers to privacy with respect to the sensitive personal information contained in tax returns. For these reasons, OTA (and our legal counterparts in the Office of Tax Legislative Counsel) worked closely with IRS and Katherine Wallman and her staff in OMB (and through them, the statistical agencies) on drafting the amendments to Section 6103(j) of the Internal Revenue Code that are the companion to the proposed “Statistical Confidentiality Act.” Section 6103(j) governs the disclosure of tax return information for statistical use. We also prepared a “Commentary” on the proposed 6103(j) amendments, which is meant to spell out more fully the intent of the legislation.

The proposed amendments to Section 6103(j) would allow access to certain tax return information to the six “Statistical Data Centers” designated in the proposed “Statistical Confidentiality Act” that currently have no access (the only Centers that currently have access are the Bureau of the Census and the Bureau of Economic Analysis in the Department of Commerce). The Division of Research and Statistics of the Federal Reserve Board would also be given access.

¹ Director for Individual Taxation, Office of Tax Analysis, Department of the Treasury. The views expressed are those of the author, and do not necessarily reflect Treasury policy.

However, the current access to individual income tax return information by the Bureau of the Census would be conformed to that of other agencies', and thereby limited relative to current law.

From my perspective, the guiding principle behind the proposed amendments to Section 6103(j) was to balance the potential benefits of lower agency costs, improved quality and timeliness of data, and reduced respondent burden against the potential costs of lower tax compliance, adverse public perception of IRS, IRS's monitoring of shared data use, and loss of taxpayers' right to privacy. This principle lead to a "rifle shot" rather than a broad approach -- to give limited and closely controlled access to a limited number of agencies and only for constructing sample frames for censuses and surveys and related statistical purposes authorized by law. Only the minimum amount of information needed will be releasable -- minimum in items included, years covered, and number of taxpayers covered, and the data will be in categorical form with the largest possible categories. Further, sharing beyond the receiving agency can only be among the other eight designated agencies, only for purposes authorized by law, and only with the approval of the IRS.

Comments on Papers

Let me turn now to the two papers presented at this Session. I found both papers informative and useful. Thomas Mesenbourg's paper, "Realizing the Promise of Data Sharing," makes the very important point that the proposed legislation, both the Statistical Confidentiality Act and the amendments to Section 6103(j), won't be self executing. The Section 6103(j) amendments will require IRS to develop regulations (as under current law) that describe precisely which tax return information items, for which returns, and on what schedule, each agency will have access to. This will require complete documentation from agencies covering the legal basis for their requests, alternatives to tax return information, security arrangements for the data, and related issues. In addition, data sharing agreements will have to be approved by the IRS. To complete all of this regulatory apparatus will take time and effort on the part of agencies as well as the IRS.

Douglas Kleweno's paper, "Data Sharing -- A USDA/NASS Perspective," notes the transfer of responsibility for the agricultural census from the Bureau of the Census to NASS, but without implementing legislation (including the proposed amendments to Section 6103(j)) which would allow NASS access to tax return information. The IRS, Bureau of the Census, NASS, and OMB are working together to see whether we can solve the problem for now administratively. But it is worth noting that the proposed amendments to Section 6103(j) were developed to handle such problems, and this situation gives us a concrete example of why the legislation is necessary.

I also found Doug's paper very useful in explaining in some detail the mission of NASS, and in particular the extensive data sharing arrangements that NASS has entered into. As a cautionary note, NASS and other agencies may have to rethink some of these arrangements under the proposed Section 6103(j) amendments if they begin to receive tax return information and integrate it with data from their own surveys and other sources. However, there clearly are ways to work out many arrangements for data sharing, and I was impressed with the effort NASS has put into finding ways of sharing data while maintaining confidentiality.

The related point I would like to make is that the "Commentary" on the proposed Section 6103(j) amendments suggests that agencies and the IRS should search carefully for ways to limit, or even eliminate, the need for disclosure of confidential tax return information, while still achieving the agencies' statistical purposes. One approach, which I think might give us the best balance, would be to work out arrangements whereby agency personnel would be detailed to IRS and be given access to certain tax return information while under the direct control of the IRS. This would allow the agency personnel to do preliminary statistical analyses on the data, and to determine the minimum amount of data that would be necessary to leave IRS in confidential form under 6103(j). In some cases, perhaps no confidential tax return information would need to be disclosed to the agency. In this way, the agency would effectively have greater access to tax return information than they would otherwise have, while IRS would maintain maximum physical control of all confidential tax return information.

Concluding Remarks

I would like to conclude by urging all statistical agencies to look very carefully once more at the proposed amendments to Section 6103(j), with an eye to making sure the legislation is workable for them. While no one involved (I think I can speak for all) wants to reopen this proposal, we would much rather do that than find in a year or two (assuming the legislation is introduced and passed) that it doesn't work.