Is Data Redundancy the Price Archivists Will Pay for Adequate Documentation

by Margaret Hedstrom[1]
New York State Archives & Records Administration

Federal, state, and local government agencies in the United States are major repositories of important soical, scientific, and economic data. By automating many of their basic record keeping functions, agencies at all levels of government have become stores of vast quantities of data on citizens and on public programs. Unfortunately, data acquisition, preservation, and dissemination, especially by state and local government archives, have not developed apace. Warnings about the loss of important contemporary and historical records at the federal level should be multiplied many-fold when considering state and local government records. Few states have addressed the issue of preserving records in machine-readable form, while not a single local government archives program preserves electronic records.[2]

This article examines some of the implications of increased data sharing among local, state, and federal agencies for the acquisition, preservation and dissemination of data by archives. For social science data archivists and others not familiar with government archives, it is important to point out that traditional archives differ in some respects from social science data archives and libraries. Government archives identify, preserve, and make available public records with enduring value for historical or other research. Traditional archives are concerned with maintaining records of how government agencies performed their mandates as well as records of the individuals, organizations, and other phenomena that were influenced by that mandate.

Government archives acquire and preserve administrative data mostly in the form of traditional paper files. In the process of regulating a myriad of activities and providing a wide rage of direct services, government agencies collect a wealth of data (an increasing portion of it in machine-readable form) on almost every aspect of social activity. The vast majority of the data collected by state and local government agencies is compiled to administer programs while research use remains, at best, a secondary consideration. The federal government is similar to state and local governments in this respect, but state and local

[1] Presented at the International Association for Social Science Information Service and Technology (IASSIST) Conference held in Washington, D.C., U.S.A. on May 26-29, 1988

[2] For a discussion of the problems of preserving electronic records from federal government agencies, see the committee on the Records of Government, Report (Washington, D.C., Council on Library Resources, 1985). In recent years several states have begun to address the problems of machine-readable records, most notably New York, Kentucky, Washington, Ohio, Delaware, and Wisconsin. However, no state has a fully developed program for acquisition, preservation, and dissemination of data. Archival programs for local governemnt records are even less developed.

governments rarely conduct original research and they do not maintain independent statistical agencies.

Since the 1960s, federal, state, and local government agencies have developed new information systems which collect data from private citizens, business, and other government agencies and distribute data to all of these constituencies. No level of government exists as an isolated island today. Increasingly, information flows between levels of government using transmission methods that range from the primitive shipment of paper forms to real–time transfers through complex, interactive computer networks.

Intricate information flows pose new problems for archivists who work in traditional government archives because they challenge a parochial vision that is bounded by the limits of a single government's jurisdiction. Traditionally, the National Archives has preserved federal records, state archives have preserved state records, and local archives (where they exist) have preserved local government records. This structure is inadequate for identifying and preserving valuable information from shared functions, because the systems designed to process and facilitate data exchanges exceed the boundaries of one level of government, while provenance and data ownership become unclear.[3]

Different levels of government exchange data for specific purposes. Data sharing occurs when federal, state, and local agencies jointly regulate activities, administer basic functions, or share in the delivery of services. Social welfare and transportation programs in which states and localities provide services that meet uniform standards in exchange for modest amounts of federal funding are examples of this type of data exchange. Different levels of government also exchange data when there is a willingness to share information that is necessary for administering functions which exceed a single jurisdiction. Data available at the local, state, and federal level to track and apprehend criminals is perhaps that best example of this type of data sharing. Public agencies may also exahnge data when the use of publicly available data is more expedient or less expensive than independent data collection. In the area of educational statistics — a function with a limited federal funding or regulatory role — it is more expedient for the federal government impose reporting requirements on states, and for states to impose reporting requirements on localities, than it is to conduct independent research surveys on educational programs.

The increasing transfer of data between levels of government reflects basic changes in the administration of govoernment programs and the delivery of public services. In the last two decades, state and federal agencies have decentralized responsibility for most direct service delivery. At the same time, government agencies have responded to a real or perceived demand for greater accountability. This dual goal of decentralization and enhanced accountability is handled operationally by passing large volumes of information between regulating or funding agencies and the agencies that directly perform government functions. A more recent, but parallel trend is subcontracting with the private sector for a growing portion of the direct services. The technological capability to transfer data between systems is another factor in the growth of data exchanges, but is not the primary reason for the increasingly complex data transfers in the last two decades.

Two examples illustrate the complexity and the volume of the information flows between levels of government. The Medicaid Management Information System (MMIS) — a state–level system found in most states — is used to determine eligibility, monitor fees, process

[3] Provenance is the principle of grouping public records according to their origins in the administrative structure.

claims, and evaluate the program's costs and effectiveness. In Utah's MMIS, the claims processing portion has more than 100 machine-readable master files and it produces 316 different output reports. The system produces six truckloads of paper and nearly 20,000 sheets of computer output microfiche each month. Many of the output reports are transferred to the federal government because they are mandated by reporting requirements. But information is also exchanged among local social service agencies, public hospitals and clinics, insurance companies, and private providers.[4] The Medicaid system may be one of the largest and most complex examples of a jointly administered program, but it exemplifies the complicated information needs of many public health and social welfare programs.

A second example is the national criminal records system, initially designed and funded by the Law Enforcement Assistance Administration (LEAA) in the late 1960s and 1970s. A complex network with data on criminal histories, criminal identities, warrants, and the like allows the transfer of data vertically between local, state, and federal law enforcement officials, and laterally among criminal justice agencies within and between states. In addition to identification, socio-demographic background, and criminal history data on millions of offenders and suspects, the system contains data on significant actions taken by police agencies, district attorneys, courts, probation departments, correctional institutions, and parole boards.[5] A major impetus for the system was the recognition that fragmented information residing with the Federal Bureau of Investigation, fifty state police organizations, and a thousands of local police agencies was ineffective for tracking and apprehending criminals who have little respect for municipal boundaries of state borders.

Data sharing is not limited to vertical transfers between the various levels of government. Responsibility for specific government functions is not always confined to a single agency. Criminals, for example, are handled by local police, district attorneys, courts, correctional institutions, and probation departments. They may also have a history of substance abuse, a family on public assistance, chronic health or mental health problems, and a need for vocational education. Similarly, transportation, environmental conservation, parks, land use planning, and taxation departments all perform functions that affect a single geographic area. Some local governments, that have computerized recently find that the major advantage to automating is the ability to combine data on clients who are served by many different programs or on one specific geographical.[6]

In spite of many automated systems for data exchange, the potential for data sharing far exceeds what is current practice today. Bureaucratic, administrative, and technical

[4] Ken White, "We Have the Program, Now We Need Federal Approval," unpublished paper presented at the annual meeting of the Society of American Archivists, Sept. 5, 1987. For another example of an inter-governmental information system, see Robert H. Crowley and James J. Heaphey, The Welfare Management System in New York State: A Case Study of Management Information Systems in Government, (Albany, NY: Rockefeller Institute of Government and the Governor's Office of Employee Relations, Oct. 1984).

[5] Alan Kowlowitz, "Hands Up, You're Under

[5](cont'd) Arrest: Appraising Criminal history Data in the Age of the Electronic Case File," unpublished paper presented at the annual meeting of the Society of American Archivists, Sept. 6, 1987. (Forthcoming in Archival Informatics Technical Reports, 1989).

[6] Rob Gurwitt, "The Computer Revolution: Microchipping Away at the Limits of government." Governing the State and Localities 1 (May 1988), pp. 34-43. Interest in recombining disparate data sets has spawned the development of geographic information systems at the local and state level. See Government Technology, special issue on computerized mapping, Vol. 1, #5 (Sept./Oct. 1988).

obstacles to data sharing create barriers to the free flow of information. The existence and availability of administrative databases is generally not well known even to those working within one level of government. Some agencies are possessive of their data and prefer not to exchange data for a variety of reasons. Moreover, administrative data systems usually are designed for a very specific purpose with idiosyncratic data structures, unique data definitions, and poor documentation.[7] Although an administrative data set might contain data related to a secondary application, it may not be specifically useful for another purpose. Finally, data exchanges are technically difficult because data interchange standards have not been widely adopted. All of these factors inhibit data sharing and lead to redundant data collection.

The records and data that document functions shared by federal, state, and local governments create several problems for data archives. One problem is data redundancy. When local authorities report to the state authorities, more often than not, they maintain copies of the data (or the hard copy records) that they transmit. When state agencies report to a federal agency, they are likely to also maintain copies of the information they transmit. In some cases, such duplication is actually required by federal and state regulations. Redundancy is even more apparent when information flows in the other direction. Policy directives and statistical data from a federal agency may be duplicated in all fifty states, and duplicated again in thousands of local government agencies. Seen from this perspective, the greatest problem facing archivists appears to be the overabundance of machine-readable data — little of which is unique.

Data redundancy, however, is not the most challenging problem of complex, intergovernmental data flows. It is a tremendous waste of limited resources for archives to preserve identical data sets at the local, state, and federal levels when much unique and valuable data is lost. Yet some duplication of data may be the price that archivists will have to pay if we want to preserve usable data and comprehensive documentation of shared functions. A more challenging problem for archives is the need to develop approaches to the analysis, appraisal and selection of data that transcend the boundaries of a single level of government. Government archivists who analyse large administrative data systems recognize that appraisal must begin with a comprehensive overview of the system, its basic functions, the general types of data it handles, and the types of output it provides — through both hard copy reports and potential on-line queries.[8] Such an overview allows the archivist to recognize the basic logic of the system and to identify key areas for more detailed appraisal. An archivist who approached the Medicaid Management Information System by systematically analyzing each of the 316 output reports, would be hopelessly lost before gaining even a slight semblance of why these reports were created or how they were used.

To develop an overview of a system like MMIS, which is designed in part to transmit information between levels of government, archivists must gain a perspective that accounts for the flow of information. Unfortunately, no mechanisms exist yet for approaching appraisal in this way. Although archivists may consult other levels of government to determine whether the records they are appraising are being preserved elsewhere, this information is

[7] New York State Criminal Justice Information Systems Improvement Program, Measurement Issue in Prison and Jail Overcrowding (Albany, Division of Criminal Justice Services, May 1988).

[8] Thomas Elton Brown, "Appraisal in the Information Age," paper presented at the annual meeting of the Association of Canadian Archivists, June 1987.

not readily available even for small sets of traditional records. More exchange of appraisal information among archives is an essential first step, but appraisal of complex information networks will require far more than simply sharing information about records preserved at various levels of government.[9] Joint analysis and appraisal projects need to be carried out simultaneously by archivists working at the federal, state, and local levels. The objective of such projects is not necessarily to avoid duplication of data between levels of government. Rather, it is to select data that adequately documents how each level of government performed its responsibility for a shared function. If local and state administrators used the same data, but used it in different ways with quite different implications for social service recipients, for example, adequate documentation might necessitate some duplication.

Another concern with shared functions is that no single level of government maintains a comprehensive collection of data on a particular program or its recipients. Some database management systems provide centralized sources of data, but the networks that support large, shared government functions do not centralize all of the information in one place. Rather, selected pieces are passed between different levels of government with no single point of data compilation. For functions that are carried out primarily by local authorities, the richest and most detailed case–level information is likely to remain at the local level. While detailed case information may be of greatest interest to sociologists, economists, social historians and other researchers, local governments have demonstrated little capability to preserve the information in machine–readable form. Moreover, the records maintained by any single locality cannot provide a comprehensive picture of statewide or national programs. Data maintained by federal agencies may provide the necessary breadth, but lack the detail that is essential for social science and policy research.

Responsbility for preserving and disseminating data from large networks that transcend a single governmental jurisdiction is also unclear. Is data, collected by local governments to administer local social assistance programs and reported to a state or federal agency to meet reporting requirements, the responsibility of the local government that collected it originally or of state and federal agencies? This may seem like a relatively simple bureaucratic question, but as long as the issue of data ownership remains unresolved, archives may lack a clear mandate for collecting it or may be unwilling to assume responsibility for its preservation. With large information systems that are used to administer major social, educational, or regulatory programs, archives at each level should preserve some pieces of the system, but without cooperative approaches to appraisal and preservation, archivists will never know which pieces to preserve.

Data interchanges also raise problems of data integrity and data quality. Very large databases that collect data from hundreds of sources often have very high error rates. In spite of well–intentioned and elaborate efforts to mandate standards for data quality, there are few effective mechanisms to monitor data providers or to maintain quality standards. Moreover, local officials are unlikely to invest much effort in providing high quality data as long as they view their data contributions as little more than meeting mandated reporting requirements for which they receive little useful information in return. This is especially true when their primary responsibility is to provide direct services to clients under increasing fiscal

[9] A project, funded by the U.S. National Historical Publications and Records Commission, will explore the use of the Research Libraries Information Network (RLIN) for the exchange of appraisal data among 15 state and local governmental archives, plus the U.S. National Archives.

constraints.[10] The ability to download parts of a database, combine it with other sources of data, and manipulate the information for new purposes also threatens data integrity. Even if archives develop the capability of preserving selected pieces of these databases, archivists may not be able to document the data accurately and precisely enough for secondary use.

The enhanced capability to exchange data in the past few years has also increased concern over privacy and access to information. Although privacy and access considerations have been an ever-present theme in data archives, the enactment of privacy protection provisions and the establishment of guidelines and procedures for handling confidential data quelled some of the concern over unwarranted invasions of privacy during the 1970s and early 1980s. This issue has surfaced again for several reasons. With the recent increase in automation at the local government level, local governments are beginning to amass large quantities of personal data in machine-readable form. Local governments may or may not have guidelines in place to administer such data, but they generally lack the experience of federal and state agencies with this problem. There is also a growing interest in the commercial sector in acquiring, linking, repackaging, and selling information from public records at all levels of government.[11] Finally, some types of linkage and data interchange that were technically challenging or too expensive to consider in the 1960s, are quite feasible today.

The response to new data linkage capabilities may be new efforts to restrict access to public records. Records that are innocuous and pose no threat to privacy, may become restricted simply because they have the potential to threaten personal privacy when linked with other records. Access restrictions are also problematic for systems that exchange data between different levels of government. Because federal, state, and local freedom of information and privacy laws are not identical, archivists or public records custodians who administer the data must determine which restrictions apply. This is challenging problem when issues of ownership remain unresolved.

The growing use of proprietary software for large integrated networks may also threaten access and compromise the ability of archives to preserve data. Many large integrated networks use software that is protected by copyrights or licensing agreements. Because some data cannot be used separately from software, special agreements may be necessary to allow for its preservation and access in an archives. Finally, the transfer of public sector functions to private facilities may limit access to data unless policies are developed to clearly define such records as part of the public record.

Just as the exchange of information among local, state and federal agencies poses new challenges for data archivists in the public sector, this development may also foster positive changes in the field. The need to exchange data and documents for administrative purposes may hasten the development and adoption of data interchange standards. Federally mandated reporting requirements already impose some degree of standardization on the data collected to document a wide variety of program activities, and these federally mandated standards make it possible to identify fairly consistent data sets in localities across the country. The widespread adoption of data interchange standards is also essential if data archives are to preserve software-dependent data

[10] For discussion of the conflicting interests of state and local administrators, see James J. Heaphey and Robert H. Crowley, Standardizing Welfare Management: The State Versus the Counties (Albany, NY: Rockefeller Institute of Government and The Governor's Office of Employee Relations, Oct. 1984).

[11] Massachusetts Office of the Secretary of State, Public Records Division. Report of the First National Conference of Issues Concerning Computerized Public Records, Boston, Mass., 1987.

without the need to also preserve hundreds of different, non-standardized software systems. But if the demand for standards comes solely from data archives, it is unlikely that software and hardware vendors will respond. The need among administrative agencies to exchange data creates a myriad of new problems in the identification and selection of data. But this need may also result in simpler and more uniform ways of exchanging information which may ultimately make it easier for data archives to preserve complex data sets.⌑