

When the practices across all 22 organizations were taken together, the most interesting and significant examples fell into eight categories. These categories provide a way of organizing the results in a way that highlights the implications for electronic access generally. Other practices, occurring at a lower frequency among these organizations, are described separately.

Proactive Acquisition

Among the organizations studied, several engaged in systematic practices to identify and recruit information providers as well as shape the content received. That is, they were far removed from passive receptors for information that simply respond to the needs or requirements of information sources. Instead, these organizations employed a variety of methods to manage or change the information flowing to the repository by initiating interactions with the sources to influence what kinds of information was produced. This is treated as a different strategy from shaping the information inputs to a repository by some kind of filtering mechanism at the point of reception. A filtering or passive strategy can influence the actions of information providers indirectly by adjusting the contingencies under which information will be accepted. Proactive acquisition, by contrast, means that repository staff (or agents) become involved with information providers in the planning and development of information resources.

Of the organizations included in this study, five reported substantial activities that can easily be considered proactive acquisition. The two that appeared to be the most active in this regard were the Zentral Archive of the University of Cologne and the USDA, particularly the Economics and Statistics System. The Zentral Archive based their proactive acquisition on decisions about the preferred content of the Archive. The Archive's mission is to promote research and understanding of the social and economic conditions in Europe, both current and historical. Based on this mission, the Archive staff identifies gaps or weaknesses in their holdings and takes action to remedy these shortcomings. They provide criteria and best practice standards for the conduct of future survey research that will produce potential content for the Archive. Staff members are active in networks of researchers to keep in touch with emerging issues and identify data sources in the planning stages. The Archive surveys approximately 10,000 researchers annually to identify work in progress, methods, and potential publications underway or planned. This survey and active networking provides a form of "early warning system" through which the Archive staff can identify where to influence research prior to data collection. These methods can shape the research products in their formative stages rather than waiting to filter unacceptable work when finished.

As sponsors of research and participants in a research community, the USDA is able to influence the directions and priorities of the research community in a proactive way. They are able to influence the direction of new studies through financial support and by participating in research conferences, publications, and sponsoring new research projects. This is discussed in more detail in the section on communities of practice below.

A narrower aspect of proactivity in acquisition was reported by NCES. This agency has an active data collection and research program of its own, which is not considered proactive acquisition in this sense. It is able to influence the actions of other data providers, however. One mechanism is to use the results of the user surveys and focus groups it conducts to identify their needs and issues. This information shapes NCES data collection and is shared with other data collection agencies to influence these other sources. NCES is also a participant in decisions on statistical policy for educational data and the kinds of information flows required by Federal education policies and programs.

Somewhat less extensive activities are found in the ICPSR approach to proactive acquisition. Since this repository maintains a longitudinal data series, it is active in seeking out the results from organizations that conduct data collection at regular intervals. ICPSR also employs a filtering approach by establishing format standards and content criteria for accepting statistical data. ICPSR also takes an active role in developing or improving the quality of metadata supplied by data providers. As with NCES, this repository takes an active role in professional and government groups that work with statistical policies and standards, all of which can influence the kinds of data sets offered for acquisition. In addition, ICPSR sponsors educational programs for researchers, covering research and statistical methods.

Proactive relationships with data providers and users are central to NYDCJS's role in state government. Due to its prominent position in the governance of the justice system in NY State, the agency has some considerable authority relative to data providers and users. The primary focus of this authority is the repository consisting of criminal histories and related crime data. The kinds of data collected, collection and processing procedures, and controls for access and use are all established by state and Federal policies and regulations. The Commissioner in charge of DCJS also has oversight authority for the other justice agencies in the executive branch of state

government: State Police, parole and probation agencies, and corrections. This agency staff is thus engaged in the overall policy making and operations that control the collection, storage, and access to most state and local justice information resources.

A current example of this role for DCJS is evident in efforts to support transitions to a new incident-based reporting system. Incident-based data, reporting individual crimes, is distinct from but closely related to the case-based data that makes up the existing criminal history and court data repositories. Criminal history repositories are well-established and highly standardized among justice agencies, but the old incident-based systems are in the process of revision to conform to a revised and expanded Federal system.⁽³⁾ NYDCJS has been proactive in promoting implementation of the new system and in seeking funding for local justice agencies to do the same.

When comparing the agencies with a record of proactive acquisition to the others, a clear pattern is evident. The agencies with proactive acquisition practices are much more likely to have a specific program or policy mission. The Zentral Archive and ICPSR have specific social science research agendas. NYDCJS has a public safety mission. The Federal agencies all have a recognizable policy or program domain, such as education, agriculture, space exploration, etc. They treat influence on the flow of information into their repositories as one of the ways to pursue the goals in that domain. It is also more likely that the staff of these organizations have training and professional experience similar to that in the provider organizations. They may be part of that larger community of practice and thus be equipped to express preferences and influence data collection. Of course not all of the agencies with specific policy or program missions reported substantial proactive acquisition. But none of the more general repositories appear to invest resources in this kind of external activity.

Collaboration

Several forms of collaboration can be found in the activities of these repositories. These forms can be roughly divided into four groups: collaboration predominately with users, with data providers, with other repositories, or with more integrated communities involving many types of agencies. The practices in each form of collaboration are sufficiently different to deserve separate discussion.

Collaboration with users can itself take several forms. NASA involves users in collaboration to provide access to its own and many other repositories of related data. This is done primarily through maintaining and improving metadata in its Global Change Master Directory (GCMD). The GCMD is NASA's directory of earth science and climate change data and services, which indexes contents and provides metadata for approximately 2000 data centers throughout the world. The linked centers maintain their own listings and metadata on line, through a database management application. This helps ensure more accurate and current listings. A different form of collaboration with users is found in the Zentral Archive's European Data Laboratory. The Archive obtained EU funding to establish an access and analysis laboratory in which researchers can collaborate with each other and Archive staff on projects. The laboratory effort led to the establishment of the Collaborative Study of Electoral Systems, involving researchers from over 30 countries. Collaboration of the New Zealand Ministry of Justice takes the form of participation with users and data providers on the development and maintenance of systems and data sets. This form of collaboration appears to fit well with the justice domain, since the users and providers are typically the same agencies. Collaboration in the governance and operation of the repository itself is also seen in the ICPSR, but in a voluntary setting. The Consortium is a member-governed organization of over 500 colleges and universities. The operation is governed by a 12-member board (Council) of researchers elected from the consortium members, and working committees. This arrangement represents collaboration in a research domain in which the providers and users are often the same individuals and organizations.

Collaboration among providers was illustrated in work among Federal agencies on matters of data privacy and security. The NCES collaborated with other Federal statistics agencies to sponsor legislation facilitating data sharing among Federal agencies. However as of this writing, the legislation has not passed. NCES also has taken a leadership role in collaboration with other Federal repositories on policies and practices to deal with privacy and confidentiality concerns. Those practices are described in more detail in the section below on data confidentiality.

Collaboration among the users, providers, and information repositories linked to the USDA is unique among the agencies we studied. There is a very high level of many kinds of collaboration reported among the organizations involved in agriculture research and data activities. The various forms of collaboration we found are part of long-standing institutional relationships among the USDA units, agriculture extension agencies at the state and local levels, universities and research centers, and agriculture producers. These institutional relationships have their roots in the creation of the land grant colleges under the 1862 Morrill Act, with their agricultural research and education mission to the farming community. The Federal relationship with those colleges and their local

extension services was formalized in 1914 and has grown and developed since.⁽⁴⁾ The result is a complex network of legal structures, professional and scientific policies and practices, interorganizational relationships, and financial flows. Therefore collaborative processes relating to information access include joint development of programs, research agendas, and data requirements among research and government partners. Formal and informal communication across users, providers, and repositories is supported by frequent professional and research conferences, staff flow across organizations, shared professional and educational backgrounds, and numerous advisory boards. These relationships are described in a bit more detail below in the discussion of the community as a provider type.

The primary result of this collaborative environment is a high level of interaction among users, producers, and custodians of agriculture information. These interactions include an annual user survey, cross agency and unit collaboration on creating new data sets, and shared funding of new access facilities and research. USDA staff from the Economics and Statistics System and other units spend substantial time with state and local offices and user organizations, traveling extensively to meetings, conferences, and other opportunities to stay in contact. As described by one interviewer, the collaboration is part of the day-to-day fabric of how these organizations and individuals work, not a separate activity engaged in for an occasional project or event.

The level of collaboration on information issues seen among agriculture organizations, based on such a long history, cannot be readily or fully duplicated in other domains. However, some of the specific aspects of the collaborative behavior can be employed elsewhere. These include regular surveys of users, attendance at professional and research conferences, instituting advisory boards, and other mechanisms for users and colleagues to participate in decision making.

Confidentiality

Providing access to many of the information resources involved in this research requires maintaining the various levels of security and confidentiality. In this research we did not concern ourselves with the aspects of information security required to protect any electronic repository from attacks or intrusions by malicious persons or organizations. These security concerns are generic to all electronic repositories. Instead, we concerned ourselves with practices to maintain various levels of privacy and security in relation to access by authorized users. This is a particular issue for repositories of electronic information that provide access for diverse or general populations of users, but must limit access or use according to some regulatory framework. For these repositories, the practices of interest deal with controlling the conditions of access as well as controlling use of information subsequent to access.

The most elaborate set of confidentiality and security provisions in our research were reported by Federal agencies, particularly the NCES. This agency's repositories contain some data about individuals (students, teachers, etc.) that is protected by law. Yet the agency must provide some access to these data sets to fulfil its mission to support research and policy analysis for education. To do so NCES maintains both public use and restricted use files and a Disclosure Review Board. Before a data set can be released to a public use file, the agency's Disclosure Review Board must review it and make a recommendation to the agency head (Commissioner). Data in public use files do not identify individuals; data in some restricted use files may have such identifiers when judged necessary for research. To obtain data from a restricted use file, the user must obtain a Restricted Use Data License from NCES. These licenses, which are legally binding, specify the conditions of use and access that must be maintained by the user. The NCES employs inspectors who perform unannounced inspections at user sites to ensure that the terms of the license are enforced. ⁽⁵⁾

A restricted use strategy is also employed by the Census Bureau and Bureau of Labor Statistics (BLS). They employ licensing procedures to control use of restricted or confidential information. But their procedures and regulations are not as detailed and elaborate as NCES. Some BLS data is time sensitive so procedures are in place to monitor its release according to these sensitivities. However, most BLS data are available only in aggregated form and not suitable for identifying individuals. Some census data is collected at the individual level and is confidential by law (13 USC). In one particularly innovative approach to preserve anonymity, the Census Bureau has developed techniques to create synthetic data at the individual level. The technique transforms data from real individual records into new artificial records that do not represent any real person or household, but retain the statistical characteristics of the original data. The synthetic data could then be released for research without violating confidentiality requirements.

A voluntary approach to controlling use is employed for part of the Urban Institute's repositories, the Assessing the New Federalism and National Center for Charitable Statistics data sets. The Institute requires users of the public use files from these sources to register before gaining access. As a private organization, the Institute has

no statutory authority to control external user's actions, but can use their registration information to communicate with users if a problem arises concerning how data are used. For the files in the Institute's Federal Justice Statistics Research Center, no registration is necessary, since there are no confidentiality requirements for accessing the crime and court files. The same applies to data sets on state welfare policies in the TANF Typologies database.

A different confidentiality issue is faced by the NYDCJS. The criminal histories in their repository are a potentially highly valuable research resource, unavailable elsewhere. Studies based on these histories could provide useful new insights into criminal behavior and aid in prevention and rehabilitation. However, the legal restrictions in place on the use of these histories prevent such research by outside researchers. The agency has tried, thus far unsuccessfully, to have legal restrictions changed to allow some research of this type. In this case the agency's mission of public safety is aligned with the research interests of scholars. So the agency is in a position to advocate for both interests and attempt to establish a collaborative research relationship through changes in confidentiality policies.

Information Management

Noteworthy information management practices in this research fell into three types. Some practices dealt with improving access through changing the way the structure and organization of information resources was managed. A second group of practices dealt with maintaining and updating the content of repositories. The third set dealt with the location of the information management activities themselves, seeking improved access and operations through moving from centralized to some distributed management model. Thus the concept of information management that we use here is broader than would apply to the content of repositories alone.

Changes in the structure and organization of information provided ways to improve access. Two organizations, the FDIC and the New Zealand Ministry of Justice (MoJNZ) developed data warehouses as new ways to consolidate information from multiple files into a single system. The warehouse method of organizing multiple data sources was used in part to provide users with access to multiple data sources from a single access point or application. A warehouse can also be structured to help users combine data from various sources for analysis and reporting.

The data warehouse approach was also described as a way to improve maintenance and updating of information bases. The MoJNZ also reported that its data warehouse made those processes easier and more efficient, helping ensure up to date information for users. Another methods for maintaining and updating information sources, planned and partially employed by NASA, was creating mirror data sets on separate systems. These mirrored sets were set up so that changes in one would be automatically made in its mirror. Since NASA maintains or links to so many dispersed databases, mirroring would be an effective way to keep them in sync and up to date.

Other forms of information integration, different from warehousing, were also seen as effective paths to improved access and use. The NESTAR tool employed by the UK Data Archive can provide for integrated searching and compilation of data from distributed sources, using metadata to search and compile, and XML for data interchange. Using other tools, the MoJNZ plans to integrate justice data with related social policy and demographic data to support research and program planning. An extraction strategy for data integration is pursued by the Zentral Archive, in which they are bringing indicators from over 300 sample surveys to create a merged file describing international and intergenerational mobility. The theme of improving access through integration was overall an important one appearing in these examples and many others related to collaboration and interactivity, described below.

Distributed management of information resources provides a related mechanism for improving the maintenance and currency of information resources. NASA's collaborative Global Change Master Dictionary, described above, is an example of this kind of design. The activity of managing metadata takes place locally, done by the custodians of the various distributed databases. But the results of that management activity are facilitated by and accessible in a central system. A similar approach was described by the UK data archives, but not fully implemented at the time of the interviews.

A broader concept of distributed management was described by the UK Data Archive staff. Though not a fully established practice, it was sufficiently interesting to deserve mention. It can best be described as a multi-tiered system of data collection, storage, access, and use. It would be based on localized information management mixed with global access. A government agency, for example, could be responsible for managing the collecting and storing, and access to information for a tier of users, such as other government agencies and their

stakeholders. Organizations and agencies in another tier, such as a group of commercial users, might manage a different access and use a structure that would obtain and analyze and annotate the same information, along with other sources, for their own purposes, creating different repositories or information products. Access to any particular resource would be controlled by local systems that would impose rules and conditions on access (fees, licenses, etc.). This is similar to the current structure of some information access management arrangements, such as the management of Justice statistics by various private or non-profit organizations (e.g., Urban Institute, universities), with access through the internet, and to some degree the NASA GCMD site. However, in the ICPSR concept, there would be common search and analysis tools to navigate access across tiers in highly flexible ways. The NESTAR tool employed by the Archive approximates this kind of tool. But it requires compatible infrastructure, standards, protocols, and a data interchange medium (e.g., XML) to operate as conceived. This appears to be a development direction for the UK Archive, ICPSR, and possibly other repositories.

Interactivity

Methods to enable interactive access to and use of data were the most frequent notable practice described in the study. Altogether, twelve of the agencies reported one or more practices that provide users the opportunity to work with information on the site beyond simply accessing and downloading records or files. In all but one case, the medium of interaction was some form of Web-based tool or application. The exception was a BLS telephone-based voice response system that allows users to request specific data tables or other extracts from files to be faxed or sent out. Since the interviews were conducted, the BLS has created Web-based interactive tools to provide that service, as well as additional features similar to the others described here.

While the goal of interactivity for users was common across these organizations, the methods and styles of interaction varied considerably. The capabilities available in these interactive systems can be described in terms of three main types:

- ad hoc queries that enable users to extract information in structures and combinations that don't exist in native form in the repository,
- visualization tools to present images based on processing and analysis of information from one or more sites,
- complex analyses of information extracted from one or more sites, with results presented to the user on line, and
- searching and indexing tools to support user exploration of the contents of one or more sites in ad hoc ways.

In some cases, the interactive capability offered by a repository included just one of these types, while others involved complex combinations. These interactive features were often described as representing the direction of planned future developments. In the time since the interviews were conducted, a quick survey of the Web sites of these organizations showed considerable expansion of the capabilities described in the interviews. Where appropriate, these newer capabilities are included in the details below.

The most advanced and complex capabilities we found combined tools for ad hoc queries, searching, data integration, and analysis. These are combined in a single tool set known as NESSTAR, developed at the University of Essex (UK) and Norwegian Social Science Data Services in Bergen. It is a combination of a browsing tool to locate data in a distributed data set and analytical tools to carry out simple analyses and download data to local files for further work. It works off a central server and uses standard syntax and data exchange tools (XML and DDI) to link across diverse systems. The central server maintains metadata about accessible data sets, filtering and authentication mechanisms, and the operational tools. The data reside in distributed organizations that collaborate in the overall system.⁽⁶⁾ NESSTAR is employed by both the UK Data Archive and The Zentral Archive. The NESSTAR system is used in the Zentral Archive's Eurolab, described above.

Similar capabilities, though not in an integrated tool set, are offered by the NCES. That Center's Web site provides 19 features that interact with one or more data sets. These features range from searches for data about individual schools and school districts to building a table from existing variables, to filtering through higher education data to find a set of institutions that are statistical peers within a chosen institution. In addition, NCES provides an online analysis engine (the Data Analysis System) to perform correlation analysis with selected variables. It also provides a mechanism to request more complex analyses by submitting requests online, with results returned by email. This last service requires registration, but involves no confidentiality control, since only unrestricted data are available. In addition, the NCES sponsors the International Archive of Educational Data, to support comparative and institutional research, which offers the Data Analysis System and similar interactive tools. The primary distinction between the NCES access and analysis tools, compared to NESSTAR, is the data

source. NCES accesses its own and closely linked Federal sources; NESSTAR can access any data set linked to the server.

The ability to process ad hoc queries for specific data, and to create tabular output was reported for several other repositories. The FHWA, BLS, Annie E. Casey Foundation, and Urban Institute have Web-based tools on their sites that support that type of interaction. The same is true for the Census Bureau, though census data available by this means have been purged of elements that would violate confidentiality requirements. The Urban Institute system also has added analytical capability that allows the user to generate cross tabulations of data from some of its data sets. The Kids Counts repository developed by the Annie E. Casey Foundation provides for extraction of longitudinal data at the national, state, and county levels. The Web site provides for creating charts showing trends in these data over multi-year periods (most for 1993-2000), and comparisons across several localities (e.g., comparing two or more states or counties with those states). This repository also provides color-coded maps of the US or counties within states, showing values for indicators chosen from the database (e.g., percentage of children living in poverty).

Ad hoc query capabilities to generate some kind of visual display were reported in government repositories as well. Map referencing to data through a geographic information system (GIS) was described by the US Census Bureau, NCES, and the Minnesota Data Center. The Minnesota Center provided users with CD's that combined data and map-based interactive display functions. That Center's Web site also provides static maps displaying various economic and social data does not provide interactive mapping capabilities. The NCES Web site allows searching for some data and reports through map references, in particular with the Data from the National Assessment of Educational Progress. The Census Web site, by contrast, offers a very high level of interaction with map referenced data, linking a sophisticated GIS with Census and other data files. That site allows users to select from a range of variables and display their distribution on highly detailed maps, with resolution down to the local political unit and census tract level. Users can change the resolution of the mapping and the variables mapped, and overlay combinations of some variables. This level of interactivity in a GIS display was the most advanced of those reported.

Other forms of visual displays were available to a limited degree. When extracting data about individual school districts, the NCES Web site displayed pie and bar charts of selected variables. The displays were not interactive to the extent of changing the content or display type, only selecting individual institutions to examine. Scatter plots of crime statistics are available from the Urban Institute's Federal Justice Statistics Research Center site. That site plots frequencies of the crimes and prosecution by crime type for several years and geographic levels. The user can choose the data to be displayed, but not the characteristics of the chart. The NESSTAR tool used by the UK Data Archive and Zentral Archive also has the capability to generate charts from analysis of data extracted through that system. The user can choose from a menu of chart formats to generate a display.

Some type of searching capability was reported for all the Web sites with interactive features. Key word searches were common, providing search engine-type access to material on the site. A geographically referenced gazetteer was reported by the UK data Archive, through which a user could search for data availability by map location. The NCES site provides a more structured search facility, with options to narrow searches by parameters matched to the contents of NCES databases. Identifying desired data sources through metadata files was the search strategy used in the NESSTAR system and the ILSIS, developed at the Zentral Archive. Metadata-based search capability has the virtue of creating a type of virtual catalog of data sources according to the search parameters established by the user. A related method for providing access to data resources via user searches was under development by the BLS. They were exploring automatic tagging of text and other content to facilitate indexing and efficient searching. However, results of this effort were not available for this report.

There has been significant development in the availability and power of Web-based interactive tools since these interviews were completed. The Web sites of all the participating organizations have been expanded and new features added since then. It was clear from their plans described in the interviews and the evidence of recent development completed that Web interactivity is a high priority. In describing these plans, many of the respondents made clear the reasons for this priority. One was the desire to provide users with easier, more efficient access and enhanced analytical power. The other was the potential for increased efficiency and cost savings for the repositories by automating access and analyses, with the user directing the processes. By investing in user-guided or controlled access and analyses, the organization could provide the same or enhanced services at lower costs to their budgets. Most of the respondents mentioned budget pressures as a constraint on responding to increased user demands in any other way. Given this combined incentive for interactive functionality, development along that path is very likely to continue.

Metadata

The quality and completeness of metadata are key factors in access practices of all kinds. The search and interactivity capabilities described above depend in large part on the metadata resources available to the searchers, the applications, and engines that do the work. The same applies to methods for integrating information from diverse sources. Managing and sharing information resources depends on the ability to describe and interpret the contents of data repositories and is also a direct function of metadata resources. However in spite of the centrality of metadata to these access programs, there were two distinct types of metadata practices reported in the research. The first had to do with improving the quality and usefulness of metadata for structured data sets, primarily statistical in nature. The other consisted of ways to create metadata for data resources that lacked it altogether or had substantial gaps in the available metadata. The strategies differ markedly between these and so are discussed separately.

The repositories that were concerned primarily with structured statistical data sets devoted more attention to the quality and completeness of metadata resources. Part of the proactive acquisition discussed for the central archive above, involves working with principal investigators who are developing new data resources. By working with these investigators prior to data collection, the staff of the central archive could insure the quality and completeness of metadata provided with those new data sets. A similar proactive approach was used by the UK data archives. These archives developed metadata standards for use by providers of data for their repository. They also worked closely with high CPS in developing the standards and applying them to development of the NASA program. Part of the effort to provide adequate metadata to users of statistical databases was directed to the problem of multiple languages in use. The central archive and the UK data archives both deal extensively with researchers from many countries. This raises the problem of translation of metadata to make it accessible internationally. The UK data archives are working with the European Community to develop a multilingual thesaurus for metadata and to develop automatic indexing capabilities. They are also working to develop what they referred to as "contextual metadata." This type of metadata would provide information to the user about the circumstances surrounding the data collection.

Standardizing and ensuring adequate metadata is a particular problem for repositories. It is a special problem for those that except datasets from a wide variety of sources. The ICPSR reported investing substantial staff resources in reviewing the metadata received with datasets. The staff will require additional documentation from suppliers when necessary. Standardized metadata is also important for repositories that provide search capability based on metadata files. This is true of the NASA Global Climate Change Archive and Federal justice statistics maintained by the Urban Institute. For the global climate change archive, NASA relies on the many suppliers of datasets to maintain the accuracy and currency of metadata on the NASA system.

Complete and high-quality metadata is much less likely to be available for data sets that come from administrative processes, collections of text, and other archival material. Metadata for these kinds of resources is typically created through indexing or tagging processes. For small volumes of material, indexing and tagging can be done manually. But that is infeasible for large volumes of information. Automatic indexing is a form of computer-based text analysis that assigns Index term, or tag, to a section of text or other material. Systems to do this kind of indexing automatically can be very valuable, but also very difficult to develop and maintain. For a general-purpose Library, such as the Washington State Library, the variety of material submitted is very large, making the indexing problem even more difficult. The Washington State Library reported success to some degree in indexing up to 40,000 current documents using their automated system. They also described efforts to work with information providers in order to have them contribute to that indexing process. They are attempting to provide support and standards for the originators of information to provide adequate indexing and other metadata to the repositories.

Migration & Preservation

The electronic information resources that were the focus of attention in this research exist in a very wide range of formats and storage media. The formats and storage medium for any information set is a result of decisions made about technology by those responsible for creating the information in the first place. Since information technology changes rapidly, new formats and storage media are becoming available at a rapid pace, and older methods and materials become obsolete just as quickly. This process presents the repository with the problem of deciding on a format and storage medium for its content that will preserve access for as long as necessary. All of the repositories in this study had developed methods for dealing with this problem. The methods differed, however, due to the nature of the information they store, the technologies and needs of their users, and the time frames for maintaining such storage and access.

The research revealed two basic approaches to solving the problem of maintaining long-term access to electronic information in multiple formats and storage media. The repositories developed policies for the kinds of formats and storage media that they would except, in an effort to reduce the variety to maintain or migrate over time. The repositories also developed strategies for migrating information from older to newer formats, according to the nature of the information and the needs of the users. The problem of receiving data in multiple formats is more severe for repositories that accept data from a wide government or research community. To reduce the variety in formats, the ICPSR accept data only in a limited set of very common formats. A similar strategy is followed by the Federal Justice Statistics Research Center. However, these repositories to make exceptions for data sets that represent substantial value, even if they are in an unusual or obsolete format. The ICPSR staff reported excepting data sets recently on punch cards, although they did not currently have equipment on-site that could read the cards. So they had to go to an equipment warehouse to find punch card readers in order to create electronic version of the data set. They noted that the same problem could occur for equipment used to create current formats. It may be necessary to maintain some obsolete equipment in working order for the purpose of processing or migrating old data sets that come to light.

Once a data set is accepted in a particular format, it will still be necessary to refresh or migrate as information as technology changes. The repositories reported systematic conversion projects and schedules for migrating to new formats. They engaged in risk analysis to better understand the consequences of alternative conversion strategies and formats. One principle coming from the risk analysis which several mentioned was to convert to the most frequently used formats, since they were likely to persevere in use over longer periods. The large number of users for these common formats would provide an incentive for developers to create migration technologies and methods for them.

User support and services

Some of the repositories devoted substantial resources to support and assistance for users. The ICPSR conducts extensive training programs for researchers, both in research methods and data storage and preservation issues. Several of the repositories conduct regular user surveys to identify user needs and areas where support can be improved. The Zentral Archive created a user laboratory in order to facilitate access to the data resources and collaboration among their research community. The NCES provides online educational materials on its Web site, published training materials, and training courses for users. The repositories that deal with the most diverse user populations appear to have the most extensive educational and help facilities on their Web sites and in their programs. These would include the NCES, the Environmental Protection Agency, and the USDA.

(3) The FBI's Uniform Crime Report system, started in 1929, was expanded into the National Incident Based Reporting System (NIBRS) in the late 1980's. Since then, the US Department of Justice has been working with states and localities to bring them into NIBRS. Currently 34 states have certified systems, with work toward certification under way in all but two of the remaining states.

(4) A brief history of the agricultural extension service can be found at <http://www.csrees.usda.gov/qlinks/extension.html>

(5) The legal framework that applies to NCES data is described at <http://nces.ed.gov/statprog/confid3.asp>. Under current law, violation of these confidentiality regulations is a Class E Federal felony.

(6) Additional information on NESSTAR (Networked Social Science Tools and Resources) can be found at <http://www.nesstar.com>.