

Part of the research goal was to examine the practices of these repositories for patterns or commonalities among them. Even though as organizations they are quite different in size, mission, and overall structure, we expected some similarities in the way they handle their information access responsibilities. Some reasonably consistent patterns did emerge in the relationship between some important features of the organizations themselves and their information access practices. A review of the similarities provides some useful insights into how organizational arrangements should be taken into account in seeking to enhance the provision of information access.

The patterns and consistencies we saw in these organizations can be described in terms of three main characteristics.

- Nature of the relationships between the repository, users, and information providers, including whether or not their interactions were routine and institutionalized or ad hoc and episodic, whether there were formal authority relationships, how long the relationships existed, and whether they extended beyond information access matters.
- The relationship of the information storage and access to the overall mission of the organization, whether providing access was the central mission of the organization or just one among many functions.
- The role of the repository in the overall life cycle of the information, whether the organization was simply a repository, or was involved in other aspects of data collection and processing.

In reviewing the activities reported by these organizations, it appeared that describing them in these terms led to reasonably clear groupings. The groupings could be named according to the common characteristics they shared. There were five in all: communities, advocates, libraries, lone rangers, and composites. Each is described in more detail below

### Communities: FDIC, Minnesota Data Center, MoJNZ, USDA-Cooperative State Research Education, and Extension Service & Economics and Statistics System

The primary characteristics of these communities are long-standing institutional relationships, shared mission, and identity. In each, most if not all of the organizations involved in data acquisition, use, and access activities have a formal or legal relationship, often based in statutes. They also work in the same policy domain, such as health, public safety, etc., which results in shared understanding of their overall mission and professional identity. The USDA community has roots and relationships going back over a century, as well as many professional and educational linkages. The FDIC deals primarily with the banking community and is closely related to it legally and professionally. The New York State agencies are tightly coupled with the public safety and public health communities respectively, and so forth. These are long-term, interdependent relationships. The kinds of information involved may be highly varied, but the conceptual and institutional frames are very similar within a community.

As a result of this pattern, some of the problems faced by other groups are less severe. Metadata and standards issues are generally less serious than with other groups, due in part to the legal frameworks available to support standards and consistencies, though they are far from fully effective. The focus on a more-or-less common mission means that the overall variety in types of information to be dealt with is less than elsewhere. Less variety in the nature of the data means fewer formats to deal with. The members of such a community are more likely to share the same assumptions about priorities and overall goals as well.

The relatively hierarchical structure and legal status of the relationships in such a community can be troublesome as well. Because the relationships and practices are often embedded in a policy and legal framework, change can be difficult and resources scarce. Government agencies must deal with annual budget cycles that inhibit long-term planning. For this group, their information activities are performed in the service of specific policy objectives. Providing information access is a means to an end for the repository, not an end in itself. Therefore, information services often must compete for resources with other programs and priorities. And unanticipated changes in policies and political priorities can interfere with information system developments and investment. The tight relationships have both positive and negative impacts.

### Data Libraries: Zentral Archive, ICPSR, NASA, WA State, UK Data Archive

The most important distinguishing characteristic of the repositories in the group is their primary, dominant function as a provider of storage and access services. While their content priorities and user communities may vary, all of

these organizations exist to acquire, preserve, and make available some class of information. This characteristic shapes most relationships with users and information providers, as well as the access practices employed. The library that is part of larger government (Washington) does have long term institutional relationships with many users and providers. However, it does not share the same mission and program goals of the other government agencies and the staff generally does not share professional identity and educational backgrounds. The relationships among the other libraries and their user/provider environments are much less institutional in nature, and may be ad hoc or short term in many instances, and may be strictly commercial. That is, in some cases, the library is simply a vendor, providing data for sale. Thus the relationships can be more market-like than a community or network of organizations. Where long term relationships are developed and maintained, they tend to be for the purpose of enhancing the quality of the information received or for improvements in the access processes rather than the pursuit of some policy objective.

The NASA Global Change Master Directory (GCMD) is a library of a narrower kind, but still similar to the other members of this group. The GCMD exists to acquire and provide access to information—in this case metadata. It has long and short-term relationships with the providers of this information, but the relationships are seldom statutory. And while potentially very large, there is no particular common identity or organizational linkage to the users of this resource, or for the very wide range of data sets accessible through the GCMD.

The range of information acquired by these libraries varies considerably. The Washington State Library is the most eclectic of the group, with its very broad mission to serve as, "the corporate library for Washington State Government, ... deliver information services to the legislature and state government entities as they develop and carry out public policy; and [as] a leader in information policy, ... partner with libraries and other entities to provide ready and equitable public access to information."<sup>(7)</sup> This broad mission means that this library deals with the highly varied materials produced in the course of Washington State government, both historically and currently. By contrast, the other members of the group have narrower missions in support of specific research communities and constituencies, dealing primarily with scientific and demographic information and statistical data sets.

The organizational structures and mission of these libraries require responses somewhat different from the community-type arrangements described above. The large number and diversity of potential suppliers of information to these libraries presents problems of acquisition management. These library-type repositories lack the close linkages and controls in a strictly government or institutional context. These libraries therefore must devote resources to managing the way suppliers present information to reduce problems arising from missing or low quality metadata, problematic formats, and other data quality and usability factors. As a result this group was the most heavily invested in proactive acquisition strategies. Libraries must also accommodate a variety of users with wide-ranging skills and technology resources, as well as disparate goals and information needs. Since providing access is central to their mission, however, as a group they devote substantial attention to working effectively with various user needs and capabilities.

### Comprehensive Operations: BLS, Census, FHWA, NYSDCJS, NYSDOH

Of the repositories selected for this study, five conducted comprehensive operations, consisting of data collection, analysis, storage, and access provision. That is, their repository function was integrated with a role as the major or exclusive originators of the information resources, or taking a substantial part in that data collection process. Some had large-scale data collection operations internal to the organization, as in the case of the Census Bureau, BLS, and NYSDOH. The others had a major role in administering or sponsoring the processes that resulted in acquisition of information. All of them had primary responsibility for providing access to these resources, including policies on confidentiality and use.

As comprehensive repositories, these organizations maintained a somewhat different set of relationships with users and those involved in supplying information. Much of the supply of information is from units within the organizations, other government agencies with which it has functional relationships, or contracted data collection by external firms or other government agencies. The intake of information is thus largely under the control of the repository or regulated by policy, especially for statutory collection and reporting requirements such as the decennial census, quarterly inflation indicators, crime statistics, or educational assessments. In addition to required information collection, these agencies can be proactive with respect to additional research programs that generate new information flows. They may conduct these studies with internal staff, or contract for the data collection through commercial organizations, other government agencies, or research organizations, such as universities.

The flow of information into and out of these repositories is largely regulated by the agency and its legal and policy framework. In the case of the NY State agencies, much of the information flows both into and out of the

repositories is confidential and thus limited to specific, legally sanctioned users and uses. This also applies to some of the data in the FHWA, BLS and Census files. For the BLS, confidentiality is in some cases a temporary constraint, since certain economic statistics (e.g., inflation and employment indicators) have important financial and political implications. That information is embargoed until the regularly scheduled release time. Premature release is illegal and punishable.

Providing appropriate access is complicated by the mix of users, running from lay persons seeking small targeted items of information for personal use (e.g., parents seeking information about a school system), to policymakers working on national issues, to researchers seeking large data sets, ad hoc queries, or new sophisticated analyses. The ones in this group that serve the general public, primarily the BLS, and Census Bureau, had therefore invested heavily in interactive access capabilities and online analysis and query tools. These provide efficient ways of supporting large volumes of user interactions with a limited staff. They reported focusing professional staff resources more on responding to requests from policy makers and the research community. Their attention to the needs of their user community is also reflected in substantial investments in user support.

This group of comprehensive repositories also paid considerable attention to the problems of multiple data formats and migration. Some of these concerns are a direct result of these agency's roles in long-term retention of government records and statistics. Even if the agency has direct control over the formats of data at the collection stage, the need to deal with both emerging new and obsolescent old formats remains. This is a particular problem for large agencies, such as these, that support many diverse Web sites, each dealing with a particular program or policy area. The current FHWA Web presence includes 40 separate Web sites, each with distinctive information content and format requirements. The FHWA's Highway History Web site, for example, includes an html version of the first issue of **Public Roads**, (Vol. 1, No.1) from May 1918. There is the additional need to provide data in digital formats to other government agencies with different formatting requirements, such as in the case of the Census Bureau and the New York State agencies. In spite of the exemplary practices these agencies have developed to deal with multiple formats, the problems will most likely persist, due to the combination of technology change and increasing conversion to digital formats.

### Advocacy or Policy-focused Repositories: EACF, Urban Institute ANF & FJSRC

This repository type consists of private, non-profit organizations that exist to pursue a specific set of policy objectives. The Annie E. Casey Foundation (AECF) states its mission as "to foster public policies, human service reforms, and community supports that more effectively meet the needs of today's vulnerable children and families." With a similar but somewhat broader mission, the Urban Institute states its purpose as, "to examine the social, economic, and governance problems facing the nation." The provision of information to policy makers and the various stakeholders in their respective domains is a central part of these missions. The Urban Institute's mission statement is explicit, i.e., to provide "information and analysis to public and private decision makers to help them address these challenges and strives to raise citizen understanding of the issues and tradeoffs in policy making." Both organizations maintain Web-accessed repositories of information, including statistical data sets that can be used to advance their respective missions.

There are, however, important differences between these organizations, in terms of funding, overall operations, and relationships with other organizations. With respect to funding, the Urban Institute is supported to some degree through contributions and primarily through grants and contracts for specific policy-related research projects. The AECF is a private foundation with an endowment (approx. \$4Billion), the income from which it uses to award grants and operate programs, including the Kids Count data sets and other data repositories. The Urban Institute's repository and research program related to Assessing the New Federalism is in fact supported in part by grants from the AECF. As a sponsor of that program, the AECF is in a position to influence the nature of the repository, including the kinds of information and research products it generates. Any of the Institute's information programs, repositories, and research efforts reflect the merger of sponsor's influences with the Institute's mission and the expertise its staff. The relationships with government agencies differ as well. The AECF is independent of government, but directs much of its effort at influencing government policy and programs. The Foundation's repositories draw heavily from government data sets as well (e.g., the US Census). The Institute is more directly connected to some Federal agencies through grants and contracts to operate repositories and conduct research on their behalf. Overall, then, the AECF is in what could best be called a patron-client relationship with its grantees, and in a community relationship with its users. The Urban Institute is in more of a client-patron relationship with its foundation and government sponsors.

In terms of access to stored information, the differences between these organizations have at least one major consequence. That is, access to AECF information is structured in a much more coherent and focused way on the

core mission of the Foundation. There is a balanced mix of access to statistical data and analyses along with indirect access through research reports. By contrast, the Institute's repositories cover a much wider range of issues and are consequently less focused. There is much more indirect access to information through research reports created for sponsors, than direct access to the statistical data on which reports are based. The FJSRC databases are, of course, available directly for download. But the interactive analytical capabilities available directly through the repository are at a lower level and do not provide trend analysis.

### Composite (mixed) Operations: GISP-OR, NCES

These repositories characterized as composite or mixed operations differ from the others primarily in the combinations of roles they play in the overall acquisition, storage, and access provision for information. The mix is such that they do not fit well with the other types. The NCES is similar in many respects to the Federal repositories in the comprehensive group. The Center is a receiver of government statistics about education, for which it provides storage and access, as well as a proactive agent in influencing what data are to be collected and by what methods. NCES is also an originator of data for its repositories, through both in-house data collection and contracting for data collection and research with other government agencies and other research organizations. As a part of the Department of Education, the Center is active in information policy formation as well, for education and for Federal statistics generally. The Center conducts in-house research and has an extensive publication program for research reports and statistical material. In this respect it is similar to the BLS and Census Bureau. However, unlike these other agencies, it also provides a rather wide range of training, research grants, and collaborative research programs with related government and private organizations (e.g., the American Educational Research Association). In addition, the institutional relationships in the education sector extend from the Federal level, to state education departments, to local school systems. This makes much of NCES's information work part of the governance of this national system.

The mix of information roles in the GISP repository is much smaller and less diverse. The focus of this repository is much narrower, namely fostering collaboration and sharing information internationally about alien invasive biological species. It is a combined repository of index and linking information about related databases together with research reports and periodical publications related to this theme. The links to and involvement of international and non-US agencies are extensive. In this respect the GISP site is similar to the NASA GCMD, though not part of a comprehensive agency or providing for localized update of metadata. What is most notable about the repository is that it has developed from a largely voluntary effort and is heavily dependent on international collaboration. It illustrates the capability of Web-based resources to support collaboration among widely dispersed and diverse organizations with a common concern or goal.

For both organizations, information access is central to their mission. Therefore electronic access to their content is a high priority. In both cases the content is both digital and paper-based, so multiple formats and delivery mechanisms are required. For NCES, however, the publications are developed largely in-house, while the GISP publications are compiled from many external sources. Therefore the requirements of administration and vetting of content are different. NCES has mostly hierarchical or contractual relationships with information providers or creators, and thus more control over content and format. The GISP organization is largely voluntary, with more network relationships and informality governing interactions.

### Patterns of Practice in Types of Repositories

In reviewing the practices reported in the research interviews, it became apparent that there were some marked variations in the kinds of practices across these types of organizations. In order to track these variations, the text of the interviews was coded according to the kinds of practices mentioned. A large number of practices were described and coded this way. However, many of them were mentioned only once or twice over all interviews, so they were not useful for comparison across types. For the analysis discussed here, only the practices with several occurrences were used.

Using the coded text material, it was then possible to tally the references to particular practices and relate that tally to the type of organization. These tallies can then be considered a rough indicator of the prevalence or importance of that type of practice in that organization. Such a counting is at best an approximation of prevalence of a practice, since a single mention may in fact involve a substantial effort, and many mentions merely embellishments of a small effort. Overall, however, the differences in where the particular practices are mentioned does provide some insight into the possible relationships between access practices and the organizational setting in which they occur. The results of this analysis are shown in Figure 1 below.

This figure shows the percentage of the total occurrences of the practice for each type of organization in which it occurred. That is, the height of the bars for each of the types of practice in the figure add to 100%. If a bar does not appear for a type of organization in the space for a practice that means that practice was not reported for that type of organization. For example in Figure 1, practices related to migration and formats have two equal height bars (50%), one each for Community and Library type organizations. In the interviews, practices related to migration and formats were reported 10 times, five each for Libraries and Communities and none for the others. This way of recording the results normalizes for the different number of organizations in each category.

In spite of the roughness of this type of tally, the patterns in Figure 1 do suggest some useful observations. The most obvious is that practices do vary considerably across these types. The library-type organizations appear to have the most prevalence of these notable practices overall. Efforts with respect to proactive acquisition, metadata, and understanding user demand seemed particularly valuable. Considering that they face a very wide range of problems of multiple user types, heterogeneous inputs, complex environments, and growing demands, it follows that they should have created many creative responses. The Community types are a close second in this kind of indicator. Practices related to integration and information management are frequent for this type. This may be a result of the needs of the community for a variety of information products, drawing on inputs or analyses from multiple sources. The community organizations tend to be concentrated in policy domains where the desire for integrated analyses for policy purposes are stronger. The only practice types that were reported by all types of repositories were interactive access and user support and friendliness. Since these repositories all share a common mission to provide access to information, concern for users would be expected. And given the growth of Web access and technology generally, this is not surprising. This may also be a result of budget pressures. Many of the interviews described interactive access efforts as ways to reduce costs or improve services without increasing expenditures.

### Figure 1 - Notable Practices by Type of Repository

The practices related to confidentiality show an interesting pattern as well. The high bar for confidentiality in the mixed group is primarily from the NCES repository, which reported many practices of this sort. Confidentiality concerns for comprehensive and advocate organizations are indicated as well, which is consistent with their contents and organizational relationships. The lack of confidentiality concerns for communities and libraries also seems consistent with their content and mission. Most of the community organizations in this study do have confidentiality needs, but no particularly notable practices in that regard were reported.

These types of organizations were recognized in the analysis of the interview data, after data collection was complete. So it was not possible to explore the implications of this kind of consistency with the organization's staff. With the information from these kinds of patterns now available, it would be potentially valuable to revisit these organizations, and others that fit the categories, to explore in more depth the origins and implications of these patterns.

---

(7) <http://www.secstate.wa.gov/library/aboutus.aspx?c=mission>