Robbin, A. (1995). SIPP ACCESS, an information system for complex data: A case study in creating a collaboratory for the social sciences. *Internet Research: Electronic Networking Applications and Policy*, *5*(2), 37-66.

SIPP ACCESS, an information system for complex data: a case study creating a collaboratory for the social sciences

Alice Robbin

The author

Alice Robbin <arr@cunyvmsl.gc.cuny.edu> is Associate Professor in the Public Policy Program of the School of Public Affairs at Baruch College, New York City, New York, and Adjunct Professor in the PhD Sociology Program, Graduate Center/City University of New York. Previously, she was senior scientist at the University of Wisconsin-Madison, where she was a Co-director and Co-principal Investigator of the *SIPP* ACCESS project at the Institute for Research on Poverty.

Abstract

The "collaboratory" concept has recently entered the vernacular of the scientific community to reflect new modes of scientific communication, cooperation and collaboration made possible by information technology. The collaboratory represents a scientific research center "without walls" for accessing and sharing data, information, instrumentation and computational resources. The principal applications of the collaboratory concept have been in the physical and biological sciences, including space physics, oceanography and molecular biology. Discusses the attributes of the collaboratory, and applies the concept developed by computer and physical scientists to the design and operation of the *SIPP* ACCESS prototype information system for complex data to be used through the Internet by sociologists, demographers and economists. Examines obstacles to collaboratory development for the social sciences. Concludes that four major obstacles will inhibit the development of collaboratories in the social sciences.

Internet Research: Electronic Networking Applications and Policy
Volume 5 · Number 2 · 1995 · pp. 37–66
© MCB University Press · ISSN 1066-2243

Introduction

Computers and telecommunications have dramatically altered the conduct of science during the last two decades. Reflecting on these developments at a 1989 workshop that addressed the relationship between information technology[1] and scientific research, William Wulf coined the metaphoric phrase "collaboratory" to represent new modes of communication, cooperation and collaboration ("c-cubed") that would improve the efficiency and effectiveness of the scientific enterprise. He wrote that a collaboratory represented a "center without walls, in which the nation's researchers can perform their research without regard to physical location – interacting with colleagues, accessing instrumentation, sharing the data and computational resources, [and] accessing information in digital libraries" (Wulf, 1989, cited by Wulf, 1993). More recently, he suggested that not only would scientific productivity be increased by leveraging technology, but technology also had the potential to "qualitatively change the kinds of questions we ask and, hence, what we know about nature" (Wulf, 1993, p. 854).

Wulf would agree that his was not a new conception of how computer and telecommunications technologies would modify scientific work[2]. The particular relevance of Wulf's collaboratory concept was, however, that it focussed the attention of science elites on the fact that scientific information infrastructure development was not principally a problem of technology which necessitated engineering solutions – problems to which the High Performance Computing Act of 1991 was designed to respond (Committee on Physical, Marketing and Engineering Sciences *et al.*, 1992; National Research Council, 1994a; US Congress, 1994)[3]. As the computer scientist Wulf himself noted, "The bottleneck to the achievement of such a vision is not hardware"(1993, p. 854)[4].

Thanks go to an anonymous reviewer who made constructive observations about current trends in the electronic network environment. The *SIPP* ACCESS project was a truly collaborative venture and, as such, my colleagues Martin David and Thomas Flory at the University of Wisconsin-Madison are (silent) co-authors of this article.

During the 1980s and early 1990s, national policy and planning initiatives in biology, neuroscience, oceanography, and space physics began to articulate the critical importance of information and data as resources necessitating human structures to organize and manage them (e.g. Greenstadt, 1981; Lander *et al.*, 1991; Cinkosky *et al.*, 1991; National Research Council, 1990, 1991a, 1991b, 1994b; Pechura and Martin, 1991)[5]. Reports on information technology and scientific research published by the US Office of Technology Assessment, prestigious scientific panels, and National Science Foundation-sponsored workshops, between 1989 and 1993, further reinforced Wulf's perspective (e.g. French *et al.*, 1990; National Academy of Sciences *et al.*, 1989; National Research Council, 1993; Scheuermann *et al.*, 1989; Silberschatz *et al.*, 1990; US Congress, 1988; US Congress, 1991b). These reports underscored the scientific community's increasing attention to the need for systematic organization, management and retrieval of data and information, and the management of instrumentation.

An important perspective on human resources was being developed, as implied by the words "organization" and "management". For example, molecular biologists commented that:

> Effectively storing data and its associated information is likely to be among the major challenges confronting biologists over the coming decades.... Making all this information easily accessible to distributed users while effectively dealing with errors, conflicts, and updates presents a challenging research problem of the utmost urgency (Lander *et al.*, 1991, p. 34).

A new emphasis on human capital placed the importance squarely on the behavioral, the cognitive (in terms of information processing and retrieval capabilities of the scientist) and the communicative (the interpersonal, group, political and economic structures and processes required for organizing the scientific community, information and data)[6]. Neuroscientists wrote that:

> The scientific enterprise is composed of men and women who generate ideas, design ways to test those ideas, collect data, and communicate the ideas and data in a variety of ways. The communication of ideas and results is as important to the growth of knowledge as the data themselves.... One of the major goals of computer network development is to create a communication environment that is as free of barriers as possible – an environment that can support the rapid communication of ideas and images at every stage of experimentation and discovery (Pechura and Martin, 1991, p. 84).

The discussion of collaboratory developments, as recorded in the authoritative reports issued by the National Academy of Sciences (NAS) and National Research Council (NRC) panels, does, however, leave the unfortunate impression that information infrastructure building activities occurred only in the physical and biological sciences during the 1980s. This article is, however, designed to rectify that impression by discussing an example of information infrastructure development in the social sciences.

The first section describes the collaboratory concept and capabilities that were identified in the NRC reports and other documents on which the NAS and NRC scientific panels relied to develop their recommendations for maximizing information technology to enhance the scientific enterprise. Although there are critical technical and engineering attributes that must be in place for a collaboratory to exist, such as the physical architecture of an information infrastructure (see McClure *et al.*, 1994; National Research Council, 1994a), this discussion emphasizes a particular model of an electronic community system that influenced the final recommendations for collaboratory development by the various scientific panels. Bruce Schatz's vision of an information system, which was designed to create an electronic community of molecular biologists who study various aspects of the nematode worm *Caenorhabditis elegans*, appears to have been instrumental in formalizing the attributes of a collaboratory. Schatz wrote that "an electronic community system encodes and manipulates the range of knowledge and values necessary to function effectively in a community or organization" (1992, pp. 87-8). The encoded knowledge represents institutionalized organizational memory (see also Simon, 1991). The task of information infrastructure development is, therefore, to create an integrated environment of tools and technologies that permanently records socially-generated, historical and current, formal and informal knowledge shared by a scientific community.

The second section presents a case study of *SIPP* ACCESS, a project to develop a prototype of an information system for complex data used by social scientists and policy analysts using the Internet, which began in late 1984 and was completed in December 1991 at the University of Wisconsin-Madison (David and Robbin, 1992). The discussion emphasizes: the theoretical basis on which the different components of the information system rested; the system's environment, functionality and structures; and the activities designed to create an integrated knowledge base of an electronic community as represented by the information system. Selective administrative data collected by the *SIPP* ACCESS project illustrate the ways in which information technology was applied to achieve improvements in efficiency and effectiveness. It is important to note at the outset that SIPP ACCESS was created at the beginning of a dynamic and unstable period of considerable technological change and standards development in the environment, including applications software, computer architecture, computational capacity and telecommunications. Consequently, some of the enabling information technology that is currently in widespread use either had yet to be "invented" or was in the early stages of development during the principal years of *SIPP* ACCESS operation and therefore "matured", and became widely accessible and affordable, only toward the end of the project (and thus too late to be incorporated in the information system)[7].

The third and final section of the article discusses some of the obstacles and problem areas related to the collaboratory concept and effective use of information technology identified by the NRC panels in terms of their applicability to the social sciences. We rely on empirical data collected by the *SIPP* ACCESS project, as well as our observations of social scientists conducting work in the electronic network, that both support and augment the discussion of the problem areas raised by the NRC reports. We discuss:

- the conceptual/technical problems of creating integrated information systems for heterogeneous scientific databases;
- cognitive and other attributes of users;
- collaboration between computer scientists and social scientists; and
- cumulative advantage of institutions.

Part one: the collaboratory concept and capabilities

The scientific research enterprise includes the major activities of data collection and analysis, communication and collaboration among scientists, and information storage and retrieval (National Academy of Sciences *et al*., 1989, p. 2). According to the *National Collaboratories* report, eight mutually reinforcing changes have affected the conduct of science and have promoted an interest in collaboration (National Research Council, 1993, pp. 5-6, 8).

These eight factors, shown in italics, are: *the increasingly complex phenomena and problems selected for study*, made possible by *developments in instrumentation and facilities that provide the capability for making more precise measurements* but which are *increasingly expensive*. The nature of the scientific activity has resulted in an *enormous growth in the volume of data and information that must be stored, accessed, analyzed and reported*, which requires *advanced information technology to manage the complexity of the phenomena under investigation (e.g. to collect, process, analyze and share the massive amounts of data being generated from observations and experiments)*, and *highly trained and specialized personnel to perform the tasks in structurally more differentiated work/production units* (see also US Congress, 1991b; US National Institutes of Health, 1993). There has also been *an increased interest in interdisciplinary research due to the recognition that "many pressing scientific problems transcend the boundaries of individual disciplines"* (National Research Council, 1993, p. 7) *and require "the meshing of different specializations to advance a research area"* (US Congress, 1991b, p. 35). Finally, fueling the interest in modifying organizational arrangements is the *more restricted funding environment to support the scientific enterprise*. *Collaboration is thus viewed instrumentally as a way to address both complexity and the "stretching and leveraging of available dollars"* (National Research Council, 1993, p. 7).

The *National Collaboratories* report (National Research Council, 1993) provides both abstract and operational definitions of a collaboratory for scientific research. Abstractly, the collaboratory is Wulf's "center without walls" (Wulf, 1989, cited in Wulf, 1993). Less abstractly, a collaboratory is an integrated environment of knowledge-generating tools and technologies designed to enhance scientific activity wherever it may take place. "Knowledge" is more precisely defined by Schatz as comprised of:

> both formal data and literature and informal results and news [that are manipulated by the scientist]. Manipulation includes...browsing through the available knowledge and recording and sharing interrelationships between the items [in a] software environment to manipulate this knowledge...[by] people with common interests and shared values (Schatz, 1992, pp. 87-8).

Translated into operational terms, the collaboratory concept is an information infrastructure that "provides a technological base specifically created to support interaction among scientists, instruments, and data networked to facilitate research conducted independent of distance" (National Research Council, 1993, p. 7). Collegial relationships are defined by area of specialization and not geographic location (National Academy of Sciences *et al.*, 1989, p. 20; see also Lederberg, 1978). The physical components of the information system are independent of the geographic location of the scientist.

Five criteria for this technological base must be satisfied: interoperability, transparency, customizability, integrity, and extensibility (National Research Council, 1993, pp. 28-9)[8]. A complete environment for collaborative scientific activity comprises:

- a distributed computer system;
- networked laboratory instruments and data-gathering platforms;
- tools to enable a variety of collaborative activities;
- financial and human resources for maintaining, evolving, coordinating, and assisting in the use of computer-based facilities;
- digital archives and libraries that include tools for organizing, describing and managing data, including images, to enable large-scale sharing of data; and
- digital libraries which include tools for organizing, describing and managing the information derived from the analysis of the shared data.

To make these activities possible, a collaboratory must support four basic capabilities:

(1) *Data and information sharing*: scientists who work on the same project should be able to obtain data and information quickly and easily from within and across databases.

(2) *Software sharing*: scientists should be able to share and exchange software conveniently for data analysis, visualization, modeling, information retrieval, etc.

(3) *Controlling remote instruments*: scientists should be able to control instruments (located in difficult-to-access regions on Earth or in space).

(4) *Communicating with remote colleagues*: scientists should be able to interact effectively with one another, despite being separated in space and/or time (National Research Council, 1993, p. 56).

Data and information sharing

Data and information sharing require three essential, dynamic, and interconnected components:

(1) Electronic libraries contain published and unpublished data(bases), formal and informal literature and unpublished findings that contribute to the knowledge base (Schatz, (1992, p. 95) calls the latter "lore"), and analysis and other applications software.

(2) The collaboratory maintains accessible archives of data.

(3) The facility provides a comprehensive system of "metadata" and "finding aids" to support browsing, filtering, retrieval and sharing of data and information.

Without exception, all disciplines placed top priority on integrated libraries for accessing the literature and helping scientists to locate and understand information and data. For example, oceanographers commented that data were frequently separated from information about the data, but it was critical that "information about the algorithms used for a derived product, quality control procedures, comparisons with independent measurements, reviews by outside experts, and so on, be an inseparable part of the data so that the user could judge the reliability of the product for a particular application" (National Research Council, 1990, p. 238).

Massive amounts of data are generated daily, with even greater amounts anticipated in the near future; the concern is that the inaccessibility of data has reached crisis stage (see also Clery, 1993; Marshall, 1993; NRC, 1991a, 1994b). A well-functioning information system preserves data, and long-term stewardship of research data is essential. Associated with the archival function of a collaboratory are the functions of data distribution, data integration and new product development, data documentation provision, data quality assurance, data identification and acquisition, selective data retrieval to meet user needs, and standards development for procedures (National Research Council, 1990, p. 235).

A comprehensive system of finding aids was recommended in order to support uniform manipulation (i.e. with one set of commands) of the heterogenous body of data and information generated by the information system. These include a variety of: "resource discovery" tools that help scientists locate and retrieve relevant data sources and applications programs[9]; and tools that systematically and efficiently relate and create the conceptual linkages across and between multiple, diverse, distributed and complexly-structured databases, and allow the individual user to "implement logical linkages between related items in different sources" (National Research Council, 1994b, p. 58; for extended discussions of these concepts, see Bright *et al*., 1992; National Research Council, 1990; Pechura and Martin, 1991; Pool, 1993; Schatz, 1987, 1992; Wiederhold, 1992).

Software sharing

The Committee on a National Collaboratory emphasized a collaboratory capability for software related to data analysis: sharing of software, application of external (i.e. not local) software to data, and application of local software to external data (National Research Council, 1993, pp. 58-9). The software capability must include visualization for graphical display of complex multi-dimensional data. A prior step requires, of course, that data be organized and managed in distributed databases, and accessible through similar protocols, as noted above. Scientists also expressed the desire to have at a wide-area-network level the same multi-functional software that exists in the personal computer environment, which provides the capability to integrate databases, word processing, electronic mail and spreadsheets, as well as other types of software for computer-supported cooperative work, for access to shared data and computing resources (National Research Council, 1993, p. 40).

Controlling remote instruments

Physical scientists, such as oceanographers and space and earth physicists, require a collaboratory capability for remotely controlling instruments and collecting data. For example, oceanographers employ remotely operated vehicles, satellite-borne sensors, trace chemical measurement, acoustic techniques, long-life buoys and floats, and sea-floor seismometers to gather time series measurements (National Research Council, 1994b, p. 8; see also National Research Council, 1993). Space and earth physicists collect delayed and real-time viewing data from multiple sensors in aircraft, spacecraft, satellites, and ground-based facilities worldwide in a distributed system that operates internationally (see Green and King, 1986; Marshall, 1993; National Research Council, 1993).

Communicating with remote colleagues

The fourth essential capability for collaboration identified by the NRC is communication with colleagues. The report commented that, "The support of interpersonal interaction among a group of collaborators may be the most challenging aspect of collaboratory construction" because it "potentially includes access to remote data, programs, and instruments, as well as to multimedia work-group communications systems" and "also requires an understanding of the complexity and vicissitudes of human behavior" (National Research Council, 1993, pp. 60-1). The basic tasks of scientific communication are multi-faceted and include:

- project organization and management;
- conduct of experiments;
- discussion and evaluation of findings;
- analysis of data;
- scientific review and commentary of discoveries;
- authoring, editing, publication, and review of documents; and

- organization of and participation in scientific conferences.

Consequently, applications to support these activities are diverse, such as electronic mail, bulletin boards, conferencing systems, file and document storage and retrieval systems, and computer-supported groupware (National Research Council, 1993, p. 61).

Human resources to design a collaboratory

The collaboratory concept derives from a set of assumptions by science elites regarding the nature of scientific communication and how scientific advances take place. Three aspects of a communication relation link scientists.

The first is that scientists collaborate because they are members of a subspecialty or subdiscipline that shares a similar intellectual framework, values and a common language, and agrees on the central problems facing the members and what constitutes the array of legitimate methodologies for problem solving (a "paradigm")[10]. Collaboration may occur intermittently, be short term, or take place over a long period of time. The dynamic that defines cohesion and the temporal conditions for the relationship is problem driven. Scientists, for example, may organize around:

- an organism as in the case of molecular biologists (Schatz, 1992);
- the development of a new technology, as in the case of the computer scientists who designed the Bitnet electronic network (Cotter, 1988);
- the development of a new application, such as the programming language COMMON LISP (Steele, 1984); or
- the design of a new product in a cooperative university-industrial venture.

Thus, Schatz suggests that two fundamental requirements must be met by members of a subspeciality in order to form an electronic scientific community: they must "have a large amount of data, both formal and informal, and a real need to manipulate these data extensively" (Schatz, 1992, p. 92).

A second aspect of the communication relation is that the boundaries of disciplines and subspecialties are typically not permeable, and groups tend to operate in isolation from one another. Language and cognitive understandings play very important roles in maintaining the boundaries of, and impeding communication between, disciplines or subspecialties. For example, Pechura and Martin (1991) observed that, when neuroscientists and computer scientists were brought together, "Beyond the expected difficulties of language (i.e., certain words mean one thing to neuroscientists and something entirely different to computer scientists), there were often fundamental differences in the perceptual frameworks used by the two groups". One outcome of the impermeable nature of the boundaries is that similar scientific discoveries may be independently derived or that inefficient solutions to a problem are developed because sufficient expertise is lacking (National Research Council, 1993).

A third aspect of the communication relation as it relates to scientific advances and interdisciplinary or intersubspeciality interaction is that members of the subspecialty must recognize that the information they need to solve their particular problem might be obtained by communicating outside their area of specialization. The motivation to seek out other scientists may be internally derived by a desire for originality (see Crane, 1969), an understanding about relationships between fields, or externally driven by agents who provide an intellectual or monetary incentive[11]. The assumption is that the action of communicating outside the area of specialization influences the development of new specialties that span boundaries and which are linked by critical individuals whom we call boundary-spanners (Allen, 1970) who provide the linkages. In this way, research will become more interdisciplinary, and a research area will advance by integrating different specializations (see US Congress, 1988, p. 35). Different organizational models of research then develop, and that research takes place in units other than academic departments and includes more highly specialized work groups and shared infrastructure.

The National Research Council reports underscore science elites' particular vision of advances in the scientific enterprise: that the intellectual boundaries of the disciplines must become more permeable in order to advance science and expand our knowledge base. Coupled with these assumptions about the communication relation is also a premise about scientific productivity: interdisciplinary research enables scientists to address more complex problems more successfully, and thereby enhances their productivity. The collaboratory concept reflects the scientific establishment's most explicit statement to date that broad benefits derived from the application of information technology, i.e. new knowledge and productivity, require a different conception of how science is conducted. Specifically, the collaboratory requires a "more explicit partnership between scientists in general and computer scientists in particular" (National Research Council, 1993, p. 1), a different reward structure for discipline scientists who engage in its infrastructure design and implementation activities, and new interdisciplinary training (see also US Congress, 1988).

The organization of human resources to maintain and develop the collaboratory

While advances in technology certainly allow scientists to probe more complex problems more deeply, these advances also create a demand for greater human resources (US Congress, 1988). Tasks associated with design, maintenance, evolution, coordination, and assistance in the use of computer-based facilities alter the size and composition of the traditional research project approach[12]. Operating a collaboratory more closely resembles an industrial model of research: production units with substantially varied and specialized competencies, including library management, information science, scientific database design and administration, computer science, engineering, as well as the particular research specialty, research program administration, and program evaluation.

Human resources to support education of scientists and users

Scientists do not come with a toolkit of technical competencies or expertise into which they reach to build and use collaboratories that will make possible effective and efficient data collection and analysis, communication and collaboration, and information storage and retrieval. The panel reports identify education as an important element of collaboratory design and functioning (National Academy of Sciences *et al.*, 1989; National Research Council, 1993). Two types of educational programs are needed. The first is designed to educate and train the people who will build the collaboratory. The second program of education is designed to educate and train the scientists who will use the collaboratory for their research projects.

Part two: *SIPP* ACCESS, an information system for complex data for the 1984 *Survey of Income and Program Participation*

Two events coincided during the 1980s that made the development of SIPP ACCESS possible. The first was the initiation of a new national longitudinal panel survey conducted by the US Bureau of the Census, the *Survey of Income and Program Participation (SIPP)*, and the second was a recognition that longitudinal panel surveys had become intractable and very difficult to manage. This section of the article:

- Discusses the mission, goals and objectives of *SIPP* ACCESS infrastructure development.
- Identifies the outcomes of an information system for complex data (ISCD) that were anticipated by the project.
- Describes the information infrastructure that was created to organize statistical data for dynamic analysis and to share the accumulated formal and informal knowledge about complex data by an electronic community of economists, sociologists and demographers.

Next we selectively describe the human resources applied to designing and maintaining *SIPP* ACCESS and the activities related to supporting the education of *SIPP* ACCESS users and other social scientists. This section concludes with an evaluation of selected aspects of one implementation objective: achieving gains in scientific productivity through a central shared data facility.

Introduction

Following intensive research, planning and development during the 1970s and a delay due to severe governmental budget reductions in the early 1980s, the US Bureau of the Census began fielding a major new government statistical series in the fall of 1983, the *Survey of Income and Program Participation*. The *SIPP* was designed as a series of longitudinal panel surveys to study income distribution, economic well being, and eligibility for and participation in government social welfare programs (see Citro and Kalton, 1993; Ryscavage, 1987). Great enthusiasm for the survey – accompanied by a certain amount of hyperbole – was manifested by the statistical and social science research and policy communities.

Nevertheless, immediately from the survey's inception, throughout the decade of the 1980s and into the early 1990s, *SIPP* was subjected to a barrage of critical public evaluation by social science researchers, policy analysts and the federal statistical community itself (see e.g. Aaron, 1985; Citro & Kalton, 1993; David, 1983, 1985; Doyle and Dalrymple, 1988; Farley and Neidert, 1989; Flory *et al.*, 1989; Jabine, 1990; King *et al.*, 1988; Marquis and Moore, 1990; National Research Council, 1989). The criticism was on a broad scale; topics examined included whether its original objectives had been met, its usefulness as a vehicle for quickly responding to new public policy issues, quality of the *SIPP* data, the Bureau's design and management of its data processing system and data distribution procedures, the complex structure and usability of the Bureau's public data use products, timeliness of public data distribution, and responsiveness of the federal agency to the survey's many stakeholders and user communities.

Concurrently, increasing national concern was articulated, particularly by social science elites and federal funding agency program managers, that a substantial national investment had been made in longitudinal panel surveys whose use had been significantly less than anticipated. The Bureau's serious problems with designing and managing a modern data processing system for the *SIPP*, the extensive delays in obtaining the data files experienced by analysts both inside and outside the Bureau, and the great difficulties that analysts had in understanding and processing the data prior to performing statistical analysis were also problems experienced by everyone who had used large longitudinal panel surveys (e.g. the *Panel Study of Income Dynamics (PSID)*, the *National Longitudinal Survey of Labor Market Experience (NLS)*, the *National Survey of the High School Class of 1972*, and the *Retirement History Survey*).

It was also clear to some social scientists that the complexity of these data and the conditions under which most social scientists operated would continue to preclude use of longitudinal panel data. First, the intellectual and capital investment required for exploiting these data had been lacking. Second, the size, scope and complexity of these data were significant impediments to timely and efficient access and retrieval. Third, appropriate technologies for efficient and low-cost data reorganization and retrieval, communication of scientific information, and exchange of data were largely unavailable to the social research community. And, fourth, the existing social science infrastructure for conducting research and policy analysis was not designed to respond optimally to the dynamic environment of data production, distribution and utilization.

Goals and objectives of *SIPP* ACCESS infrastructure development

In late 1984, we proposed a prototype of an information infrastructure to stimulate the production and sharing of knowledge about a complex dataset by social scientists in a research network to the National Science Foundation[13]. Weaknesses in the models for distributing statistical data led us to conceptualize an integrated information system that would enhance scientific productivity and efficiency. *SIPP* ACCESS, an information system for complex data, was designed as a model for timely and efficient access to and retrieval of complex data, and as a solution for communicating about data by mobilizing computer and telecommunications technology and creating a scientific infrastructure in an electronic network environment. The model was consciously designed to be generalized and applied by members of the social research community to complex datasets other than the two datasets that were ultimately incorporated into the information system. Four implementation goals were identified:

(1) efficient access to data;
(2) enhanced scientific communication and feedback among data producer, project staff and analysts;
(3) diffusion of information about and adoption of new technologies for analyzing complex data; and

(4) application of newly available software, computer and communication technologies to large-scale, complex data.

Anticipated outcomes included:

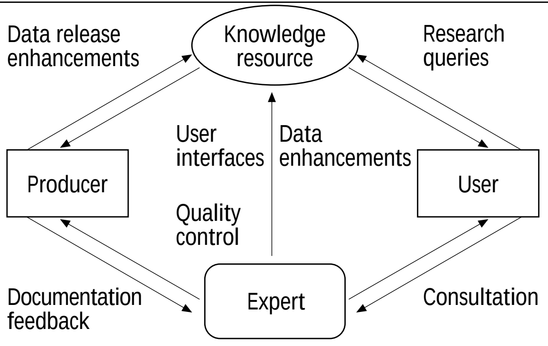
- improvements in data quality;
- a new generation of social researchers trained to employ new technologies for managing large-scale, complex data;
- increased scientific productivity because the cost of access would be reduced, the scientific design of the data would be clarified, and research results would be more rapidly produced; and
- new knowledge about social-science infrastructure development through our own research and experimentation.

An electronic research network to stimulate scientific communication, cooperation, and data sharing for an electronic community of social researchers and policy analysts in government, universities and the private sector was created in April 1985, and over the next several years more than 200 people participated. The critical relationship between the data producer and user community via *SIPP* ACCESS was represented by the Bureau of the Census's membership in the research network.

Design of an information system for complex data

The ISCD's infrastructure was designed to integrate information, statistical data, computers, software, and communication networks. Discoveries made by analysts, project staff and data producers about the statistical data became part of the knowledge resource of the ISCD. Figure 1 shows: the knowledge production system and relationships between and among the knowledge resources (data and information about data); the data producer responsible for producing and updating the data and preparing the documentation; the project staff who served as the expert intermediary between the users, the data producer, and the knowledge resource, and who were responsible for maintaining quality control, developing user interfaces, and enhancing the data; and the user/analyst who might be a sociologist, economist or demographer. The information system incorporated: capabilities of data and information management, search and retrieval; library and archival storage for data and information; and communications of news related to results and problem identification and solving.

Figure 1 Knowledge resources and knowledge-producing components of the ISCD



Key: ⟶ = High speed and other communication links

Two longitudinal panel surveys, the I*ncome Survey Development Program (ISDP)* and the 1984 *Survey of Income and Program Participation (SIPP)*, produced by the US Bureau of the Census were selected as test datasets[14]. The *SIPP* continues to be viewed as perhaps the most complicated set of survey data ever created. The first release of the complete panel of the 1984 *SIPP* contained about seven million observations and more than 20,000 variables (attributes) that represent 36 months of data for sample members. Later reorganization of the dataset, which also included additional variables by the Bureau of the Census, increased the size of the dataset by about a third. Multiple versions of the data files were released, reflecting errors that were identified and corrected. (Although subsequent annually-fielded panels were smaller in size, the aggregated quantity of data reached well over 12 GB of data in less than four years.) In size, of course, the 1984 *SIPP* dataset is far, far smaller than the terabytes of data routinely collected by space science projects, but it is structurally more complex because, in part, of the heterogeneity of and variations in measurements, analysis units, aggregations, and the multiple concepts of time embedded in the data.

Technical infrastructure

SIPP ACCESS operated in a "distributed" computing environment which linked a variety of technologies. These included electronic networks, mini- and microcomputers, optical storage devices, and commercially-developed minicomputer and microcomputer relational database management systems. The system included the primary node of the central data facility located at the University of Wisconsin-Madison Physical Sciences Laboratory (PSL) and secondary nodes where members of the research network of analysts were located. The primary node maintained the archival data on laser disk, a relational database management system (RDBMS) with the complete and sample databases, statistical packages, applications programs designed to make complex operations transparent to the end-users, libraries of reference documents and data documentation, menu-driven help files that reported new developments in the database and user discoveries about the data, and electronic mail for communicating with the project staff. The primary and secondary nodes were linked via the Internet for electronic communication and file transfer. The secondary nodes cloned the database management capabilities of the principal node with a microcomputer version of the relational database management system. The RDBMS contained four relational databases:

(1) *ISD*PRUN (complete panel data),
(2) *ISD*PTEST (2 percent sample of complete panel data),
(3) *SIPP*RUN (complete panel data), and
(4) *SIPP*TEST (2 percent sample of complete panel data).

These data are further explained below.

Information storage and retrieval: integration of metadata and statistical data

The information system integrated statistical data and metadata (information about the data). The metadata included information on the database design, logical structure, and contents (e.g. characteristics of the variables);

- survey design;
- field operations and processing procedures;
- reference materials; and
- the results of data analysis.

It provided the linkage between attributes in the database to the variables in the original public use files and helped users locate relevant variables for cross-sectional or longitudinal analysis. Data dictionary utilities, accessible from inside the database, provided details on the meaning of attributes and tables.

Data management

Applications programming was devoted to developing a relational database and user utilities to facilitate access to and retrieval of information and data. That is, the project relied on off-the-shelf tools, and the project did not allocate resources to design and develop technology. Instead, staff resources were spent designing the intellectual structures required to access complex data.

A database schema was developed to exploit the semantic principles implicit in the data collection process (for an extended discussion, see David and Robbin, 1990, 1992). The cross-sectional *SIPP* data were reorganized and transformed into time-series, longitudinal summaries, events, and new units of analysis in response to analysis needs. Relational principles also led to large-scale reductions in data storage. New data structures were designed to facilitate an understanding of the complexity of longitudinal panel surveys, reduce conceptual errors, and obtain large reductions in the amount of time required to prepare data for longitudinal panel analysis. Summary data were developed to facilitate dynamic analysis of panel data.

We designed and prototyped a sample (test) database called *SIPP*TEST, a 2 percent representative sample of sample units in wave one (first interview), including all subunits (e.g. households, families, and persons) and attributes associated with these units from waves one through nine extracted from the complete panel dataset. Although only a 2 percent sample of a very large database, *SIPP*TEST contained more cases than most national sample surveys. *SIPP*TEST was also used to prototype the design of a relational database for the complete panel database *SIPP*RUN, whose logical structure was identical to that of *SIPP*TEST. The only difference between *SIPP*TEST and *SIPP*RUN, was that *SIPP*TEST contained the metadata and a *sample* of statistical data, whereas *SIPP*RUN contained only the complete panel statistical tables. The *SIPP*TEST concept was later extended to a microcomputer version that was distributed internationally, in order to reduce reliance on remote access to the central data facility and to remove institutional constraints on using a "foreign" or remote computer.

Data storage and retrieval
*SIPP*TEST resided online 24 hours a day. In contrast, the large complete panel database *SIPP*RUN was maintained offline, located on an optical archive store (OAS) laser disk. Analysts independently retrieved data from the OAS into the database.

Communication
The *SIPP* ACCESS information system provided a variety of informational tools about the data. These included:

- interactive, menu-driven help utilities;
- bulletin boards signaling new additions to the databases and errors discovered in the data by project staff and users; and
- libraries of work in progress, reports, publications, and reference materials prepared by the data producers, analysts and *SIPP* ACCESS project staff.

The electronic network provided the principal vehicle for analysts, project staff, data producer, and technical staffs from multiple sites to communicate with one another. The online project staff consultant (*SIPP*ASSIST) assisted in problem solving, and the online liaison with the Bureau of the Census served as an invaluable source for understanding *SIPP* and tracking down answers inside the Bureau of the Census. (See Robbin (1992) for a description of the communication flows that took place.)

Human resources to design, maintain and develop *SIPP* ACCESS

The principal staff of project directors, post-doctoral level, doctoral and pre-doctoral students; and the database administrator were what computer scientists call "domain specialists". The core staff consisted of social scientists with specialties in income and wealth distribution, demography, geography, information science and organizational behavior. Our primary scientific training and experiences as social scientists pertained to the substance of the data managed in the ISCD. The social scientists were supplemented by programming, clerical and library management staff.

The project staff were responsible for designing and implementing the information system and identifying user needs for the *ISDP* and *SIPP* data in a relational database management system environment. The conceptual issues about data related principally to their longitudinal nature and representation in a RDBMS (e.g. how to enhance the data to assure the linkage and integrity of longitudinally relevant samples over time). Database implementation issues related to the large number of observations and variables in the surveys that needed to be processed (e.g. how to process quantities of data on an *ad hoc* basis and at the same time maintain the integrity of the database). The *SIPP*TEST database was used by the project staff to learn about the *ISDP* and *SIPP* and the RDBMS environment, test queries, develop new table structures, debug command files, and so on. For example, developmental work was carried out to identify missing data and to design efficient procedures for identifying and linking individuals across waves and for identifying the set of persons relevant for any type of longitudinal analysis. Selected data were reformatted as event tables and as time series to enhance the longitudinality of the data. We developed a procedure for converting measurement (survey) time to real (calendar) time because they are not identical in the *ISDP* and *SIPP* (owing to the structure of the questionnaire and the staggered interviewing of four subsamples).

Our conceptual framework underlying *SIPP* ACCESS identified the analyst as the central focus of information system and database design. These needs included identifying appropriate hardware and software for the user community, creating retrieval tools and a networking capability, handling production requests for the complete sample after users had completed their experimentation on the test database, and accessing the PSL and retrieving *SIPP* data from remote sites. We prepared tools which facilitated system access, information and data retrieval and communication between the project staff and users and among users, and which also reduced the amount of detail users were required to know in order to understand the operating system and RDBMS. *SIPP*TEST was designed as an inexpensive tool to learn about the *SIPP* and RDBMS. Menu-driven help was the available. Online consulting was provided through electronic mail.

Experimentation with technologies also took place and later led to a variety of tools to make data transfer and retrieval simple and transparent to users. An automated recordkeeping system was developed to monitor use of the *SIPP* ACCESS facility, databases and computational resource expenditures. Electronic communications (electronic mail) between users and the project staff were archived for future analysis (see Robbin, 1992). Audit trails were designed to serve as a permanent record of the decisions made during the design phase. In 1988, our evaluation of the design, implementation, use and cost of the *SIPP* relational database and *SIPP* ACCESS led to a redesigned logical structure for the *SIPP* database and the development of a RDBMS microcomputer version of the *SIPP*TEST database. This objective extended our experimentation to additional technology and to newly-available applications software. PC-*SIPP*TEST was released to the public after beta testing and extensive preparation and revision of documentation. The design requirements for a 1985 *SIPP* panel database were also analyzed; our primary interest was testing whether metadata developed for the 1984 panel could "drive" the loading of a new panel database – a fundamental modification of the procedures for developing a database. We also devised a strategy for permanent archiving of the history of the project in a form which would be available for public review. The *SIPP* ACCESS project was formally completed in December 1991, although use of the *SIPP* databases continues.

Human resources to support education of *SIPP* ACCESS users and other social scientists

The sample database *SIPP*TEST was designed with multiple learning objectives in mind, such as:

- clarifying fuzzy research questions;
- understanding the scientific design of the *SIPP*;
- exploring the metadata and data; and
- testing understandings and hypotheses about the policy problem under investigation *before* analysts prepared extracts from the *SIPP*RUN complete panel database.

The sample database provided rapid interaction with the data at low cost and reduced the future costs of using the complete panel database *SIPP*RUN.

Diffusing information about the *SIPP* and *SIPP* ACCESS and encouraging adoption of the new relational database technology were implemented through a teaching program and reference and consulting services for the *SIPP* and the *SIPP* ACCESS project. We recognized that, if researchers and policy analysts were to take advantage of *SIPP* ACCESS, we needed to teach the intricate design of the *SIPP*, the skills of remote computing and communication, and the special theory of relational data management. Three- and four-day workshops to communicate knowledge about the *SIPP* and the new technologies of the *SIPP* ACCESS information system were held in Madison and at other sites throughout the USA between 1985 and 1991. We also served as a national resource for researchers who planned to use the *SIPP* data or wanted information about the RDBMS and database design, and also made a concerted effort to publicize the *SIPP* ACCESS data facility through participation in national and international conferences and workshops.

Achieving gains in efficiency, scientific productivity and collaboration through an ISCD

The first implementation objective of *SIPP* ACCESS was to achieve gains in scientific productivity by improving access to large-scale, complex data. While measures of gains in scientific productivity are elusive (see US Congress, 1986)[15], our evaluation focusses on performance, that is, scientific activity, and outputs that resulted from the *SIPP* ACCESS data facility. These outputs are multi-dimensional and reflect direct and indirect products of the scientific research process, which is stimulated by data facilities like *SIPP* ACCESS[16]. This section discusses three tangible outputs that are evidence of *SIPP* ACCESS performance: data products, client use of the facility, and intellectual networks. This section evaluates the complete panel *SIPP*RUN relational database in terms of the performance dimension of "feasibility", describes results from some of the intensive monitoring of the information system that we carried out, and examines the intellectual network of scholars created by *SIPP* ACCESS.

Data products of SIPP *ACCESS: complete panel relational database* SIPP*RUN*

The data products of the *Survey of Income and Program Participation* were evaluated according to four performance dimensions: feasibility, portability, user demand and cost; however, this discussion addresses only the dimension of feasibility for the complete panel relational database *SIPP*RUN. Data quality and enhancements that occurred through *SIPP* ACCESS are then briefly examined.

Feasibility was strongly questioned at the outset of the project; some analysts argued that RDBMS were inappropriate for large datasets (Doyle, 1989; Doyle *et al.*, 1987). The production of *SIPP*RUN as a relational database demonstrated that it was feasible to use RDBMS technology for longitudinal panel data (see also Wells, 1991). The relational model yielded two important efficiencies: large-scale reduction in data storage and the ability to retrieve small subsets of data on an as-needed basis. Normalization reduced the more than 2.2GB of the original public use data files by 75 percent. The logical structure of the database partitioned the vast quantity of statistical data into a few categories which organized the variables (attributes) by their semantic meaning in individual tables; thus, only a small subset of the entire dataset had to be searched and retrieved in a database query. This reorganization also reflects the way in which analysts work: analysis is typically carried out on a subset of the population and a very small subset of the very large number of variables. Reorganizing the database to include time aggregations resulted in a net reduction in the volume of data requested by each analyst. These enhancements greatly reduced the computation required to estimate dynamic models using the data, as well as the chance for analysts' errors arising from the ambiguous statements about relevant universes that appear in official Bureau of the Census documentation.

Table I illustrates the relationship between retrievals of data and efficiencies gained in reorganizing the original public use files into a relational database. Columns (1)-(4) describe two different kinds of information on income that are collected by the *SIPP*: income and amounts from social programs (including food stamps, social security, veterans, and welfare for head of the program unit and its members), and assets income type (labeled B through F) and amounts for individuals and couples. Columns (5) and (6) describe the number of retrievals made on these data and the relationship between retrievals and number of tables which contained the data.

The original public data files were distributed by the Bureau of the Census in two formats: one called "RECtangular" and the second called "RELational" (a misnomer because the file structure was basically hierarchical). The *SIPP* ACCESS project obtained the files in a RELational format, which contained all program income on one record (labeled the G1 record by the Census Bureau) and all assets income on another record (labeled the G2 record). During Phase One of database development, we loaded all the program receipt and income data for heads and membership into nine tables (one table per interview) and all the asset income for individuals and couples into nine tables (each table represented one interview, of which there were nine during the life of the survey) – just as these records were distributed by the Census Bureau. Column (1) shows the size of the files as distributed by the Census Bureau in phase one in the RECtangular format ($N \cong 400{,}000$) and in the RELational format ($N \cong 161{,}000$). Column (1) then shows the result of restructuring the data in phase two. The significant difference between phase one and phase two databases is that the restructured phase two database contained information only for sample relevant units of analysis (e.g. of the total sample of more than 161,000 persons, there were only 12,090 food stamp units and 2,577 mothers receiving WIC assistance). Column (2) records the number of tables that were created for each unit. Column (3) indicates the number of variables associated with the unit of analysis (e.g. 26 variables for 12,090 heads of a food stamp unit; 15 variables for 2,577 WIC recipients). Reading across to column (4) shows that the phase one program data in the RECtangular format occupied about 98.40Mb and the RELational public use file occupied about 39.61Mb. (The significant difference between the two phase one structures is explained by the fact that the

Table I Density of access for selected phase one and phase two tables

	Sample size[b]	**Number of tables**	**Number of variables**	**Number of Mb[c]**	**Number of retrievals**	**Density (col 5/col 2)**
Table type[a]	**1**	**2**	**3**	**4**	**5**	**6**
Program receipt income (G1 record)						
Phase one	Total sample:					
	N = 400,000 (REC)			98.40		
	N = 161,000 (REL)	9	999	39.61	46	5.1
Phase two:	Sample relevant	11	232	10.72	61	5.5
Food stamps (head unit)	12,090	1	26	0.58	5	5.0
Food stamps (members)	2,808	1	14	0.08	4	4.0
Social security (unit head)	57,356	1	57	5.60	6	6.0
Social security (members)	887	1	9	0.02	2	2.0
Veterans (head unit)	5,824	1	26	0.34	7	7.0
Veterans (membership)	3,825	1	13	0.11	4	4.0
Welfare (head unit)	7,696	1	25	0.44	6	6.0
Welfare (members)	3,896	1	13	0.11	3	3.0
Other welfare (head unit)	60,546	1	19	3.35	10	10.0
Women, infants and children (WIC) (head unit)	2,577	1	15	0.06	4	4.0
Social security (couples)	877	1	15	0.03	10	10.0
Asset receipt/income (G2 record)						
Phase one	Total sample:					
	N = 400,000 (REC)			192.40		
	N = 222,000 (REL)	18	1,926	106.78	47	2.5
Phase two:	Sample relevant:	10	119	16.15	40	4.0
Part B (individuals)	192,421	1	16	9.27	7	7.0
Part C (individuals)	19,326	1	16	0.90	7	7.0
Part D (individuals)	36,182	1	17	2.15	7	7.0
Part E (individuals)	13,279	1	20	0.90	7	7.0
Part F (individuals)	13,279	1	17	0.86	7	7.0
Part B (couples)	55,277	1	5	0.91	1	1.0
Part C (couples)	4,947	1	5	0.90	1	1.0
Part D (couples)	7,139	1	7	0.08	1	1.0
Part E (couples)	6,664	1	7	0.15	1	1.0
Part F (couples)	2,051	1	5	0.03	1	1.0

Notes:

[a] The database was developed in two stages (phases), with the phase one logical structure most closely resembling the original public use data files. Phase two represents the results of semantic analysis and normalization. The Bureau of the Census adopted a naming convention for these two record types: the "G1 record" represented program income receipt and amounts and the "G2 record", assets income receipt and amounts

[b] Two sample sizes are calculated (rectangular format, relational format). We use *N* = 400,000 as an estimate of total number of sample persons less attrition and noninterviews that appear in the rectangular (REC) format public use files. We use *N* = 20,000 as an estimate of sample persons less attrition and noninterviews for the relational (REL) format. Calculations take account of reduced sample sizes in relevant waves. The much smaller sample size of the relational (REL) format reflects the fact that the public use data file was organized by record types (G1, G2)

[c] Megabytes for phase one data were calculated by multiplying the number of observations by the number of characters in the G1 (*N* = 246) or G2 (*N* = 481) record

program income and assets income items in the questionnaire are relevant to only a very small part of the entire sample because most people do not receive program support and most people have little assets income. Most of the space in the RECtangular files was filled with "not-in-universe" (NIU) values.) Column 4 also shows the result of the phase two normalization for program income: size decreases from 39.61-10.72Mb. Column (5) shows the number of recorded retrievals by researchers in phases one and two. For example, researchers made a total of 46 retrievals to the G1 record in phase one, and, after restructuring in phase two, made a total of 61 retrievals. But the real significance is that researchers were actually retrieving data for a *particular* program type (e.g. food stamps, Aid to Families of Dependant Children (AFDC), etc.). The ratio of retrievals to table increases (the term "density"), as shown in column (6).

The lower half of Table I also shows similar differences for the asset receipt and income data. In phase one the original public use data occupied 18 database tables. Both the rectangular and relational public use files had large amounts of NIUs because the sequentially organized public use files had to reserve space for a particular asset income question, whether or not that question was relevant to a person. We reorganized the asset income data according to asset type and significantly reduced the size of the tables: from 192.40Mb for the RECtangular and 106.78Mb for the RELational formats to 16.15Mb for the restructured data tables.

In other words, restructuring had many benefits because only a very small amount of data had to be retrieved to study a very small subset of the welfare or assets-holding population. The effect of reorganizing (normalizing) the original data was a significant reduction in the size of a table and a significant decrease in physical storage size of the database. Table I supports the contention that different parts of the data were retrieved differentially, from which we infer that only subsets of the data were of interest to the analysts (the analyses described in the publications produced by *SIPP* ACCESS users confirm this fact as well).

Users of the *SIPP* databases were largely unaware of the role that *SIPP* ACCESS played in establishing the quality of the *SIPP* data. At numerous points in the history of the 1984 *SIPP* Panel, *SIPP*RUN generated different statistics than control totals, and it was later determined that the Bureau of the Census public use files were in error. *SIPP*RUN also uncovered inconsistencies among counts; that is, totals generated from different parts of the scientific design were not logically possible. Feedback to the Bureau of the Census resulted in revisions to the public use datasets in most cases. Subsequent updating of *SIPP*RUN (and *SIPP*TEST) provided immediate correction to research underway. Researchers were informed of changes in the database in bulletins posted at the time that researchers logged on the system. The results of monitoring quality were cumulative and embedded directly in the *SIPP*RUN database. Information about all known corrections and known inconsistencies was online, available to everyone who logged into *SIPP* ACCESS. *SIPP* ACCESS provided a clearinghouse for purging bad data from research use. This level of monitoring and public cumulation of information did not occur in future *SIPP* databases or public use files released by the Bureau of the Census.

Getting results: how was SIPP *ACCESS used?*

The *SIPP* ACCESS information system was extensively monitored, including system logons, accesses to the databases, and retrievals of data from optical storage. The cost of using *SIPP* ACCESS was monitored through the central node's accounting system. Computer mail directed to *SIPP* ACCESS was archived. In this part of the article use of the *SIPP* ACCESS system is described, with primary emphasis on logons to the information system, accesses to the *SIPP*TEST and *SIPP*RUN databases, and retrievals of data from the optical archive. Table II provides an overview of how the information system was used.

Over the life of *SIPP* ACCESS there were 203 individuals who were authorized to use SIPP ACCESS at the central node (the PSL) (row (A) of Table II). These were project holders whom we assigned a unique user name. Sixty analysts (user names) were associated with the 41 accounts established at the PSL to use *SIPP* ACCESS after workshop training (row A′). Typical users would first learn how to use *SIPP* ACCESS through *SIPP* ACCESS-funded workshop accounts. They would then establish

Table II Measures of *SIPP* ACCESS activity and their interrelationships

Dimension	Measure of use (1)	Project user counts (2)	Density of use (3)	Intensity per access (1) / (2) Ratio	Intensity per access Definition (4)
Clients					
(A) PSL project holders[a]		203			
(A′) With own account[b]	60				
PSL system access[c]					
(B) System logons	19,650	163	121		
Database access					
(C) All (PSL, IRP)[d]	20,834	151(248)	138		
(D) *SIPP*TEST	6,379	105(193)	61	0.31	D1/C1
(E) *SIPP*RUN	14,455	46 (55)	314	0.69	D1/C1
Optical storage access (PSL, IRP)					
(F) OAS accesses	2,743	45	61	0.19	F1/E1
Communication					
(G) E-mail (client, data producer, *SIPP* ACCESS staff)	1,646	121	13.6		

Notes:

[a] Includes everyone who ever established a computer account at the central node (PSL) – 131 were workshop participants

[b] Includes analysts who established a PSL computer account independently of having attended a training workshop

[c] System logons include only those clients who used *SIPP* ACCESS through PSL. Use of PSL by Institute for Research on Poverty (IRP) analysts would be reflected in these system logon counts if an IRP analyst attended a training workshop or bought time on the PSL computer when the IRP computer became clogged. Analysts could have project accounts associated with both PSL and IRP between 1986 and 1990, depending on the time period they were associated with IRP

[d] All *SIPP*RUN users are *SIPP*TEST users. The *SIPP*TEST counts exclude 88 persons (workshop attendees and others who used the *SIPP*TEST database before April 1987 and *SIPP* ACCESS staff for the entire project). We have recorded *SIPP*RUN accesses for 46 individuals, but a total of 55 analysts, including *SIPP* ACCESS staff, used the *SIPP*RUN database; density is calculated on a base of 46. Counts indicated in parentheses reflect total number of *SIPP*RUN users

their own project accounts. A particular research investigation on a project account could span several funding sources or be a small component of a single project account. Accesses to the *SIPP* ACCESS system are a clearer measure of use. Row (B) in Table II shows that the 163 users generated an average number of 121 accesses to the system. (Note that, although 203 people were authorized users, logons were recorded for only 80 percent of them.)

We obtained a more precise measure of *SIPP* ACCESS use by monitoring access to the two databases *SIPP*TEST and *SIPP*RUN. Database accesses is a minimal measure of the role of *SIPP* ACCESS, since much statistical analysis based on the system derived from, but did not make use of, the databases because it was carried out at the analyst's site. The number of accesses to the databases exceeds the number of logons. This result may seem paradoxical. The explanation lies in the fact that a database could be accessed several times during one system access; that is, for example, during one session at the PSL, queries were first developed and tested in *SIPP*TEST database, after which a production job was submitted using the *SIPP*-RUN database.

Table II shows that only half of the unique users ever logged on to the databases after we began recording database accesses. The primary reason is that 131 of the 203 *SIPP* ACCESS authorized users were associated with workshop activity, but a total of only 60 people were associated with user-funded project accounts. Some workshop participants never accessed the databases (they worked in groups and delegated data entry to someone in the group). In addition, some principal investigators were authorized *SIPP* ACCESS users but assigned all database work to research assistants. The average number of database retrievals per database user is 138. We will refer to the ratio of retrievals to use as a measure of *density* (e.g. 151 project users made a total of 20,834 retrievals from the databases for a density of use equal to 138). This density is higher than the average number of system logons per user (a density equal to 121), because the number of database users is smaller than the number of system users. More than two-thirds of database usage represented calls to *SIPP*RUN. Row (F) of Table II shows access to the OAS archival storage where the *SIPP*RUN tables were stored as files. As might be expected, the density of OAS use is lower than *SIPP*RUN use: multiple tables were retrieved once from the archival store, but, once loaded in the database, researchers made multiple retrievals from these tables. This fact is quantified in the intensity column – column (4) – as the ratio 0.19.

Monitoring retrievals from the OAS was helpful in deciding which portions of the original public use data files warranted reorganizing to improve the efficiency of user queries. Our monitoring also confirmed that analysts were principally carrying out longitudinal analysis of the *SIPP* data and that only a small subset of the total database was frequently retrieved by users. Many tables in the *SIPP* database were seldom addressed; thus the relational organization segregated data that were not needed and provided easy access to data of research interest[17].

Electronic mail generated 1,646 messages to or from the *SIPP* ACCESS staff. That count includes messages with substantive content pertaining to the *SIPP*, the databases, use of *SIPP* ACCESS information services, and research strategy. Those messages were analyzed by Robbin (1992). Their existence is reported in Table II to underscore the fact that monitoring activity required monitoring communication as well as the use of databases.

Client output and scientific activity

The design of *SIPP* ACCESS as a research network implied the fostering or reinforcing of an intellectual network of scholars. Table III examines the research output in terms of publication authorship and the institutions that participated in research through 1990 using the *SIPP* ACCESS ISCD. The total number of publications in this table includes *SIPP* ACCESS project staff papers if they resulted from secondary analysis of the data; we have removed 22 staff papers that deal largely with the design of *SIPP* ACCESS, resulting in a base of 147 papers. We coded private, for-profit and not-for-profit institutions as “private”; disaggregating the private institutions would give us a total of 22 different institutions that used *SIPP* ACCESS, rather than the 19 that we show in the table.

We derive several insights from this table. More than 50 percent of the research output was produced by individual authors; 29 percent were jointly authored at the same institution; and 14 percent jointly authored at different institutions. Four institutions, University of Wisconsin-Madison, Brown University, Bowdoin College, and Brandeis University, account for nearly three-quarters of the total research output during the first six years of *SIPP* ACCESS. As might be expected, because of the location of the Institute for Research on Poverty and a long history of poverty research, the University of Wisconsin-Madison produces slightly more than one third of the total number of papers. Researchers at nine of the 22 institutions had intellectual ties to the Institute, either as graduate students or affiliates who participated in poverty research through grants or contracts.

We conclude that the *SIPP* ACCESS facility reinforced an existing national network of

Table III Research output and intellectual networks

Institution[a]	Sole author[b]	Joint author[c]	Joint author outside author's own institution[c]	Number of publications
University of Wisconsin-Madison	38	10	5	55
Brown	3	16	5	24
Bowdoin	15	1		16
Brandeis	7		5	12
Sub total	63	27	15	107
Percentage grand total				73
Private[d]	3	4		7
Maryland	5		1	6
Columbia	1	2	1	4
Indiana	2	2		4
University of North Carolina	2	2		4
Duke	1	2		3
Government/agency	2	1		3
Cornell		2		2
Central Florida	2			2
University of California, Davis	2			2
Northwestern	1			1
Oregon State	1			1
South Adelaide			1	1
Vassar	1			1
University of California at Los Angeles		1		1
Totals	86	43	18	147[e]
Percent	57	29	14	100

Notes:

a Institution is where author was first located if publication is produced within one year after first publication; if publication is dated two or more years after date at first location, the author is coded as located at the second institution. This situation occurred most frequently with graduate students who completed their work and took positions elsewhere. Authors on temporary leave from their home institutions are coded as members of their home institutions even if they produced a publication while on leave

b Production of a publication took place at the sole or joint authors' home institution

c Publication is coded as "joint authorship with author at another institution" if second or third author was formerly a member of another institution

d Includes four institutions

e The total numbers of papers produced by users of *SIPP* ACCESS is 169. We include papers produced by *SIPP* ACCESS project staff in this table if the paper resulted from secondary analysis of the *SIPP* data but exclude 22 staff papers that dealt largely with the design of the information system poverty researchers. At the same time, however, Table III also shows that the *SIPP* ACCESS network succeeded in enlarging this small network of poverty researchers by providing a resource to a larger group of institutions, none of which had previous ties to the University of Wisconsin-Madison. (See also Hesse *et al*. (1993) for evidence of the importance of the electronic network for researchers at the "periphery" of scientific activity.)

Part three: *SIPP* ACCESS: a model of a collaboratory for the social sciences? Obstacles to collaboratory development

Did *SIPP* ACCESS, an ISCD, meet the operational definitions and criteria for a collaboratory for scientific research? The answer is affirmative for the conceptual framework and model that led to the design of the information system, but the complete range of knowledge and functionality in an all-inclusive network system discussed in the first section of this article and advocated by Schatz (1992) was not achieved. The latter was, in part, a function of timing, of a historical moment – the limited availability of appropriate telecommunications, computational and applications software, and its occurence in the 1990s, an era of declining funding for social scientific infrastructure research, development and experimentation. Certainly, an adequate funding base was a critical obstacle to achieving a *SIPP* ACCESS collaboratory.

Still, even in its six years, *SIPP* ACCESS did not cause/achieve the sociological change advocated by the science elites. Computer scientists were not directly involved in the project (one served on the National Board, however), and an electronic community model along the lines advocated by Schatz was not explicitly adopted as an organizing principle and institutionalized for the conduct of research by others in the social science community[18]. Part three examines some of the conceptual/technical, cognitive, social-psychological, and sociological obstacles to developing an institutionalized collaboratory for the social sciences:

- integrating heterogeneous databases;
- cognitive, experiential knowledge, and other attributes of users;
- collaboration between computer scientists and domain specialists; and
- cumulative advantage of institutions.

What social scientists have learned from research into innovation diffusion and adoption is relevant here.

Conceptual/technical obstacles to integrating heterogeneous databases

Social scientists face problems similar to those faced by space physicists, brain scientists or oceanographers, in that they want to relate and integrate in a common information system multiple databases whose information representation and structures are different. Wiederhold explains that:

> For single databases, a primary hindrance for end-user access is the volume of data that is becoming available, the lack of abstraction, and the need to understand the representation of the data. [But] when information is combined from multiple databases, the major concern is the mismatch encountered in information representation and structures (1992, p. 38).

For example, the National Library of Medicine would like to provide social scientists with Internet access to its large and complex series of annual and biannual health surveys which have digitized X-rays associated with the medical examinations that doctors carried out and the statistical data that derive from the questionnaires administered to nationally representative samples of households. There are, of course, technical problems of data storage, transfer and management accompanying digitized X-ray data and gigabytes of historical data. But it is the creation of an integrated information system of X-rays, statistical data, and the textual materials associated with multiple designs, implementation, processing, and analysis of the surveys and other data – the sum total of raw, calibrated, validated, derived and interpreted data – that is fundamentally, conceptually difficult.

The problems associated with the task are very subtle and concern the idiosyncratic cognitive reasoning process that establishes a relationship between data and knowledge resources, which may well go beyond what can reasonably be obtained from integrated libraries. For example, what organizing structure for the information should be used? How do we represent the complexity and multi-dimensional aspects of the data? How do we specify the theoretical connections among specific areas of interest, which will certainly be idiosyncratic given the diversity of interests of social and behavioral scientists and the potential of these data to support many different research applications? It is clear that new concepts, techniques and functionalities must be developed to support the interoperability and integration of autonomous, semantically heterogeneous database systems with unique conceptual schema in order for a diverse community of scientists to operate effectively in a distributed network environment. Solutions are not immediately obvious to the complex array of severe problems of fusion and semantic inconsistency, although work is being carried out on "mediators" that embody administrative, technical and scientific domain knowledge (see Wiederhold, 1992).

Different cognitive, experiential knowledge, and other attributes of users

Effective use of the *SIPP* ACCESS information system required extensive knowledge in a large number of general and specific content domains. One set of content domains included the policy problem under investigation, survey design, statistics, programming, software and computation (knowledge acquired in formal training through course work, although social scientists rarely study programming formally and typically learn to use statistical software and the computer as part of their academic training in their disciplines). Proficiency in using data required a second set of content domains that included knowledge of data structure, data management, and the grammar of a query language for operating on the data (knowledge formally taught in computer science and management information systems courses; social scientists typically rely on programmers for this kind of expertise). A third set of knowledge was specific to the datasets, the software and computational environments in which the data were located, and the complex interrelationships of data, software and computational environment (knowledge gained through direct experience, although only a few social scientists have a detailed understanding of the computational environment and the relationships between data and software, and software and the computer).

The *SIPP* ACCESS project staff instituted an instructional program which assumed that users had prior knowledge and experience in the subject matter of the first set of knowledge domains. The program was designed to assist users in acquiring facts and procedural skills in the second and third sets of subject domains. We recognized that a workshop setting was inadequate for integrating the vast amount of background information required for using the *SIPP* ACCESS system and, instead, concentrated on communicating essential information (organized in a handbook) that would provide users with a basis for acquiring the skills necessary to learn on their own after the workshop, supplemented with assistance in problem solving provided by the online consultant *SIPP*-ASSIST. The instructional strategy was a "scaffold" to foster autonomous learning skills: to teach students (faculty, researchers, programmers, project assistants and policy analysts) what they needed to know and how to acquire and process information, that is, effective strategies and procedures for locating information and for problem solving.

Contrary to our initial expectations about the skill level of workshop attendees, however, we discovered that social scientists came to the ISCD with widely varying and considerably different analytical, methodological, statistical and technical competencies, and these deficiencies could never be overcome. But there are also other attributes that influenced whether or not researchers adopted the *SIPP* ACCESS innovation.

A cognitive capacity to process new and particularly complex information was an important factor. Complexity inhibited use by novices who had attended our workshops and was an even more severe impediment for users who had not participated in these workshops. Only the most highly sophisticated data analysts could be expected to use the *SIPP* ACCESS facility. Researchers who were successful had a repertoire of compiled knowledge that was extensive: substantial knowledge about large datasets, survey design, statistics, and policy issues, and also extensive experience in using computers and statistical and other software. For them, the information system created greater efficiencies and improved effectiveness over standard statistical and data management systems (see Baker, 1989).

Motivation to use the *SIPP* ACCESS facility outside of the controlled environment of the workshop also influenced acceptance and use of the innovation. Using *SIPP* ACCESS required that the individual be flexible and playful about adopting an innovation. Analysts who went on to use the *SIPP* ACCESS facility after a workshop approached the idea of a relational database and the learning that was required as a "new and interesting game". Using *SIPP* ACCESS and the relational database were "fun", and *SIPP* ACCESS was an opportunity to add to their storehouses of knowledge. We did not appreciate the extent to which people were resistant to altering their past behavior and how difficult it was to adopt an innovation when new mental models and new ways of working were required.

Using *SIPP* ACCESS also required a very large investment of time, and the learning curve could be steep and long. Analysts who were prepared to make the investment in learning went on to use the complete panel database after a workshop. As one graduate student, an independent learner, said, "You know that you have a lot of work to do when you start something you've never done before". Only a few participants were willing to invest time and energy to do so. It was not that the *SIPP* ACCESS system was too complicated to learn. Rather, it was a rational calculation that the new technology carried certain high costs in learning and information overload relative to the potential benefits of increased productivity, improvements in understanding the data, greater precision in data retrieval, and higher quality statistical analysis. Analysts "minimized loss" in the short-run by making the decision not to use the new technology. We did not anticipate this benefit-cost calculation, however informal, associated with learning.

Having a well-formulated research problem was also a strong incentive to use *SIPP* ACCESS. Workshop participants who had a well-defined research problem for which the *SIPP* was applicable went on to use *SIPP*RUN. A well-defined research problem meant that prior planning had already taken place, even if complete knowledge of the *SIPP* data was not yet in place. Many workshop participants came to a workshop because they had heard or read about the *SIPP*; they did not come with a specific policy problem worked out in advance for which the *SIPP* could be used, and only a very few succeeded in working out a problem in the months following a workshop.

The ability to work autonomously without a face-to-face social network appears to be important factor in using an electronic network. Although the electronic network provides significant capabilities for enhancing the analytical capabilities of the end-user, the distributed client/server system and services approach places considerable responsibility on the end-user to navigate successfully. This environment requires not only a high level of computer sophistication, but it also requires people who rely minimally on others for solutions to problems. Nearly all the successful *SIPP* ACCESS users from outside of the University of Wisconsin-Madison were like this. Their e-mail messages indicated that they relied primarily on their own resources to solve problems and called for help only after they had tried a variety of alternate procedures to solve their problems. We did not appreciate the importance of self-reliance until we observed that our Wisconsin graduate students, only two of whom had attended a workshop, learned about *SIPP* and the relational technology from one another. Every day, graduate students spent many hours together in the computer room. They taught the *SIPP* to each other; more experienced *SIPP* users helped novices solve problems; and they shared their programs and scripts with one another. They created their own support network. We concluded that autonomous individuals will perform effectively in an electronic network environment. For others, peer reinforcement to learn about and use the *SIPP* after the workshop will be an important factor that influences use of a collaboratory-type network facility. For the less autonomous individual, the electronic network environment may not successfully provide this important social-psychological component to learning. As such, an electronic network and frequent communication with project staff may not be able to substitute for the traditional modes of scientific work where geographical proximity is perceived as necessary.

Collaboration between computer scientists and social scientists

The reports of the National Academy of Sciences, National Research Council, National

Science Foundation, and Office of Technology Assessment emphasize how critical it is to bring together domain specialists, computer scientists and informatics specialists in order to "realize the information future" and respond effectively to its challenges. The *SIPP* ACCESS project directors recognized prior to the inception of the project, that development of the information system would require the collaboration of computer scientists. During the six-year period we attempted to interest both faculty members and graduate students in computer science in the project but were ultimately never successful. Most often, the problems we identified were summarily dismissed as "uninteresting". There was a certain irony in hearing this, particularly because at about the same time, but in isolation, computer scientists were also acknowledging a variety of problems that social scientists had already attended to or had written publicly about for more than a decade: locating and accessing heterogeneous scientific data, archiving and preservation, development of metadata, adequate description and documentation, development of standards for data, data citation and proprietary datasets, and structures to represent appropriately time and spatial data, (see e.g. French *et al.*, 1990; Scheuermann *et al.*, 1989; Silberschatz *et al.*, 1990). The irony is reinforced when reviewing the *National Collaboratories* (National Research Council, 1993); which completely ignored any collaboratory-enhancing work carried out by social scientists[19].

The resistance of computer scientists is a critical obstacle to implementing the recommendations of the science panels for the creation of collaboratories. It has, as we discovered from our *SIPP* ACCESS experience, its origins in the different languages used to define what constitutes important problems of the different disciplines. Similarly, we also learned that few social scientists could identify their data management and other related computational problems as being amenable to solutions which computer scientists were eminently suited to develop. During the six years we also heard regularly from social science colleagues that we were "wasting precious time that should be devoted to 'doing research'"; building infrastructure was not rewarded by the social-science community. From our experience it can also be said that we discovered few computer scientists or social scientists who were interested in attempting effective communication across the boundaries of the two disciplines. There were indeed, as Pechura and Martin (1991) observed when neuroscientists and computer scientists were brought together, "fundamental differences in the perceptual frameworks" because each discipline approaches its work differently. It is doubtful that external agents such as funding agencies, which are ostensibly committed to programs across disciplinary boundaries, will be able to alter the situation. The impermeability of specialty boundaries is great, differences in the reward structures are great, and boundary spanners will have to be highly trained in the domains of social science and computer science in order to recognize how solutions derived from different disciplines can be brought to bear on the problem at hand.

Cumulative advantage

We have already noted that *SIPP* ACCESS users were differentially advantaged with regard to knowledge and social-psychological attributes. Successful *SIPP* ACCESS users were those who had accumulated significant advantage related to the knowledge domains required for manipulating complex data in a relational database and networked environment, and they were similarly advantaged in their approaches to adopting an innovation. There is, however, another aspect to cumulative advantage that is a significant obstacle to any widespread collaboratory development in the social sciences and also in the other sciences.

Institutions also accumulate advantage and disadvantage. This cumulative advantage means that a variety of resources are institutionally supplied to the researcher so that high scientific productivity is made possible (see Allison and Long, 1990, for statistical evidence of the effect of institutional infrastructure). Merton writes that these resources include: "access to needed equipment, an abundance of able assistance, time institutionally set aside for research, and, above all else perhaps, a cognitive microenvironment composed of colleagues at the research front who are themselves evokers of excellence, bringing out the best in the people around them (1988, p. 615)." High-impact papers are produced by institutions which have extensively invested in infrastructure (see *Science*, 1994b), and it is these institutions which continue to attract, as Merton noted, “far larger resources of every kind, human and material” (1988, p. 617). (See also annual reports of the National Science Foundation that discuss the allocation of academic science and engineering research and development funds for evidence that only ten institutions account for about 25 percent of federal research funding.)

During the six years of the *SIPP* ACCESS project, we discovered that an institutional infrastructure that has the capacity to support the conduct of research was a critical factor in whether analysts could or did use the *SIPP* data or the *SIPP* ACCESS facility. Institutional policies and rules governing who had access to these resources also influenced subsequent use of *SIPP* ACCESS. Most workshop participants who logged on to PSL after the workshop and used the complete panel database *SIPP*RUN worked at an institution whose infrastructure made research on complex data possible. These were institutions with a research mission or with a physical plant that facilitated the research activity. Computational facilities were extensive, communications systems were modern, computing was either free or low cost, and personnel dedicated to supporting the research mission were available. In addition, these researchers had access to either external or internal funds for purchasing data, carrying out research, buying time on another computer, and making long distance telephone calls to access PSL (prior to Telnet access). Without this institutional infrastructure and access to either internal or extramural support in real dollars, workshop participants could not be expected to use *SIPP* ACCESS in the future.

This infrastructure could be in place, but whether the workshop participant had access to it was a critical determinant of future use of *SIPP* ACCESS. Institutional policies and rules also played an important role. For example, informal and formal policies typically provided university funds only to faculty, which excluded use by graduate students. Policies also prohibited use of external facilities when these facilities “duplicated services” at one’s own institution; improved productivity and efficiency offered by *SIPP* ACCESS were deemed irrelevant. At the beginning of the project, we did not appreciate the very large effect that an individual’s institutional environment would have on future use of *SIPP* ACCESS. Probably more than any other factor for many analysts, deficiencies in the social setting in which everyday research and teaching activities take place explain why interest in *SIPP* ACCESS did not translate into use.

The message seems clear: institutional information infrastructure development and the provision of appropriate facilities for educating a student body are so extraordinarily investment-intensive that only a few elite institutions can be expected to have the resources to do so. Can countervailing processes reduce the institutional monopoly or domination by a few institutions, to use Merton’s language (1988, p. 619)? Merton argued that institutional advantage also meant “being located at strategic nodes in the networks of scientific communication that provide ready access to information at the frontiers of research”. We have already shown that electronic networks can compensate somewhat for the lack of these institutional resources. Although there appears to be some evidence that individual departments can achieve improvements in their intellectual capacity in a relatively brief time, it does not appear that ready access to information, which in itself requires infrastructure investment, can compensate for the lack of institutional resources which Merton identified as critical for the conduct of scientific activity on the frontiers of research. As the National Research Council Computer Science and Telecommunications Board noted in its report, “significant numbers of higher-education institutions remain with limited or no Internet access” (National Research Council, 1994a, p. 121).

Finally, cumulative institutional advantage interacts with the culture of the various specialties of the social sciences and the rewards for social scientific activity by science elites. Evidence of the social-science community’s capacity to develop electronic communities-cum-collaboratories might be assessed by examining the extent to which collaborative, interinstitutional research is carried out. The National Science Foundation Division of Social, Behavioral, and Economic Research grant list for fiscal year 1994 was reviewed. While comparative data on physical scientists and the nature of their research activities are needed to assess with some degree of certainty that conditions similar to those of physical scientists who have formed electronic communities also exist for a critical mass of social scientists[20], some information about the extent of interinstitutional collaboration by social scientists is available. Table IV, displays the grant awards by program area, the total number of awards, the number of collaborative, interinstitutional awards, and their percentage of the program's total awards. Table IV shows that the social sciences are dominated by single-institution-investigator awards. Of the 1,270 awards made by the Division, only 88 or 6.9 percent of the awards were made for interinstitutional collaborative research. Of the total of 18 program areas, five made no awards for interinstitutional collaboration. Even though the economics program dominates, only 6.1 percent of its awards were made for collaborative, interinstitutional research. One program area, archaeometry, made 30.7 percent, and four other programs made between 10 and nearly 22 percent of their awards for interinstitutional research activities. Disciplines like the decision risk, management science and political sciences, where there is a greater dependence on larger and more complex datasets, have a higher number of interinstitutional collaborative research activities, whereas cultural anthropology and linguistics, which have been disciplines dominated by sole investigators conducting small-scale research, rank very low in the number of collaborative awards. Innovation adoption requires a certain critical mass of individuals before widespread adoption takes place. But whether certain specialties of the social sciences, such as the decision sciences, archaeometry, or political science, have reached that stage and are "ripe" for collaboratory development is unknown at this point.

Table IV National Science Foundation Division of Social, Behavioral, and Economic Research Grant Awards, fiscal year 1994

Programs	Total (N) awards[a]	Total (N) collaborative[b]	Percentage of total (col 2/1)
Economics	262	16	6.1
Archaelogy	87	0	0.0
Human cognition and perception	83	8	9.6
Sociology	82	6	7.3
Cultural anthropology	78	2	2.6
Geography	76	10	13.2
Decision risk & management science	75	13	17.3
Physical anthropology	74	2	2.7
Political science	69	15	21.7
Linguistics	68	2	2.9
Social psychology	66	2	3.0
Ethics and values studies	63	2	3.2
Science and technical studies	63	0	0.0
Law and social sciences	58	6	10.3
Methodology, measurement and statistics	42	0	0.0
Research on science and technology	18	0	0.0
Archaeometry	13	4	30.7
Systemic anthropological collecting	3	0	0.0
Totals	1,270	88	6.9

Notes:
a Programs rank ordered by number of awards (highest to lowest)
b Indicated in list as "collaborative research" that is interinstitutional

Source: National Science Foundation (1994c)

Conclusion

Just prior to retiring from the National Science Foundation in 1990, our program manager Murray Aborn urged us to prepare a complete document that reported the conceptual framework and implementation of all aspects of the design of the *SIPP* ACCESS information system for complex data. His view was that this prototype would be the first and only one of its kind that would be developed for the social sciences because the obstacles for institutionalizing an information system like *SIPP* ACCESS for other types of data stores would be too great to overcome. While his pessimism is acknowledged and confirmed, particularly with regard to altering the natural conservatism of the various disciplines that are needed in order to mount this type of effort, other changes have been observed in the intervening years that create optimism for the future, although these changes have largely been in the technical area.

For example, database designers have now developed nearly transparent and standard interfaces between statistical packages and RDBMSs, and the major statistical package SAS now incorporates the capability to create relational databases. Graphical user interfaces greatly enhance the user's capability to access and use data. Relational database management system designers have begun to understand the deficits of the relational model which impede use of scientific databases.

Designers of data repositories and data libraries have begun to recognize that client/server approaches require a new end-user philosophy: the database must support whatever type of interface, system or analysis the end-user desires. Managers of data repositories have automated the evaluation of data quality, the development of documentation, and the distribution of their datastores. End-users are being provided with powerful tools for automated access, processing, modification and display of data on their desktop computers. Advanced systems even provide cooperative computation services, so that the host will help the end-user's computer, working collaboratively to process data for display on the end-user's machine, in the end-user's chosen format or application.

To respond to deficits in the infrastructure of higher education, some federal agencies and private foundations are providing support for infrastructure development and enhancements, as well as providing support for integrating computer technology into the curriculum and for educating computer-literate students. There appears to be a growing recognition that literacy and national economic renewal and global competitiveness are intimately related to each other. Increasing constraints on available dollars to support the information infrastructure are leading to a recognition that, just as with library collections, public data collections can be multi-site, geographically distributed and shared through an "interlibrary loan" capability that the Internet provides naturally.

Informatics specialists can take a leadership role in thinking about how to make the collaboratory a reality, thinking hard about organizing and managing the preservation of, access to, and widespread diffusion of data and information resources. Policy recommendations made by science elites regarding K-12 and higher-educational needs related to improving science education, as well as research into the behavioral and cognitive attributes of the user community and into how information is valued and used, need to be taken seriously and implemented. We need to focus more on communication flows and social networks and the posited data-to-information-into-knowledge processes, particularly if we are to benefit from the potential of applications that enable the mass diffusion of processed information. A sea-change in thinking about, and the application of cost-efficiency to, academic infrastructure must take place by senior administrators in primary and secondary schools and in higher education. As William Wulf acknowledged, the "bottleneck to the achievement of vision [of a collaboratory] is not hardware" (Wulf, 1993).

Notes

1 The National Academy of Sciences *et al.* (1989, p. 1) defines information technology as the set of computer and telecommunications technologies that makes computation, communication, and storage and retrieval of information possible.

2 The Arpanet was established in 1969 as a computer network-supported collaboration (National Research Council, 1993, p. 9). It is viewed, for example, that the programming language COMMON LISP could never have been possible without the computer network (National Academy Sciences *et al.*, 1989, p. 20). For some of the many published discussions about collaborative work using information technology that preceded Wulf's elaboration on the collaboratory concept, see Baker and Zwicki (1984), Bush (1945), Comer (1983), Green and King (1986), Greenstadt (1981), Lederberg (1978), President's Science Advisory Committee (1963), Rees *et al.* (1986), and Steele (1984).

3 See my abbreviated critique of the failure to address information infrastructure issues in Robbin (1992). The central importance of the content of information, information resource management, as well as the community of users, is also underscored by McClure *et al.* (1991).

4 Lederberg commented that we had to acknowledge that the human dimension of computers and communications was a "task of equal priority to engineering the hardware" (1978, p. 1318).

5 For the historical record it is also important to record that as early as the 1950s, at the height of the Cold War,

geophysicists in the international community were persuaded that the future of their research agenda and the creation of new knowledge lay in formalizing a collaborative international infrastructure that linked people and instruments. (I am grateful to Marie Dvorzak, director, Geology-Geophysics Library, University of Wisconsin-Madison for this historical note on the international geophysical year.)

6 I use the word "new" in terms of the particular importance placed on cognitive capabilities and social structure as they relate to the content of what is produced by scientific activity, for which considerable research has been conducted on the scientific community by behavioral, social and information scientists, but not by physical scientists. Clearly, the concerns registered by the hard sciences about professionalization of their disciplines, careers and job training reflect attention to "human capital" issues. See, for example, the annual issue of *Science* magazine that is devoted to discussions of the future of the young scientist (e.g., *Science* 1994a). Similarly, efforts to measure the outputs of science, as part of the study of scientific productivity, reflect another dimension of interest in the study of "human capital" for which considerable research has been conducted. See, for example, US Congress (1986), Small (1990), Hesse *et al.* (1993).

7 Major advances that particularly affected the infrastructure development included information and data storage, management, retrieval and other applications software, local- and wide-area networks, distributed computing, and a variety of client/server approaches. The project staff and users would have benefitted from, for example: multiple platforms for relational database management systems (RDBMS) residing in a distributed computing environment; multiple-gigabyte CD-ROM and magnetic tape cartridge storage; powerful personal computers or workstations on which complete sample tables or the entire database itself could reside; seamless interfaces between the RDBMS and statistical packages; multimedia, including hypermedia; graphical user interfaces to facilitate user access to all components of the information system; and cooperative computational services.

8 The term "interoperability" was broadly applied by the 1993 National Research Council report and used in the context whereby, for example, experiments, data, models, graphical and tabular interpretation, textual summaries, documentation, and publications would be integrated and portable. "Transparency" was emphasized because of the physically distributed nature of the information, data, software, technologies, and human resources. The attribute of "customizability" was used to reflect the fact that users would want to be able to create "familiar" interfaces and impose restrictions on access to their own work. "Integrity" was recognized as a critical attribute because system and data security had to be maintained, but also because the quality of the information being submitted to the system had to be verified. "Extensibility" was required so that the system would continue to evolve and be able to incorporate dynamic changes in the technological environment as well as new services and features that would meet the scientists' research needs.

9 Examples of resource discovery browsing and display tools include Archie, Gopher, World Wide Web (and tools such as Mosaic, Netscape and Lynx), wide area information servers, and tools that are currently being prototyped, such as Essence (Hardy and Schwartz, 1993).

10 Chubin comments that the "key to conceptualizing the structure of specialties may be the nature of the communication relation used to link scientists" (1976, p. 451). Hagstrom suggests that members of a specialty can be identified by characteristics such as frequent and intense communication, exchange of preprints and reprints, and citation of one another's work (1970, p. 91).

11 A mathematician acquaintance once said to me, "The best scientists are those who see the relationships between fields – that they share concepts. That's how new fields develop, too". For example, the renaissance of geometry in the USA reinforces these insights (see the historical analysis of the differential geometer Osserman (1989)). The communication relation may also be motivated by some incentive provided by agents external to the specialty. External agents attempt to influence innovation adoption (i.e. interdisciplinary collaboration) by providing isolated groups with intellectual and monetary incentives for interacting with one another. Federal funding agencies, such as the National Science Foundation, and professional organizations, such as the American Association for the Advancement of Science, play a "nurturing" role and organize interdisciplinary conferences where scientists from different disciplines can talk to one another. The granting agencies use their budgets as instruments for mobilizing collaborative undertakings (see, e.g. the Directorate for Education and Human Resources at the National Science Foundation). External agents may also include private-sector firms which support R&D as a means to develop new products and services.

12 Considerable gaps in data collection about personnel make it difficult, however, to confirm that production units have actually been altered (US Congress, 1988), although anecdotal evidence suggests that technical tasks of programming and information management are handled by discipline-based scientists rather than computer scientists. The results of a study of funding support and the workforce for investigator-initiated research grants (R01) and program project grants (P01) by the US National Institutes of Health (NIH) (1993, pp. 1-2) for fiscal years 1983, 1985, 1987 and 1990 are instructive in this regard. The report states that the single most important factor in budget increases was personnel costs. Between 1983 and 1990, the average number of full-time equivalents (FTEs) per grant supported by the NIH increased from 2.99-3.33. But whereas, for example, estimated FTEs for more senior doctorates increased from 0.84 FTE to 1.08 per grant, there were decreases in the estimates for technical support (0.66 FTE per grant to 0.46) and other categories of staff (from 0.15-0.13 FTE) per grant.

13 Martin David and Alice Robbin served as co-principal investigators and co-directors of the SIPP ACCESS project, and Thomas Flory as database administrator. Portions of the second part of this article have been excerpted from our *Final Report* to the *NSF* (David and Robbin, 1992). The *SIPP* ACCESS project was supported in part by the National Science Foundation (SES-8411785, SES-8921213 and SES-8701911), the Sloan Foundation (B1984-25 and B1987-46), the Social Science Research Council, the Bureau of the Census (through the National Science Foundation), and the University of Wisconsin-Madison. Administrative support was provided by the Institute for Research on Poverty and Center for Demography and Ecology (P30 HD05875) at the University of Wisconsin-Madison. We remain indebted to our first NSF program officer Murray Aborn for his wisdom, foresight and nurturing of this project.

14 Although our principal efforts (and most of the discussion in this article about these data) were devoted to the *Survey of Income and Program Participation*, a second set of data – the precursor to the *SIPP* – the *Income Survey Development Program*, was also incorporated into the information system and was used by researchers and policy analysts.

15 The Office of Technology Assessment (OTA) (US Congress, 1988) discusses the difficulties of relating R&D expenditures to scientific productivity or other economic benefits. It notes that the principal benefits of research, new and unexpected knowledge, cannot be assigned a direct economic benefit. Scientific activity has indirect effects, such as "spillovers" and "spinoffs", which are difficult to quantify. Our discussion does, however, utilize bibliometric measures, which are a principal non-economic indicator of the "output of science", and follows the OTA's recommendation that it is necessary to attend to the process between inputs and outputs, that is, intermediate steps in the research activity. Andrews (1979) and his colleagues do, however, establish a broad set of indirect quantitative and qualitative "performance-effectiveness" indicators in their international study of scientific productivity.

16 A data facility may also produce indirect outputs in the form of scientific communication, improvements in data quality, the learning of new tools to be used for scientific investigation, and new knowledge about conceptual frameworks, methodologies, etc., that are applied elsewhere.

17 For example, during our evaluation of the phase one database structure (which resembled quite closely the structure of the original public use relational file), we noted that the heaviest retrievals were being made of 18 tables which contained interview status and demographic information on the sample. Our monitoring led to the creation of a table called *retention*, which summarized interview status and sample relevance, and a table called *constants*, which summarized fixed demographic information (e.g., birth date, gender, education), over the entire panel for all sample members. Thus the 36 tables that had to be retrieved were reduced to two tables in the redesigned (phase two) database.

18 This is not to say that inchoate electronic communities do not exist in social science. There are informal collectives of scientists who interact regularly because of their research on similar problems. The Internet stock market exchanges organized by economists over the past several years certainly meet the requirements for forming an electronic community: "a large amount of data, both formal and informal, and a real need to manipulate these data extensively" (Schatz, 1992, p. 92).

19 But there is also another explanation as to why social scientists' work was ignored, which has to do with the hierarchical nature of science and which fields do and do not have prestige and the resources associated with that prestige. The committees of the National Research Foundation, as well as science panels of the National Research Council and National Academy of Sciences, are dominated by the physical or "hard" sciences. In general, there is widespread disdain for "soft" or social and behavioral sciences.

20 Some indirect evidence of the extent of multi- or interinstitutional collaboration in the physical and life sciences is available in a recent report on the increase of multi-author publications (Regalado, 1995).

References and further reading

Aaron, H. (1985), "Comments on 'Evaluation of Census Bureau procedures for the measurement of noncash benefits and the incidence of poverty', by Barry Chiswick", *Proceedings of the Conference on the Measurement of Noncash Benefits*, *Vol. 1*, US Bureau of the Census, Washington, DC, pp. 57-62.

Allen, T.J. (1970), "Roles in technical communication networks", in Neson, C.E. and Pollack, D.K. (Eds), *Communication among scientists and engineers*, Heath Lexington Books, Lexington, MA, pp. 191-208.

Allison, P.D. and Long, J.S. (1990), "Departmental effects on scientific productivity, *American Sociological Review*, Vol. 55, pp. 469-79.

Andrews, F.M. (Ed.) (1979), *Scientific Productivity: The Effectiveness of Research Groups in Six Countries*, Cambridge University Press, New York, NY.

Baker, S. (1989), "University scientists crack high-tech welfare data shell", *Computerworld*, Vol. 22 No. 21,May 23, p. 18.

Baker, D.N. and Zwicki, R.D. (1984), "Meeting report: NASA data systems users", *EOS*, Vol. 65 No. 6, February 7, p. 46.

Bright, M.W., Hurson, A.R. and Pakzad, S.H. (1992), "A taxonomy and current issues in multidatabase systems", *Computer*, Vol. 25 No. 3, pp. 50-9.

Bush, V. (1945), "As we may think", *Atlantic Monthly*, Vol. 176 No. 1, pp. 101-8.

Chubin, D.E. (1976), "The conceptualization of scientific specialties", *The Sociological Quarterly*, Vol. 17 No. 4, pp. 448-76.

Cinkosky, M.J., Fickett, J.W., Gilna, P. and Burks, C. (1991), "Electronic data publishing and Genbank", *Science*, Vol. 252, May 31, pp. 1273-7.

Citro, C.F. and Kalton, G. (Eds) (1993), *The Future of the Survey of Income and Program Participation*, National Academy Press, Washington, DC.

Clery, D. (1993), "ERS-1: a cautionary tale of data overload", *Science*, Vol. 261, p. 847.

Comer, D. (1983), "The computer science research network CSNET: a history and status report", *Communications of the ACM*, Vol. 26 No. 8, August 13, pp. 747-53;

Committee on Physical, Mathematical, and Engineering Sciences; Federal Coordinating Council for Science, Engineering, and Technology; and the Office of Science and Technology Policy (1992), *Grand Challenges 1993: High Performance Computing and Communications, The FY 1993 US Research and Development Program (1992)*, supplement to the President's Fiscal Year 1993 Budget, Committee on Physical, Mathematical, and Engineering Sciences; Federal Coordinating Council for Science, Engineering, and Technology; and the Office of Science and Technology Policy, Washington, DC.

Cotter, H. (1988), "Birth of a network: a history of BITNET", *CUNY/University Computer Center Communications*, Vol. 14, pp. 1-10.

Crane, D. (1969), "Social structure in a group of scientists: a test of the "invisible college" hypothesis", *American Sociological Review*, Vol. 32 No. 5, pp. 335-52.

David, M.H. (Ed.) (1983), *Technical, Conceptual, and Administrative Lessons of the Income Survey Development Program (ISDP): Papers Presented at a Conference*, Social Science Research Council, New York, NY.

David, M.H. (1985), "The distribution of income in the United States: implications for the design of the SIPP panel", *Journal of Economic and Social Measurement*, Vol. 13 Nos. 3-4, Special Issue, pp. 305-17.

David, M.H., and Robbin, A. (1990), "Database design for large-scale, complex data", in Berk, K. and Malone, L. (Eds), *Computing Science and Statistics. Proceedings of the 21st Symposium on the Interface Between Statistics and the Computer*, American Statistical Association, Alexandria, VA, pp. 205-14.

David, M. and Robbin, A. (1992), *Building New Infrastructures for the Social Science Enterprise. Final Report to the National Science Foundation on the SIPP ACCESS Project, November 1984 - December 1991, 2 Vols* Institute for Research on Poverty, University of Wisconsin-Madison, *Madison, WI,* January.

Doyle, P. (1989), "Data base strategies for panel survey", in Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M.P. (Eds), *Panel Surveys*, John Wiley & Sons, New York, NY, pp. 163-89

Doyle, P. and Dalrymple, R. (1988), "The impact of imputation procedures on distributional characteristics of the low income population", *Proceedings of the Third Annual Research Conference*, US Bureau of the Census, Washington, DC, pp. 483-508.

Doyle, P., Citro, C. and Cohen, R.L. (1987), *Feasibility Study of Long-term Access to SIPP*, Mathematica Policy Research, Inc, Washington, DC, July.

Farley, R. and Neidert, L.J. (1989), "Analyzing the characteristics of blacks: a comparison of data from SIPP and CPS", *1988 Proceedings of the Social Statistics Section*, American Statistical Association, Alexandria, VA, pp. 103-8

Flory, T.S., Martini, A. and Robbin, A. (1989), "Spells of AFDC receipt in the 1984 SIPP Panel", *1988 Proceedings of the Social Statistics Section*, American Statistical Association, Alexandria, VA, pp. 267-72.

French, J.C., Jones, A.K. and Pfaltz, J.L. (Eds) (1990), *Scientific Database Management (Final Report)*, Report of the Invitational NSF Workshop on Scientific Database Management, March 1990, Charlottesville, VA (Computer Science Report No. TR-90-21), Department of Computer Science, University of Virginia, Charlottesville, VA, August.

Green, J.L. and King, J.H. (1986), "Behind the scenes during a comet encounter", *EOS*, Vol. 67 No. 9, March 4, pp. 105, 11.

Greenstadt, E.W. (1981), "Data systems users working group", *EOS*, Vol. 62 No. 5, February 10, p. 50.

Hagstrom, W.O. (1970), "Factors related to the use of different modes of publishing research in four scientific fields", in Nelson, C.E. and, Pollock, D.K. (Eds), *Communication among Scientists and Engineers,* D.C. Heath, Lexington, MA, pp. 87-124.

Hardy, D.R. and Schwartz, M.F. (1993), "Essence: a resource discovery system based on semantic file indexing", *1993 Winter USENIX,* San Diego, CA, January 25-29, 1993, pp. 361-73,

Hesse, B.W., Sproull, L.S., Kiesler, S.B. and Walsh, J.P. (1993), "Returns to science: computer networks in oceanography", *Communications of the ACM*, Vol. 36 No. 8, pp. 90-101.

Jabine, T.B. (1990), *Quality Profile. Survey of Income and Program Participation*, US Department of Commerce, Bureau of the Census, Washington, DC, May.

King, K.E., Petroni, R.J. and Singh, R.P. (1988), "Data quality of the Survey of Income and Program Participation", *1987 Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, pp. 221-6.

Lander, E.S., Langridge, R. and Saccocio, D.M. (1991), "Mapping and interpreting biological information", *Communications of the ACM*, Vol. 34 No. 11, pp. 33-9.

Lederberg, J. (1978), "Digital communications and the conduct of science: the new literacy", *Proceedings of the IEEE*, Vol. 66 No. 11, pp. 1314-19.

McClure, C.R., Bertot, J.C. and Zweizig, D.L. (1994), *Public Libraries and the Internet: Study Results, Policy Issues, and Recommendations*, US National Commission on Libraries and Information Science, Washington, DC, June.

McClure, C.R., Bishop, A.P., Doty, P. and Rosenbaum, H. (1991), *The National Research and Education Network (NREN): Research and Policy Perspectives*, Ablex Publishing Corporation, Norwood, NJ.

Marquis, K.H. and Moore, J.C. (1990), "Measurement errors in SIPP program reports", *Proceedings of the Sixth Annual Research Conference,* US Department of Commerce, Bureau of the Census, Washington, DC, pp. 721-45.

Marshall, E. (1993), "Fitting planet Earth into a user-friendly database", *Science*, Vol. 261, August 13, pp. 846, 848.

Merton, R.K. (1988), "The Matthew effect in science, II: cumulative advantage and the symbolism of intellectual property", *ISIS*, Vol. 79 No. 299, pp. 606-23.

National Academy of Sciences; National Academy of Engineering; Institute of Medicine; Committee on Science, Engineering, and Public Policy (1989), *Information Technology and the Conduct of Research: The User's View*, Report of the Panel on Information Technology and the Conduct of Research, National Academy Press, Washington, DC.

National Research Council (Committee on National Statistics) (1989), *The Survey of Income and Program Participation – An Interim Assessment*, National Academy Press, Washington, DC.

National Research Council (Committee on Global Change – US National Committee for the IGBP – of the Commission on Geosciences, Environment, and Resources) (1990), *Research Strategies for the US Global Change Research Program*, National Academy Press, Washington, DC.

National Research Council (Committee on Arctic Solid-Earth Geosciences; Polar Research Board; and Commission on Geosciences, Environment, and Resources) (1991a), *Opportunities and Priorities in Arctic Geoscience*, National Academy Press, Washington, DC.

National Research Council (Panel on a Global Network of Fiducial Sites. Committee on Geodesy; Board on Earth Sciences and Resources; Commission on Geosciences, Environment, and Resources) (1991b), *International Network of Global Fiducial Stations: Science and Implementation Issues*, National Academy Press, Washington, DC.

National Research Council (Committee on a National Collaboratory: Establishing the User-Developer Partnership) (1993), *National Collaboratories: Applying Information Technology for Scientific Research*, National Academy Press, Washington, DC.

National Research Council (NRENAISSANCE Committee) (1994a), *Realizing the Information Future: The Internet and Beyond*, National Academy Press, Washington, DC.

National Research Council (Commission on Geosciences, Environment, and Resources; Ocean Studies Board) (1994b), *The Ocean's Role in Global Change: Progress of Major Research Programs*, National Academy Press, Washington, DC.

National Science Foundation (1994c), *Division of Social, Behavioral, and Economic Research Grant List. Fiscal Year 1994 (October 1993 - September 1994 (draft)),* National Science Foundation, Arlington, VA.

Osserman, R. (1989), "The geometry renaissance in America: 1938-1988", in Askey, R., Duren, P.L. and Merzbach, U.C. (Eds), *A Century of Mathematics in America. Part II*, American Mathematical Society, Providence, RI, pp. 513-26.

Pechura, C.M. and Martin, J.B. (Eds) (1991), *Mapping the Brain and its Functions: Integrating Enabling Technologies into Neuroscience Research*, National Academy Press, Washington, DC.

Pool, R. (1993), "Beyond databases and e-mail", *Science*, Vol. 261, August 13, pp. 841-3.

President's Science Advisory Committee (1963), *Science, Government and Information*, US Government Printing Office, Washington, DC, January 10.

Rees, D., Perla, I., Meredith, N.P., Green, J. and Van der Heijden, N. (1986), "Networking ground-based images of Comet Halley during the Giotto Encounter", *EOS*, Vol. 67 No. 50, December 16, pp. 1385-7.

Regalado, A. (1995), "Multiauthor papers on the rise", *Science*, Vol. 268, April 7, p. 25.

Robbin, A. (1992), "Social scientists at work on electronic research networks", *Electronic Networking: Research, Applications and Policy*, Vol. 2 No. 2, pp. 6-30.

Ryscavage, P. (1987), *SIPP: Filling Data Gaps on the Poverty and Social Welfare Fronts* (*Survey of Income and Program Participation* Working Paper No. 8705), US Bureau of the Census, Washington, DC.

Schatz, B.R. (1987), "Telesophy: a system for manipulating the knowledge of a community", *Proceedings of the IEEE Globecom '87*, Tokyo, November, Computer Society Press of the IEEE, pp. 1181-6.

Schatz, B.R. (1992), "Building an electronic community system", *Journal of Management Information Systems*, Vol. 8 No. 3, pp. 87-107.

Scheuermann, P., Yu, C. and Program Committee (1989), *Heterogeneous Database Systems*, Report of the NSF Workshop on Heterogeneous Database Systems, December 11-13, Department of Computer Science, Northwestern University, Evanston, IL.

Science (1994a), "Science careers: playing to win", (1994). *Science*, Vol. 265, September 23, pp. 1905-39.

Science (1994b), "Crème de la crème (cont'd)", *Science*, Vol. 266, December 16, p. 1811.

Silberschatz, A., Stonebraker, M. and Ullman, J.D. (Eds) (1990), *Database systems: Achievements and Opportunities*, The "Lagunita" Report of the NSF Invitational Workshop on the Future of Database Systems Research, February 22-23, Palo Alto, CA (TR-90-22), Department of Computer Sciences, University of Texas at Austin, Austin, TX.

Simon, H.A. (1991), "Bounded rationality and organizational learning". *Organization Science*, Vol. 2 No. 1, pp. 125-34.

Small, H. (1990), *Bibliometrics of Basic Research*, prepared under contract to the Office of Technology Assessment, (PB 91-166 579), Department of Commerce, National Technical Information Service, Springfield, VA, July.

Steele, G. (1984), *COMMON LISP: The Language*, Digital Press, Bedford, MA.

US Congress (Office of Technology Assessment) (1986), *Research Funding as an Investment: Can We Measure the Returns? – A Technical Memorandum* (OTA-TM-SET-36), US Congress, Office of Technology Assessment, Washington, DC, April.

US Congress (Office of Technology Assessment) (1988), *Mapping our Genes – The Genome Projects: How Big, How Fast?* (OTA-BA-373), US Government Printing Office, Washington, DC, April.

US Congress (102nd Congress) (1991a), High Performance Computing Act of 1991, Public Law 102-194, 105, *Stat*, December 9, pp. 1594.

US Congress (Office of Technology Assessment) (1991b), *Federally Funded Research: Decisions for a Decade*, (OTA-SET-490), US Government Printing Office, Washington, DC, May.

US National Institutes of Health (Office of Science Policy and Regulation) (1993), *Staffing Patterns of the National Institutes of Health Research Grants*, US Department of Health and Human Services, National Institutes of Health, Washington, DC, March.

Wells, J. (1991), “Processing and analyzing large files with small computers: new technology developments”, 1990 Proceedings of the Section on Survey Research Methods, American Statistical Association, Arlington, VA, pp. 121-6.

Wiederhold, G. (1992), “Mediators in the architecture of future information systems”, *Computer*, Vol. 25 No. 3, pp. 38-48.

Wulf, W.A. (1989), “The national collaboratory – a white paper”, in “Towards a national collaboratory (appendix A”, unpublished report of a National Science Foundation invitational workshop held at Rockefeller University, New York, 17-18 March 1989 (Lederberg, J. and Uncapher, K., co-chairs).

Wulf, W.A. (1993), “The collaboratory opportunity”, *Science*, Vol. 261, August 13, pp. 855-6.