e-Science Investments in the Social and Behavioral Sciences at the National Science Foundation: An Overview of Projects, Programs, and Policy Issues

Lawrence Burton
National Science Foundation
United States of America
lburton@nsf.gov

Julia Lane
National Science Foundation
United States of America
jlane@nsf.gov

Abstract. This paper presents an overview of recent National Science Foundation (NSF) efforts to formulate a strategic framework for e-science investments, with specific emphasis on e-social science programs and how they fit into the larger NSF framework. It identifies a set of existing funding programs that have already invested in e-social science projects and capabilities, and highlights a few examples from those past investments. It also briefly introduces two new program calls that are being competed in summer 2005. Finally, it draws on the results of three workshops to highlight the special potential and problems for e-social science in the U.S.

Introduction

Advances in CyberInfrastructure[1] are making revolutionary changes in science and engineering research possible. At the same time, they are increasing the relevance of research to important national goals in the U.S. Social scientists hold a unique position in this rapidly evolving area precisely because the efficient adoption, deployment and utilization of CyberInfrastructure requires not just investing in the appropriate physical technology but also understanding how that technology becomes embedded in society and the economy. Social scientists also will gain by having new tools and resources, a new vital focus, and expanded knowledge. These two streams of activity (technical and social) are linked and occurring simultaneously – the challenge to social scientists is to be full and effective partners in this new opportunity.

The Directorate for Social, Behavioral and Economic Sciences (SBE), and the National Science Foundation (NSF) more broadly, are currently developing both a short-term and a

[1] "e-Social Science" is a European term; the approximate U.S. equivalent is "CyberInfrastructure."

long-term vision of e-science for science and engineering research.[2] As part of this process, SBE sponsored two workshops with participants from the full spectrum of the social and behavioral sciences disciplines in Fall, 2004, and in March 2005 co-sponsored a third workshop with the Directorate for Computer and Information Sciences and Engineering (CISE) attended by both computer/information scientists and social/behavioral scientists. This paper reports on what was learned at these workshops.

Three unique aspects of e-social science were identified in these workshops. First, the efficient development and deployment of e-science requires social and behavioral science input. In other words, the social and behavioral sciences bear responsibility for studying the human activities of implementing and using CyberInfrastructure across the domains of science and engineering and throughout society. Second, as noted by Brady (2004), SBE scientists use a uniquely wide range of data and data collection modes to describe human behavior – ranging from sensors and imaging to geospatial and economic data to audio and audio-video recording of human interactions. Archiving, indexing, curating, and providing metadata across such an array of important sources pose on-going challenges to the SBE sciences. Third, since SBE research is often based on human subjects, SBE scientists face unique privacy and confidentiality challenges in implementing and extending e-social science.

Overview of the Issues

In this section, we discuss in more detail the insights on each of the three unique features. We draw heavily on the input provided by participants at the three recent SBE workshops.

Understanding human use of CyberInfrastructure

The seminal document for U.S. discussions of e-science is the NSF-sponsored report of the Blue Ribbon Advisory Panel on CyberInfrastructure—familiarly referred to as "the Atkins report" (Atkins, 2003). This report recognized the importance of understanding how humans use CyberInfrastructure. The panel envisioned a layered CyberInfrastructure for e-science. The top-most layer consists of communities of researchers across all of the disciplines. These communities access a lower level of generic applications (including middleware) that are largely the responsibility of the computer/information science domain. Finally, the generic applications rest atop a stratum of both technological and social systems, responsibility for which is to be shared by the CISE and SBE communities. The schematic conceptualizing these layers of e-science is reproduced from the Atkins report in Figure 1.

[2] NSF is organized into 7 research directorates: SBE, Biological Sciences, Computer and Information Science and Engineering, Education and Human Resources, Engineering, Geosciences, and Mathematical and Physical Sciences. In addition 2 specialized offices also sponsor research: International Science and Engineering, and Polar Programs.

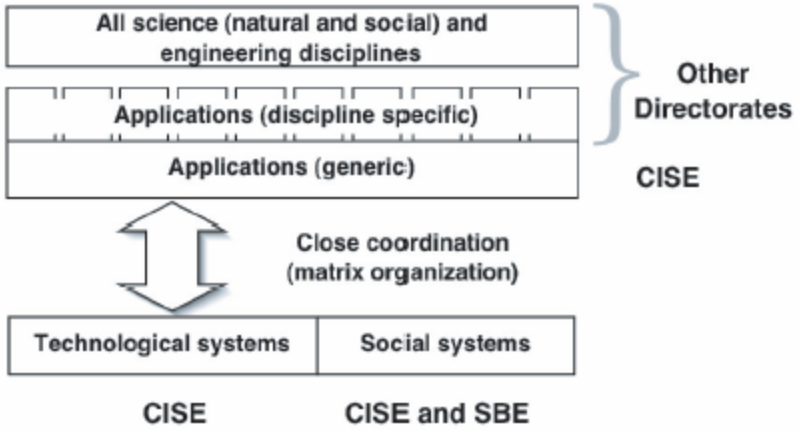

Figure 1. Assignment of responsibility for the vision and governance of cyberinfrastructure to NSF directorates

Source: Atkins, Daniel E. et al. (2003), figure 4.6.

Figure 1 illustrates the first of the three unique aspects of the social/behavioral sciences vis-à-vis e-science; by identifying the SBE sciences as part of the bedrock of e-science for all disciplines, it emphasizes that investments in physical CyberInfrastructure can be wasted if there is insufficient attention paid to the ways in which humans adopt and implement it. The sense of participants in both SBE workshops in 2004 reinforced the view of the Blue Ribbon Panel that effective investments in CyberInfrastructure pose clear challenges that must be addressed by social science researchers.

These challenges include understanding how humans adopt and use CyberInfrastructure, identifying the organizational and market mechanisms that create the best incentive structure for the development of CyberInfrastructure, and the impact of CyberInfrastructure on the economic, social and behavioral framework of the United States. Participants also emphasized that investment in CyberInfrastructure for the social sciences held great promise for promoting U.S. social science to the rank of big science. CyberInfrastructure developments offer an opportunity to answer fundamental scientific questions about the complexity of human behavior by collecting, linking, and accessing data on a scale transcending what is now available. Participants acknowledged that with that opportunity came a number of difficult tasks and trade-offs, ranging from protecting the privacy of those who provide the data, to the difficulty of designing infrastructure whose maintenance costs could be kept under control, to standardizing data collection instruments.

The final report of the SBE/CISE Workshop on "CyberInfrastructure for the Social and Behavioral Sciences" (Berman and Brady, 2005) reiterated this theme as follows:

> Yet implementing CI is not easy. The high failure rate of large IT projects and their tendency to overrun budgets indicate that we must pay attention to implementation. Strong empirical evidence demonstrates that the sources of failures and cost overruns typically are organizational and managerial rather than technological. The higher level of complexity of CyberInfrastructure demands a commensurably higher level of organizational and managerial knowledge and expertise. (p29)

That same report stresses that the interactions between the SBE and CISE communities are of necessity beneficial to both, and summarized these interactions as in Table I.

Table I – Roles of SBE and CISE for Each Task

TASKS	ROLE FOR EACH GROUP	
	Social, Behavioral and Economic Sciences (SBE)	**Computer & Information Science and Engineering (CISE)**
CISE Developing CyberInfrastructure to support and enable the SBE sciences	*Identify SBE needs and requirements for* • Data collection • Computational analysis, simulation and modeling • Tools for data comparison and measurement • Data storage, management, and preservation • Communication and collaboration, etc.	*Work with SBE researchers to identify needs and requirements and frame them as technology problems* *Design, develop and deploy technical solutions for SBE problems* *Work with SBE community to target CyberInfrastructure tools and technologies to community needs, and to assist the community to use them effectively*
SBE Helping CISE design scientific infrastructure	*Work with CISE community to structure CyberInfrastructure organizations and infrastructure to promote coordination, functional social dynamics, effective decision-making, conflict resolution, etc.* *Work with CISE community to define effective incentive and allocation structures that promote stability, efficiency, and usability of CyberInfrastructure* *Work with CISE community to better define and develop mechanisms for discouraging malevolent behavior*	*Incorporate models, frameworks, incentives, policies, and other mechanisms from the social science community into tools and technologies comprising CyberInfrastructure* *Target enabling CyberInfrastructure tools and technologies to social, behavioral and economic science applications* *Develop CyberInfrastructure solutions to enable SBE* • Data collection • Data analysis and modeling • Data comparison and measurement • Data archiving • Communication and collaboration
CISE and SBE Assessing the societal impact of CyberInfrastructure	*Describe and assess social impacts of CyberInfrastructure on:* • Human interaction • Jobs and income • Privacy • Social and institutional frameworks, etc.	*Develop coordinated instrumentation, tools and technologies for assessing the social impacts of CyberInfrastructure* *Assess the technical vulnerabilities of CyberInfrastructure*
Source: Berman and Brady (2005).		

Data Resources

The second theme emerging from the workshops concerned data resources. A major result of the advances in cybertechnology has been expanding abilities to collect information from a wide variety of different sources, as well as to measure human behavior in very different ways. This has posed new challenges for the collection, indexing, archiving, curation, and preservation of a wide variety of data important to the social and behavioral sciences research communities. In other words, realizing the remarkable potential of long-lived data requires not only that they be preserved, but also that they be discoverable by others (using various search tools) and available in a usable format, including essential metadata describing the nature, quality, and history of the data. Among the CyberInfrastructure resources necessary to provide for preservation and access are data centers, digital libraries, and electronic repositories. These resources may be centralized or distributed, may serve global, community, or single-institution functions, and may be operated by a university, a library system, a professional society or other non-profit organization, a corporation, or a government agency. These resources share common features: digital storage hardware, network access, software tools, and expertise (both information science and domain-specific). Largely as a result of this phenomenon, the number of data centers in all areas of science and engineering is increasing rapidly because the products of research are more frequently digital in nature and because the opportunities presented by robust digital data sets are recognized more broadly in all areas of science and engineering. As the number of observatory networks increases, so too will needs for data storage and curation.

A national opportunity for social scientists is clear: CyberInfrastructure provides a new capacity to discover and integrate data that in turn enables social scientists to understand human and organizational behaviors, ranging from career path and retirement decisions made by individuals, to organizational decisions about collaboration and interaction versus competition and isolation, to land use choices made by institutions and organizations affecting ecosystem sustainability and environmental change, and even further, to the efficient human adoption and implementation of CyberInfrastructure investments across a variety of disciplines.

These new capacities can also contribute to meeting critical national security needs. The major security threat to the United States is inherently human and an improved ability to understand and predict behaviors can provide means for addressing that threat. Meeting these challenges requires support for a variety of activities including tool development, more intelligent search and discovery methods, data mining, workforce training, and catalyzing change in the reward structure of academe. The challenge has been broadly recognized: a recent report by the National Science Board has identified this as a priority area (National Science Board, 2005).

Confidentiality Protection

The third central issue for the social/behavioral science communities resulting from the revolution in CyberInfrastructure is confidentiality protection. As Margo Anderson pointed out at the SBE/CISE workshop:

> The extraordinary growth of electronic infrastructure, capacity, and use in the past decade has posed a profound new set of questions about the control,

> dissemination, power and use of information. On the one hand the high speed Internet and the World Wide Web, email, electronic shopping, and cell phone use have opened up extraordinary new worlds of communication and are changing the way we work, play, and learn. On the other, as the electronic world enters our daily lives, the private space untouched by the intrusions of cyberspace and information seekers shrinks - for individuals, firms, and organizations. …There is also another challenge. The need to build more efficient surveillance networks to combat potential terrorist attack argues for less privacy for the individual person or firm to guarantee the security of the society in general. It is in this environment that citizens, business and technology leaders, and policy makers have to figure out how to understand, manage, and regulate the new cyberworld. (SBE/CISE workshop March 15-16, 2005, http://vis.sdsc.edu/sbe/About)

In sum, the challenge for social scientists is to use information derived from vast streams of human related data without breaching confidentiality. This is a serious challenge. Breaches of confidentiality that result from the actions of one researcher can threaten the ability of scientists everywhere to collect and use data. Preserving access to high quality scientific data is essential to the empirical replication that is at the core of good science. As George Duncan indicated at the same workshop:

> In important cases social scientists [as well as scientists in the biological and environmental sciences] require the rich and sensitive data that makes confidentiality consequential. The technological factors that have led both to the explosive growth in the capability of providing such data, as well as in the capability of data snoopers to breach confidentiality, are in the domain of computer scientists. (SBE/CISE workshop, March 15-16, 2005. http://vis.sdsc.edu/sbe/About)

There is already an existing community that has focused on the importance of confidentiality – and NSF sponsored a workshop on the topic in 2003[3]. In particular, U.S. federal statistical agencies have devoted substantial resources to both statistical and technical ways to protect confidentiality (see Doyle, et al. 2001), and the Social and Behavioral Research Working Group recently drafted a report entitled "Achieving Effective Human Subjects Protection and Rigorous Social and Behavioral Research" for the Human Subjects Research Subcommittee of the Committee on Science, National Science and Technology Council. In addition PITAC recently issued a report on cybersecurity that addressed some confidentiality issues and numerous studies have been undertaken by the National Academy of Sciences and the Committee on National Statistics.

Current Programs

E-science awards in the SBE directorate at NSF have been made across a variety of programs, subdisciplines, initiatives, etc. Notable recent cross-directorate initiatives with e-social science awards have included the Information Technology Research (ITR) program that ended in 2004 and the Digital Libraries Initiative. Another cross-Foundation program (and managed by SBE) is the Human and Social Dynamics initiative, now entering its second year of full implementation. These initiatives and programs are summarized in Table II, which

[3] http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5421&org=SES&from=home

also lists specific projects that are described on the NSF website at http://www.nsf.gov/awardsearch/.

Within SBE, virtually all of the disciplinary research programs have invested in e-social science projects. In addition, special competitions like Enhancing Infrastructure for the Social and Behavioral Sciences—last competed in 2000—provide occasional and focused opportunities for e-social science investments. On-going data projects that have been used by hundreds of researchers are also part of the SBE e-social science portfolio. These include: the Panel Study of Income Dynamics (PSID), the American National Election Study (ANES), the General Social Survey (GSS), and the special data programs on science and engineering in the Congressionally-mandated Division of Science Resources Statistics (SRS).[4] (These have counterpart studies in other countries.)

E-social science investments by SBE to date have been specific to the SBE domains (but see next section on Future Programs), that is, the directorate has not funded projects that were intended, as their primary purpose, to be shared across the disciplines. This does not mean that some of the information technology research being funded within the directorate might not have potential applications outside of the SBE sciences.

The funding data in Table II are derived from internal NSF figures and reflect a conservative estimate of the SBE investments in e-social science. The authors have added columns for programs and selected examples of individual projects.[5] The projects noted are examples only and are all still underway as of this writing, but some will have ended by start of FY2006 in October, 2005.

[4] SRS data programs and access tools are available at http://www.nsf.gov/statistics/.

[5] Brief descriptions of these projects can be viewed at www.nsf.gov/awardsearch/ by typing the project number from Table II into the field "Search Award For."

Table II. Directorate for Social, Behavioral, and Economic Sciences
FY 2006 Investments in CyberInfrastructure
(Dollars in Millions)

SBE Cybertools and Services	FY 2006	Programs	Project Examples
Infrastructure and Instrumentation (The SBE directorate supports major infrastructure and instrumentation awards, often through special competitions. This may include large scale data collection and analysis projects including both large scale surveys and collaboratories.)		SBE Infrastructure Program; Human and Social Dynamics	Adams-0094908 Aristar-0094934 Beck-0094928 Holt-0094800 Kidd-0096588 Mutz-0094964 Bowker-0433369
Infrastructure Component	33.33%		
Instrumentation Component	9.80%		
CyberInfrastructure Innovation (To increase the capacity for doing research in the social, behavioral, and economic sciences, specific projects will be funded that will provide new and innovative infrastructure)	7.35%	All SBE basic research programs; Information Technology Research (ITR)	Stanton-0312078 Danziger-0121232
SBE Centers (includes facilities, data depositories, and often the development of novel hardware, software, and communications systems to support the particular requirements of the center)	4.41%	National Center on Violence Research; Environmental Social & Behavioral Sciences	Not applicable
SBE Domain Specific Data Sets (the creation and support of datasets and infrastructure that will be shared across specific SBE disciplines)	3.92%	Various	Panel Study of Income Dynamics-0094942
Modeling Activities (supports the development of new and innovative models across SBE)	2.45%	Various	Friedman-0351801
Spatial Social Science (includes the development of shared infrastructure within the social science research community)	4.90%		Goodchild-9978058
Basic Infrastructure (support for the CyberInfrastructure requirements for proposals in the basic research programs across the directorate)	15.69%	All SBE basic research programs	Rao-0452033 Chen-0079001 Bolton-0351408
Statistical Resources (cyber facilities, instrumentation, and analysis developed by the Division of Science Resources Statistics)	0.00%	Division of Science Resources Statistics programs	Not applicable
Facilities Component	2.45%		
Instrumentation Component	14.22%		
Analysis Component	1.47%		
TOTAL, SBE CyberInfrastructure	**100%**		
NOTE: To see project descriptions, insert project number at “Search Award For” at http://www.nsf.gov/awardsearch/			

New Programs

Recognizing the importance of both human interaction with CyberInfrastructure and data resources, the SBE Directorate has collaborated with other NSF directorates in two current CyberInfrastructure solicitations – CyberInfrastructure Team (CI-Team) and Next Generation Cybertools (NGCT).

Human Interaction with CyberInfrastructure

The preparation of a workforce knowledgeable about e-science is an essential component of human interaction with CyberInfrastructure. In recognition of this, the NSF research directorates have initiated the CI-Team program, which solicits promising demonstration project proposals from partnerships of organizations committed to the preparation of a diverse e-science workforce.[6] Examples of such projects may include:

- Mobilization of a community of interest or practice around existing or future CyberInfrastructure resources or services;
- Preparing faculty in junior colleges to use CyberInfrastructure effectively to promote and advance learning and discovery;
- Developing new science and engineering curricula based on CyberInfrastructure capabilities such as simulation, modeling, and data driven science;
- Preparing individuals in organizations traditionally underrepresented in science and engineering research activities to participate more fully;
- Preparing CyberInfrastructure professionals to develop and support CyberInfrastructure services;
- Engaging new communities, organizations, groups and/or individuals in science and engineering through the improved effectiveness of collaboratories; and
- Exploring the complementary roles that for-profit and not for-profit organizations play in supporting and sustaining CyberInfrastructure and the CyberInfrastructure workforce.

Data Resources

Another important solicitation, Next Generation Cybertools (NGCT), which was developed jointly by SBE and CISE, is intended to build a synergistic relationship between researchers in the social and behavioral sciences and researchers in the computer and information sciences.[7] The focus of the solicitation is on the development and utilization of data. The new program will take advantage of CyberInfrastructure advances that have permitted social and behavioral scientists to find new ways to create and analyze data. Likewise, CyberInfrastructure advances have enabled computer and information scientists to conduct research that yields new ways to improve both domain-specific and general-purpose tools to analyze and visualize scientific data—such as improved processing power, enhanced interoperability of data from different sources, data mining, data integration, information indexing, and data confidentiality protection.

The solicitation invites groups of researchers from each of the two domains to work together in advancing research on "information infrastructure testbeds" in the development of the next generation of cybertools applied to data from various sources collected at two levels of

[6] http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=12782&org=SCI
[7] http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=13553&org=CISE&from=fund

analysis: organizations and individuals. The tools that are developed on these platforms must not only change ways in which social and behavioral scientists research the behavior of organizations and individuals, but also serve the sciences more broadly.

The solicitation has required the following:

- Development of tools that facilitate the integration of qualitative and quantitative information from heterogeneous sources and in multiple media;
- Investment in basic research that addresses the protection of the confidentiality of respondents in computerized, widely accessible databases; and
- Development of incentives, standards and policies for collecting, storing, archiving, accessing, and publishing research results.

Barriers to e-social science and proposed solutions

Progress in e-social science in the U.S. will confront several obstacles identified by the attendees at the workshops.

Misalignment of current incentive structures

Participants in the workshops noted that there was substantial misalignment both in assignment of ownership rights and in how academic credit is granted. Ownership rights in data generated in a collaborative project are difficult to assign, yet the data themselves may have substantial financial value. Likewise, some social science communities and departments do not have a tradition of granting academic credit to tool builders or researchers who share their data widely. Workshop participants suggested that scientific credit be shared with the originator of the data. This could be instituted as simply as a citation in journal articles, just as references are made to other scientific contributions; failure to cite the data provider/creator/documentor could be treated as plagiarism. Users could be required to agree to such a provision as part of the normal user agreement when data were accessed. Biographical sketches could include "Data creation/production/documentation" as an element, together with a metric such as a count of how many times the data were used. Other suggestions included establishment of a national award for data sharing or innovation in collaboration, and by building into grant documents a requirement that—and funding for—project data be curated, stored and/or widely disseminated.

Fostering collaboration across domains

Workshop attendees pointed out that there were few established models for collaborative work between computer/information scientists and social/behavioral scientists, and it was this topic that dominated the discussions in the joint SBE/CISE workshop in March 2005. (The New Generation Cybertools solicitation explicitly addressed this historical lack of collaboration by *requiring* interdisciplinary teams in building the testbeds in that program.) Workshop attendees noted that another way of learning about the needs for social and behavioral science research in other scientific domains was to attend the large national research conferences of non-social scientific societies. It was suggested that NSF consider establishing a new national conference featuring e-science projects, or sponsoring e-science sections in existing conferences. Founding a new scientific research journal on the topic of e-science was also discussed.

Ensuring sustainability

Because the evaluation and maintenance of data tools and products are both costly and not a natural part of research culture, they are unlikely to happen without a coordinated strategy to develop a critical mass of resources and appropriate incentive systems. The principal near-term opportunity is to survey existing mechanisms of hardening and sustaining CyberInfrastructure at various levels and test the most promising approaches. Candidates for testing the hardening process might be drawn from recent ITR or related investments. Possible examples of successful hardening might be found in the Digital Government program (see http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5459).

Human Resources

SBE scientists expressed substantial concern about sufficient numbers of trained individuals for the full exploitation and maintenance of e-social science investments. Extensive thought needs to go into devising the most effective management for CyberInfrastructure projects. A cadre of paraprofessionals may be needed to supplement Ph.D. researchers. It was noted that the actual learning of the new technologies is not time consuming; rather, it is their adaptation for specific uses in the laboratory that requires great amounts of (expensive) principal investigator time. One participant thought that exposure to basic research methods, and use of existing large databases, should begin as early as elementary school. Another foresaw a need to educate the public in the "public service" of serving as a human subject.

Conclusion

Investments by the National Science Foundation in e-science have heretofore been largely decentralized across the research-sponsoring directorates, although some programs have anticipated the development of e-science across domains, notably CISE's Middleware Initiative in 2002.[8]. However, the planning and outreach programs by the directorates, similar to the SBE workshops described in this paper, may lead to a new organizational structure for funding NSF e-science investments. However, both domain specific and cross-directorate funding modalities are likely to continue.

While modest in size, NSF's e-social science investments have taken on new importance as a result of these planning exercises and of the widespread agreement about the unique role of the SBE sciences in contributing to the effective development of a CyberInfrastructure for all of the sciences and engineering. In addition to this role as an enabler for all of e-science, the new research possible in the emerging e-social sciences will also enhance the visibility of the social and behavioral sciences overall in the United States. In sum, advances in CyberInfrastructure, together with both national and international linkages and collaborations, present enormous opportunities to advance social science research and increase its impact on national priorities.

Acknowledgments

We gratefully acknowledge the comments of Rick Lempert, Rolf Lehming and Wanda Ward on earlier drafts of this paper.

The views presented in this paper do not necessarily represent the position of the National Science Foundation.

[8] See http://www.nsf.gov/news/news_summ.jsp?cntn_id=103046&org=NSF.

References

Atkins, Daniel E. et al. (2003). *Revolutionizing Science and Engineering Through CyberInfrastructure.* Arlington, VA: National Science Foundation. Available at http://www.nsf.gov/cise/sci/reports/atkins.pdf.

Berman, Francine and Henry Brady. *"Final Report: NSF SBE-CISE Workshop on Cyberinfrastructure and the Social Sciences."* Forthcoming 2005 at http://vis.sdsc.edu/sbe/About.

Brady, Henry. Speech to the ACLS Commission on CyberInfrastructure for the Humanities and Social Sciences, Doe Library, University of California – Berkeley, Berkeley, California, August 21, 2004. http://www.acls.org/CyberInfrastructure/cyber_meeting_notes_august.htm#brady_summary Accessed February 13, 2005.

Doyle, Pat, Julia Lane, Laura Zayatz and Jules Theeuwes,*Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, North Holland, 2001.

National Science Board *"Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century"* March 30, 2005, http://www.nsf.gov/nsb/meetings/2005/LLDDC_draftreport.pdf Accessed May 3, 2005.

President's Information Technology Advisory Committee "*PITAC Report to the President on Cyber Security: A Crisis of Prioritization"*, February 2005, http://www.itrd.gov/pitac/reports/20050301_cybersecurity/cybersecurity.pdf Accessed May 3, 2005.