TRANSACTIONS ON DATA PRIVACY 1 (2008) 2 —16

Data Access in a Cyber World: Making Use of Cyberinfrastructure

Julia Lane*, Pascal Heus and Tim Mulcahy*****

* National Science Foundation
** Open Data Foundation
*** NORC/University of Chicago

Abstract. The vast amount of data now collected on human beings and organizations as a result of cyberinfrastructure advances has created similarly vast opportunities for social scientists to study and understand human behavior. It has also made traditional ways of protecting social science data obsolete. The challenge to social scientists is to exploit advances in cyberinfrastructure to develop new access modalities that not only provide access but preserve data and create scientific communities. This paper outlines an approach that draws on both advances in the social science and the computer science literatures.

1 Introduction

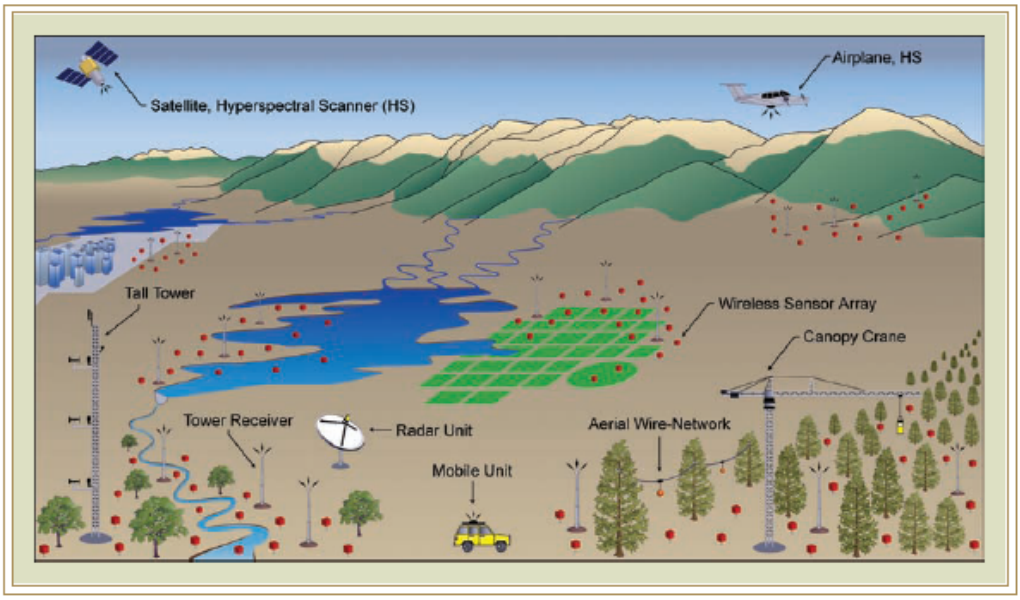
"So voluminous is the data surrounding CIOs that it could almost be likened to a tsunami that will engulf business. In fact an IDC study entitled 'The Expanding Digital Universe' estimated that the current size of that universe was 161 billion gigabytes. Moreover, it anticipated that the volume would grow six fold by the end of this decade to 988 billion gigabytes. In effect, that represents a CAGR (compound annual growth rate) in global data of 57 percent in over four years." Surviving the Ditigal Tsunami, CIO Magazine, Dec 13, 2007[1]

The vast amount of data now collected on human beings and organizations, such as businesses, illustrated by both the picture and the quotation has made traditional ways of protecting social science data obsolete. The availability of data due to data security breaches exacerbates the challenge: PrivacyRights.org estimates that over 223 million data records of U.S. residents have been exposed

[1] CIO refers to Chief Information Officer; for more information on IDC see www.idc.com

due to security breaches since January 2005.[2] At the same time, the data collection provides an invaluable opportunity to develop understanding of human behavior to an extent unimaginable even a decade ago.

Figure 1. Source National Science Foundation: Cyberinfrastructure Vision for 21st Century Discovery



An artist's conception (above) depicts fundamental NEON observatory instrumentation and systems as well as potential spatial organization of the environmental measurements made by these instruments and systems.

Social scientists could now reap the benefits of these transformational opportunities in a manner already realised by biological scientists with the Human Genome Project and by astronomers with the advent of whole sky surveys. The rich information now available on humans, and hinted at in Figure 1, ranges from biological information, through biomarkers, geospatial information, through RFID[3]'s and sensors, brain activity through MRIs[4] and social interactions, through virtual lives. Biological scientists were able to create a complete parts list of all human genes such that the component parts of the complex system could be delineated. This led to a view of biology as an information science

[2] www.privacyrights.org Accessed April 11, 2008

[3] RFID refers to radio frequency identification device; for more information see http://www.rfidjournal.com/

[4] MRI refers to magnetic resonance imaging

and created a *systems approach* to understanding health and disease. Astronomy was transformed by the advent of virtual astronomy: astronomers were able to discover significant patterns by analyzing rich image/catalog databases, and understand complex, astrophysical systems by using integrated data and large numerical simulations. Social scientists could potentially use the new data to transform their science: the challenge is not only to extract insights from this complex and heterogeneous set of data but also to share the information to ensure that the analytical work is generalizable and replicable.

The most trenchant challenge is, of course, providing access to and information about the new range of data. This paper argues in the following section that relying on one approach, like statistical disclosure limitation, is no longer feasible, and that it is necessary to develop new access modalities. The next section of the paper describes how cyberinfrastructure advances can be used to structure a portfolio approach to protecting data: combining technical, organizational, statistical and legal protections so that data on human subjects can be protected but so that access is available to the greatest possible number of researchers. Advances in cyberinfrastructure can also be used to archive, index and curate microdata, and this is described in the subsequent section. The paper concludes by describing how access modalities could also be structured to permit the development of virtual organizations that promote the sharing of knowledge about data.

2 Background

The creation and analysis of high-quality information are core elements of the scientific endeavor. No less fundamental is the dissemination of such data, for many reasons. First of all, data only have utility if they are used. Data utility is a function of both the data quality and the number and quality of the data analysts. The second is replicability. It is imperative that scientific analysis be able to be replicated and validated by other researchers. The third is communication. Social behavior is complex and subject to multiple interpretations: the concrete application of scientific concepts must be transparently communicated through shared code and metadata documentation. The fourth is efficiency. Data are expensive to collect – the U.S. 2010 Census alone is projected to cost over $15 billion, so expanding their use, promoting repurposing and minimizing duplication is fiscally responsible. Another reason is capacity-building. Junior researchers, policy makers and practitioners need to have the capacity to go beyond examining tables and graphs to a point where they may develop their understanding of the complex response of humans to rapidly changing social and legal environments. Access to micro-data provides an essential platform for

evidence based decision-making. Finally, access to micro-data permits researchers to examine outliers in human and economic behavior – which is often the basis for the most provocative analysis[1].

These arguments are not simply theoretical. The value added of access to micro-data is confirmed empirically. Illustrative examples include the deeper understanding of business dynamics made possible by examining the contribution of firm births and deaths, as well as expansions and contractions to net employment growth[2]. Similarly, a landmark 1954 study of survey data on doctors' smoking habits matched to administrative data on their eventual cause of death was critical in establishing the link between smoking and both cancer and coronary thrombosis [3], cited in [4]. And data on individual financial transactions are now routinely used to model and limit losses due to defaults on loans.

There are two major barriers limiting access to individual-level data. The first is that information on human subjects that have been collected by national statistical agencies, private organizations or researchers is limited by legal and ethical protections. Violations of such protections can have severe adverse consequences to data collectors in terms of reputation, response rates and legal action. Another barrier to dissemination is that there is insufficient recognition of the ownership and intellectual property rights relating to data production and sharing.

The approaches that have been used by national statistical institutes to provide access are often unsatisfactory. Probably the most well-known approach is to provide access via Research Data Centers, but the cost in time and money for researchers to access the data has led to serious underutilization and hence a reduce return on the agency's investment [5][4][4]. Remote buffered access, like the Luxembourg Income Study, has the advantage of providing access to many users but does not provide direct access to micro data. In addition, the delays entailed by the layers of review before any output is seen places a high burden on the statistical agency and often results in unacceptable delays for decision-makers.

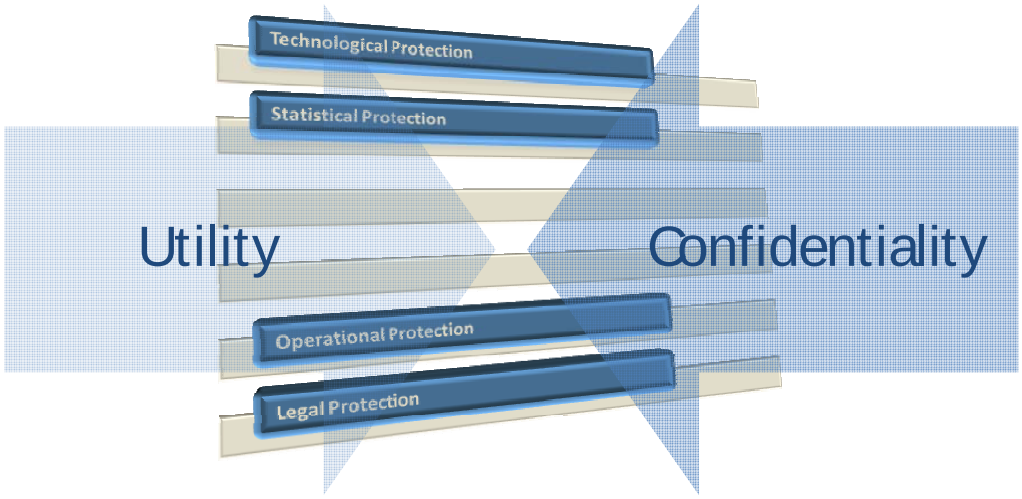
The investment in research to promote access to microdata has been encouraging [6], as has the movement to online remote access systems. These approaches use modern computer science technology, together with researcher certification and screening, to replace the burdensome, costly and slow human intervention associated with buffered remote access. For example, Statistics Denmark has found that remote access arrangements are now the dominant mode of access to micro-data. Statistics Sweden has increased the accessibility of microdata for external users at the same time that it has increased security precisely because the client's computer functions like an input/output terminal. Statistics Netherlands makes use of biometric identification – the researcher's fingerprint – to

ensure that the researcher who is trying to connect to the facility is indeed the person he or she claims to be[7].

3 Data Access

A critical feature of data access is that the data be protected, as far as possible, from disclosure in the host, disseminating, institution. In earlier work, we have argued that this should be a multi faceted, or portfolio, approach[7]. It should include technical security, i.e., appropriate IT security protocols must be implemented. There also should be some statistical protection of the data that researchers use, as well as disclosure limitation review of researchers' output prior to publication. Researchers should be trained to develop an understanding of both confidentiality principles, as well as the basics of the data, such as sampling strategy, frame and weights. Finally a set of legal protocols should be implemented, to ensure that only authorized researchers, from trusted institutions, be permitted access, and that the research is done for statistical purposes only. A graphical description of this view is provided in Figure 2.

Figure2. The portfolio approach to confidentiality protection



The host disseminating institution should implement a portfolio approach to provide technical, educational, operational, statistical and legal tools to minimize the chance of a data security breach, while maximizing data access. The approach of using "sneaker-net", or physical enclaves, to permit access is not only costly and inefficient, but is somewhat embarrassing for social scientists, given the widespread use of cyberinfrastructure advances to permit remote ac-

cess by entities using far more sensitive data, such as the Department of Defense and the Department of Homeland Security, let alone financial institutions.

3.1 Technical approach

Remote access has begun to be more widely adopted because it allows geographically dispersed individuals to access the data in a controlled manner over the internet[5]. VPN technology for example, can structure access to prevent an outsider from reading the data transmitted between the researcher's computer and the host network. For example, the user in addition to using a pre-defined user id and password can also be required to use Smart Card technologies or biometrics so that users must validate his/her identity in real time. Other components of the VPN technology allow control to be established over which network resources the external researcher can access on the host network. Physical constraints on the researcher site can also be imposed, such as webcams, secure rooms, and electronic card entry.

Once the VPN connection has been established, the host can apply a variety of Web-based technologies to access the applications and data available on the host machine. One example is Citrix technology, which is based on thin-client, terminal/host based approach to computing. With this technology, widely adopted in the U.S. federal government, although all applications and data run on the server at the data enclave, the researcher still interacts with a full Windows graphical user interface. This means that the researcher never has to install any complex software applications on his/her remote computer. It also means that the host can prevent the user from transferring any data from data enclave to a local computer. For example, the host can configure Citrix so that data files cannot be downloaded from the remote server to the user's local PC. Similarly, the user cannot be able to use the "cut and paste" feature in Windows to move data from the Citrix session into an Excel spreadsheet sitting on the local computer. Finally, the user is prevented from printing the data on a local computer.

Through the Citrix technology, researchers are provided with a variety of powerful data analysis tools, such as statistical data processing packages (e.g. SAS, SPSS, or STATA). Researchers are also provided with visual tools that support online data and metadata searching, browsing and analysis. To enable high-quality use of the data, all data are delivered in tandem with the related metadata, providing the researchers with the necessary context for data analysis.

[5] See the cybertrust initiative at the National Science Foundation for a list of recent research in this area.

Although these approaches are beginning to be accepted by statistical agencies, there is ongoing work to advance the state of the art [8; 9; 10].

3.2 Operational Approach

The operational aspects are similarly important. Regardless of the type of data that is being disseminated, the data custodian is typically interested in ensuring that only trusted, approved and authorized researchers have access to the data – both because that reduces the risk of malicious disclosure and because it reduces the risk that the humans who are the source of the data will have negative perceptions associated with data access. So the host institution needs to establish a set of operational procedures to ensure appropriate access.

In addition, because providing access also results in a support burden, appropriate operational incentives should be put in place that accurately reflect the cost of support. Examples of such operational incentives could include charging the marginal cost of statistical disclosure review, charging for excessive storage costs, and charging a small weekly access fee to ensure that there is an incentive for projects to be completed in a timely fashion.

Similarly, since most data custodians want to provide access to data in order to promote data analysis, operational incentives should be put in place to promote analysis. This could include highlighting the work of particular researchers, instituting a working paper series, or, as discussed in a subsequent section, actively promoting the development of a virtual organization around a particular dataset.

3.3 Researcher training

Researcher training needs to be a core component of the portfolio approach for three reasons. One is to properly instill a "culture of confidentiality" within the researcher group, which is difficult to instill remotely. The second is so that the researchers and data producers get to know each and begin to develop the trust that is necessary to the development of a community of practice. The third is to develop their knowledge about the data, which reduces the likelihood that they will make errors resulting in incorrect analysis.

The confidentiality training should provide an overview of the legal background of each data source, and discusses the various disclosure definitions, as well as the principles of disclosure control, together with exercises to bring the ideas home. The training should focus on the importance of having safe projects (approved projects), safe people (i.e. authorized researchers), safe settings (ie. remote access) and safe conduct (care with handling and releasing data). The data producers should provide training on the data itself. This includes information about the data design, including information about frame and the correct use of weights.

Probably the most critical part of the training is to train researchers to do a preliminary disclosure review of their research prior to requesting formal release. This part of the training has several advantages. It reduces the most time and cost intensive part of providing access: the application of disclosure limitation protocols. It helps the researcher understand what type of information is needed before release is permitted. Finally, it sensitizes researchers to how much time is required to do a thorough disclosure review, and reduces the number and volume of requests.

3.4 Statistical Approach

There has been a great deal of research in this area, probably best summarized at a recent conference [11]. However, some best practices have been evolving. Even within an enclave, it is clear that at a minimum, the host should protect every data set by constructing a set of unique identifiers that can substitute for variables that are explicit personal/organizational identifiers, such as name, address, phone number, Social Security Number and Taxpayer Identification Number. The host should also be able to limit researchers' access to the data they need for their specific research questions if necessary. To accomplish this, the host should set up the capacity to create custom analytic data files that contain a subset of the columns (and even rows) contained in the master data set.

Most data producers typically remove some geographic detail, although that clearly limits the quality of the potential analysis. Memo fields also are typically removed since it is possible that individuals or companies may have conveyed identifiable information. The data producer should be made aware that each of these approaches can have an effect on the validity of social science analysis. In addition, the decision to apply them does not fully capture the costs to society of reduced data quality. A good discussion of the issues is provided in [7; 12] .

A major challenge to the data privacy community is the development of disclosure limitation techniques that are flexible enough to be used in a wide variety of situations. Considerable effort has gone into developing disclosure limitation methods for tabular data that effectively lower disclosure risk and provide products with high utility to legitimate data users. These techniques include cell suppression, local suppression, global recoding, rounding, and various forms of perturbation. However, protecting the confidentiality of qualitative, biomarker, brain scan and sensor information while providing useful information is still an under researched area[13].

3.5 Legal Framework

The legal and ethical issues are complex[14; 15]. In practical terms, the host should develop a Memorandum of Understanding with the data producer that describes the set of available parameters for providing external researchers with

access to the data set. The depositing agency in turn should determine which of the host's service offerings it wishes to utilize. Responses to the following questions lead to the development of business rules, for example:

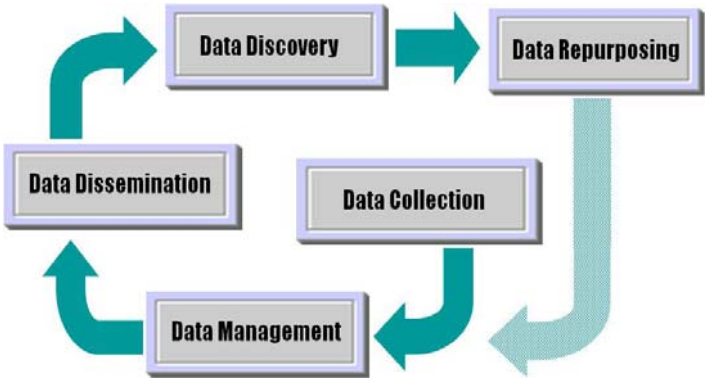
- Will the data enclave need prior approval from the agency before granting access to an external researcher who meets certain pre-defined requirements?
- Who at the agency will be the data enclave's representative responsible for approving all requests for data access?
- Will external researchers be able to create new data sets by combining data from multiple datasets housed in the enclave (including both data sets from a single agency and those from more than one agency)? What rules will govern this process?
- Will external researchers be able to create new data set by combining enclave data with data from other sources and, if so, what are the rules will govern this process?

4 Data Preservation

Another important feature that is a necessary component of data access, and one too often neglected, is that data should be stored in a way that preserves their life by making sure that they are indexed, archived and documented. A classic paper by Smith and Forstrom highlights the problems associated with trying to retrieve 35 year old survey data on American reactions to the Kennedy Assassination in order to make a comparison with reactions to the 9/11 tragedy -- the paper reveals the very real potential loss to social science of important datasets [16].

This point is illustrated by examining Figure 3. The collection and dissemination of data is only one part of the lifecycle of data – data need to be managed before they are disseminated, analysed and repurposed.

Figure 3. The Data Stewardship Lifecycle



Again, the social science community can take advantage of cyberinfrastructure advances in this field. There is an entire community devoted to this effort: the International Association for Social Science Information Service and Technology, and substantial investments in cyberinfrastructure to preserve the digital deluge [17; 18]. Probably the best known approach is the Data Documentation Initiative (DDI) standard which has been developed to index, archive and document all producer data files [19]. DDI, which originated in the Inter-university Consortium for Political and Social Research (ICPSR), and is now the project of an alliance of about 25 institutions in North America and Europe, has established an international standard for the content, presentation, transport, and preservation of micro-data documentation.

The DDI specification was designed to encompass various kinds of micro-datasets. It provides a comprehensive set of elements to be used to record and communicate in detail the characteristics of data obtained from sample surveys, censuses, administrative records and other systematic methodologies for generating empirical measurements.

The DDI is expressed as an XML Document Type Definition to maximize exchangeability and take advantage of the internet technology to share data and metadata,. In other words, the DDI encodes the metadata elements into a database following a standard structure and specification language. The DDI therefore facilitates interoperability as codebooks marked up using the DDI specification can be exchanged and transported seamlessly, and applications can be written to work with these homogeneous documents.

These elements—some of them mandatory, most of them optional— are structured into five sections:

- **Section 1.0** - Document Description consists of bibliographic information that can be considered as the header whose elements uniquely describe the full contents of the compliant DDI file.

- **Section 2.0** - Study Description consists of information about the data collection. This section includes information about who collected and who distributes the data, about the scope and coverage, sampling (if relevant), data collection methods and processing, citation requirements, etc.
- **Section 3.0** - Data Files Description provides information about the data file(s).
- **Section 4.0** - Variable Description provides a detailed description of variables, including (when relevant) the variable type, variable and value labels, literal questions, computation or imputation methods, instructions to interviewers, universe, descriptive statistics, etc.
- **Section 5.0** - Other Study-Related Materials allows for the inclusion of other materials related to the study such as questionnaires, user manuals, computer programs, interviewer manuals, maps, coding information, etc.

5 Data Community

Setting up a remote access environment also creates the opportunity to develop an environment that allows the sharing of information about data in the same fashion as that adopted by the physical and biological sciences, namely creating virtual organizations[6]. [20; 21]This not only serves the function of ensuring the generalizability and replicability of work that is fundamental to high quality research, but also promotes a healthy interaction between producers and researchers.

The opportunity is clear from the way in which ubiquitous information technologies have transformed many facets of human interaction and organization. Tools such as the Grid, MySpace, and Second Life have changed how people congregate, collaborate, and communicate. Increasingly, people operate within groups that are distributed in space and in time that are augmented with computational agents such as simulations, databases, and analytic services which

[6] Is a group of individuals whose members and resources may be dispersed geographically, but who function as a coherent unit through the use of cyberinfrastructure. A virtual organization is typically supported by, and provides shared and often real-time access to, centralized or distributed resources, such as community-specific tools, applications, data, and sensors, and experimental operations.

interact with human participants and are integral to the operation of the organization.

The study of virtual organizations is attracting attention in its own right as a way of advancing scientific knowledge and developing scientific communities. As Cummings et al. note [22].

"A virtual organization (VO) is a group of individuals whose members and resources may be dispersed geographically and institutionally, yet who function as a coherent unit through the use of cyberinfrastructure. A VO is typically enabled by, and provides shared and often real-time access to, centralized or distributed resources, such as community-specific tools, applications, data, and sensors, and experimental operations. A VO may be known as or composed of systems known as collaboratories, e-Science or e-Research, distributed workgroups or virtual teams, virtual environments, and online communities. VOs enable system-level science, facilitate access to resources, enhance problem-solving processes, and are a key to national economic and scientific competitiveness." (p1).

The social science community could potentially transform its empirical foundations if it moved away from the current practice of individual, or artisan, science, towards this more generally accepted community based approach adopted by the physical and biological sciences. It would provide the community with a chance to combine knowledge about data (through metadata documentation), augment the data infrastructure (through adding data), deepen knowledge (through wikis, blogs and discussion groups) and build a community of practice (through information sharing).

This opportunity to transform social science through adopting the type of organizational infrastructure made possible by remote access could potentially be as far-reaching as the changes that have taken place in the biological and astronomical sciences. It could lead to the "democratization of science" opening up the potential for junior and senior researchers from large and small institutions to participate in a research field. Already some such examples exist, such as nanoHUB (http://www.nanohub.org/) where the tools and collaboration approaches are determined by the community itself. It is, however, an open research question for the social science data community as to how such an organization should be established, how data should be accessed, how privacy should be protected, and whether the data should be shared on a central server or distributed servers. Some approaches are centralized, like the approach taken by the UK's ESRC in creating a specific call for a secure data archive[7] or decentral-

[7] http://www.esrc.ac.uk/ESRCInfoCentre/opportunities/current_funding_opportunities/ads_sds.aspx?ComponentId=25870&SourcePageId=5964

ized, like the U.S. National Science Foundation approach which lets the community decide. [8]. Certainly both the users and the owners of the data, whether the data be survey, administrative, transactions based, qualitative or derived from the application of cybertools, would need be engaged in the process

Similarly, it is an open research question as to the appropriate metrics of success, and the best incentives to put in place to achieve success[23]. However a recent solicitation[9] as well as the highlighting of the importance of the topic in NSF's vision statement[10], suggests that there is substantial opportunity for social science researchers to investigate the research issues.

6 Summary

This paper has argued that the new cyberworld has and continues to provide unprecedented opportunities for social scientists to capture and analyse data on human beings. The next immediate challenge is to provide access to and information about the new range of data and enable social scientists with an eye toward realizing a transformation in their science similar to that which has occurred in the biological and astronomical sciences. Two open research questions remain. One is: what are the appropriate technical, organizational, statistical and legal approaches that can be applied so that data on human subjects can be both protected and appropriately accessed? The second is: how can a virtual organization be developed and structured to permit knowledge sharing about data and the development of a data infrastructure?

References

[1] H. Brady, S. Grand, A. Powell, and W. Schink, Access and Confidentiality Issues with Administrative Data. in: M. VerPloeg, R. Moffitt, and C. Citro, (Eds.), Studies of Welfare Populations: Data Collection and Research Issues, Committee on National Statistics, National Research Council, Washington DC, 2001.

[2] J. Abowd, J. Haltiwanger, and J. Lane, Integrated Longitudinal Employee-Employer Data for the United States. American Economic Review 94 (2004) 224-229.

[8] http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141

[9] www.nsf.gov/pubs/2008/nsf08550/nsf08550.htm

[10] NSF Cyberinfrastructure Vision for 21st Century Discovery, March 2007

[3] R. Doll, and A.B. Hill, The mortality of doctors in relation to their smoking habits. British Medical Journal 4877 (1954) 1451-5.

[4] L. Calderwood, and C. Lessof, Enhancing longitudinal surveys by linking to administrative data, Centre for Longitudinal Studies Working Paper, University of Essex, 2006.

[5] T. Dunne, in the Establishment and Management of Secure Research Sites. in: P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, (Eds.), Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, North-Holland, Amsterdam, 2001.

[6] D. Boneh, J. Feigenbaum, A. Silberschatz, and R.N. Wright, PORTIA: Privacy, Obligations, and Rights in Technologies of Information Assessment. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 27 (2004) 10-18.

[7] J. Lane, Optimizing the Use of Micro Data. Journal of Official Statistics 23 (2007) 299--317.

[8] B. Mungamaru, H. Garcia-Molina, and C. Olston, Configurations: Understanding Alternatives For Safeguarding Data, Stanford InfoLab Publication number 2005-41, 2005.

[9] S. Forrest, J. Balthrop, M. Glickman, and D. Ackley, Computation in the Wild. in: K. Park, and W. Willinger, (Eds.), The Internet as a Large-Scale Complex System. SFI Studies in the Sciences of Complexity, Oxford University Press, 2005.

[10] G.S. Manku, Balanced Binary Trees for ID Management and Load Balance in Distributed Hash Tables, ACM Symposium on Principles of Distributed Computing, 2004, pp. 197-204.

[11] S.E. Fienberg, A. Anton, E. Bertino, C. Dwork, E. Viegas, and L. Zayatz, Workshop on Data Confidentiality, Arlington Virginia, 2007.

[12] J. Smith, Data Confidentiality: A Researcher's Perspective, American Statistical Association, Section on Social Statistics,, 1991.

[13] L. Zayatz, Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update Journal of Official Statistics 23 (2007) 253-265.

[14] A. Barth, A. Datta, J.C. Mitchell, and H. Nissenbaum, Privacy and Contextual Integrity: Framework and Applications, 27th IEEE Symposium on Security and Privacy, IEEE Computer Society, 2006.

[15] T.M. Weber, Values in a National Information Infrastructure: A Case Study of the US Census, 14th International Conference of the Society of Philosophy and Technology, Delft, The Netherlands, 2005.

[16] T. Smith, and M. Forstrom, In Praise of Data Archives: Finding and Recovering the 1963 Kennedy Assassination Study. IASSIST Quarterly Winter (2001).

[17] C. Humphrey, e-Science and the Life Cycle of Research, 2008.

[18] National Science Board, Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century, Arlington, VA, 2005.

[19] OECD, Data and Metadata Reporting and Presentation Handbook, OECD, Paris, France, 2007.

[20] L. Pang, Understanding Virtual Organizations. Information Systems Control Journal 6 (2001).

[21] I. Foster, C. Kesselman, and S. Tuecke, The Anatomy of the Grid: Enabling Scalable Virtual Organizations. International Journal of Supercomputer Applications 15 (2001) 200-222.

[22] J. Cummings, T. Finholt, I. Foster, C. Kesselman, and K. Lawrence, Beyond Being There: A Blueprint for Advancing the Design, Development and Evaluation of Virtual Organizations, 2008.

[23] J. Cummings, and S. Kiesler, Coordination costs and project outcomes in multi-university collaborations. Research Policy 36 (2007) 1620-1634.