Administrative Transaction Data

Lane, Julia

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
SSG Sozialwissenschaften, USB Köln

German Council for Social
and Economic Data (RatSWD)

www.ratswd.de

RatSWD
Working Paper Series

Working Paper No. 52

Administrative Transaction Data

Julia Lane

January 2009

Federal Ministry
of Education
and Research

Working Paper Series of the Council for Social and Economic Data (RatSWD)

The *RatSWD Working Papers* series was launched at the end of 2007. Since 2009, the series has been publishing exclusively conceptual and historical works dealing with the organization of the German statistical infrastructure and research infrastructure in the social, behavioral, and economic sciences. Papers that have appeared in the series deal primarily with the organization of Germany's official statistical system, government agency research, and academic research infrastructure, as well as directly with the work of the RatSWD. Papers addressing the aforementioned topics in other countries as well as supranational aspects are particularly welcome.

RatSWD Working Papers are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, and locally specialized journals. The *RatSWD Working Papers* are not available in bookstores but can be ordered online through the RatSWD.

In order to make the series more accessible to readers not fluent in German, the English section of the *RatSWD Working Papers* website presents only those papers published in English, while the the German section lists the complete contents of all issues in the series in chronological order.

Starting in 2009, some of the empirical research papers that originally appeared in the *RatSWD Working Papers* series will be published in the series *RatSWD Research Notes*.

The views expressed in the *RatSWD Working Papers* are exclusively the opinions of their authors and not those of the RatSWD.

Administrative Transaction Data

Julia Lane

Contact:

National Science Foundation
Science of Science & Innovation Policy
Social, Behavioral and Economic Sciences
4201 Wilson Blvd
Arlington, VA 22230
USA
jlane@nsf.gov

Abstract

The value of administrative transaction data, such as financial transactions, credit card purchases, telephone calls, and retail store scanning data, to study social behaviour has long been recognised. Now new types of transactions data made possible by advances in cyber-technology have the potential to further exland social scientists' research frontier.

This chapter discusses the potential for such data to be included in the scientific infrastructure. It discusses new approaches to data dissemination, as well as the privacy and confidentiality issues raised by such data collection. It also discusses the characteristics of an optimal infrastructure to support the scientific analysis of transactions data.

Keywords: transactions data; administrative data; cybertechnology; privacy and confidentiality; virtual organizations

IMPROVEMENTS AND FUTURE CHALLENGES FOR THE RESEARCH INFRASTRUCTURE:

"ADMINISTRATIVE TRANSACTION DATA"

JULIA LANE

The New Astronomy

"All astronomers observe the same sky, but with different techniques, from the ground and from space, each showing different facets of the Universe. The result is a plurality of disciplines (e.g., radio, optical or X-ray astronomy and computational theory), all producing large volumes of digital data. The opportunities for new discoveries are greatest in the comparison and combination of data from different parts of the spectrum, from different telescopes and archives."[1]

INTRODUCTION

The value of administrative transaction data, such as financial transactions, credit card purchases, telephone calls, and retail store scanning data, to study social behavior has long been recognized (Engle & Russell, 1998). Now new types of transactions data made possible by advances in cyber-technology have the potential to further expand social scientists' research frontier. For example, a person's interest and networks can be uncovered through the online behavior documented by the major search engines, such as Yahoo! and Google, as "data collection events"[2] Geographic movements can be tracked by cellphones which include GPS location information[3] Health, work and learning information can be tracked by the use of administrative data from hospital records, employment records and education records.[4] In sum, the new cyber-enabled ability to collect information from a wide variety of sources, which has transformed many disciplines, ranging from astronomy to medical science, can potentially transform research on social behavior.

Certainly the use of some transactions data for research and statistical purposes is becoming routine[5]. The Handbook of Survey Research will include a chapter on linking administrative records to survey data. The United Kingdom's Economic and Social Research Council has established An Administrative Data Liaison Service to link the producers of administrative data to

---

[1] *Links: NVO: http://www.us-vo.org/; IVOA: http://www.ivoa.net/*

[2] http://bits.blogs.nytimes.com/2008/03/09/how-do-they-track-you-let-us-count-the-ways/?scp=17&sq=privacy%20yahoo!&st=cse accessed Sept 19, 2008.

[3] http://www.nytimes.com/2008/06/22/technology/22proto.html?scp=3&sq=gps%20privacy&st=cse accessed Sept 19, 2008

[4] (Jones & Elias, 2006)

[5] The term "transactions data" is broadly used in this chapter to include administrative records which are "information that is routinely collected by organisations, institutions, companies and other agencies in order that the organisation can carry out, monitor, archive or evaluate the function or service it provides" (p2) (Calderwood & Lessof, 2006). The term as used here also includes the enormous amount of transactions datasets that are becoming available from, for example, credit card records, and stock trading, as well as the location information stored from cellular telephone and the clickstreams derived from online activity.

the academic community. And both the OECD and the Conference of European Statisticians are examining ways to use administrative data for the production of official statistics.

The opportunities are immense. Social sciences could be transformed by access to new and complex datasets on human interactions. The impact of social science on policy could be transformed as a result of new abilities to collect and analyze real time data. In addition, the funding exists: the United States has invested heavily in cyberinfrastructure[6] and the United Kingdom has established a National Centre for eSocial Science[7]. A good review of European Union activity is provided in a recent report by (Barjak, Lane, Procter, & Robinson, 2007)[8]

A number of important issues remain.

- What is the potential for new data (e.g. citation tracking, web-scraping, biomarkers, geospatial information, through RFID's and sensors, web-based social interactions) to be included in the scientific data infrastructure? How can such data be validated, analyzed, matched and disseminated?
- How have new approaches to data dissemination (e.g. protected remote access, combined with organizational, educational and legal protocols) advanced the potential for using transactions data in scientific research?
- What is the optimal infrastructure to promote the scientific analysis of administrative data – so that research can be generalized and replicated? What can we learn from the study of virtual organizations?

BACKGROUND

There has long been recognition in the research community of the value added of administrative data (Hotz, Goerge, Balzekas, & Margolin, 2000). The study of medical outcomes, for example, has been transformed by the use of administrative records[9]. The potential to examine the employment and earnings outcomes of low-wage workers is vastly expanded[10]. Of course, there are a number of challenges: a detailed discussion of the issues associated with using administrative data is provided in Lane (Lane, 2009).

Increasingly, statistical agencies are also using administrative records, because of the considerable pressures to keep costs down at the same time as creating new information. Indeed, the Public

[6] The Office of Cyberinfrastructure was established at the National Science Foundation in 2006.

[7] http://www.ncess.ac.uk

[8] http://ww3.unipark.de/uc/avross/

[9] (Skinner & Wennberg, 2000)

[10] (Autor, 2009)

Policy Program of the Washington Statistical Society, in partnership with the Federal Committee on Statistical Methodology's Subcommittee on the Statistical Uses of Administrative Records, is pleased has launched a seminar series on "Administrative Data in Support of Policy Relevant Statistics." More concrete examples are provided by the LEHD program in the United States[11], and the LEED program in New Zealand [12]. Because an infrastructure based on administrative records created a new sample frame for economic dynamics, it has been used in its own right to create new measures of workforce dynamics at detailed geography and industry ranging from earnings for incumbent workers, new hires, and separated workers to the number of quarters of non-employment of separated workers and measures of job retention and stability.

Another reason that the approach has been attractive is that administrative data have a breadth of information that is simply unattainable from other sources. For example, outside of manufacturing industries, the US. Census Bureau's measurement of inputs does not even distinguish between production and supervisory employees. After the implementation of the LEHD program, economic entities in all sectors (establishments or enterprises, as appropriate) , were used to create detailed summaries of the distribution of observable (demographic) and unobservable characteristics of the workforce in terms of earnings, external earnings potential and mobility.

Finally, administrative records shed new light on new economic structures. For example, using the LEHD program as an illustrative example, such data can be used to create new ways of classifying firms into particular industries based on worker activities (Benedetto, Haltiwanger, Lane, & McKinney, 2007); new ways of identifying the changing structures of firm mergers, acquisitions, births and deaths, based on worker flows(Benedetto et al., 2007); new approaches to providing place of work and industry coding on demographic surveys such as the American Community Survey (Freedman, Lane, & Roemer, 2008), more accurate and complete coding of individual outcomes (Abowd & Vilhuber, 2005) and new measures of demand side factors on household and individual surveys. Statistics on individual and household income and income mobility now include factors like whether the employer was growing or shrinking, whether the employer was profitable, and what other kinds of employees were also at the employer. (Andersson, Lane, & McEntarfer, 2005)

[11] http://lehd.did.census.gov/led/ accessed Seot 20, 2008

[12] http://www.stats.govt.nz/leed/default.htm

WHAT IS THE POTENTIAL FOR TRANSACTIONS DATA TO INFORM RESEARCH?

In 2006, the amount of digital information created, captured, and replicated [worldwide] was 1,288 x 1018 bits. In computer parlance, that's 161 exabytes or 161 billion gigabytes. This is about 3 million times the information in all the books ever written. [13] The sheer magnitude of this information means that this paper can only provide an illustrative, rather than exhaustive review of the types of data that can be collected and used to describe human behavior: here we describe what can be captured using RFID's, web archiving, web-scraping and datamining of electronic communications.

The potential to describe minute by minute human interactions with the physical environment became reality with the development of RFID (radio frequency identification devices) and video technologies. RFID's can be produced for pennies a unit and emit a wireless signal that enables the bearer to be tracked. Businesses now use the technology routinely to track employees (e.g. to ensure that night guards do their assorted tours at the assorted times) and to track their customer behavior (see Figure 1). The potential for social science research is clear – ranging from tracking time use information in a far more granular fashion than from survey data, to the environmental impacts on social behavior to measuring the number and quality of human interactions. In fact similar technologies are already being used for research purposes to great advantage. For example, Schunn uses video data collected from a recent highly successful case of science and engineering, the Mars Exploration Rover, to study the way in which human interactions contributed to the success of the project. While the project both wildly exceeded engineering requirements for the mission and produced many important scientific discoveries, not all days of the mission were equally successful. Schunn uses the video records to trace the path from the structure of different subgroups (such as having formal roles and diversity of knowledge in the subgroups) to the occurrence of different social processes (such as task conflict, breadth of participation, communication norms, and shared mental models) to the occurrence of different cognitive processes (such as analogy, information search, and evaluation) and finally to outcomes (such as new methods for rover control and new hypotheses regarding the nature of Mars).(Schunn, 2008)

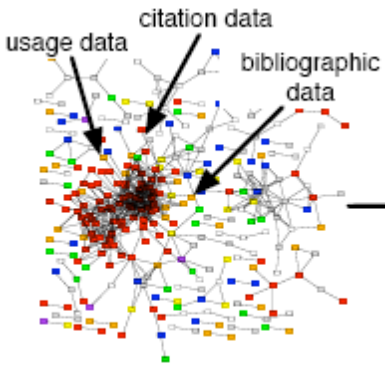
PARIS: Thousands of garments in the sprawling men's department at the Galeria Kaufhof are equipped with tiny wireless chips that can forestall fashion disaster by relaying information from the garment to a dressing-room screen. The garments in the department store, in Essen, Germany, contain radio frequency identification chips, small circuits that communicate by radio waves through portable readers and more than 200 antennas that can not only recommend a brown belt for those tweed slacks but also track garments from the racks, shelves and dressing rooms on the store's third floor. .. But the rapid development of RFID technology is also being regarded cautiously by the authorities in the European Union, who are moving quickly to establish privacy guidelines because the chips – and the information being collected - are not always visible. Their goal is to raise awareness among consumers that the data-gathering chips are becoming embedded in their lives - in items like credit cards, public transportation passes, work access badges, borrowed library books and supermarket loyalty cards.

Source: International Herald TribuneMarch 2, 2008

[13] The Expanding Digital Universe, March 2007, IDC White Paper sponsored by EMC Corporation

The Wayback Machine: http://www.archive.org/index.php

Of course, human behavior is increasingly captured through transactions on the internet. For example, most businesses, as well as registering with the tax authority, also create a website. It is now entirely possible to use web-scraping technologies to capture up to date information on what businesses are doing, rather than relying on administrative records and survey information. Historical records on businesses can also be created by delving into the repository of webpages on the Wayback Machine (see Figure 2 for an example of the webpages for Citibank). This archive takes snapshots of the web every two months and stores them in the manner shown, providing a rich archive of hundreds of billions of web pages. Individual as well as business behavior can be studied using this archive. Indeed, major NSF grants, such as the Cornell Cybertools award[14], have funded the study of social and information networks using these very large semi structured datasets.

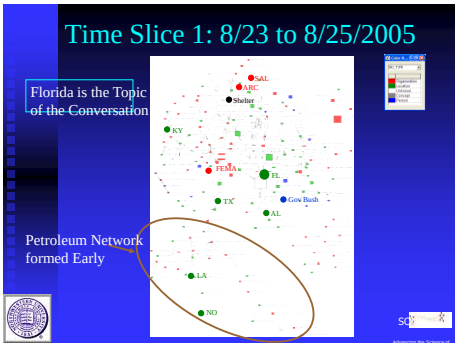


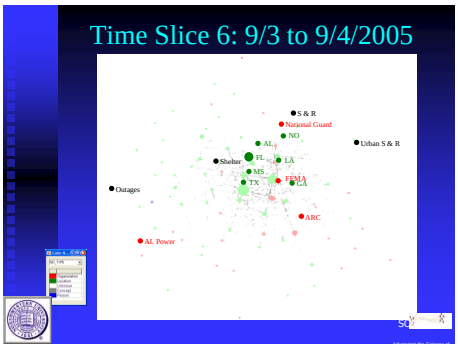
Other ways of collecting information on human behavior from the web include capturing clickstreams from usage statistics. The MESUR project[15], for example, has created a semantic model of the ways in which scholars communicate based on creating a set of relational and semantic web databases from over one billion usage events and over ten billion semantic statements. The combination of usage, citation and bibliographic data (see Figure 3) can be used to develop metrics of scholarly impact that go well beyond the standard bibliometric approaches used by academics. (Bollen, Rodriguez, & Sompel, 2007)

[14] Very Large Semi-Structured Datasets for Social Science Research, NSF award 0537606 http://www.infosci.cornell.edu/SIN/cybertools

[15] MESUR: Metrics from Scholarly Usage of Resources http://www.mesur.org/MESUR.html

A final illustration of the value of capturing transactions data, is evident from the work of Noshir Contractor. He studies a variety of ways in which humans interact with each other, including cell phone and email interactions. In a recent study he examined the emergency response of key agencies and individuals to Hurricane Katrina. The first slide in Figure 4 shows the result of analytical work based on the Data to Knowledge application at the National Center for SuperComputer Applications at the University of Illinois. This is a rapid, flexible data mining and machine learning system which allows automated processing by creating itineraries that combine processing modules into a workflow. This procedure was first applied to the body of communication between 8/23/2003 and 8/25/2005 (as Katrina was approaching Florida). An examination of the top panel of Figure xx shows the American Red Cross on the top. FEMA interactions only exist at FEMA Administration (Middle Left). Florida and Palm Beach have many mentions. At the bottom of the figure, it is clear that Oil and Power groupings are quite important, as is the pocket of National Parks in the middle. The location flags are heavily based in Florida, except for the Petroleum Network. New Orleans is very much on the fringe at the bottom.





The second slice of time that was examined was 9/3/2005 to 9/4/2005 – as the hurricane was hitting LNew Orleans. As is evident from the pictoral description of the analysis, Mississippi and Louisiana are the most frequently mentioned stats. Urban Search and Rescue has joined the network as a key concept. The topic of power has changed to Outages, Alabama Power is stil at the margin, and Shelter has moved back to the middle. FEMA and ARC have essentially swapped positions and the National Guard is moving towards the center. (Contractor, 2008).

This vividly illustrates how new approaches to capturing information could transform social scientists ability to provide information to policy makers. Imagine a similar exercise being done in the study of financial markets, for example. Real time data collected from the web analysis

of online blogs and newspaper articles could have picked up clusters of concern about "Lehman Brothers", "Goldman Sachs" and "Bear Stearns" and potentially described the information cascades that transformed the financial infrastructure in September and October of 2008. Or, in another example, new data could be collected on the innovation processes that generate competitive advantage within firms. [16]

Of course, together with new data, new analytical techniques need to be developed. Standard regression analysis and tabular presentations are often inadequate representations of the complexity of the underlying data generation function. There are a variety of reasons for this inadequacy. First, the units of analysis are often amorphous – social networks rather than individuals, firm ecosystems rather than establishments. Second, the structural relationships are typically highly nonlinear, with multiple feedback loops. Third, theory has not developed sufficiently to describe the underlying structural relationships, so "making sense" of the vast amounts of data is a substantive challenge. There has been substantial effort invested in developing new models and tools to address the challenge, however. For example, since a major national priority is understanding the formation and evolution of terrorist networks through the internet and other communication channels, substantial resources have been devoted to the field of visual analytics. Their research agenda aligns very closely with a potential research agenda for social scientists, focusing as it does on the science of analytical reasoning, visual representations and interaction techniques, data representations and transformations, as well as the production, presentation and dissemination of complex relationships. (Thomas & Cook, 2005) It is also worth noting that new partnerships are being formed to address the nontrivial computing challenges[17].

The Effect of New Data Dissemination Protocols

Both transactions and administrative data are often highly sensitive. The dissemination of such data is, however, critical for a number of reasons. The first is that 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 building a collective knowledge base, particularly with new data whose statistical properties are unknown. The fifth is capacity building. Junior researchers, policy makers and practitioners need to have the capacity to go beyond examining tables and graphs and develop their understanding of the complex response of humans to rapidly changing social and legal environments. Access to complex micro-data provides an essential platform for evidence based decision-making. Finally,

---

[16] http://www.conference-board.org/nsf. Carol Corrado "Workshop on developing a new national research data infrastructure for the study of organizations and innovation"

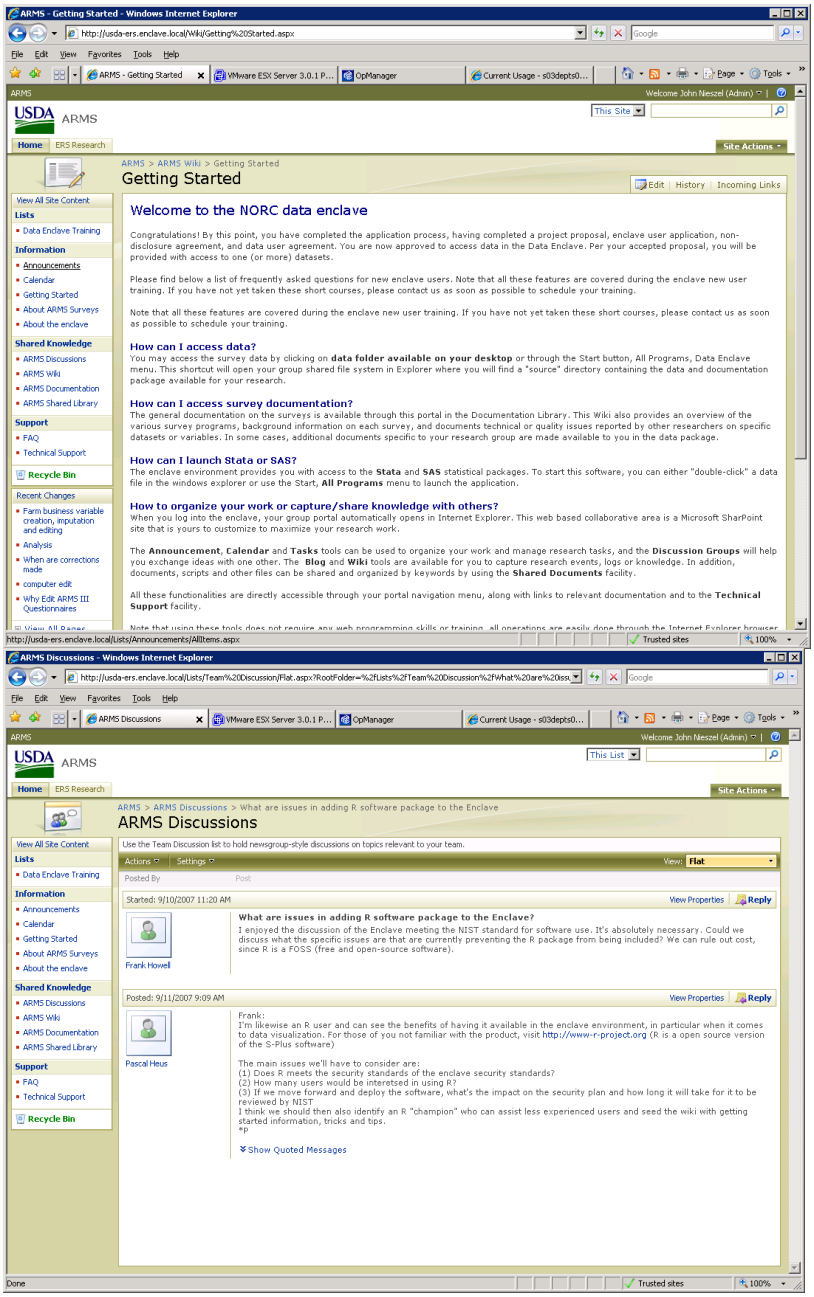
[17] http://www.nsf.gov/news/news_summ.jsp?cntn_id=111470

access to micro-data permits researchers to examine outliers in human and economic behavior – which is often the basis for the most provocative analysis.

A major barrier to the use of administrative data is the difficulty of getting permission to use administrative data for purposes other than which it was collected. This is an extremely time consuming process: since the data are collected to administer programs and not for research purposes. Legal, ethical and financial issues similarly act to restrict access.

However, new data dissemination protocols are being developed. Remote access 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 (Lane, Heus, & Mulcahy, 2008). The Office for National Statistics (ONS) (Ritchie, 2005) for example, instituted a full "remote laboratory" service in January 2004. Their approach is to use a thin client service, which means there is no data transfer at the user end. They have also centralized data management operations, which makes it much more efficient to work across different sites. Statistics Denmark(Borchsenius, 2005) has found that remote access arrangements are now the dominant mode of access to microdata. Statistics Sweden's system for remote access to microdata (MONA(Söderberg, 2005)) provides users with secure access to databases at Statistics Sweden from almost any place with internet access. In this manner, 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. All application processing is done in the server. Statistics Netherlands(Hundepohl & de Wolf, 2005) has gone even further in terms of its remote access. It has begun a pilot project, called the OnSite@Home facility[i] which 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.

The NORC data enclave has taken the remote access approach one step further. It recognizes that a remote access environment also permits the development of 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. (Foster, Kesselman, & Tuecke, 2001; Pang, 2001). Tools such as the Grid, MySpace, and Second Life have changed how people congregate, collaborate, and communicate: the NORC enclave offers social scientists the same opportunities. Promoting virtual collaboration 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 data collectors, data producers and data users. In particular, the NORC enclave allows multiple people on a team access to the data, and team members are set up with individual workspaces that are complemented by team workspaces. Each workspace allows the user to save their result sets and related notes. NORC supports the ongoing collaborative annotation of data analysis and results through wikis and blogs and discussion spaces. There is also a group portal environment that enables the collaborative development of research deliverables such as journal articles. Figure 5 gives a visual idea of the enclave approach



The social science community could potentially transform its empirical foundations if it adopted such a collaborative framework. It could use remote access to a common dataset to move away from the current practice of individual, or artisan, science, towards the more generally accepted community based approach adopted by the physical and biological sciences. Such an approach 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). 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 astronomical

sciences, and cited in the opening section. 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.

However, it is worth noting that the establishment of a virtual community to advance the development of a data infrastructure is a social science challenge in its own right: indeed, 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. (Cummings, Finholt, Foster, Kesselman, & Lawrence, 2008) note

"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).

It is clearly 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[18] or decentralized, like the U.S. National Science Foundation approach which lets the community decide. [19]. 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(Cummings & Kiesler, 2007). However a recent solicitation[20] as well as the highlighting of the importance of the topic in NSF's vision statement[21], suggests that there is substantial opportunity for social science researchers to investigate the research issues.

---

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

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

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

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

3) ETHICS AND PRIVACY ISSUES

A related social science research challenge that the new cyber-technologies pose, as well as potentially help to solve, is the ethical issues raised by the new capacities to collect data on human beings, particularly a focus on the privacy and confidentiality issues raised by collecting data on the interaction of human subjects.

The philosophical issues are well summarized by Madsen (Madsen, 2003). He identifies a "privacy paradox" in confidentiality research – which occurs when data managers, in interpreting the right to privacy very narrowly, results in less social benefit, rather than in more. Two factors contribute to this paradox. One is the fear of a pan-opticon society, in which an all-seeing few monitor the behavior of many, which has been exacerbated since Sept 11, 2001. The second is a fundamental uncertainty about data ownership – whether data constitute private or public property. It is possible that the tension in the core paradox results from a framework which simply includes rights and responsibilities into the decision-making mix, rather than including social utility. But much more research must be done in this area.

The second set of issues is economic in nature(Lane, 2003). Given the clear public good aspects of data collection and dissemination, how can the costs and benefits of the social investment in data be tallied to identify the optimal level of data collection? A partial list of the social benefits would include: improved decision making, avoidance of the moral hazard associated with monopoly government control of information, and improved data quality. A similar list of the social costs would include legal sanctions, the cost of breaches of confidentiality (which might substantially reduce data quality), and support costs. Simply refusing to collect and analyze data which could inform public decision making – and have tremendous public benefit, may not be a socially optimal decision.

Also of interest is how to convey the quality of such confidentiality measures to the humans who are the subject of study. Social scientists could expand their current interest in confidentiality to develop approaches that ensure the collaboration and engagement of individuals and organizations in providing data to the research community, as well as permit the data to be shared so that empirical analyses can be generalized and replicated.

It is worth noting that there is increasing interest by computer scientists in ways in protecting confidentiality so that sensitive data can be collected and analyzed without revealing individual identities – and so that researchers can generalize and replicate scientific results[22]. This interest includes policies for the anonymization and sanitization of the data, retention and storage protocols, transformation prior to dissemination and retaining usability.

[22] http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5033268&org=CNS

4) RECOMMENDATIONS

The social science community should act to address these challenges. Some work is already being done, such as the work by Peter Elias on behalf of a number of international agencies to establish the International Data Foundation. However, specific, targeted, activities could be undertaken to develop a new social science data infrastructure capable of answering new scientific and policy issues.

Recommendation 1: Invest in new methods of collecting transactions data

The community should take advantage of the interest of funding agencies in funding cyberinfrastructure for the social sciences to collect new data sources. These would include clickstream information, data from webarchives, email transactions, firm administrative records, social interactions in cyberspace (such as Facebook and MySpace) and video data. The social science community should partner with data collectors, such as Google, Yahoo!, Facebook and the business community to create joint value.

Recommendation 2: Invest in new ways of analyzing transactions data

The social science community should recognize that while new units of interest to social scientists can now be studied, such as social networks, there are a number of analytical challenges. The units of analysis are amorphous and change rapidly over time. The information that is collected is no longer precisely measured: there is a high noise to signal ratio. There are large amounts of heterogeneous data. The social science community should partner with other disciplines to develop new analytical techniques. Computer and behavioral scientists have substantial expertise in creating analytical datasets in this environment; the visual analytics community has and is developing experience in "making sense" of such data.

Recommendation 3: Invest in new ways of disseminating transactions data

In order to develop the scientific basis for studying transactions data, the social science community needs to develop an open and transparent data infrastructure. A scientific dialogue needs to be developed about the establishment of a scientific frame, the integrity of the data and the validation of results. In other words, social scientists must join the "hard" sciences in ensuring that their work is generalizable and replicable (i.e. scientific). A number of remote access sites are being established by leading data disseminators, such as the NORC data enclave, the UK ESDS and CESSDA that promote the development of virtual organizations around data. These new access modalities offer the social sciences a way of creating virtual organizations that have new ways of collecting, accessing and analyzing transactions microdata.

Recommendation 4: Invest in new ways of conveying complex information

The social science community should invest in new ways of conveying complex information to the broader policy making and lay communities. Tabular techniques may no longer adequately provide

sufficient clarity: further investment in such visualization techniques as maps and and graphs is warranted

REFERENCES

Abowd, J., & Vilhuber, L. (2005). The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers. *Journal of Business and Economic Statistics, 23*(2), 133-152.

Andersson, F., Lane, J., & McEntarfer, E. (2005). Successful Transitions out of Low-Wage Work for Temporary Assistance for Needy Families (TANF) Recipients: The Role of Employers, Coworkers, and Location, *http://aspe.hhs.gov/hsp/low-wage-workers-transitions04/index.htm* (Vol. http://aspe.hhs.gov/hsp/low-wage-workers-transitions04/index.htm). Washington DC: U.S. Department of Health and Human Services.

Autor, D. (2009). *Labor Market Intermediaries*: NBER/University of Chicago Press.

Barjak, F., Lane, J., Procter, R., & Robinson, S. (2007). *Accelerating Transition to Virtual Research Organisation in Social Science (AVROSS)*. Brussels, Belgium: European Union.

Benedetto, G., Haltiwanger, J., Lane, J., & McKinney, K. (2007). Using Worker Flows in the Analysis of the Firm. *Journal of Business and Economic Statistics, 25*(3), 299-313.

Bollen, J., Rodriguez, M., & Sompel, H. V. D. (2007). *MESUR: usage-based metrics of scholarly impact.* Paper presented at the Joint Conference on Digital LIbraries, Vancouver.

Borchsenius, L. (2005). New Developments In The Danish System For Access To Microdata, *Joint UNECE/Eurostat work session on statistical data confidentiality*. Geneva, Switzerland.

Calderwood, L., & Lessof, C. (2006). Enhancing longitudinal surveys by linking to administrative data, *Centre for Longitudinal Studies Working Paper*: University of Essex.

Contractor, N. (2008). CI-KNOW: A Tool for Understanding and Enabling the Transformative Power of Cyberinfrastructure in Virtual Communities, *Presentation at the National Science Foundation, September 15, 2008*.

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

Cummings, J., & Kiesler, S. (2007). Coordination costs and project outcomes in multi-university collaborations. *Research Policy, 36*(10), 1620-1634.

Engle, R., & Russell, J. (1998). Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data. *Econometrica, 66*(5), 1127-1162.

Foster, I., Kesselman, C., & Tuecke, S. (2001). The Anatomy of the Grid: Enabling Scalable Virtual Organizations. *International Journal of Supercomputer Applications, 15*(3), 200-222.

Freedman, M., Lane, J., & Roemer, M. (2008). New Data for Economic Geographers. *Journal of Official Statistics*.

Hotz, J., Goerge, R., Balzekas, J., & Margolin, F. (2000). *Administrative Data for Policy-Relevant Research: Assessment of Current Utility and Recommendations for Development*. Evanston, Illinois: Northwestern University/University of Chicago Joint Center for Poverty Research.

Hundepohl, A., & de Wolf, P.-P. (2005). OnSite@Home: Remote Access at Statistics Netherlands, *Joint UNECE/Eurostat work session on statistical data confidentiality*. Geneva, Switzerland.

Jones, P., & Elias, P. (2006). *Administrative data as research resources: a selected audit*. Swindon.

Lane, J. (2003). The Uses of Microdata: Keynote Speech, *Conference of European Statisticians*. Geneva, Switzerland: UNECE.

Lane, J. (2009). Administrative and Survey Data. In P. Marsden & J. Wright (Eds.), *Handbook of Survey Research*. Oxford: Oxford University Press.

Lane, J., Heus, P., & Mulcahy, T. (2008). Data Confidentiality in a Cyber World. *Transactions in Data Privacy*(1).

Madsen, P. (2003). The Ethics of Confidentiality: The Tension Between Confidentiality and the Integrity of Data Analysis in Social Science Research, *NSF Workshop on Confidentiality Research*. Arlington, Virginia.

Pang, L. (2001). Understanding Virtual Organizations. *Information Systems Control Journal, 6*.

Ritchie, F. (2005). Access to Business Microdata in the United Kingdom, *Joint UNECE/Eurostat work session on statistical data confidentiality*. Geneva, Switzerland.

Schunn, C. (2008). Integrating Social and Cognitive Elements of Discovery and Innovation. In N. S. Foundation (Ed.) (Vol. Award number 0830210).

Skinner, J., & Wennberg, J. (2000). Regional Inequality in Medicare Spending: The Key to Medicare Reform? In A. Garber (Ed.), *Frontiers in Health Economics*: MIT Press.

Söderberg, L.-J. (2005). MONA – Microdata On-Line Access At Statistics Sweden, *Joint UNECE/Eurostat work session on statistical data confidentiality* Geneva, Switzerland.

Thomas, J., & Cook, K. (2005). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*.

[i] Hundepohl, Anco and Paul-Peter de Wolf "OnSite@Home: Remote Access at Statistics Netherlands", paper presented at the Joint UNECE/Eurostat work session on statistical data confidentiality (Geneva, Switzerland, 9-11 November 2005)