DATA SERVICES
SHARE AND SHARE ALIKE? DATA-SHARING PRACTICES IN DIFFERENT DISCIPLINARY DOMAINS

JoAnn Jacoby

INTRODUCTION

In this paper, I review the literature on the data-sharing practices of social science researchers to examine when and how data is shared within a particular research community. Data-sharing practices are analyzed in the context of broader disciplinary norms, participation in data archives (which can be interdisciplinary, multidisciplinary, or discipline-specific), and new data-sharing policies instituted by funding agencies. Drawing on three case studies (anthropology, economics and population studies), a model of factors shaping norms is proposed to predict when and how data is shared within a particular research community. The implications different data-sharing practices have for interdisciplinary scholarship – and for libraries serving researchers across these domains – are also considered.

My interest in this topic derives from my involvement as a Library Liaison in a multi-institutional grant (“Investigating Data Curation Profiles Across Multiple Research Disciplines” IMLS/LG06-07-0032). Bringing together librarians and researchers from the University of Illinois Library, Purdue University Libraries, and the Graduate School of Library and Information Science at the University of Illinois, the Data Curation project examines scholarly practice related to data collection, management, publication and preservation, including “at which point in the research cycle are researchers willing to share data, with whom, and under what conditions?” The study also explores the role of academic libraries in supporting e-science activities, by studying how librarians can interact with scientists to make their research output available, identifying practices and tools to support further metadata development and capture workflow. The present paper draws on interviews with social scientists conducted as part of this study, but primarily serves as a literature review and analysis of previously published literature as a backdrop for interpreting current local practice.

DATA SHARING IN THE SOCIAL SCIENCES: AN OVERVIEW

Data-sharing is a topic of growing interest in the Library and Information Science (LIS) literature as libraries seek to define their role with regard to data stewardship and preservation as part of broader e-scholarship and scholarly communications initiatives (Association of Research Libraries, 2006; Research Information Network, 2008). Attention to these issues has been also spurred by data-sharing
policies recently instituted by funding agencies seeking to ensure that their investment in research has the greatest possible impact. Many funding agencies, both public and private, are starting to either encourage applicants to specify a plan for sharing their data (e.g., the National Science Foundation, OECD), require such a plan for grants of a certain size (e.g., the National Institute of Health, for grants over $500,000/year) or mandate deposition in a centralized repository for awards of all sizes (e.g., Wellcome Trust). These requirements are intended to make data produced in the course of sponsored research more widely available, facilitate the validation of results, and enable secondary analysis and meta-analysis—thereby ensuring that the research investment has broad and sustainable impact. The policies, however, must be backed by an appropriate infrastructure in order to produce the intended result. As some leading data archivists have cautioned, “data sharing rules may create more problems that they solve, because they can lead to a proliferation of web sites for self dissemination by researchers...which place long term access and preservation into question” (Gutmann, Schüer, Donakowski and Beedham, 2004, p. 219).

The existing infrastructure to support data sharing is uneven: highly developed in some areas and completely lacking in others. The greatest challenge, however, is cultural—open and systematic data sharing is simply not a part of the norms and habits of researchers in many fields:

Many research funders are putting policies in place to ensure that datasets judged to be potentially useful to others are curated in ways that allow discovery, access and re-use. But there is not a perfect match between cultural norms in some research disciplines and funder requirements. Some disciplines are well ahead of funding bodies in that they have had a culture of sharing data for a long time and have developed the infrastructures and methods for doing this. In other disciplines, data sharing is not commonplace and therefore funder policies may imply significant modifications to researchers’ attitudes and behaviour.” (Research Information Network, 2008, p. 12)

As an example of a research area where funding agency policies and scholarly practices do not yet mesh, a biological anthropologist at the University of Illinois reports that he is trying to figure out how to best comply with the U.S. National Science Foundation’s (NSF) strong encouragement to include a data sharing plan in grant applications. While he had previously self-archived his data on a server in his lab and shared it informally among his graduate students and a small circle of colleagues, the thought of someone pulling the plug on the local server has long been worrisome. He is eager for assistance as he begins to explore the options for a systematic and robust approach to sharing and archiving his research data. He recognizes that a more systematic approach to sharing his data will make his research more prominent, help the field by allowing for knowledge to be built incrementally and resolve his concerns about the vulnerability of his local server. Without an established path, however, he is faced with navigating a host of decisions about how and where to deposit his data on an individual, ad hoc
basis. No existing repository maps to his needs: the biometric data falls outside of the Inter-University Consortium for Political and Social Research’s (ICPSR) collecting profile (and ICPSR is not known in his field, so data deposited there would not be visible to his colleagues) and none of the existing health-related repositories are well-aligned with his specific needs. Noting that older colleagues in this research area generally do not share their data, he finds himself at the vanguard of a new practice with patchy infrastructure and little to guide him beyond the funder requirements (Personal Communication, Spring 2007).

In contrast, other social science fields have a long established history of sharing research data (Feinberg, Martin and Straff, 1985). Researchers in disciplines like political science and sociology, particularly those engaged in quantitative research, routinely deposit data in large central repositories such as ICPSR, which was founded in 1962 and is now the world’s largest archive of digital social science data. Repositories like ICPSR, United Kingdom Data Archive (UKDA), and the Council of European Social Science Data Archives, in turn, have established routines for dealing with ingest, metadata (including a newly updated Data Document Initiative, standard, DDI)\(^{n}\) where \(n=2,\) sampling, confidentiality and permissions throughout the data lifecycle. The solutions established by repositories like ICPSR have not been adopted by all social scientists, however, nor are they necessarily suited to the needs of all researchers. Repositories like ICPSR are well-equipped to deal with numeric data collected via survey or census, but there are many other forms of social science data with different infrastructure requirements. Not only do these data come in a wide array of formats (with the attendant technical considerations), they also come from distinctly different epistemological contexts that have implications for confidentiality, contextualization and interpretation of the data.

To Share or Not to Share: Publication and Quality Assurance of Research Data Outputs, reports the results of a large-scale study of data sharing practices in the United Kingdom (Research Information Network, 2008). This project’s underlying goal was to “discover what motivates researchers to publish their data and, for those who choose not to, what factors inhibit them” (Research Information Network, 2008, p. 13). They conclude that the “lack of uniformity across different research disciplines in terms of behaviour, policies or needs [means that] any solutions to the problems we identify, therefore, will need to be tailored to the requirements and practices of each individual research discipline. Interdisciplinary research needs especially careful consideration in this light.” (Research Information Network, 2008, p.13).

The RIN study found that “there are two main ways of storing and curating data – using large, centralised national or international data centres; or using a distributed array of local data stores (based on or in research institutions, researchers’ own resources, or formal publication outlets such as journals)” (Research Information Network 2008, p. 12). Examples of the former mentioned, as discussed above, include ICPSR, United Kingdom Data Archive (UKDA),
and the Council of European Social Science Data Archives. Distributed, ad hoc data sharing includes dissemination via journals (*World Cultures, Economic Modeling*), websites focusing on a particular field (The Paleobiology Database <www.paleodb.org>) or stored on a local or institutional server, or on diskettes, zip-drives or other media, in a researcher’s office. Carlson (2006) suggests that there is more data in small, heterogeneous “repositories” (or on zip drives, floppies, etc.) than in centralized repositories. This long tail is one of the biggest challenges facing librarians grappling with data curation and stewardship.

Gutmann et al. (2004) have identified some of the key factors mitigating against the deposition of social science data in a shared repository:

− Reluctance to share data with potential competitors; proprietary attitudes
− Lack of time or expertise to prepare data for sharing
− Confidentiality issues
− Rewards accrue more to new research than to preserving or reusing existing data
− Data producers unaware of where or how to archive data
− Data are forgotten or set aside once funding ends and publications are written

**WHAT IS SOCIAL SCIENCE DATA?**

Social science data comes in many forms. In a succinct overview of the issues surrounding selection and retention of digital social science data (Gutmann et al., 2004, p. 210) delineate the major types of social science data, while Altman (2008) suggests a slightly different taxonomy (Table 1).

Table 1: What is Social Science Data?

<table>
<thead>
<tr>
<th>Gutmann et al. 2004</th>
<th>Altman 2008</th>
<th>Additional Types of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Survey data:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>− Categorical or close-ended responses</td>
<td>Raw measurements</td>
<td>Fieldnotes (Silverman &amp; Parezo, 1995)</td>
</tr>
<tr>
<td>− “Full-text” responses to open-ended questions</td>
<td>Numeric tables</td>
<td>Texts</td>
</tr>
<tr>
<td>− Administrative records (&amp; email)</td>
<td></td>
<td>Models (e.g., <em>Economic Modeling</em>)</td>
</tr>
<tr>
<td><strong>Nonsurvey data:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>− Images</td>
<td>Video &amp; audio interviews, transcripts, blogs</td>
<td>Artifacts, samples, specimens, other physical objects (Gould &amp; Handler, 1989)</td>
</tr>
<tr>
<td>− Sound</td>
<td></td>
<td></td>
</tr>
<tr>
<td>− Video</td>
<td></td>
<td></td>
</tr>
<tr>
<td>− Multimedia, etc.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Ranging from abstract models to physical objects, the data types listed in Table 1 provide some indication of the heterogeneity of data sources, but this list is far from exhaustive. Indeed, what counts as data across the social sciences is almost boundless. While physical objects (or even sound and video recordings) may challenge the usual categorization of data as either numerical tables or textual documents, these artifacts are quite familiar to museums, archives and special collections who have established mechanisms for appraising, cataloging and archiving these materials. They are, however, quite distinct from numeric data and present unique challenges in terms of data sharing, storage, and preservation. Clearly, the types of “data” listed in Table 1 will not easily fit into the same container, be it physical or metaphorical.

CASE STUDIES: DATA IN ANTHROPOLOGY, ECONOMICS, AND POPULATION STUDIES

To illustrate the diversity of data types and how data creation and dissemination fits into research and scholarly communications practices, three case studies are examined in depth: anthropology, economics and populations studies.

Anthropology

Anthropology encompasses four distinct subfields: sociocultural, linguistic, archaeological and biological anthropology. The varying foci of these four subfields are united by a holistic approach to the study of humankind, as well as the hallmark of anthropological research – a methodology grounded in fieldwork or direct, in situ, observation. In this respect, anthropologists closely resemble ecologists who also go into “the field” to gather data and whose work is inextricably grounded in a particular place and time. Thus Zimmerman’s (2003) insights into data sharing practices among ecologists also pertain to anthropologists:

Fieldwork performs an important function in shaping ecologists’ formal and informal knowledge, which carries over to their reuse of data. The informal knowledge ecologists acquire as collectors of their own data in the field or laboratory plays the most important role in their reuse of data. The secondary use of data on a large scale requires a greater emphasis on standardization, peer review, and quality control, which alters the extent of reliance on informal knowledge. However, a formal system offers only some of the information that scientists require to reuse data, and there is a danger in thinking that informal knowledge is easily replaced and is no longer necessary or important. (Zimmerman 2003, p. 220).

The importance of context and informal knowledge in these fields mean that data are not always easily portable. In addition, issues of pattern and scale come into play when attempting to synthesis data from multiple field sites.
Types of data collected and used by anthropologists include fieldnotes (which may include discursive accounts, censuses, genealogies, taxonomies, lexical compilations, sound recordings, maps, texts or other documents) objects/artifacts, images, spectrographs, samples, or specimens (Silverman and Parezzo, 1995; White, 1991). What counts as data in anthropology is both diverse and particular – virtually anything relevant to a particular human context.

In general, anthropologists (especially sociocultural anthropologists) are not in the habit of sharing their original data beyond a trusted network of colleagues and graduate students. Factors influencing this proprietary attitude toward data include the fact that context and informal knowledge is essential in interpreting the data, as was the case with the ecologists Zimmermann (2003) studied. In addition, confidentiality (in the case of sociocultural anthropologists working with living informants) and the need to handle material remains with sensitivity and respect (in the case of archaeologists and biological anthropologists) complicates open data sharing. Finally, gathering anthropological data is laborious and time-consuming which can contribute to a strong feeling of personal ownership (Gould and Handler, 1989; White, 1991; Silverman and Parezo, 1995).

Data sharing is becoming increasingly common in archaeological and biological anthropology, mostly in small-scale repositories serving well-defined research areas, such as ALFRED (Allele Frequency Database < alfred.med.yale.edu>) and Paleobiology Database <www.paleodb.org>). The National Archaeological Database <http://www.nps.gov/history/archeology/tools/nadb.htm> brings together information on archaeological sites on public lands, and may help establish standards and protocols that facilitate sharing data more broadly. It is worth noting that fieldwork in archaeology and biological anthropology is often conducted by a team, necessitating common standards for collecting and recording data, whereas sociocultural anthropologists tend to work solo. Moreover, the primary vehicle of communication among sociocultural anthropologists, the ethnography, inextricably interweaves observation and analysis. This complicates the status of data as something that can be isolated from the interpretive context and then shared in any meaningful way. It is not surprising, then, that sociocultural anthropologists generally do not share data beyond a small circle of close collaborators, with the notable exception of the small community of scholars engaged in cross-cultural research who contribute to repositories like World Cultures and eHRAF (White, 1991).

Economics

Economists in the academy use data published by local, regional and national governments (Dee, Evans & Murray, 1991), as well as data from financial exchanges and banks. Corporate data is also of interest, but is not generally available in the public domain (Gould & Handler, 1989). Economists are voracious
users of data gathered by others, in clear contrast to anthropologists who almost exclusively rely on data they have gathered themselves through fieldwork. While economists do not generally gather original data, they may generate unique “micro-data” or data disaggregated from large-scale data sets and broken down by finer level detail such as, expenses by household or imports of specific models of cars by income, education level, or neighborhood (Gould & Handler 1989, p. 10-11). They also derive models from data – these models are shared in publications and preprints, and can, themselves, be considered a form of data.

Growing interest in replication led some economics journals to mandate that complete microdatasets be made available in the public domain, and the requisite repository system was developed in cooperation with the NSF and ICPSR starting in the late 1980s. Prior to this, the norm was not to share datasets resulting from secondary analysis. Thus economists have long made use of data that was freely available, but it was not until the late 1980s that they began to systematically share the microdata generated through individual analysis (Kane, 1984; Fienberg, Martin & Straf, 1985; Freese, 2007).

But even now, many researchers are not in the habit of sharing microdata, especially those files not tied to a published study in a journal that requires data deposition. An economics professor at a small liberal arts institution in the D.C area reports posting his datafiles on his campus course management system to share with his graduate students and stores his archived data on floppies, zip drives and disks of various vintages that can be read on a computer he keeps “somewhere around here, I’m not sure exactly. I think the secretary knows” (personal communication, May 2008).

In economics, as in other fields trying to predict the outcomes of complex phenomena or discern laws or patterns (e.g., physics, atmospheric science), models themselves can be form of data. Indeed, these models are widely published and disseminated in peer-reviewed publications such as Economic Modeling. It is interesting to note that like physics, economics has developed a robust system of disciplinary repositories, including the Social Science Research Network <http://ssrn.com/> and Research Papers in Economics (RePEcN) <http://repec.org/>, where most papers in the field are first published as preprints prior publication in peer-reviewed journal.

**Population Studies (Demography)**

The field of population studies or demography focuses on fertility, migration, and household composition. Evans (1991) describes population studies as a “bridge discipline that organizes the creative efforts of diverse scholars into large scale data collection.” The field is methodologically diverse, involving qualitative as well as quantitative data and drawing together researchers from health and epidemiology, social psychology, anthropology and demography.
Like economists, researchers in population studies make use of data gathered by national governments and international organizations. Commonly used data sets in population studies include the World Fertility Survey, General Social Survey (GSS) and census data (Gould and Handler, 1989). Researchers in this area may also be involved with data gathering and sampling for large-scale population studies. Demographers, in particular, have a long tradition of data sharing – they both cooperate to gather data to be used for administrative purposes and they share microdata derived from secondary analysis. More complicated, however, is sharing data across these communities of practice involved in population research, which can involve distinct methodological approaches and domain knowledge from areas as diverse as community health, anthropology, and demography (Evans, 1991).

DATA SHARING PRACTICES: A MODEL

The case studies discussed above suggest that the different research practices – how information is gathered, analyzed and communicated – in these three fields form a constellation of factors that shape the data sharing norms in those disciplines. The question then becomes, what are the common patterns shaping these informal and formal practices? Can these patterns be used to better understand the factors likely to influence data sharing norms in different fields? As a first step toward discerning these patterns, I stepped back to take a broader look at the factors defining different disciplinary domains, following Becher’s (1989) taxonomy of “knowledge territories.” As Parry (2007, p. 18-19) explains:

According to Becher’s (1989) classification, pure science, as exemplified by physics, is described as hard-pure, reflecting the nature of its knowledge base as cumulative, atomistic and concerned with universals, quantities and simplification; resulting in discovery or explanation. The humanities, as exemplified by history, and the pure social sciences, as exemplified by anthropology, are described as soft-pure, being reiterative, holistic (organic, river-like), concerned with particulars, qualities and complication; resulting in understanding and/or interpretation. The technologies, as exemplified by mechanical engineering, are described as hard-applied, being purposive and pragmatic, producing know-how via hard knowledge and concerned with mastery of the physical environment; resulting in products and techniques. The applied social sciences, as exemplified by business studies or education, are described as soft-applied, being functional, utilitarian, producing know-how via soft knowledge, and concerned with the enhancement of professional practice; resulting in protocols and procedures.

Using Becher’s (1989) “knowledge territories” as a framework, Figure 1 proposes a model of the factors influencing data sharing writ large.
The model proposed in Figure 1 suggests that in the humanities and “pure” social sciences (upper right quadrant), where the goal is to understand the whole of a particular phenomena, the epistemological and methodological framework discourages data-sharing. In contrast, in the physical sciences (lower right quadrant), where the approach is defined as “cumulative, atomistic, and concerned with universals,” there are no inherent barriers to data sharing. Engineering and technological sciences (lower left quadrant), which result in products and techniques, bring a concomitant concern about the proprietary aspect of data – issues of ownership, patents, and licensing mitigate against unfettered data sharing. For the applied social sciences like education and social work (upper left quadrant), regulatory issues constrain data sharing, including ethical concerns related to confidentiality and the right to privacy (Etten & Petrone, 1994; Duncan, Jabine & Wolf, 1993). Health and medical researchers also share these concerns. The ubiquity of genetic data and geospatial data, which make it easy to pinpoint individuals, have heightened these concerns (Austin, Harding & McElroy, 2003; Van Wey et al., 2005; Sherman et al., 2007).

Looking specifically at the range of research practices in the social sciences, Figure 2 maps methodological approach (on the horizontal axis) and epistemological orientation (on the vertical axis) that define the different research prac-
tices in the social sciences to create a model of the factors that influence data sharing norms.

In Figure 2, the left to right axis moves from methodologies involving large-scale, team projects to individual, small-scale (or from research involving work across the disciplines, such as health or economic development, to discipline specific practices such as literary analysis or economic modeling). The top to bottom axis moves from epistemologies involving interpretive, gestalt analysis to comparative studies seeking to uncover regularities or generate predictive algorithms.

The data sharing model for the social sciences (Figure 2) suggests that data sharing is likely to be encouraged in research domains defined by large-scale and/or interdisciplinary studies that seek to generate results that can be generalized through replication or comparison across cases (lower left quadrant). This is in accordance with the case studies of population studies and secondary analysis in economics, in which data sharing and re-use is part of the norms and practices of researchers in these fields. Preceding clockwise, research involving large-scale or interdisciplinary teams and holistic or interpretive analysis poses methodological challenges to data-sharing. Sharing data across research domains is complicated by the importance of implicit or informal knowledge and the inherent difficulty of separating knowledge from the context in which it was also generated limits the portability of data. More to the point, the interpretive frame calls into question the very notion that data can or should be understood in isolation from its particular context. As Zimmerman (2003) found in her study of ecologists, the importance of informal knowledge and issues of scale and pattern present inherent methodological challenges to data sharing. For individual scholars pursuing interpretive analysis in areas like cultural anthropology (upper right), these issues are even more prominent. Within the humanities tradition of the individual scholar, data sharing is an unfamiliar concept and the interpretive frame resists the reduction of knowledge to data elements or data points that can be isolated. Such decontextualization is anathema to scholars working in this domain. Finally, individual or small-scale research seeking to generate results that can be replicated or generalized, or to create models with predictive value, faces logistical challenges to data sharing. The methodology generates data that is portable, but as we saw with economic modeling, sharing data requires a system of scholarly communication that encourages individual researchers to incorporate data deposition into their workflows. In this case, the barriers are technical rather than cultural and data sharing can easily be facilitated by creating the requisite infrastructure. This is in contrast to the upper half of the diagram, in which the obstacles to data sharing are intrinsic to the way knowledge is constructed and communicated in these fields.
Figure 2. Factors Affecting Data-sharing Norms in the Social Sciences

DATA SHARING ACROSS DOMAINS:
ONE SIZE DOES NOT FIT ALL

What implications do these different data-sharing practices have for interdisciplinary scholarship – and for libraries serving researchers across these domains? Sharing data across disciplinary domains where knowledge is built incrementally through replicative or comparative studies and that have a similar approach to data collection and analysis present few if any challenges, except those of awareness and discovery, e.g., researchers doing quantitative sociology, demography and political science are all well served by ICPSR. But what of sharing across communities of practice with distinct methodological and epistemological approaches? Or in areas where a more interpretive, gestalt approach to knowledge building pertains?

Schröder (2007, p. 1) cautions, that “large scale data policies may have unplanned effects of homogeneity.” According to Schröder, throughout history of science, data were an inextricable part of the research process, all but useless when removed from the context of a particular research trajectory. Similarly, as Zimmerman (2003) found in her study of ecologists, data are not always easy to disaggregate from the process and context in which knowledge is generated.
In a panel discussion on “Data Discovery and Dissemination” at the 2008 International Association for Social Science and Information Service and Technology Conference (IASSIST), Myron Gutmann (2008) discussed the “fear of miscegenation” as variables from one data set are recombined with other in “mash-ups” for meta-analysis. Not only does this challenge some assumptions about methodology and rigor, but it also presents new challenges as the variables circulate independently and in novel combinations. Who then owns the recombined dataset? How do you cite and track sources? Can we develop provenance metadata and citation standards for variable-level data—and promulgate these standards so that they are actually used?

Parsons and Duerr (2005, p. 32) offer some cogent insights into the need to balance the paradoxical imperative to “maintain data and documentation in a way that facilitates broad but appropriate use so that it continues to be useful to specific group of users, but can also be used by other, perhaps unanticipated user communities with very different needs and ways of approaching analysis.” Drawing on Lakoff and Johnson’s (1980) work on importance of metaphor as a basis for conceptual understanding, Parsons and Duerr (2005) call attention to how easily terms and concepts can be misunderstood when transplanted from one context to another. The example they discuss is the term “metal” which in astronomy denotes any substance with an atomic weight heavier than hydrogen, which includes oxygen, carbon, and other materials not considered to be metal in the common definition. Another example is “animal,” which, for zoologists includes distinctly non-furry creatures like rotifers and nematodes. Perhaps more relevant to the present discussion is “race” which to demographers evoke a discrete category with a precise definition, while biological anthropologists might argue that race is cultural invention that masks a fluid range of biological variation across populations. As (Parsons and Duerr 2005, p. 32) argue, “these metaphors vary from discipline to discipline and are bound to change over time, even within a given discipline…[therefore] we must challenge our assumptions regarding the knowledge of our user community and provide suitable context for users to understand the data.”

A recent study of the research behaviors of scientists and graduate students at the University of Minnesota (Marcus et al., 2007) found that the lack of a shared vocabulary was one of the primary challenges to interdisciplinary collaborations. But the real barrier may run deeper than terminological differences to the deep semantic knowledge that underlies the different terms, a less easily resolved challenge.

As Zimmerman (2003) concludes, formal structures (standard research methods, metadata, storage formats) alone are not sufficient for large-scale data integration:

Standard research methods, metadata standards, and common storage formats make it possible to integrate data on a large scale, but this power comes from leaving out information that is necessary to secondary data use. Ecology teaches
us that there are multiple sides to issues of trust, standards, understanding, and judgments about data quality. To be effective vehicles of data sharing, digital libraries and data repositories must capture public and private knowledge and must find ways to document the implicit knowledge that ecologists recognize and can articulate.

It is interesting to reflect on the fact that the conditions that foster data sharing also foster collaboration across and among the disciplines. These conditions include a common set of standards (for metadata, systems interoperability, documentation, terminology, etc.) as well as a shared framework for understanding (Lakoff & Johnson, 1980).

The challenge then becomes building repositories that make the implicit explicit and incorporate some means to capture and communicate informal knowledge in ways that researchers in those domains recognize and can articulate. In this regard, librarians may have a key role to play in this discussion. We have long served as envoys, building collections and information discovery mechanisms that balance the needs of multiple audiences and allow people to navigate unfamiliar information terrain. We also have long experience in balancing the need for short-term access with long-term stewardship.

Most social science disciplines support a range of practices from lone researchers doing highly contextualized research, to those doing comparative or predictive studies whose units of analysis can easily be transported to other contexts. The latter have been in the vanguard of developing large-scale data repositories, but what of the former? Does data sharing have any relevance to those whose intellectual work involves directly engaging a corpus of texts and/or artifacts and whose methods generate knowledge that is inextricably grounded in this constellation of texts/artifacts? What is needed, in this case, are ways to assemble a constellation of texts and primary sources, to create repositories that allow scholars to create idiosyncratic collections of texts and, artifacts or other manifestations of the human experience, to interweave commentary and analysis with the primary sources, and to share and recombine these collections in order to answer a particular research question, akin to the “Multidimensional Framework for Academic Support” proposed in the University of Minnesota’s Mellon-funded study (University of Minnesota 2006, p. 50) or the American Council of Learned Societies’ report of cyberinfrastructure for the humanities and social sciences (2006). Creating this sort of framework is the next challenge for data-sharing in the social sciences.

NOTES

1 SHERPA-JULIET <www.sherpa.ac.uk/juliet> provides an index of the open access policies, including data archiving requirements, of various funding agencies.
2 DDI or “the Data Documentation Initiative is “an international effort to es-

tablish a standard for technical documentation describing social science data”
< http://www.ddialliance.org/>. DDI 3.0, a XML-based metadata standard,
was ratified by the DDI Alliance in May 2008.

3 The long tail is a statistical term associated with distribution also known as
80-20 rule, wherein 80% of the effects are attributable to 20% of the causes
(e.g., in Pareto’s eponymous study, he found that 80% of income in Italy
went to 20% of the population). The term “Long Tail” (note capitalization)
was popularized in a 2004 article in Wired magazine (Anderson 2004) to
describe the business strategy of online retailers of video and book products
that focus on capturing a niche market and thereby sell a wide assortment of
items in small quantities to a large number of people.

BIBLIOGRAPHY

Altman, M. (2008). Towards a data network for integrated social science re-
search. Paper presented at the Annual Meeting of the Association for the
Advancement of Science (AAAS), Boston, MA. Retrieved March 21, 2008,
from http://maltman.hmde.harvard.edu/presentations/aaas_08.ppt

American Council of Learned Societies. (2006). Our cultural commonwealth:
The report of the American Council of Learned Societies Commission on Cy-
berinfrastructure for the Humanities and Social Sciences, Retrieved April 24,


stewardship of digital data sets in science and engineering. Arlington, VA.

and social issues in developing population genetic databases. Genetics in
Medicine 5(6), 451-457.

cultures of disciplines. Bury St. Edmunds, Suffolk: The Society for Re-
search into Higher Education & Open University Press.

tion 52(42), A35

Dee, T. S., William E. N., & Murray, S. E. (1999). Data watch: research data in
the economics of education. Journal of Economic Perspectives, 13(3), 205-216.

public policies: Confidentiality and accessibility of government statistics.

Etten T., & Petron R. (1994). Sharing data and information in juvenile justice:
Legal, ethical and practical considerations. Juvenile and Family Court Jour-
nal, 45(3), 65-90.


Schröder, Peter. 2007. Possible downsides to data sharing in the research commons: Assets and liabilities, opportunities and risks. Data Science Journal, 6, 67-70.


**BIOGRAPHICAL STATEMENT**

JoAnn Jacoby recently took on a new role as Coordinator, New Service Model Programs for University of Illinois Library, where she previously served as Anthropology & Sociology Subject Specialist. Her research interests include library assessment and scholarly work practices.