ETHICS OF SHARING SCIENTIFIC AND TECHNOLOGICAL DATA: A HEURISTIC FOR COPING WITH COMPLEXITY & UNCERTAINTY

J.E. Sieber, Ph.D.
California State University, East Bay
Hayward, CA, 94542 U.S.A.
Joan.sieber@csueastbay.edu

ABSTRACT

Data sharing poses complex ethical questions for data management. Manifold conflicting and shifting values need to be reconciled in pursuing viable data-management policies. For example, how does one make data available in useable form to stakeholders including scientists, governments and businesses worldwide, while assuring confidentiality, satisfying one's research ethics committee, protecting intellectual property and national security, and containing costs? Increasingly, ethical problem solving requires integration of ethics with technological "know how" and empirical research on the presenting problem. Each problem is highly contextual; broad application of general ethical principles such as always practice openness, or prepare all data for sharing, may have harmful unintended consequences. Chaos theory provides a heuristic or vision for understanding and coping with complexity and uncertainty. It does not provide answers to problems of data management, but frames the issues, and provides appropriate expectations and heuristics for considering data management problems.

Keywords: Data sharing, Data management, Confidentiality, Intellectual property, National security, Cost containment, Chaos theory

1.   INTRODUCTION

The ethics of data concerns their validity and net benefit to society. The validity of data when they are first gathered and analyzed, is a topic of immense complexity that varies depending on the discipline one considers, the design and analytical methods employed, whether one is concerned with external as well as internal validity and the logic of the claims that flow from the data and analysis. In turn, the benefit to society depends on how the data, once analyzed and reported, are used, if indeed they are used at all. Data are useful or beneficial depending on what is done with them, and the context in which those transactions occur. In turn, contexts are defined by scientific, technological and social events. A subset of this puzzle is the issue of how to manage data for purposes of sharing. Utilitarian ethics is about maximizing the amount of good over harm that accrues to all. Preparing and preserving all data for sharing is perhaps even more wasteful of scientific and social resources as destroying all data after their first use. The ethical dilemma, then, is to discover the most intelligent course through a thicket of ever-shifting circumstances surrounding the preparation, storage, and ultimate secondary usefulness of data.

Contexts change, giving rise to new ethical issues, new clashes of values, and new dilemmas for scientists and others to resolve. One can inadvertently do harm, as an unintended side effect, while seeking to do good. “Doing ethics of data” in the trenches of science, then, means ferreting out the possible outcomes of a given course of action, and minimizing risks while promoting benefits. It is a problem solving process; however as the nature of the problem changes, today's solution may not work tomorrow. More disturbing still, today’s presumed solutions may be too short-sighted to solve even today’s problem.

As an example of the changing context of data sharing, prior to about 1975, openness in most areas of science was manifested by the way research methods were described and data were presented in publications; replication of results established the validity and generalizability of results. There were only a few areas of science in which actual raw data were widely and openly shared. Meteorology, geology, oceanography, astronomy, demography and economics were notable areas where sharing of major, government-sponsored, costly data sets occurred openly and systematically between scientists, organizations and nations. Some other sharing occurred informally among colleagues. By the 1970s, some federal funding agencies, most notably the National Science Foundation, began to encourage more formal sharing of documented data in *all* areas of science. However, certain barriers to data sharing continue to slow its progress. Norms of sharing have not developed evenly across disciplines. Scientific societies and their journals vary with respect to their data-sharing requirements. Sharing of human-subjects data hinges on development of techniques for protecting confidentiality. Such techniques are being developed, but their dissemination has been slow. These barriers can be overcome, but new problems will probably always continue to arise, as the following recent events would suggest:

Unanticipated events have recently impacted data sharing policies and practices. Privacy issues are increasingly salient; kinds of data previously considered innocuous are increasingly regarded as *sensitive*. Electronic storage and transmission of data amplify concern that unanticipated uses of sensitive data will jeopardize the economic, legal and social standing of persons. The advance of high technology has raised unresolved problems of protecting intellectual property when sharing data. The terrorist attack of September 11, 2001 produced fears that terrorists might use openly shared scientific data (e.g., genetic analyses of deadly pathogens) to devastate entire societies. This concern has given rise to federal requirements concerning data that are "sensitive but not classified." Such data are not to be shared openly, but only on a "need to know" basis, leaving scientists to wonder just how such a concept is to be operationalized and what its effect on science will be (Kennedy, 2003).

This litany of recent events impacting data sharing suggests what is to come -- an unending procession of new and unanticipated problems as long as there are humans around to create conflict. If an act as simple, and *prima facie* good, as the sharing of data among scientists has become so problematic, what other data management issues will confront science and society? One need not be able to look into the future to see that questions of when to publish (what is too early or too late?), whether journals should publish null results (Rosenthal, 1979; Scargle, 2001), what constitutes plagiarism (see Loui, 2002, for some puzzling cases), or how to determine and adjudicate cases of data fraud (Kennedy, 2002) will probably be around for a long time and will take on ever new meanings and implications for data management and science. As new methods are developed for assuring the security of data files or the confidentiality of data some of the headaches of data management and sharing will vanish, and new headaches will take their place. The definitions of what it means to publish have changed with the advent of electronic publication and the escalation of hard copy publications. The definitions and accompanying problems connected with publication will no doubt continue to change with emerging technologies, and to bewilder those who are concerned about the quality and honesty of published scientific work. For example, the concept of *open access,* that is, the making available of scientific results on the internet, free to anyone anywhere, becomes a possible threat to the very existence of peer review if it causes libraries to cancel subscriptions to journals whose articles are available to everyone at no cost. This, in turn, motivates a search for solutions that foster open access without jeopardizing the financial basis for publication of peer reviewed carefully edited scientific papers (Zerhouni, 2004).

Each set of data an investigator obtains is a slice of reality that may or may not be valid or reliable. Consequently, scientists must find ways of pooling data to determine the robustness of individual findings. Increasingly, scientists recognize the importance of empirically based models that combine data points. For example, three major developments in the social sciences are meta-analysis (Rosenthal & Rubin, 1986), Empirical Implications of Theoretical Models, or EITM (Granato & Scioli, 2004), and the Statistics Canada LifePaths model (Wolfson & Rowe, 2004). These developments challenge older assumptions about the usefulness of data, but some have argued that they take science farther from direct observation and into realms of mathematics and statistics that may do as much to confuse as to enlighten. These debates and conflicts are only beginning as the models for interpreting empirical phenomena will surely proliferate. Another part of the debate will rage around questions about the quality of the data that are used in modeling. Other debates will arise about the accuracy or risk involved when we use the predictions of these models to make consequential decisions about the lives of individuals and the welfare of societies.

There are ethical issues at every turn as we seek to gather and disseminate valid data, to translate it into accurate and useful models, to share data and knowledge throughout human society, and to create a better world. The ethical issues to be resolved are kaleidoscopic – changing with each new turn of events. What, then, is the process of being ethical as regards data? Is there some heuristic or theory to guide this process?

The process of ethical data management and sharing involves reconciliation of diverse conflicting values. How does one achieve such reconciliation? Traditional scientific approaches assume that general principles can guide mastery of problems. However, theories of normative ethics do not tell us what we should do in a direct way; they require intelligent interpretation. Hence, given that life is lived on a slippery slope, except in the most well-understood and unchanging contexts, normative ethical theory often turns out to be of limited use or even a source of confusion and misinterpretation. Different contexts involve different value priorities, different nuances of values, and invention of new technologies and social norms to achieve optimal outcomes.

2. WHAT ISSUES SHOULD BE CONSIDERED AS PART OF THE ETHICS OF DATA SHARING?

It is easy to enumerate some of the major issues that should be considered as part of the ethics of data sharing. However, such enumeration only demonstrates the complexity and uncertainty surrounding these decisions:

1. Effective data analysis, transformation, dissemination, archiving, storage, retrieval and sharing is about anticipating the future of science and society. How does one anticipate the nature, problems and methods of science in the years to come? For example, useful data include those that enable us to understand cycles of climate, economics, and health and disease. How does one piece together data such as ice core sample data and silt samples from 100,000 years ago with the 150 years of temperature and precipitation data that have been gathered worldwide? How do the foods of one era or one culture compare to those of another, and what are the impacts on human development? What future technologies, theories and discoveries will dictate new uses of old data? What data do we need to assemble to answer such important questions? What models or methods will provide useful assemblies of these data?

2. What ethical principles should be fostered in data management? Two kinds are to be fostered:

- **Micro-ethics** - the principles or values that scientists are taught, such as validity, transparency of methodology, appropriate use of statistics, and data sharing.
- **Macro-ethics** - values relating to public policy and the larger society in the nation and the world, and to the responsibility of scientists and data managers to decide, for example: What data should be shared, at what cost, to whom? What mode of research financing and data sharing will maximize world health? ... save the oceans? ... foster world peace and prosperity? ... help control pollution? ... educate scientists and laymen about science? And so on.

In the United States, intelligent integration of micro-ethics and macro-ethics of data management is important to scientists who seek federal funding. The National Science Foundation requires data sharing where appropriate, and funds only those proposals that adequately address macro-ethical, as well as micro-ethical issues. NSF's web site www.nsf.gov/pubs/2002/nsfo22/bicexamples.pdf provides many examples of broader social impacts that research might strive to attain. Briefly, the five main criteria are:

- Advance discovery and understanding while promoting teaching and learning at all levels.
- Broaden participation in science over groups, cultures, nations.
- Enhance infrastructure via distributed user facilities, digital libraries, databases, internet-based facilities.
- Disseminate broadly via diverse media; partnering with industry and public institutions.
- Benefit society at large; present results in formats useful and understandable to nonscientists.

3. Why do normative ethical theories play a minor role in the practical ethics of data sharing? There are various theories of normative ethics. Each provides some insight into what is good and how good might be produced in the world, though they do not necessarily agree either about the nature of good or how it is produced. Three salient normative ethical theories are very briefly described here to give some flavor of the limitations of normative ethical theory with respect to practical questions about the ethics of data:

- **Deontology** - seek to do that which is right or good under any circumstances. What might that be? Promote validity of data? ... openness of science? Deontological ethics begs the question: *How?*
- **Rule Utilitarianism** - follow the rule most likely to lead to the most good for the most people. Which rule is that? Do what produces the most creative science? ... what benefits society most? ... what best supports scientific innovation? Does any of these values trump all others?
- **Act Utilitarianism** - do what seems, in the ***particular*** case, most likely to lead to the most good for the most people. This would tend to vary with circumstances. Act-utilitarian ethics is sometimes called situation ethics. The decision rule may be subject to the personal interests of the decision maker; it may be short-sighted or even unethical. Or it may provide needed flexibility.

The application of these theories is confusing because the values to be fostered are many, complex and changing. The list of values grows as circumstances change, and priorities change with context. For example, before terrorist attacks on the United States of September 11, 2001, most persons concerned with the democratization of data sharing via the internet were unabashedly concerned with openness; after September 11th national security became an issue. Should gene-sequencing data for some human pathogens be freely shared? Should on-line maps hide features such as dams? As another example of shifting priorities, 15 years ago universities were eager to support research and encourage innovation through licensing. Today some university Offices of Technology Licensing are spending more in legal fees than they generate in royalties. Their supposed cash cow eats profits, and restrictions on data and technology make the academic environment less intrinsically rewarding to productive scientists.

The technology and knowledge needed to achieve important data-related values evolve rapidly. Today's expensive archive may be obsolete or inaccessible tomorrow; conversely what is impractical today may become cost effective tomorrow. The meaning of some values, e.g., protection of intellectual property, changes with scientific context, with whether the data were produced in academe, a public lab, or private industry, and with which nation's laws govern intellectual property (Reichman & Uhlir, 2003). The value and meaning of intellectual property and whether it should yield to broader rights of scientists to have access to information also changes as we experiment with new incentives for innovation and sharing (Samuelson, 2002). In a different realm, new approaches to assuring confidentiality are continually being devised and the robustness of existing approaches is continually being tested (e.g., DeWolf, 2002).

Even the terms ***data***, and ***data sharing*** have manifold meanings, and various ethical issues attach to them accordingly. In any research project, data go through many transformations from raw data to cleaned and perhaps digitized data. There are many kinds of data: quantitative data, descriptive data, cell lines, samples of rock or tissue; data gathered in labs or in natural settings; experimental and observational data; human subjects data. There is also "know how" that needs to be shared so that others can understand or replicate research. For those interested in investigating scientific fraud or fiscal responsibility, the financial data of projects are of interest. Which kinds of data do we archive? At what expense? How does one estimate the usefulness of each kind of data? Moreover, there are various kinds of archiving and sharing. For example, there are:

- Public data in public archives.
- Privatized data.
  -- public data, with value-added features that make it user friendly, sometimes sold at marginal cost.
  -- data produced by private industry, then shared in part or whole with others, sometimes for profit.
- Data of individual scientists shared via the invisible college with close colleagues.
- Data of individual scientists prepared for sharing in an organized archive.

Each kind of archive brings with it somewhat differing issues.

4. What values are to be reconciled*?* Each kind of data brings with it different concerns about transparency of method, kind and amount of documentation appropriate, sensitivity of the data, cost of archiving and sharing, and preservation. Presumably an ethical analysis should take into consideration all values that would have a material impact on the outcome of data management. The following are some of the main values that might be considered:

Important data should become known through publication, but how?
- Hard copy or electronic publication?
- Early and incomplete? Or later, after elaboration and replication?
- File drawer problem: Publish null results so others won't repeat our errors?
- Roles of peer review, especially of null results and electronic "publication?"

Useful data should be shared in most usable form, but how?
- What issues will be important in the future, and what archives will become "data graveyards?"
- What methods of data integration will be employed in the future? Some current examples:
  -- Meta-analysis
  -- Empirical Implications of Theoretical Models; see *http://www.nsf.gov/sbe/ses/polisci/eitmreport.htm*
- What configurations of data are most useful, and have highest priority for sharing?
- What stage(s) of data: (e.g., raw, ..., digitized) should be shared?

Responsibilities for sharing should be appropriate, raising questions such as:
- How early should data be released, by whom, for use by others?
- Who operates the archive, answers the user's questions, and updates the archive?
- How much data documentation is appropriate?
- What provisions are there for teaching required "know how" to data users?
- What resources should be allocated to documentation? ... to assisting users?
- Who pays for these services? What pricing formulas are appropriate?
- What are the trade-offs between funding new research versus quality archiving and sharing?

5. **Does protection of intellectual property (IP) interests raise special complexities?** Creators of IP concepts need to be clear about what advances knowledge, research and human well-being (Samuelson, 2002; Reichman & Uhlir, 2002). What works in one context may be counter-productive in others. Protection of IP interests should be administered with a sense of proportionality. What is appropriate depends largely on the main goal or interest of the sharing institution:

Institution	**Main Goal**	**Open or Privatized**
University	Education	Open (small amount of privatization, if any)
Government (public)	Serve society	Open
Private industry	Production	Mix of open and privatized, as fosters productivity

Openness of science serves education and society, and advances knowledge for use in the private sector.
Hence scientists should experiment with different approaches, consider gradations of protection of interests, experiment with different incentives for productivity, and guard against over-regulation. Some fields have immediate application of IP protection. For example, some gene sequencing work involves large investment and immediate pharmaceutical application. Some privatization is appropriate to production. However, most fields do not have immediate, or even obvious, application. Astronomy has no immediate applications, hence total openness of data and technology seems appropriate. Many areas of natural and social science advance understanding of ourselves and our world but have little commercial application and did not involve large investment.

6. How can one make sound ethical decisions about data given this kaleidoscopic array of variables to try to optimize*?* In one sense, the problem is unsolvable; there can be no stable correct solutions to this dynamic problem. However, successful ventures into the unknown require vision—a sense of what is possible and what is worth pursuing. There may be many visions on which to make decisions about data management and sharing. *Vision* means the kinds of goals that a committee of scientists might articulate when trying to decide how to allocate resources, what data to prepare for sharing, how to orchestrate the sharing, and why. Of course, any vision may turn out to be a Quixotic attack on windmills. But without vision there can be little intelligent progress in dealing with complex problems. With vision, that is a relevant heuristic, one can open one's mind to possibilities and pursue them. Chaos Theory offers a relevant heuristic for embracing and coping with change and uncertainty concerning values that should dictate data management and sharing (Briggs & Peat, 1999):

Science is man made, but it has a life of its own that we cannot predict or control. The issues are complex, changing and ambiguous. They shift in a kaleidoscopic fashion and call for new paradigms of ethical problem solving. Nothing seems to describe the course of modern science or the ethical challenges of data sharing better than chaos theory. Briggs and Peat's (1999) seven principles of chaos can provide concepts, expectations and heuristics that may help to guide ethical data management and sharing:

- Chaos is complex, changing, uncertain, ubiquitous, and a natural aspect of the world.
- In chaos, small changes can have huge impacts, e.g., gullies can become canyons; inventions such as the transistor and internet can revolutionize how people communicate and interact.
- Chaotic structures tend to self-organize.
- Complex chaotic structures may contain simple subtleties.
- Chaotic self-organization can be elegant.
- Time may be regarded as a process of developing – not as units on a clock or calendar.
- Reductionistic notions of dissection and control are not the only way to understand the world.

3. WHY CHAOS THEORY MAY OFFER USEFUL HEURISTICS TO ETHICAL PROBLEM SOLVING REGARDING DATA SHARING

Ethical data management and sharing appears to require a model or paradigm that takes account of the complexity, change and uncertainty inherent in these data-management problems. Such a model could not provide simple, reductionistic or permanent answers, but it could suggest guidelines or expectations that would foster appropriate perceptual sets and problem solving strategies. The following are some ideas and opportunities that chaos theory presents:

Ubiquitous complexity, change and uncertainty provide opportunities for continuous creative thought and openness to new ideas. This is an open invitation to development of new approaches to problem solving. E.g., how do we assemble recent and ancient data to understand global climate changes? How to assure confidentiality of data or nondisclosure of information that could threaten national security?

Small changes, e.g., in intellectual property law, in data sharing, etc. can have huge impacts. A small change in data management can have profound and even irreversible impacts on the way science is done; e.g., consider the arguments of Samuelson (2002) and Reichman (2002) concerning effects of property rights infringements on the intellectual commons. Hence, scientists should experiment with small, local changes in data management, and beware of making sudden large changes.

Chaotic structures tend to self organize. One should avoid starting with hierarchical structures; rather one should work collaboratively and observe what works in different settings. One should observe what scientists want in order to be productive and motivated. E.g., what can we learn from observing open-source software engineering by individual hackers? Academicians? Government lab engineers? Engineers in private industry?

Complex chaotic structures can have rich, simple subtleties. The overall issues may be complex, but some simple solutions may lie within. We should beware of stereotyped ideas about data and data sharing that don't fit the realities. E.g., strict copyright laws that impede education and protect little are often violated and not prosecuted. What does this tell us?

Chaotic self-organization can have an elegance and beauty of its own. The esthetics of how different groups use and share data and ideas should not be lost on those who would hope to improve data management policies. Institutions use a wide range of resources and technologies to make sharing of ideas and data a pleasing and exciting process. E.g., among our various kinds of institutions are some that have created magnificent modes of education, collaborative research, and public service using data. These should be showcased broadly world-wide and emulated or experimented with as appropriate.

*Time may be regarded as a process of **developing** -- not as units on a clock or calendar.* The processes of understanding how to manage data most usefully may come together as we daydream about what we most want from data. Perhaps our most useful processes of figuring out how to manage data should be allowed to occur on their own time, not on a dictated time-table. Perhaps universities, funders and industry can find creative new venues for envisioning and experimenting with effective new data management practices, without forcing them into Procrustean time frames.

Reductionistic notions of dissection and control are not the only way to understand the world. New values, new priorities, and new technologies will emerge and take on a life of their own. Perhaps our most creative uses of data will emerge as we learn to accept and work within this environment that we cannot control.

4. REFERENCES

Briggs, J. & Peat, F. D. (1999). *Seven Life Lessons of Chaos*. New York: Harper.

De Wolf, V. (2002). Issues in accessing and sharing confidential survey and social science data. *Data Science Journal, 2*(17) 66-74. Retrieved 17 December, 2005 from: http://www.codata.org/codata02/11datapolicy/deWolf/deWolf-paper.pdf.

Granato, J., & Scioli, F. (2004). Puzzles, Proverbs, and Omega Matrices: The Scientific and Social Significance of Empirical Implications of Theoretical Models (EITM). *Perspectives on Politics, 2*, 313-323.

Kennedy, D. (2002). Editorial. Next steps in the Schon affair. *Science*, 298(5593) 495.

Kennedy, D. (2003). Editorial: Two cultures, *Science* 299 (21 February), 1148.

Reichman, J. H. & Uhlir, P. F. (2003). A contractually reconstructed research commons for scientific data in a highly protectionistic intellectual property environment. *Law and Contemporary problems*, 66(Spring), 315. Retrieved 17 December, 2005 from: http://www.law.duke.edu/journals/lcp/articles/lcp66dWinterSpring2003p315.htm.

Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin, 86*, 638-641.

Rosenthal, R., & Rubin, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin, 99*, 400-406. Retrieved 17 December, 2005 from: http://web.uccs.edu/lbecker/Psy590/es.htm.

Samuelson, P. (2002). Legal issues in using and sharing scientific and technical data preserving the positive functions of the public domain in science. Retrieved 17 December, 2005 from the University of California, Berkley website: http://www.sims.berkeley.edu/~pam/papers/CODATA_slides.ppt. .

Scargle, J. D. (2000). Publication Bias: The “File-Drawer” Problem in Scientific Inference. *Journal of Scientific Exploration, 14*(1), 91–106.

Wolfson, M. & Rowe, G. (2004). Disability and informal support: Prospects for Canada. In Cohen, S.B., Lepkowski, JM (eds). *Eighth Conference on Health Survey Research Methods*. Hyattsville, MD: National Center for Health Statistics.

Zerhouni, E. (2004). Information Access: NIH Public Access Policy**,** *Science, 30,* 1895