Reduce, Reuse, Recycle: Issues in the Secondary Use of Research Data

*by Margaret Law**

This paper summarizes ethical concerns about the secondary use of data and the arguments for encouraging or facilitating it. It includes some potential solutions and discusses the implications of the increased use of new technology. While it focuses on the Canadian regulatory environment, similar issues arise in other countries.

Introduction

"Reduce, reuse, recycle", a phrase familiar to us from the environmental movement, can also be used to reflect on the secondary use of research data. Secondary research refers to the use of research data to study a problem that was not the focus of the original data collection. This may be data collected for administrative, health or educational purposes, census data, or data collected as part of a previous study. This secondary analysis may involve the combination of one data set with another, address new questions or use new analytical methods for evaluation (Szabo & Strang, 1997).

Both benefits and dangers have been attributed to the secondary use of research data. Distinctions are often made between large scale data collections, particularly sample survey data collected at public expense, and smaller bodies of data collected at personal expense. There is general agreement that the first should be shared and made generally available in a "timely" fashion, but little agreement about the second. There is also no sense of agreement about what would constitute timely in this situation (Clubb, Austin, Geda, & Traugott, 1985).

While the questions surrounding the secondary use of research data have always existed, they have become more pressing with the use of new technologies. New capabilities include easier data sharing, faster and more complex analysis, and the development of large scale data banks. Previously, the ability of researchers to communicate was limited by time and distance; now data can be shared globally at the click of a mouse. As we adapt to the electronic environment there are new concerns about confidentiality and the threat of security lapses. The potential for finer data resolution becomes possible with better data collection tools and technological innovations. While the fundamental ethical issues have not changed, the possibilities created by new technologies have brought them to the forefront.

Just as governments have taken a strong leadership role in developing and supporting good environmental habits, they must be encouraged to develop and support good habits concerning the storage and use of research data.

Concerns about data sharing and data confidentiality affect researchers and data librarians across the world. While each country may have a different regulatory environment and a different research culture, the need to find an appropriate balance between the optimal use of data and the protection of individuals is worldwide. With increasing globalization, and an increase in international research, the development and articulation of appropriate guidelines becomes paramount.

Data sharing is a fundamental value for IASSIST, and individual or random decisions about data sharing stand in the way of providing the best support for researchers. By looking at the Canadian situation, data librarians may develop and share common messages as part of an overall advocacy plan to support data sharing. There must be limits, of course, to protect respondents, but these must be delineated and managed in a coherent way that not only recognizes their rights, but also those of researchers, and of the taxpayers who frequently fund the research. This advocacy effort must be aimed at regulatory bodies, funding agencies and the researchers themselves in order to change the cultural values around secondary use of research data.

In Canada, much research involving humans is governed by the three major granting councils, who have developed a shared policy statement to govern all research involving human participants done in Canada, and by Canadians outside of Canada. Section C3 of the Canadian Tri-Council Policy Statement (Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans, 1998) lays out guidelines for Research Ethics Board (REB) approval of research that proposes the secondary use of data. It is clear that REB approval is required if identifying information will be involved, but leaves it to the researcher and the REB to determine exactly what constitutes identifying data.

There is sufficient legislation in most provinces to provide a framework for researchers wishing to use data that has been collected for purposes other than research. For example, the Alberta Health Information Act (Health Information Act Chapter H-5, 2000) provides direction for researchers' use of health care data. Division 3, 'Disclosure for Research Purposes', defines the role of the ethics committee, including consideration of whether the researcher should be required to obtain consent from subjects, implying that there is some choice in this matter. The ethics committee must also decide whether the research is of sufficient importance to outweigh privacy concerns, whether the researcher is qualified and whether there are sufficient safeguards.

The picture is not so clear, however, when the data was originally gathered for research purposes. This data is not governed by legislation, and guidelines are open to interpretation. There is considerable discussion about who owns the data and decisions about whether to share data are often made by the original researcher and may depend on a number of personal factors. A requirement for all researchers to consider the potential for secondary use of their data, either by themselves of by others, would contribute to a more orderly use of data with resulting benefits for researchers, subjects and the community.

Concerns about the secondary use of data

Concerns about secondary use of data generally focus on the potential for harm to the individual subjects of the research and the lack of informed consent. Many writers are passionate about the primacy of informed consent for any type of research involving human subjects. Consent applies not only to a particular researcher, but also for an identified purpose. To quote Kalman (1994), 'the requirements to seek an individual's consent to participate and to provide data for a specific purpose must take precedence.' Since researchers generally are not able to predict potential requests for secondary use of data that they are collecting, they are unable to fully inform subjects of the primary research about potential future uses of data. As this full disclosure of information is one of the requirements of informed consent, it follows that it is not possible to get informed consent for unanticipated uses of data.

Others argue that if the second researcher were to contact subjects to ask consent to re-use data, the original researcher must first identify the individuals thereby breaching their privacy. This situation could be managed by having the original researcher contact the subjects on behalf of the second researcher. Privacy is generally defined as a personal issue, defined by the subject. The subject may have felt comfortable disclosing information to the first researcher because of their relationship or rapport, but secondary research could leave him open to actions of researchers with whom he feels less comfortable (Homan, 1992).

Technology-driven data analysis techniques also create the potential for triangulation of data: the combining of variables that allows identification of specific individuals and organizations even though identifying information was removed from the original data sets. For example, there has been concern that the combination of census data and geographic information can allow the identification of small or unique groups. The use of GIS allows for closer identification of geographic data through the availability of differing degrees of granularity (Trainor & Dougherty, 2000).

There is additional concern for vulnerable populations that could be at particular risk if their confidentiality were breached. Current North American legislation and the media have raised awareness about profiling issues, and certain populations such as those involved in criminal activities or who are HIV-positive have a high risk of harm if they are identified.

In addition, particular forms of data such as oral histories, photographs or diaries cannot be made anonymous because the identification of the respondent is a large part of the value of the data. Researchers in these situations often feel that they have given unqualified pledges of confidentiality to participants, leading them to bar access to the material unless participants can be contacted for permission (Hedrick, 1985). Considerable commitment from the first researcher would be needed to contact the subjects for get permission for them to be approached by a second researcher.

Ethical practice requires a balancing of benefits and harms when conducting research. While this may be assumed to refer to the benefits and harms that may be experienced by the subjects of the research, it could also be interpreted as requiring a consideration of the potential harm to the original researcher. The 'design and execution of data collection effort is a creative activity that sometimes involves innovative techniques.' It seems reasonable to question why a secondary analyst should benefit from someone else's work, particularly if the second research is a potential scholarly competitor (Clubb et al., 1985).

High quality data are expensive to collect, organize and store in an accessible form. If the data is to be used by someone else at a later date, additional work and documentation may be required. If this is carried out at the original researcher's expense, it would seem to create the potential for harm with no counterbalancing benefit. This is particularly true if the original researcher is not cited, as it could have a negative effect on tenure and future funding opportunities (Sieber, 1991).

Some writers have also proposed a negative effect on "good science", brought about as a result of too much data-sharing

(Stanley & Stanley, 1988). "It is a lot easier, faster, and less costly to obtain someone else's data than it is to design a study, recruit participants, collect and analyze data" (p. 178). This potentially leads to fewer original data sets, reducing the potential for multiple independent evaluations.

Methodological problems also arise. For example, rules requiring the original researcher to delete identifying information or other methods of anonymization may prevent the accurate use of data files, or interfere with their appropriate linking with other data, resulting in incorrect associations (Fienberg, Martin, & Straf, 1985). If the original data were not well collected or documented the second researcher may lack information about possible errors, the relationship of the data to the universe of responses, details about the sample, ways in which data were analyzed (Sieber, 1991) and the assumptions underlying interpretations (Fienberg et al., 1985). Any of these could lead to flawed research.

In support of secondary use of data

The arguments in favor of secondary use of data may not be as straightforward, but should also be considered within the framework of the ethical principles in the Tri-Council Policy Statement. The basis of these arguments focuses on the ethical obligations to good science, and to the benefit of the community at large.

A fundamental principle of research ethics is respect for human dignity, incorporating both the selection and achievement of morally acceptable ends and the morally acceptable means to those ends. If this is understood to mean that individuals are important and should be treated appropriately, one implication is that researchers should make as much use as possible of the data that is collected in order to reduce the burden on research subjects. This would provide an improved benefit/harm ratio for vulnerable groups who may be at risk from repeated data gathering intrusions into their lives. One might argue that there is, in fact, the potential for greater benefit if research with already collected data provides more opportunities to support these groups. Taking steps to ensure that interpretations of data are valid through encouraging multiple methodologies demonstrates respect for research subjects through accurate interpretations of their behaviour.

Secondary analysis creates an opportunity to establish relationships that were entirely unpredictable at the time of the original data collection (Dale, Arbor, & Proctor, 1988). For example, some of our understanding about the causes of disease has occurred through the secondary analysis of medical records that were not collected with the intention of making such a causal relationship (Dale et al., 1988). The ability to link data files and to create families of data creates possibilities that together they can contribute knowledge that none could contribute alone. In essence, this is a situation where the whole is greater than the sum of the parts (Johnson & Sabourin, 2001).

If a second research repeats original calculations to assure accuracy, it is not considered to be a secondary analysis. If, however, it analyzes the data from a different perspective or within a different theoretical framework it allows the findings to be challenged and debated, and creates an opportunity both for further discovery and for a deeper understanding of the interpretations of the data (Dale et al., 1988). Developing and implementing protocols for data sharing creates potential for testing the generality of research findings, and comparing analyses on different data sets across time or across locations allows us to 'generalize findings about social phenomena' (Fienberg et al., 1985).

Good science requires that data be available for scrutiny and reanalysis as part of scientific enquiry (Fienberg et al., 1985). The practice of a second researcher reanalyzing data is widespread, although this would seem to pose the same concerns about breach of confidentiality as any other access to data by a second researcher. "It seems reasonable to argue that if one is prepared to publish assertions about the nature of reality based on collected data, then one should equally be prepared to allows others to examine that data to check the validity of the assertions"(Reidpath & Allotey, 2001).

Concerns about privacy and confidentiality are the most frequently raised objections to secondary use of research data. Some writers believe that it is "likely that this obstacle is cited much more frequently than is warranted" (Hedrick, 1985 p.142). Attempts to quantify the risk of identification, particularly from anonymized records (Marsh et al., 1991) support to some extent the notion that the risk is over-stated. Other writers assert that achieving informed consent for secondary research is never truly voluntary as there are pressures on subjects to agree simply because they have already agreed once before, but this appears to not have been adequately investigated.

The issue is further confused by a discussion of what is meant when the original researcher states on the consent form that personal data will not be shared. Some researchers would interpret this in the most conservative way to mean that none of the data will be shared, ever. A more sensible interpretation might be that data can be shared as long as it is properly anonymized and all identifying characteristics are removed (Johnson & Sabourin, 2001). While the real question is how the subject interprets it, not the researcher, the subject is likely influenced by the researcher's position. This is an ethical position that must be resolved before it can be managed through improved methodology.

The Canadian Tri-Council Policy Statement allows for a breach of confidentiality in section 3.3 (c) if the individuals to whom the data refer have not objected to secondary

use. The Research Ethics Board is charged with the responsibility of evaluating the sensitivity of information, seeking consent to used the stored data, and allowing the researcher to propose an appropriate strategy. The REB is directed to pay particular attention to the possibility of "harm or stigma" that might be attached to identification. While this obviously does not preclude the secondary analysis of research data, it clearly does not take a strong position in favor of it.

The sharing of research data must also be considered under the ethical principle of balancing harms and benefits. In many situations, the original data collection was paid for through research grants, funded by the taxpayer. This 'harm' to the community of taxpayers should be balanced by an appropriate benefit; the most logical way of maximizing that benefit is to ensure that the optimal use is made of all data collected. The allocation of harm and benefit in this case needs to be extended to include all of the participants in the research process, not just the immediate subjects of each piece of research. To quote Davey Smith (1994), "data paid for by public money are public property." The additional analysis of data also provides the benefit of increased confidence in the outcomes of research to the larger community.

"Publishing the findings of research in peer reviewed journals implies a high level of confidence by the authors in the veracity of their interpretation. Therefore it stands to reason that researchers should be prepared to share their raw data with other researchers, so that others may enjoy the same level of confidence in the findings" (Reidpath & Allotey, 2001).

The ethical principle of reducing harm can also be viewed as support for the secondary use of data. Subsequent use of data already collected reduces the impact on the larger population by involving a smaller number of research subjects and subjecting them to a smaller number of tests. On a more practical level, the use of previous studies can help formulate a good research question and refine the analysis carried out in subsequent studies (Davey Smith, 1994). Good methodology is one of the primary mechanisms for reducing harm.

The Tri-Council Policy Statement articulates the maximization of benefit as a guiding principle. This strongly supports the secondary use of data as a cost-effective and convenient mechanism for the advancement of knowledge. As research money becomes more restricted, increased secondary analysis will allow for ongoing research in situations where new data collection is hampered by lack of resources (Szabo & Strang, 1997). In situations where research will have a significant impact, for example in influencing public policy, it is essential that data be considered from many directions to reduce the possibility of flawed or weak conclusions. Hedrick (1985) states that secondary analysis allows for the "reinforcement of open scientific inquiry" (p.127) by providing for evaluation of research and the opportunity to replicate or reanalyze it using the same or different methods. A critical process will increase public confidence in the value of research and reduce the incidence of faked and inaccurate results. Increased public confidence may also benefit the research community by providing support for research funding.

The sharing of research data is a logical process that maximizes the benefits of research while reducing much of the potential for harm. Many of the anticipated risks and harms can be managed through improved methodologies. Once this position is understood and widely shared, those solutions will become part of the research ethos.

Solutions for anticipated risks

A number of writers have proposed solutions for the anticipated risks stated by individuals who are not in favor of secondary use of research data. While the list below is not complete, it demonstrates the breadth and ingenuity of researchers who are committed to good science and maximum benefit to the community.

A number of the solutions focus on the requirements for confidentiality from the secondary researcher. For example, Clubb, Austin et al. (1985) recommend "a form of licensing or swearing in as a condition for access to data with the possibility of legal sanctions and penalties for breaches of confidentiality" (p. 62). The British Sociological Association, cited in Heaton (1998) recommends that researchers consider obtaining consent that at least "covers the possibility of secondary analysis."

A number of approaches to the original consent form have been proposed. In some cases the original consent form includes provision for secondary research with the requirement that the secondary study receives approval from an ethics review committee. At the very least, this raises the question of potential secondary use in the minds of both the researcher and the subject, and allows respondents the opportunity to object should they wish. While this may not strictly meet the requirement for informed consent, it demonstrates an effort to resolve the situation early in the research process. It assumes that the secondary analyst is "bound by the same confidentiality and privacy restrictions as the primary analysts"(Szabo & Strang, 1997 p.7).

Better anonymization can be built in by the original researcher as a required part of research ethics approval for gathering data concerning humans. Proposed methods include a uniform practice of removing names and substituting numeric codes, removing occasional data values that reflect rare attributes and could allow for identification of specific individuals and organizations,

aggregating data in such a way that the performance of identifiable individuals or organizations is not obtainable, and various forms of encryption (Clubb et al., 1985; Johnson & Sabourin, 2001). This requires a better understanding of which identifying items data need to be maintained to keep the data useful while protecting the respondents. While it can be argued that some forms of data such as photographs or diaries have little value without identifying information, these should be regarded as exceptions rather than the norm and general policy should not be based on them.

A mathematical solution has been proposed that adds enough uncertainty to statistical analysis to prevent the identification of individuals while not significantly affecting the outcome of the analysis. The process, known as "jittering", is defined by Johnson and Sabourin (2001) as "adding a small, normally distributed random value with a mean of zero to all fields that might be used to identify an individual by matching against publicly accessible records."

The real problem may not be a lack of potential solutions, but a reluctance to implement them. This could be encouraged through a number of means outlined by Sieber (1991) in support of secondary research:

· In appealing to enlightened self interest grant bodies could require willingness to share data as funding criterion. This is already required by some funding bodies (Davey Smith, 1994). If not a requirement, funding priority could be given to those who create and share important data files and to research which builds on upon existing data files. Note that this requirement only works if it is monitored and there is a mechanism for sharing.

· To minimize the potential harm to researchers through not having their work adequately cited, the research community could require the implementation of clear and enforced standards for citation of data files. Standards for authorship should include identification of the source of data.

· In order to reduce researcher fears about secondary use of data, research education should be enhanced to include improved understanding about the advantages, process, and barriers in data sharing.

Funding agency policy statements could be a stronger advocate for secondary use of research data by including further instructions for the original researchers. It should work from the assumption that data sharing is standard practice unless there are specific reasons for prohibiting secondary analysis. It would then include, for example, the requirement for the secondary analyst to properly recognize the original researcher and a clear stipulation of the conditions under which data sharing is prohibited. This would facilitate good science by removing the potential conflict of interest that occurs when a researcher must decide whether or not to share data.

Research ethics approval could require a process for coding, storing and providing access to the data in a uniform fashion. This could be accomplished by adding an information professional, such as a librarian or an archivist, to the research team (Humphrey et al, 2000). A uniform requirement would mean that the burden of this additional work was evenly spread among researchers and would be considered as part of the original research design.

This has primarily been a discussion of the ethical issues surrounding data sharing. There is also the practical consideration of whether researchers are prepared to voluntarily share their data, or whether they have maintained it in a form that allows it to be used by other people. (Corti, Foster, & Thompson, 1995; Reidpath & Allotey, 2001) At this time, to share or not is still largely an ad hoc decision made by individual researchers. A clearly articulated policy that evaluated the situation on scientific merit and an analysis of harms and benefits would ensure that the ethical principles were the basis for decision-making.

Conclusion

The secondary analysis of existing research data provides many exciting opportunities for the development of new knowledge. It can be aligned with the ethical principles of research in many countries by minimizing the respondent burden and maximizing the potential benefits from the data. To make a change in the research culture requires strong advocacy on the part of data librarians, to change the thinking of funding bodies, regulatory agencies and researcher.

Traditionally we have not required that the potential for data sharing be a part of every research proposal. Now technology has provided us the opportunity to 'build the corpus of knowledge, not through the frenzied winnowing that has characterized our evaluations in the past but through an orderly interlocking of the puzzle pieces contributed by the disparate sub-fields. We have the means, for the first time in our history, to begin putting together the full picture of human behaviour' (Johnson & Sabourin, 2001). It is important that we champion the changes needed to accept this challenge, and to advocate for the creation of an ethical basis that requires the development and implementation of strategies to overcome potential barriers to data sharing.

"Reduce, reuse, recycle". This phrase has shaped a generation of behaviors about environmental concerns. Governments and funding agencies have promoted changed behaviour through investment in infrastructure, and in policy directions. The same thinking can be used to shape our understanding about ways of reducing the costs and burdens of data collection, increasing the value of research, and maximizing the benefits for everyone involved in the research process.

* Margaret Law, University of Alberta, margaret.law@ualberta.ca.

With this article Margaret Law won the IASSIST Strategic Plan Publication Award in 2005.

The author wishes to acknowledge the advice received from Dr. G.Griener, Department of Philosophy, University of Alberta.

References

Clubb, J. M., Austin, E. W., Geda, C. L., & Traugott, M. W. (1985). Sharing Research Data in the Social Sciences. in S. E. M. M. E. S. M. L. Fienberg (editors), Sharing Research Data (pp. 39-88). Washington, D.C.: National Academy Press.

Corti, L., Foster, J., & Thompson, P. (1995). Archiving qualitative research data. Social Research Update, 10.

Dale, A., Arbor, S., & Proctor, M. (1988). Doing Secondary Analysis (Contemporary Social Research Series No. 17). London: Unwin Hyman Ltd.

Davey Smith, G. (1994). Increasing the Accessibility of Data. BMJ, 308(June 11), 1519-1520.

Fienberg, S. E., Martin, M. E., & Straf, M. L. (editors). (1985). Sharing Research Data. Washington: National Academy Press.

Freedom of Information and Protection of Privacy Act, RSA 2000, c. F-25, sec. 42

Health Information Act, RSA 2000, c. H-5, ss. 48 – 56, (2002). Ottawa: Government of Canada.

Heaton, J. (1998). Secondary analysis of qualitative data. Social Research Update, (22).

Hedrick, T. E. (1985). Justifications for and Obstacles to Data Sharing. in Sharing Research Data (pp. 123-147). Washington, D.C.: National Academy Press.

Homan, R. (1992). The Ethics of Open Methods. The British Journal of Sociology, 43(3), 321-332.

Humphrey, C. K., Estabrooks, C. A., Norris, J. R., Smith, J. E., & Hesketh, K. L. (2000). Archivist On Board: Contributions To The Research Team. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 1(3).

Johnson, D. H., & Sabourin, M. E. (2001). Universally accessible databases in the advancement of knowledge from psychological research. International Journal of Psychology, 36(3), 212-220.

Kalman, C. J. (1994). Increasing the Accessibility of Data. 309(17 September), 740.

Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D., & Walford, N. (1991). The Case for Samples of Anonymized Records from the 1991 Census. Journal of the Royal Statistical Society. Series A (Statistics in Society), 154(2), 305-340.

Reidpath, D. D., & Allotey, P. A. (2001). Data Sharing in Medical Research: An Empirical Investigation. Bioethics, 15(2).

Sieber, J. (1991). Social Scientists' Concerns About Sharing Data. in Sharing Social Science Data; advantages and challenges (pp. 141-150). Newbury Park, California: Sage Publications, Inc.

Stanley, B., & Stanley, M. (1988). Data Sharing: The Primary Researcher's Perspective. Law and Human Behavior, 12(2), 173-180.

Szabo, V., & Strang, V. R. (1997). Secondary Analysis of Qualitative Data. Advances in Nursing Science, 20(2), 66-74.

Trainor, T., & Dougherty, K. (2000). Selected issues concerning disclosure avoidance in the context of user-defined geography. Statistical Journal of the UN Economic Commission for Europe, 17(2), 133-139.

Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans. Medical Research Council of Canada//Natural Sciences and Engineering Research Council of Canada//Social Sciences and Humanities Research Council of Canada. August 1998.