From Card, J.J., Kuhn, T.J., 2015. Archiving: Ethical Aspects. In: James D. Wright (editor-in-chief), International Encyclopedia of the Social & Behavioral Sciences, 2nd edition, Vol 1. Oxford: Elsevier. pp. 916–920.
ISBN: 9780080970868

Archiving: Ethical Aspects

Josefina J Card, Sociometrics Corporation, Los Altos Hills, CA, USA
Tamara J Kuhn, ETR, Scotts Valley, CA, USA

Abstract

Archiving refers to the process of transferring records from the individual or organization who created the material, to a repository for appraising, cataloging, organizing, preserving, and providing access to others. Social and behavioral science research engenders by-products that are worthy of archiving for a variety of future uses, including the furthering of basic and applied research, policy making, and the development and replication of effective intervention programs. Examples of such archive-worthy by-products are research data, their associated data collection instruments (questionnaires, interview protocols), and, more recently, intervention programs shown to be effective by evaluation research. Ethical issues occasionally arise in the archiving process. These issues typically relate to protecting the integrity of the selection process; protecting respondents' confidentiality; censoring potentially controversial or offensive material; the timing of the release of information to an archive; assignment of due credit to both original producer and archivist; the tension between fidelity and usability in the archiving process; and the ownership of the research and development by-products contained in the archive. Approaches to handling and resolving these ethical challenges are discussed.

Archiving By-Products of Social and Behavioral Research

Archiving refers to the process of transferring records from the individual or organization who created the material to a repository for appraising, cataloging, organizing, preserving, and providing access to others (National Archive of Australia, 2013; Pearce-Moses, 2005).

The social and behavioral sciences produce intellectual by-products at various stages of the research process that, if preserved and organized, could further basic and applied research, aid policy making, and facilitate the development and replication of effective social intervention programs. A variety of institutions preserve such materials. They include government archives, academic data archives and libraries, and specialized organizations in both the public and private sectors. Professional organizations of social science archivists and librarians have been formed to further the field.

The National Archives and Records Administration (NARA) is the US federal agency that preserves and ensures access to those official records which have been determined by the Archivist of the United States to have sufficient historical or other value to warrant their continued preservation by the Federal Government; and which have been accepted by the Archivist for deposit in his custody (44 U.S.C. 2901). Information about NARA's electronic records holdings (most of which are data files) can be obtained from the Internet site http://www.archives.gov/.

CESSDA ERIC, while not itself a data archive, is a large consortium for social science data archives across Europe. Established in the 1970s, the organization provides access through their catalog to information and links to nearly 9000 datasets stored at CESSDA archives throughout Europe. Those member organizations serve more than 30 000 social science researchers yearly, together deliver over 70 000 data collections each year, and continue to add more than 1000 data collections annually (http://www.cessda.org/).

The Inter-university Consortium for Political and Social Research (ICPSR) is the largest academic social science data archive. Founded in 1962 at the University of Michigan, ICPSR is a membership-based organization, which provides access to a large archive of computer-based research and instructional data in political science, sociology, demography, economics, history, education, gerontology, and criminal justice. More information about ICPSR and its holdings is available from http://www.icpsr.umich.edu.

The UK Data Archive, established more than 40 years ago, is the UK's largest collection of digital research data in the social sciences and humanities. The archive contains several thousand datasets acquired from the academic, public, and commercial sectors; and some datasets are available through Nesstar so that frequency counts and questions can be viewed online (http://data-archive.ac.uk/).

Norwegian Social Science Data Services, owned by the Ministry of Education and Research, is one of the largest archives of Social Science research data. Data are organized in four main categories: Individual Level Data, Regional Data, Data About Institutions, and Data About the Political System. Many of their data sets can be accessed, analyzed, and visualized with maps, graphs, and tables using Nesstar. In addition to direct on-line access, data can be downloaded and/or exported to preferred format (http://www.nsd.uib.no/nsd/english/index.html).

An example of private sector archives, Sociometrics Corporation was established in 1983. The company's primary mission is the development and dissemination of social science research-based resources for a variety of audiences, including researchers, students, policymakers, practitioners, and community-based organizations. Sociometrics has pioneered in the establishment and operation of topically focused data, instruments, and (since the mid-90s) program archives (http://www.socio.com): (1) *Data Archives*: collections of original machine-readable data from over 500 exemplary studies, many of them longitudinal, on the American family,

International Encyclopedia of the Social & Behavioral Sciences, Second Edition, 2015, 916–920

teen sexuality and pregnancy, social gerontology, disability, AIDS and STDs, maternal drug abuse, and geographic indicators; (2) *Instrument Archives*: the questionnaires, interview protocols, and other research instruments that were used to collect the data in the data archives; (3) *Program Archives*: collections of program and evaluation materials from several dozen intervention programs that have proven effective in preventing risky behaviors such as unprotected sex and drug use and in treating psychological and behavioral disorders, such as anxiety, depression, and aggression in children. These topically focused archives synthesize research in the field in one place; facilitate further research with the best existing data and accompanying instruments; promote data-based policy-making; and help service providers and practitioners use the insights gained from research (Card et al., 2007, 2006; Card and Kuhn, 2006).

Two professional organizations of social science data archivists and librarians are the Association of Population Libraries and Information Centers (APLIC) and the International Association for Social Science Information Service & Technology (IASSIST). APLIC's membership, consisting of both individuals and organizations, represents some of the oldest population and family planning agencies and institutions in the United States. IASSIST is an international organization dedicated to the issues and concerns of social science data librarians, data archivists, data producers, and data users. This unique professional association assists members in their support of social science research. The APLIC and IASSIST membership lists provide pointers to the various social science data collections housed all over the world (http://www.aplici.org/; http://www.iassistdata.org/).

Ethical Aspects in Archiving

The development of collections such as those contained in the above archives involves a series of decisions with ethical considerations and implications.

Protecting the Integrity of the Selection Process

Given the limited nature of resources allocable to archiving, how should the contents of collection be selected? Some archives sidestep this challenge by merely cataloging and warehousing archival material (e.g., data sets) donated to them by the field. While this procedure undoubtedly results in the lowest per-capita archiving cost, the quality of the resultant archival collection is uncertain at best. A better procedure is to set objective technical and substantive standards for inclusion in the archival collection and then actively recruit material that meets or surpasses such standards. Setting standards for inclusion can be problematic and contentious. Archivists and curators may often disagree on the future research needs in the organization or in the field. There are some general guidelines that can be used to minimize these challenges, such as making sure that the process used to make data selection decisions is transparent and accountable and guided by policy and legal requirments (Whyte and Wilson, 2010). Decisions about inclusion need to be based on a set of objective criteria for assessing the long-term significance of the research data, which should be developed and agreed upon by larger research communities, institutions, or experts in the field. These guidelines are unlikely to be static. Another challenge is that over time it is likely that standards and thinking in the field will change and inclusion criteria will need to be modified. For example, several years ago criteria for inclusion in an archive of effective behavioral interventions may have required that an evaluation of the intervention be conducted and that it should demonstrate continued behavior change at a 3-month follow up time period. However current standards require a 6-month follow up to demonstrate effectiveness and merit inclusion.

The previously described data and program archives at Sociometrics have worked with Scientist Expert Panels in establishing criteria for inclusion in various collections. For the data archives, the selection criteria are scientific merit, substantive utility, and program and policy relevance of the data sets comprising the collection. For the program archives, the selection criteria are documented effectiveness in preventing the particular social problem or disease (e.g., drug use, teen pregnancy, sexually transmitted disease, HIV/AIDS) or in changing these problems' risky-behavior antecedents (e.g., delaying age at first intercourse, increasing the use of contraception and/or an STD-prophylactic at first and every act of sexual intercourse, and abstaining from or reducing the frequency of drug use). Having established objective inclusion criteria, archive staff then work with their respective Scientist Expert Panels to identify and prioritize available data sets and intervention programs for inclusion in the collections. The end result is an archival collection with integrity and credibility.

Protecting Respondents' Confidentiality

Respondent confidentiality is one of the areas of greatest ethical concern, because once data are released for public use it is impossible to monitor use and ensure that confidentially is respected (ICPSR, 2012). Data archives often contain responses to sensitive questions, some of which, for example, ask respondents to admit to illegal, immoral, or 'private' behavior, such as abortion, premarital or extramarital sexual activity, mental illness, alcohol abuse, and drug use. How can the researcher's need to know (the incidence, prevalence, antecedents, and consequences of these social problems) be balanced against respondent's rights to privacy? This ethical consideration is most often addressed by stripping all archival material of information that could be used to identify individual subjects. Direct identifiers, e.g., name, address, social security number, and exact date of birth (often only month and year of birth are included in a public-use database), are typically easily removed. However indirect identifiers, variables that make unique respondents visible when paired with another, such as a participant with a low-incidence disability attending a rural school with unique characteristics, pose a greater challenge.

A problem arises when data holders want to strip the data set of key variables, such as those measuring the sensitive behaviors listed above, prior to placing a data set in a public use archive. This desire is motivated by the fear that such information could be linked to particular respondents by malicious, hardworking sleuths, even without the help of individual identifiers such as name, address, and so forth. Such censorship restricts the range of uses to which the data set can be put by

future researchers and archives typically make an active attempt to find alternate solutions. There are several solutions to problem. Indirect identifiers can be eliminated or modified by restricting the upper range of the variables, by combining or collapsing variables, or by adding random variation or stochastic error to the variable (ICPSR, 2012; Lawrence, 2010). Data donors typically consult with the archive producers to resolve concerns about identifiers. Other solutions focus on restricting use of the data. It is becoming common to require data users to sign a data use agreement that specifies a set of highly controlled conditions for the use of data prior to being granted access. This may be something as simple as signing a confidentially agreement and pledging to use the data only for research purposes.

However for datasets that have highly confidential material, more requirements are put in place. Many agreements require that the researcher provide an abstract of their research questions and explain why access to the restricted use file is required. They also have to outline a plan for safeguarding the data, and in some cases provide evidence of Institutional Review Board approval. Researchers may be given access to the data only for a limited time period, at the end of which they are expected in good faith to destroy them. There are instances when access to the most confidential data may require even more restrictions and users are typically only able to access those materials through a data enclave. This has meant that a researcher has to access those materials in a secure physical location. These enclaves allow access to the original data in a controlled setting and have physical security measures, such as video monitoring, key card entry, and controlled access; also, the computers housing the data are not connected to the Internet. Use is monitored by archive staff who confirm that materials are not removed and that analyses do not contain any confidential information. Virtual enclaves are replacing physical enclaves in some archives. Virtual enclaves are data portals that allow remote but restricted access to the application system and the restricted data. This access is monitored and the systems restrict users from emailing, copying, or moving files outside of the environment.

The original researcher should collaborate with the archive to determine the appropriate level of protection, aiming to allow the widest access to the data while maintaining confidentiality.

Consent for Data Sharing

While data sharing has long been encouraged and even required by research funders and a data sharing plan is often a required part of a research proposal, the original researcher often has not asked for the consent of the participants for sharing the data beyond the original research project. This can present an ethical dilemma for both the data donor and the archive. While requirements for consent vary by country and funder, this issue can be most easily addressed in the research-planning process and in planning for data sharing when developing study materials. For example, the Australian government has established the position that at minimum, consent forms should not preclude data sharing, e.g., by promising to destroy data; and more specifically, the National Statement defines three levels of consent for the future use of data that must be made clear to the participant: (1) Specific: which is limited only to the project; (2) Extended: which gives permission for the use of the data in future projects in the same general area of research; or (3) Unspecified: which allows for the use of data or tissue for any future research (NHMRC, 2007). If consent was not originally given for data sharing then the archive and the developer have to weigh the benefits of distributing the data against the problem of confidentiality and consent, with one likely outcome – the distribution only of data that is fully anonymous.

Special Considerations for Clinical Trial and Biomedical Data

While most social and behavioral science research projects and archives do not include clinical trial data, there are instances in which clinical trial or biomedical data may be included within a social science archive, such as the recently released Reaching for Excellence in Adolescent Care and Health (REACH) dataset (Adolescent Medicine HIV/AIDS Research Network, 2000; Pardini and Wu, 2011) which combines data on medical visits and laboratory results with access to tissue samples and a series of behavioral questions. REACH data were placed in a social science archive by the US National Institutes of Health specifically to expand the reach and analyses of the data to those beyond the medical sciences. The archiving of clinical trial and biomedical data bring additional ethical considerations. A majority of clinical trials are conducted by biopharmaceutical companies or clinical research groups and the data are typically not made accessible to other researchers or the public, and are rarely placed within a public archive. The lack of sharing within the scientific community has reduced the ability of scientists to validate the findings and examine the methodology of these studies. In many cases, (estimated at between 25 and 50% of trials), the results of these clinical trials are not published or made public (Ross et al., 2009, 2012). For those that have been published, in comparing published articles with trial protocols, 50% of efficacy and 65% of harm outcomes per trial have been shown to be incompletely reported and biased toward the reporting of statistically significant findings (Chan et al., 2004). This lack of transparency and availability of clinical trial data have led to calls for stronger data-sharing requirements. In response, in 2013, EFPIA (which represents the pharmaceutical industry operating in Europe) and PhRMA (the Pharmaceutical Research and Manufacturers of America which represents several US biopharmaceutical research and biotechnology companies) committed to Principles for Responsible Clinical Trial Data Sharing (PhRMA and EFPIA, 2013). Under this agreement, member companies commit to enhance data sharing with qualified researchers, share results with the patients who participate in clinical trials, enhance public access to clinical study information, and reaffirm their commitment to publish clinical trial results. However, rather than committing to archiving the data in a public archive it is the companies that determine who is a qualified to access the data, which again may limit the public availability of such data. One final consideration as clinical trial data becomes available to more researchers is the question of what to do with new results that may be found for participants in the study. In order to make the data available to researchers, the data is deidentified. This makes the conveying

of new results to participants nearly impossible, without the cooperation of the original researchers.

Censoring Potentially Controversial or Offensive Material in Collections

Intervention program archives occasionally encounter an analogous censorship-related challenge. For example, several of the effective programs selected for archiving by the Scientist Expert Panel of PASHA (Card et al., 2007) contain sexually explicit material that could be viewed as offensive and inappropriate by some individuals and communities. However, because these prevention programs are targeted at high-risk, already sexually active youth, the material could also be seen as appropriate, even necessary, to drive home relevant points. In addition, these programs, like other PASHA programs, meet the collection's inclusion criterion of demonstrated effectiveness in changing sexual risk-related behavior in at least one subgroup of teens. The decision was made to include the material without alteration or censorship but to publicize and disseminate the collection as an eclectic one, with different schools and communities being encouraged to replicate programs consistent with their own values, norms, and target populations. An accompanying database of program abstracts was developed so that both the approach and the content of the program packages could be perused, prior to requesting the program from the archive.

Timing of Release of Information to an Archive

Holders of data sets and developers of effective programs often, and understandably, want to reap some payoff from their professional investments by keeping the data or programs to themselves until they have published what they wish to from the data (or have tweaked the intervention program in several implementation cycles to their satisfaction). The ethical issue arises when this 'private' or 'proprietary' period of time stretches to what the field would view as abnormally long. This is especially true when the data were collected, or the intervention program developed, with government funds. Several US federal agencies are trying to forestall the problem by building resource-sharing ground rules into the original funding award document. For example, the National Science Foundation states that researchers with quantitative data should be prepared to place their data, in fully cleaned and documented form, in a data archive or library within a year after the expiration of an award (NSF, 2013). In contrast, the US National Institutes of Health is not as specific and instead expects those receiving large awards to share their data in a timely fashion with 'timely release and sharing,' defined as no later than the acceptance for publication of the main findings from the final data set. Ideally this solution gives the original developer the fair 'head start' his or her efforts have earned, while ensuring that the data collected with government funds will be shared with the field before it gets stale. However two primary issues remain; first, unfortunately the data sharing promises or pledges included in the funded grant proposal are typically not enforced and secondly, in many instances the timely release is based on the discretion of the researcher and many are reluctant to release the data, sometimes for more than a decade, until the data are outdated.

Assignment of Due Credit to Both Original Producer and Archivist

Data sets and program materials are typically received by an archive in a format that data developers and their colleagues find workable but one not yet suitable for public use. The archivist contributes significant additional value in preparing the database for public use. For example, with the approval of the data donor, inconsistencies in the database are eliminated, or at least documented. The documentation is augmented, both at the study level (describing study goals, sampling, and data collection procedures) and at the variable level (assigning names and labels for each variable; documenting algorithms for constructed scale variables). Occasionally, the variable and scale documentation is done using the syntax of a popular statistical analysis package such as SPSS (Statistical Package for the Social Sciences) or SAS (Statistical Analysis System), facilitating future data analysis. Archivists who prepare intervention program packages for public use make analogous contributions. Program materials are edited and 'prettified' for public use. User's Guides, Facilitator's Manuals, checklists, and handouts are created so that the package is implementation-ready in the absence of the original developer. In short, the archiving process is best viewed and executed as collaboration between original developer and archivist. Care must be taken to give due credit for the final product to all individuals, teams, and institutions involved.

Tension between Fidelity and Usability in the Archiving Process

The collaborative model is productive not only for assignment of due credit but also for joint resolution of the fidelity versus usability issues that occasionally arise during the archiving process. Others will be using the materials and have the opportunity to note errors in data or methodology, which leads to several ethical questions. Should obvious errors in the data base be corrected or only documented? Should resulting concerns about methodology be highlighted for those who acquire a dataset or program? Should original program materials that were found effective in the developmental site be altered when replication sites find them unclear or when the curriculum they present is based on out-of-date data? Issues such as these are best resolved on a case-by-case basis by the archivist and original developer, working side by side in collaborative fashion.

Ownership of the By-Products of Research and Development

The purposes and procedures of the archive accepting a donation should be made clear to the donor at the outset. It should be communicated to data donors that the research by-product they are donating toward is being put in a public archive whose main goal is the preservation of the resource.

Some archives also actively publicize and disseminate their holdings. In addition, as seen above, archives vary in the extent to which they work with the donor in 'upgrading' the material for public use. The donor should be informed in advance of what to expect along these lines. Issues of credit and ownership should also be agreed to before archiving work begins. How will professional credit for the collaborative product be allocated? Will the resultant product be sold to the end user (at cost or for profit) or given away? If the product will be sold for profit, will royalties be given to the original developer? If the product will be sold at cost, free, or discounted copies be made available to the original developer?

Conclusion and Looking to the Future

Social science research yields many by-products that, if properly archived, can be used to further future research, aid policymaking, and foster the development and replication of effective prevention and treatment programs. Several challenges, some with ethical implications, arise in the archiving process as it exists today. Many are resolvable with good will, a commitment to upholding the standards established by those in the research and archiving communities, and commitment to the public good on the part of both original developers and archivists. However, the recent discussions around data transparency of research findings and publications based on unreleased clinical data and social science data collected by companies, such as Google and Facebook, are likely to lead to new discussions around the ethics of archiving and the concept of archived data as being publically available. As we move into the age of 'big data' and collections of social networking data owned by corporations, discussions around confidentiality and privacy will probably evolve. The burden will fall to the community of scientists, archivists, and funding agencies to ensure that data sharing and data archiving for public use continues and is conducted ethically.

See also: Data Bases and Statistical Systems: Archives and Historical Databases; Privacy of Individuals in Social Research: Confidentiality; Surveys and Polling: Ethical Aspects.

Bibliography

Adolescent Medicine HIV/AIDS Research Network, 2000. Reaching for Excellence in Adolescent Care and Health (REACH): 1996–2000 [Computer File]. SUNY Health Sciences Center (Producer), Brooklyn, NY. Sociometrics Corporation (Producer & Distributor), Los Altos, CA.

Card, J.J., Lessard, L., Benner, T., 2007. PASHA: facilitating the replication and use of effective adolescent pregnancy and STI/HIV prevention programs. Journal of Adolescent Health 40 (3), 275.e1–e14.

Card, J.J., Kuhn, T., Wells, T., 2006. User-centered design and innovation in the sociometrics social science electronic data library (SSEDL). IASSIST Quarterly 30 (2), 12–17.

Card, J.J., Kuhn, T., 2006. Development of online suites of social science-based resources for health researchers and practitioners. Social Science Computer Review 24 (2), 255–261.

Chan, A.W., Hrobjartsson, A., Haahr, M.T., Gotzsche, P.C., Altman, D.G., 2004. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. Journal of the American Medical Association 291, 2457–2465.

Inter-university Consortium for Political and Social Research (ICPSR), 2012. Guide to Social Science Data Preparation and Archiving: Best Practice throughout the Data Life Cycle, fifth ed. Ann Arbor, MI.

Lawrence, D., 2010. Analysis of Public Use Microdata Files: A Researcher's Perspective. National Statistical Service. http://www.nss.gov.au/nss/home.nsf/NSS/58CC18F435E0800CCA2579E20006BA4B?opendocument – Analysis of public use microdata files.

National Archives of Australia, 2013. Glossary. Retrieved 08.01.13. http://www.naa.gov.au/records-management/publications/glossary.aspx.

NHMRC, 2007. National Statement on Ethical Conduct in Human Research. National Health and Medical Research Council, Australian Research Council, Australian Vice-Chancellors' Committee. Australian Government.

National Institutes of Health, 2012. NIH Grants Policy Statement. http://grants.nih.gov/grants/policy/nihgps_2012/nihgps_ch8.htm.

National Science Foundation, 2013. Directorate for Social, Behavioral and Economic Sciences, "Data Archiving Policy".

Pardini, B., Wu, J., 2011. Reaching for Excellence in Adolescent Care and Health (REACH): 1996–2000: A User's Guide to the Machine-readable Files and Documentation (Dataset 0158). Sociometrics Corporation, Los Altos, CA.

Pearce-Moses, R., 2005. A Glossary of Archival and Records Terminology. Archival Fundamentals Series II. Society of American Archivists, Chicago.

PhRMA & European Federation of Pharmaceutical Industries and Associations, 2013. Principles for Responsible Clinical Trial Data Sharing. Retrieved 08.05.13. http://phrma.org/sites/default/files/pdf/PhRMAPrinciplesForResponsibleClinicalTrialDataSharing.pdf.

Ross, J.S., Mulvey, G.K., Hines, E.M., Nissen, S.E., Krumholz, H.M., 2009. Trial publication after registration in ClinicalTrials.gov: a cross-sectional analysis. PLoS Medicine 6, e1000144.

Ross, J.S., Tse, T., Zarin, D.A., Xu, H., Zhou, L., Krumholz, H.M., 2012. Publication of NIH funded trials registered in ClinicalTrials.gov: cross sectional analysis. British Medical Journal 344, d7292.

Whyte, A., Wilson, A., 2010. "How to Appraise and Select Research Data for Curation". DCC How-to Guides. Digital Curation Centre, Edinburgh. Available online: http://www.dcc.ac.uk/resources/how-guides.