DATA MANAGEMENT IN SOCIAL SCIENCE RESEARCH

The Data Life Cycle

Researchers should plan for eventual archiving and dissemination of research data before the data even come into existence. According to Jacobs and Humphrey (2004), "Data archiving is a process, not an end state where data is simply turned over to a repository at the conclusion of a study. Rather, data archiving should begin early in a research and incorporate a schedule for depositing products over the course of a research's life cycle and the creation and preservation of accurate metadata, ensuring the usability of the research data itself. Such practices would incorporate archiving as part of the research method."

DATA SHARING

Archives and domain repositories that preserve and disseminate social and behavioral data perform a critical service to the scholarly community and to society at large by ensuring that these culturally significant materials are accessible in perpetuity. The success of the archiving endeavor, however, ultimately depends on researchers' willingness to deposit their data and documentation for others to use.

Data sharing also allows scientists to test and replicate each others' findings. "The replication standard holds that sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information from the author" (King 1995).

There are many benefits to data sharing that go beyond replication. Fienberg (1994) argues that data sharing:

- Reinforces open scientific inquiry. When data are widely available, the self-correcting features of science work most effectively.
- Encourages diversity of analysis and opinions. Researchers having access to the same data can challenge each other's analyses and conclusions.
- Promotes new research and allows for the testing of new or alternative methods. Examples of data being used in ways that the original investigators had not envisioned are numerous.
- Improves methods of data collection and measurement through the scrutiny of others. Making data publicly available allows the scientific community to reach consensus on methods.

- Reduces costs by avoiding duplicate data collection efforts. Some standard datasets, such as the General Social Survey and the National Election Studies, have produced literally thousands of papers that could not have been possible if the authors had to collect their own data. Archiving makes known to the field what data have been collected so that additional resources are not spent to gather essentially the same information.
- Provides an important resource for training in research. Secondary data are extremely valuable to students, who then have access to high-quality data as a model for their own work.

Planning Ahead for Archiving and Preservation of Data

Data management and sharing plans should be developed in conjunction with an archive to maximize the utility of the data and to ensure the availability of the data in the future.

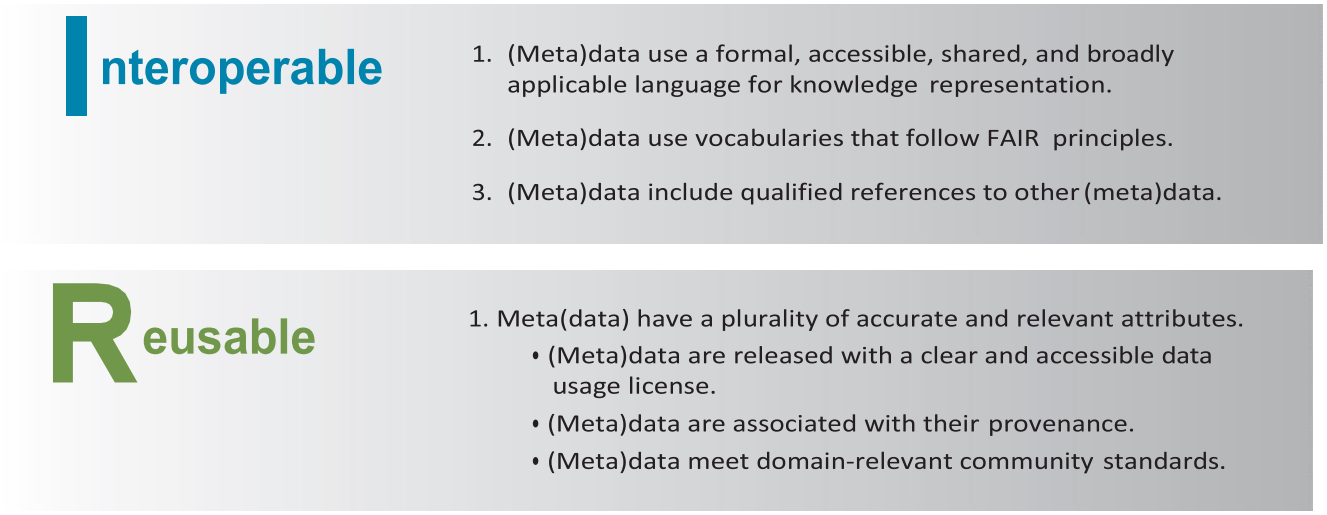
It recommends that researchers consult as early as possible with the data archive in which they plan to deposit data; this will facilitate preservation and dissemination of the research data. Archiving and disseminating derived datasets — that is, those resulting from the combination of data from more than one data source, including existing data outside the current research scope — also should be considered. See following flowchart : Depositing Data, for a more in-depth discussion.

Findable

1. Data and metadata are assigned a globally unique, eternally persistent identifier.
2. Data are described with rich metadata.
3. (Meta)data are registered or indexed in a searchable resource.

Accessible

1. (Meta)data are retrievable by their identifier using a standardized communications protocol.
   - The protocol is open, free, and universally implementable.
   - The protocol allows for an authentication and authorization procedure, where necessary.
2. Metadata are accessible, even when data are no longer available.



Metadata

Describe the metadata to be provided along with the generated data, and discuss the metadata standards used. As metadata are often the only form of communication between the secondary analyst and the data producer, good descriptive metadata are essential for effective data use.Structured or tagged metadata, such as the XML format of the Data Documentation Initiative (DDI), are optimal because of the flexibility they offer in display. XML is also preservation-ready and machine-actionable. For a more detailed discussion on metadata and documentation, please see the "Best Practice in Creating Metadata" section in Section 5: Data Collection and File Creation.

Storage and Backup

Indicate how and where you will store copies of your research files to ensure their safety, as well as how many copies you will keep and how you will synchronize them. The best practice for protecting data is to store multiple copies in multiple locations.

Security

Describe measures you will take to ensure your data are secure. This is an important consideration over the entire life cycle of the data. Raw data may include direct identifiers of study participants and should be well protected during collection and processing. Examples of good security practices include access restrictions such as passwords, encryption, power supply backup, and virus and intruder protections.

Responsibility

State who will act as the responsible steward for the data throughout the data life cycle. Researchers should describe any atypical circumstances. For example, if there is more than one principal investigator, describe the division of responsibilities between them.

Intellectual Property Rights and Data Ownership

Indicate who will hold intellectual property rights to the data and other information created by the research, and whether these rights will be transferred to another organization for data distribution and archiving. If any copyrighted material (i.e., instruments or scales) are used, how will the research obtain permission to use or disseminate it? Data archives need a clear statement from the data producer of who owns the data before they can be disseminated. However, issues of data ownership can be complex. For example, principal investigators on federally funded researchs are responsible for collecting research data and publishing their research findings, but the resulting research data are typically owned by the institution where the principal investigator is employed.

Funding organizations expect researchers to share their data. Public archives can help universities meet those expectations without requiring a transfer of copyright along with research data. A copy of the research data can be shared publicly through an archive while ownership rights remain with the copyright holder. Agreements to publicly archive data typically grant a repository permission to preserve and disseminate the data.

Access and Sharing

Indicate how you intend to archive and share your data, and why you have chosen that particular option. Mechanisms for archiving and sharing include:

- Domain repositories, such as ICPSR (social science)

- Self-dissemination through a dedicated website created by the research team. Options for eventual dissemination should be arranged through an established archive after the self-dissemination period ends. A schedule of when dissemination will be turned over to a third party should be included. The archive may want to make a preservation copy

during the period of self-dissemination for a number of reasons: (1) to develop expertise with the data; (2) to process the data while knowledgeable staff are available; and (3) for general safekeeping.

- Preservation with delayed dissemination, in which the data producer arranges with a public data repository for archival preservation with dissemination to occur at a later date, usually within a year. With delayed dissemination, the deposit may be completed when it is easiest for the depositor and the archive to manage the data, as opposed to delaying preservation activities until the time has come to disseminate the data. Issues regarding the schedule for eventual dissemination, embargo periods, and human subject protections specific to these studies will be settled prior to deposit, as will ground rules on the extent of processing by archival staff while the study remains in the "preservation with delayed dissemination" category.

- Institutional repositories at academic institutions, which have the goal of preserving and making available some portion of the academic work of their students, faculty, and staff. Not all such repositories have the capacity to accept and curate data. There are generally two types of institutional repositories: those with a focus on a particular discipline, and those without. Each type provides certain benefits and drawbacks for data producers and users that should be considered when deciding which to use.

- Restricted-use collections. In cases in which masking of sensitive data would lessen the analytic power of a dataset, a restricted-use release may be appropriate.

Access to restricted-use data can be limited to approved researchers under controlled conditions. Some archives, including ICPSR, can provide both restricted-use and public-use releases, where the public files have been altered to prevent disclosure of sensitive information about survey participants. See Section 7: Addressing Confidentiality Issues, for more on protecting respondent confidentiality.

Budget

The investigator should outline the plans for and cost of preparing the data and documentation for archiving. Ideally, this should be planned in conjunction with an archive. Some potentially costly activities are listed below:

- For quantitative data, investigators should allocate resources to create system-specific files with appropriate variable and value labeling, to supply the syntax for derived variables, etc.

- For some data, especially certain types of qualitative data, there may be costs associated with storage due to the size of the data files.

- Grant applications should allocate sufficient time and money for the preparation of high-quality documentation.

- Informed consent and confidentiality issues impact costs for archiving. For clarity, informed consent agreement forms should be drawn up at the start of the research.

- It is strongly recommended that a set period of time be dedicated to preparing and collating materials for deposit. This normally comprises the majority of the costs for archiving.

Data Backups

All relevant files, particularly datasets under construction, should be backed up frequently even more often than once a day — to prevent having to re-enter data. Master datasets should be backed up every time they are changed in any way. Computing environments in most universities and research centers support devices for data backup and storage. It is also advisable to maintain a backup copy of the data off-site, in case of an emergency or disaster that could destroy years of work.

Virtual data enclaves : These data portals allow users to obtain remote access to restricted data that would not otherwise be available for research. This often includes using a restricted access application system, getting set up with secure remote access to the restricted data (including possible on-site inspection), monitoring research behavior during data access, and having analytic results reviewed for disclosure risk before they are permitted to leave the secure environment. Such systems generally prevent users from emailing, copying, or otherwise moving files outside of the secure environment, either accidentally or intentionally.

Physical data enclaves : A physical data enclave is a secure data analysis laboratory that allows access to the original data in a controlled setting. Secure data enclaves have added security features to ensure the safekeeping of the most confidential data. They typically have appropriate physical security measures (no windows, video monitoring, key card entry) to strictly control access. Their computing environments are not connected to the Internet, but rather have their own network server (connected to a small number of work stations). Researchers using the enclave are monitored by archive staff who see to it that no unauthorized materials are removed. Any analyses produced

are scrutinized to determine that they do not include any potential breaches of confidentiality. Other policies and procedures also govern the use of restricted data in enclaves.

Secure Survey Documentation and Analysis (SSDA) :SSDA is an online data analysis program that performs bi-variate cross-tabulation, comparison of means, correlation, and regression analyses. SSDA is designed to provide a safe, reliable way to distribute restricted-use data publicly, thereby democratizing access to data that was previously unavailable or required special procedures to obtain. SSDA automates several disclosure protections that prevent the use of organization-defined high-risk variables, singularly or in combination, and restrict types of output commonly associated with disclosure risk (e.g., small unweighted sample sizes).

References

www.icpsr.umich.edu