PROTOCOL LETTER

Key Elements to Consider in Preparing a Data Sharing Plan

Research results developed with NIH funding should be broadly available to the research community for furthering research. This resource document is intended to assist applicants by outlining certain key elements that should be addressed in any data sharing plan.

While the precise content of a data sharing plan may vary depending on the data being generated and collected, addressing the basic questions of What, Who, Where, When, and How can assist researchers and research administrators in formulating a meaningful data sharing plan that communicates essential information about:

1. **WHAT** data will be shared?

To optimize the benefits of data sharing, ***final research data along with metadata and descriptors should be shared to make sharing meaningful and usable by other researchers.*** In describing what data will be shared, a data sharing plan should indicate:

- What types of data are to be collected in the study and shared (e.g., genetic, physiological, clinical, medical history, etc.)?
- Will the study include unique data that cannot be readily duplicated (e.g., large surveys that are too expensive to replicate; studies of unique populations, such as centenarians; studies conducted at unique times, such as a natural disaster; studies of rare phenomena, such as rare metabolic diseases; etc.)?
- Will individual-level data or raw data also be shared, and if so, will the whole data set be shared?
- Will aggregate data (e.g., summary statistics or tables) also be shared? Will the analytical methods used (tools and parameters) be defined?
- What data quality control measures will be implemented?
- What data documentation will be shared (e.g., metadata, descriptors, schema) so that others can understand and use the dataset and to prevent misuse, misinterpretation, or confusion?
- What commonly accepted data standards or standardized vocabularies will be used to enable others to interpret the data and improve interoperability with other data systems?
- What format will be used to encode the data? Will this format be consistent with extant, commonly used standards?
- In addition to final research data, what other data will be available?

2. **WHO** will have access to the data?

To maximize the benefits of data sharing, ***data should be shared as broadly as possible to the extent consistent with applicable laws, regulations, rules, and policies.*** In describing who will have access to data, a data sharing plan should indicate:

- Will the general public have access to some or all of the data?
- Will access to certain data or certain components of the data be restricted to qualified researchers, e.g., to address specific rules, laws, regulations or policies (e.g., IRBs, human subjects, informed consent, etc.)?
- If data access is restricted, what are the justifications/criteria for restricting access (e.g., relevant laws (local, State, Federal, etc.), regulations, rules, institutional policies, IRB approvals, and consent documents)?
- What will researchers who seek to obtain data need to do to comply with any data access restrictions?
- Are there any limitations on release of data that may be considered “sensitive”?
- What data sharing agreements will be necessary to appropriately restrict the transfer of protected, sensitive, or confidential data to others and to require that data be used only for research purposes?

- Who will be operationally responsible for ensuring that no personally identifiable information is made available (e.g., principal investigator, independent curator)?

3. **WHERE** will the data to be shared be located?

To minimize additional administrative workloads for sharing of data, ***data repositories with common standards and an established infrastructure dedicated to the appropriate distribution of data would generally be ideal for data sharing.*** In determining where data to be shared will be located, a data sharing plan should indicate:

- Will an existing database, data repository, data enclave, or archive be used to store and disseminate the data (e.g., dbGaP, National Database for Autism Research (NDAR)), and if so, how the policies and procedures in place for others to access the data are consistent with applicable NIH policies?
- Will a new repository need to be developed, and if so, who/what will maintain the repository?
- Will the data be distributed directly by an investigator to those who request it (e.g., through an electronic file)?

4. **WHEN** will the data be shared?

To optimize the timely and broadest usage of data, ***data should be made available as soon as possible and for as long as possible.*** In determining the timeframes for data sharing, a data sharing plan should indicate:

- The schedule for release of data:
    - What data, if any, will be shared prior to publication?
    - What data will be shared upon acceptance for publication?
    - If using a repository, when will data be submitted to the repository?
- Will data from ongoing longitudinal studies be released in increments as data become available?
- Will the timing of data sharing be specifically linked to other relevant policies concerning the timing of release of data (e.g., NIH GWAS policy, ClinicalTrials.gov, specific requirements in the funding opportunity announcement (FOA))?
- How will data maintenance and access be ensured after the award ends?
- Will there be support for continued sharing of data (e.g., through grant applications, administrative supplements, or other sources) or planned migration of data to another database, data repository, etc.?

5. **HOW** will researchers locate and access the data?

To optimize usage of the data, ***researchers need to be able to easily identify locations of relevant data and to be able to easily access the data.*** In describing how researchers will learn about, locate, and access the data, a data sharing plan should indicate:

- What steps will be taken to help researchers know that the data sets exist?
- Will registries, repositories, indexes, word-of-mouth, publications, and/or other approaches be used to publicize the availability and accessibility of the data?
- Will these be linked and cross-referenced so other researchers can readily find them?
- How will the data be accessed (web service, ftp, etc.)?

Human Subjects: if your project includes human subjects, please complete the information below.

a. Address the risk to subjects
b. The Adequacy of protections against risk
c. Potential benefits of the research to subjects and others
d. Importance of the knowledge to be gained
e. Inclusion of Women, Minorities and Children or Vertebrate animals
f. Indicate whether you already have IRB approval (attach to proposal) or pending IRB application.
g. If you are collecting new data and/or using existing data that are not exempt, please complete the Planned Enrollment table.

If you have any questions about IRB requirements, please contact Jan Jaeger at jajaeger@pennmedicine.upenn.edu