Eight Highly Recommended Elements of a NSF Data Management Plan

The following was included in a January 12, 2011 Webinar provided by IPSCR (Inter-University Consortium for Political and Social Research). The complete PowerPoint presentation and other valuable resources are available online at: ICPSR’s Data Management Plan Website http://www.icpsr.umich.edu/icpsrweb/ICPSR/dmp/index.jsp

1. Data description - a brief description of the information to be gathered; the nature, scope and scale of the data that will be generated or collected. Data description is important to include because it will help reviewers understand the characteristics of the data, their relationship to existing data, and any disclosure risks that may apply.

2. Access and sharing - Indicate how you intend to archive and share your data and why you have chosen that particular option. Access and sharing – is important to include in a DMP to specify how the data will be accessed and shared. Sharing data helps to advance science and to maximize the research investment. Notes re: Data sharing:
   - A recent paper [http://deepblue.lib.umich.edu/handle/2027.42/78307] reported that when data are shared through an archive, research productivity increases and many times the number of publications result as opposed to when data are not shared.
   - With respect to timeliness of data deposit, archival experience has demonstrated that the durability of the data increases and the cost of processing and preservation decreases when data deposits are timely. It is important that data be deposited while the producers are still familiar with the dataset and able to transfer their knowledge fully to the archive.

3. Metadata - A description of the metadata to be provided along with the generated data, and a discussion of the metadata standards used. Good descriptive metadata are essential to effective data use. Metadata are often the only form of communication between the secondary analyst and the data producer, so they must be comprehensive and provide all of the needed information for accurate analysis. Structured or tagged metadata, like the XML format of the Data Documentation Initiative (DDI) standard, are optimal because the XML offers flexibility in display and is also preservation-ready and machine-actionable.

4. Intellectual property rights - A description of entities or persons who will hold the intellectual property rights to the data, and how IP will be protected if necessary. Any copyright constraints (e.g., copyrighted data collection instruments) should be noted. It is important because: in order to disseminate data, archives need a clear statement from the data producer of who owns the data. The principal investigator’s university is usually considered to be the copyright holder for data the PI generates.

5. Ethics and privacy - A discussion of how informed consent will be handled and how privacy will be protected, including any exceptional arrangements that might be needed to protect participant confidentiality, and other ethical issues that may arise. Protection of human subjects is a fundamental tenet of research and an important ethical obligation for everyone involved in research projects. Disclosure of identities when privacy has been promised could result in lower participation rates and a negative impact on science.

6. Format - Describe the formats in which the data will be generated, maintained, and made available, including a justification for the procedural and archival appropriateness of those formats.
This is important because depositing data and documentation in formats preferred for archiving can make the processing and release of data faster and more efficient. Preservation formats should be platform-independent and non-proprietary to ensure that they will be usable in the future.

7. Archiving and Preservation
The procedures in place or envisioned for long-term archiving and preservation of the data, including succession plans for the data should the expected archiving entity go out of existence. This is important because digital data need to be actively managed over time to ensure that they will always be available and usable. This is important in order to preserve and protect our investment in science. Preservation of digital information is widely considered to require more constant and ongoing attention than preservation of other media. Depositing data resources with a trusted digital archive can ensure that they are curated and handled according to good practices in digital preservation.

8. Storage and Backup
A description of storage methods and backup procedures for the data, including the physical and cyber resources and facilities that will be used for the effective preservation and storage of the research data. This is important because digital data are fragile and best practice for protecting them is to store multiple copies in multiple locations.

Other Suggested Elements of a Data Management Plan:

- **Data organization** - How the data will be managed during the project, with information about version control, naming conventions, etc. It is important to describe situations in which research data are in some way atypical with respect to how they will be organized. For example, some data collections are dynamically changing and version control is central to how the data will be used and understood by the scientific community.

- **Quality Assurance** - procedures for ensuring data quality during the project. This is important because data of high quality is essential to the advancement of science, and every effort should be taken to be transparent with respect to data quality measures undertaken across the data life cycle.

- **Security** - a description of technical and procedural protections for information, including confidential information, and how permissions, restrictions, and embargoes will be enforced. Security for digital information is important over the data life cycle. Raw research data may include direct identifiers or links to direct identifiers and should be well-protected during collection, cleaning, and editing. Processed data may or may not contain disclosure risk and should be secured in keeping with the level of disclosure risk inherent in the data. Secure work and storage environments may include access restrictions (e.g., passwords), encryption, power supply backup, and virus and intruder protection.

- **Responsibility** - names of the individuals responsible for data management in the research project. Typically data are owned by the institution awarded a Federal grant and the principal investigator oversees the research data (collection and management of data) throughout the project period. It is important to describe any atypical circumstances. For example, if there is more than one principal investigator the division of responsibilities for the data should be described.
- **Budget** - The costs of preparing data and documentation for archiving and how these costs will be paid. Requests for funding may be included.
- **Legal requirements** - A listing of all relevant federal or funder requirements for data management and data sharing.
- **Audience** – the audience of users for the data. This is important because the audience for the data may influence how the data are managed and shared—for example, when audiences beyond the academic community may use the research data.
- **Selection and retention periods** - a description of how data will be selected for archiving, how long the data will be held, and plans for eventual transition or termination of the data collection in the future. This is important because not all data need to be preserved in perpetuity, so thinking through the proper retention period for the data is important, in particular when there are reasons the data will not be preserved permanently. Example language: Our project will generate a large volume of data, some of which may not be appropriate for sharing since it involves a small sample that is not representative. The investigators will work with staff of the [repository] to determine what to archive and how long the deposited data should be retained.

### Example Language for Eight Highly Recommended Sections:

#### Data Description

**Generic Example 1:**
This project will produce public-use nationally representative survey data for the United States covering Americans' social backgrounds, enduring political predispositions, social and political values, perceptions and evaluations of groups and candidates, opinions on questions of public policy, and participation in political life.

**Generic Example 2:**
This project will generate data designed to study the prevalence and correlates of DSM III-R psychiatric disorders and patterns and correlates of service utilization for these disorders in a nationally representative sample of over 8000 respondents. The sensitive nature of these data will require that the data be released through a restricted use contract.

**Example Language if one intends to archive their data in ICPSR's collection:**
[Provide a brief description of the information to be gathered -- the nature, scope, and scale of the data that will be generated or collected.] These data, which will be submitted to ICPSR, fit within the scope of the [ICPSR Collection Development Policy]. A letter of support describing ICPSR’s commitment to the data as they have been described is provided.

#### Access and sharing

**Generic Example 1:**
The research data from this project will be deposited with [repository] to ensure that the research community has long-term access to the data.

**Generic Example 2**
(Self-dissemination example: should describe and justify why choosing self-dissemination and describe plans to preserve the data):
The project team will create a dedicated Web site to manage and distribute the data because the audience for the data is small and has a tradition of interacting as a community. The site will be established using a content management system like Drupal or Joomla so that data users can participate in adding site content over time, making the site self-sustaining. The site will be available at a .org location. For preservation, we will supply periodic copies of the data to [repository]. That repository will be the ultimate home for the data.
Generic Example 3 (delayed dissemination plan):
The research data from this project will be deposited with a repository to ensure that the research community has long-term access to the data. The data will be under embargo for one year while the investigators complete their analyses.

Generic Example 4:
Designate a institutional repository on one's campus
The research data from this project will be deposited with the institutional repository on the grantees' campus.

ICPSR Example: For those who intend to use ICPSR - include this text which designates ICPSR as the intended archival home for the data. It indicates ICPSR's dissemination commitment to both public and restricted-use data files and covers ICPSR's terms of use.

The research data from this project will be deposited with the digital repository of the Inter-university Consortium for Political and Social Research (ICPSR) to ensure that the research community has long-term access to the data. The integrated data management plan proposed leverages capabilities of ICPSR and its trained archival staff. ICPSR will make the research data from this project available to the broader social science research community. Public-use data files: These files, in which direct and indirect identifiers have been removed to minimize disclosure risk, may be accessed directly through the ICPSR Web site. After agreeing to Terms of Use, users with an ICPSR MyData account and an authorized IP address from a member institution may download the data, and non-members may purchase the files. Restricted-use data files: These files are distributed in those cases when removing potentially identifying information would significantly impair the analytic potential of the data. Users (and their institutions) must apply for these files, create data security plans, and agree to other access controls. The Delayed Dissemination Policy allows for data to be deposited but not disseminated for an agreed-upon period of time (typically one year). Timeliness: The research data from this project will be supplied to ICPSR before the end of the project so that any issues surrounding the usability of the data can be resolved. Delayed dissemination may be possible.

Metadata

Generic Example 1: references this ideal (DDI standard).
Metadata will be tagged in XML using the Data Documentation Initiative (DDI) format. The codebook will contain information on study design, sampling methodology, fieldwork, variable-level detail, and all information necessary for a secondary analyst to use the data accurately and effectively.

Generic Example 2: clinical data - references the metadata standard common to clinical data which is called CDISC. The clinical data collected from this project will be documented using CDISC metadata standards.

ICPSR Example: (Language to use if depositing with ICPSR)
Substantive metadata will be provided in compliance with the most relevant standard for the social, behavioral, and economic sciences -- the Data Documentation Initiative (DDI). This XML standard provides for the tagging of content, which facilitates preservation and enables flexibility in display. These types of metadata will be produced and archived:

Study-Level Metadata Record. A summary DDI-based record will be created for inclusion in the searchable ICPSR online catalog. This record will be indexed with terms from the ICPSR Thesaurus to enhance data discovery.

Data Citation with Digital Object Identifier (DOI). A standard citation will be provided to facilitate attribution. The DOI provides permanent identification for data & ensures that they will always be found at the URL.

Variable-Level Documentation. ICPSR will tag variable-level information in DDI format for inclusion in ICPSR's Social Science Variables Database (SSVD), which allows users to identify relevant variables & studies of interest.

Technical Documentation. The variable-level files described above will serve as the foundation for the technical documentation or codebook that ICPSR will prepare and deliver.

Related Publications. Resources permitting, ICPSR will periodically search for publications based on the data and provide two-way linkages between data and publications.
**Intellectual property rights**

*Generic Example 1:* The principal investigators on the project and their institutions will hold the copyright for the research data they generate.

*Generic Example 2:* Many archives do not ask for a transfer of copyright but instead just request permission to preserve and distribute the data.

The principal investigators on the project and their institutions will hold the copyright for the research data they generate but will grant redistribution rights to [repository] for purposes of data sharing.

*Generic Example 3:* Copyright may also come into play if copyrighted instruments are used to collect data. In these cases, data producers should initiate discussions with archives in advance of data deposit.

The data gathered will use a copyrighted instrument for some questions. A reproduction of the instrument will be provided to [repository] as documentation for the data deposited with the intention that the instrument be distributed under "fair use" to permit data sharing, but it may not be redisseminated by users.

*ICPSR Example: This is the ICPSR text that can be included.*

Principals investigators and their institutions hold the copyright for the research data they generate. By depositing with ICPSR, investigators do not transfer copyright but instead grant permission for ICPSR to redisseminate the data and to transform the data as necessary to

**Ethics and privacy**

*Generic Example 1: address one of the most important considerations - whenever possible researchers collecting data should not include language in the informed consent that prohibits data sharing.*

For this project, informed consent statements will use language that will not prohibit the data from being shared with the research community.

*Generic Example 2: ICPSR recommends the following language be used in the informed consent:* The information in this study will only be used in ways that will not reveal who you are. You will not be identified in any publication from this study or in any data files shared with other researchers. Your participation in this study is confidential. Federal or state laws may require us to show information to university or government officials [or sponsors], who are responsible for monitoring the safety of this study.

*Generic Example 3: pertains to HIPPA information being collected in studies*

The proposed medical records research falls under the HIPAA Privacy Rule. Consequently, the investigators will provide documentation that an alteration or waiver of research participants’ authorization for use/disclosure of information about them for research purposes has been approved by an IRB or a Privacy Board.

*ICPSR Example: for use by those who will include their data in ICPSR – covers ICPSR procedures that minimize disclosures and protect confidentiality that are part of our routine data processing*

Informed consent: For this project, informed consent statements, if applicable, will not include language that would prohibit the data from being shared with the research community. Disclosure risk management: The research project will remove any direct identifiers in the data before deposit with ICPSR. Once deposited, the data will undergo procedures to protect the confidentiality of individuals whose personal information may be part of archived data. These include: (1) rigorous review to assess disclosure risk, (2) modifying data if necessary to protect confidentiality, (3) limiting access to datasets in which risk of disclosure remains high, and (4) consultation with data producers to manage disclosure risk. ICPSR will assign a qualified data manager certified in disclosure risk management to act as steward for the data while they are being processed. The data will be processed and managed in a secure non-networked environment using virtual desktop technology.

**Format**

*Generic Example 1:*

Quantitative survey data files generated will be processed and submitted to the [repository] as SPSS system files with DDI XML documentation. The data will be distributed in several widely used formats, including ASCII, tab-
delimited (for use with Excel), SAS, SPSS, and Stata. Documentation will be provided as PDF. Data will be stored as ASCII along with setup files for the statistical software packages. Documentation will be preserved using XML and PDF/A.

**Generic Example 2 (video format example):**
Digital video data files generated will be processed and submitted to the [repository] in MPEG-4 (.mp4) format.

**ICPSR Example:** covers ICPSR’s preference of format to receive data in, access formats ICPSR creates, and presentation formats for the data ICPRs maintains.

**Submission:** The data and documentation will be submitted to ICPSR in recommended formats. Access: ICPSR will make the quantitative data files available in several widely used formats, including ASCII, tab-delimited (for use with Excel), SAS, SPSS, and Stata. Documentation will be provided as PDF. **Preservation:** Data will be stored in accordance with prevailing standards and practice. Currently, ICPSR stores quantitative data as ASCII along with setup files for the statistical software packages, and documentation is preserved using XML and PDF/A.

**Archiving and Preservation**

**Generic Example 1 (preservation being handled by a repository):**
By depositing data with [repository], our project will ensure that the research data are migrated to new formats, platforms, and storage media as required by good practice.

**Generic Example 2: (even with self-dissemination of data, archiving and preservation should be covered):**
In addition to distributing the data from a project Web site, future long-term use of the data will be ensured by placing a copy of the data into [repository], ensuring that best practices in digital preservation will safeguard the files.

**ICPSR Example:** highlights ICPSR’s 50-year track record, commitment to migrating data to viable future formats/technologies and succession plan for our content in the event of the archive failing.

ICPSR is a data archive with a nearly 50-year track record for preserving and making data available over several generational shifts in technology. ICPSR will accept responsibility for long-term preservation of the research data upon receipt of a signed deposit form. This responsibility includes a commitment to manage successive iterations of the data if new waves or versions are deposited. ICPSR will ensure that the research data are migrated to new formats, platforms, and storage media as required by good practice in the digital preservation community. Good practice for digital preservation requires that an organization address succession planning for digital assets. ICPSR has a commitment to designate a successor in the unlikely event that such a need arises.

**Storage and Backup**

**Generic Example 1:**
[Repository] will place a master copy of each digital file (i.e., research data files, documentation, and other related files) in Archival Storage, with several copies stored at designated locations and synchronized with the master through the Storage Resource Broker.

**Generic Example 2 When the data relate to past data that have been collected:**
Data have been collected on this topic previously (for example: [add example(s)]). The data collected as part of this project reflect the current time period and historical context. It is possible that several of these datasets, including the data collected here, could be combined to better understand how social processes have unfolded over time.

**ICPSR Example:** describes ICPSR’s multiple copies held by partner organizations, mentions that the multiple copies are kept in sync.
Research has shown that multiple locally and geographically distributed copies of digital files are required to keep information safe. Accordingly, ICPSR will place a master copy of each digital file (i.e., research data files, documentation, and other related files) in ICPSR’s Archival Storage, with several copies stored with partner organizations at designated locations and synchronized with the master.