

# Data Plans in SBE and Geography

Joseph Messina Ph.D.

Professor and Associate Dean of Research  
Geography & College of Social Science

[http://www.nsf.gov/sbe/SBE\\_DataMgmtPlanPolicy.pdf](http://www.nsf.gov/sbe/SBE_DataMgmtPlanPolicy.pdf)

<http://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp>



# BCS / GSS – TOM BAERWALD

- ▶ *“Because the project generating data and other products is supported with public funds, the data and other products should be made readily accessible for others to benefit from unless there are legitimate restrictions on their dissemination.”*

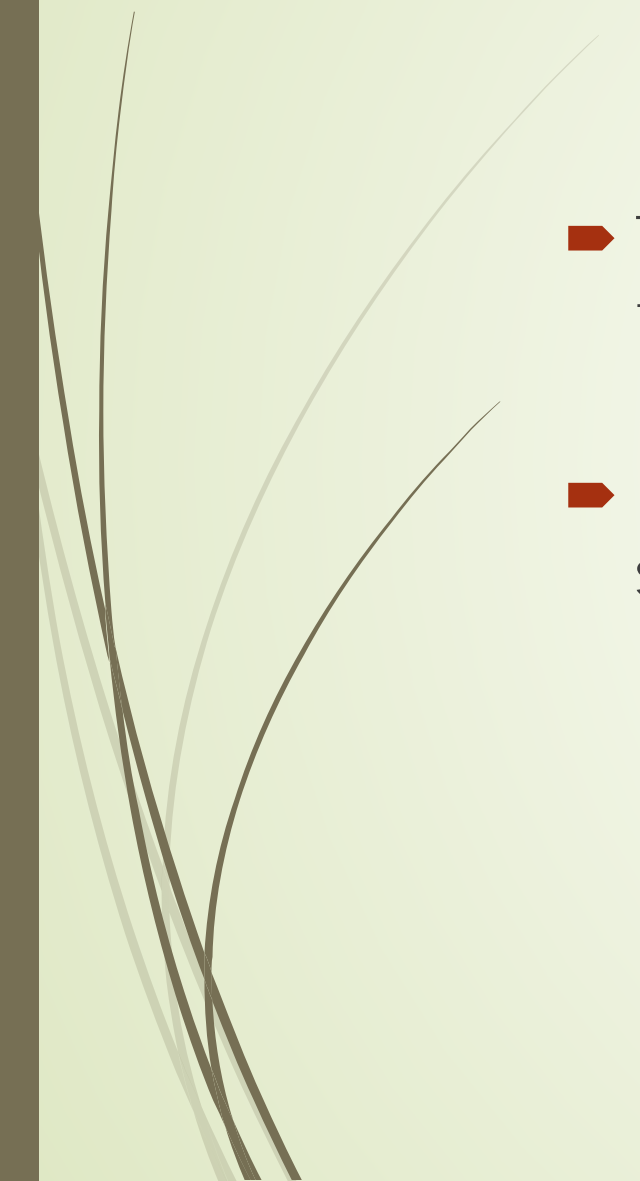


# NSF - Proposal and Award Policies and Procedures Guide (PAPPG)

- ▶ Required for all proposals submitted after January 17, 2011
- ▶ Supplementary document of no more than two pages
- ▶ The DMP is not included in the 15-page limit for proposal bodies.
- ▶ Fastlane will not permit submission of a proposal missing the DMP.
- ▶ Proposers who feel that the plan cannot fit within the supplement limit of two pages may use part of the 15-page Project Description for additional data management information; the plan may not be used to circumvent the 15-page Project Description limitation




# Review of the Data Management Plan

- ▶ The Data Management Plan will be reviewed as part of the intellectual merit and/or broader impacts of the proposal.
  - ▶ Data management activities must be reported in subsequent proposals by the PI and Co-PIs under “Results of prior NSF support.”
- 



# NSF does not attempt to define data

- Standards will be developed by members of the respective research communities.
  - The best way to determine what data to include in your plan is to consult the guidelines offered by the appropriate NSF directorate and/or division, as well as any special requirements laid out in the solicitation.
- 



# However, data...

- ▶ Research data are formally defined by the U.S. Office of Management and Budget as ***“the recorded factual material commonly accepted in the scientific community as necessary to validate research findings.”***
- ▶ This definition includes both analyzed data and associated metadata
  - ▶ “Analyzed data” include, but are not limited to, digital information that would be included in scientific publications, including digital images, published tables, and tables of the numbers used to create charts and graphs.
  - ▶ “Necessary metadata” include, but are not limited to, descriptions or suitable citations of experiments, apparatuses, raw materials, computational codes, model parameters and input conditions.
  - ▶ In general, research data are anything an investigator would need to reproduce published results.





# Data-Management Plan

1. The **types of data**, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project;
2. The **standards** to be used for data and metadata format and content (where standards are absent or inadequate, this should be documented along with any proposed solutions);
3. Policies for **access** and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements;
4. Policies and **provisions for re-use, re-distribution, and the production of derivatives**
5. Plans for **archiving data**, samples, and other research products, and for preservation of access to them



# SBE DMP Organization

*PIs should organize their data-management plans so that they address each of the five points specified in the NSF Grant Proposal Guide.*

*Although not required, PIs would make the cleanest presentation if they included separate sections for each of those five points.*





# Proposals should discuss...

- ▶ The types of data that their project might generate and eventually share with others, and under what conditions
- ▶ How data are to be managed and maintained
- ▶ Factors that might impinge on their ability to manage data, e.g. legal and ethical
- ▶ Restrictions on access to non-aggregated data
- ▶ The lowest level of aggregated data that PIs might share with others in the scientific community
- ▶ The mechanism for sharing data and/or making them accessible to others
- ▶ Project metadata

*Some kinds of data or products cannot be made available by the researchers, such as data collected in ways that their disclosure would violate the privacy or confidentiality of the research subjects or that they were proprietary when made available by someone else. Other data come from other publicly available sources. But even if those were the sources of primary data, data derived from analyses conducted using those data may be valuable for others and may be disseminated without violating promises made to research subjects or data providers.*



# *Data storage and preservation of access*

- ▶ The DMP should describe physical and cyber resources and facilities that will be used for the effective preservation and storage of research data.
- ▶ These can and often do include third party facilities and repositories.



# Does not include – per OMB definitions

- ▶ Preliminary analyses
- ▶ Drafts of scientific papers
- ▶ Plans for future research
- ▶ Peer reviews
- ▶ Communications with colleagues
- ▶ Trade secrets, commercial information, materials necessary to be held confidential by a researcher until they are published, or similar information which is protected under law; and
- ▶ Personnel and medical information and similar information the disclosure of which would constitute a clearly unwarranted invasion of personal privacy.
- ▶ For some NSF directorates, raw data fall into the category of “preliminary analyses” and are thus excluded from the DMP; **others may require a DMP for raw data.**



# Period of retention

- ▶ SBE is committed to timely and rapid data distribution.
  - ▶ However, it recognizes that types of data can vary widely and that acceptable norms also vary by scientific discipline.
- ▶ Timely access must be covered in the DMP statement.

*Researchers may delay dissemination of data for reasonable periods of time to enable them to publish initial findings based on those data, but in general, all data and other products should be made accessible within a couple of years after the data have been collected or generated.*



# Data Archiving

- ▶ For GSS, CNH, and IBSS, special emphasis is placed on having all data and other products made easily accessible to other potential users over long time periods from institutionally maintained repositories.
  - ▶ By long time periods, we're talking about decades.
  - ▶ By institutionally maintained, we mean university repositories (often based in libraries) or community-based repositories like ICPSR, NASA-supported DAACs, and the Tree-Ring Lab at the U of Arizona.
  - ▶ For distribution of software, online-based repositories may also be appropriate.
- ▶ **Unacceptable forms of providing access** are requiring users to request them from the researchers or making them available on personal or project websites.
- ▶ **Even university department-based repositories may be unacceptable** if there is no evidence there is a longer-term university commitment to ensure that data and other products will remain accessible well into the future if departmental structures or personnel change.





# Reporting

- ▶ *Annual Reports* - must provide information on the progress on data management and sharing of the research products.
  - ▶ This information could include citations of relevant publications, conference proceedings, and descriptions of other types of data sharing and dissemination of results.
- ▶ *Final Project Reports* - must discuss execution and any updating of the original DMP
  - ▶ Data produced during the award
  - ▶ Data to be retained after the award expires
  - ▶ Verification that data will be available for sharing
  - ▶ Discussion of community standards for data format
  - ▶ How data will be disseminated
  - ▶ The format that will be used to make data available to others, including any metadata
  - ▶ The archival location of data
- ▶ *Subsequent proposals*. Data management must be reported in subsequent proposals by the PI and Co-PIs under "Results of prior NSF support."





# Where is all this headed?

- ▶ Increasing emphasis on making data available over long time periods from institutionally maintained repositories.
  - ▶ The end of project or PI based data repositories
- ▶ Increasing pressure for rapid release of data and project transparency
  - ▶ The annual reports maybe used to dictate data release in the near future
- ▶ How these challenges are resolved is a core institutional responsibility. It is important that MSU be a leader in this initiative.
  - ▶ Institutions are broadly competing to establish these data repositories, and I suspect these will be a competitive indicator for proposal success in the future.