Appendix C

Current Policies on Data Sharing and Archiving

There are a variety of current policies in place at NSF and in other agencies that vary considerably in their scope and in their provisions. There is also a variety of community standards, some set by professional societies, some set by journals, and some established through community meetings.

This Appendix provides examples of existing policies, illustrates how policies can differ across the NSF and across agencies, and identifies areas where there may be a lack of adequate policy or a lack of appropriate consistency across different policies.

EXAMPLES OF DATA POLICIES

NSF Policies

This section includes examples of NSF policies, including NSF's general conditions for grants as well as the data policies of several specific programs.

NATIONAL SCIENCE FOUNDATION GRANT GENERAL CONDITIONS

NSF's Grant General Conditions include the following:

> Article 36. Sharing of Findings, Data, and Other Research Products
>
> a. NSF expects significant findings from research and education activities it supports to be promptly submitted for publication, with authorship that accurately reflects the contributions of those involved. It expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work. It also encourages awardees to share software and inventions or otherwise act to make the innovations they embody widely useful and usable.
>
> b. Adjustments and, where essential, exceptions may be allowed to safeguard the rights of individuals and subjects, the validity of results, or the integrity of collections or to accommodate legitimate interests of investigators.

These conditions are quite general, and do not address archiving of data, the duration of data collections, or requirements for providing metadata or finding aids.

Division of Environmental Biology

The Division of Environmental Biology (DEB), within the Directorate for Biological Sciences, follows general NSF policy and has developed the following statement for program announcements.

> Proposals submitted to all programs in DEB must adhere to the general NSF policy on data sharing as described in the Grant Proposal Guide… Thus, proposals should describe plans for specimen and information management and sharing, including where data and metadata, will be stored and maintained, and the likely schedule for release. These plans will be considered as part of the review process. http://www.nsf.gov/bio/deb/

Division of Ocean Sciences

The Division of Ocean Sciences, within the Geosciences Directorate, has a long-standing and detailed policy for oceanographic data. Excerpts from the policy statement are provided below.

> POLICY FOR OCEANOGRAPHIC DATA, NSF 94-126
>
> Ocean data collected under Federal sponsorship and identified as appropriate for submission to a national data center are to be made available within a reasonable time as described below.
>
> Principal investigators are required to submit all environmental data collected to the designated national data centers as soon as possible, but no later than two (2) years after the data are collected. Inventories of all marine environmental data collected should be submitted to the designated national data centers within sixty (60) days after the observational period/cruise…
>
> Data sets identified for submission to the national data centers must be submitted to the designated center within two (2) years after the observational period. This period may be extended under exceptional circumstances by agreement between the principal investigator and NSF. Data produced by long-term (multi-year) projects are to be submitted annually…

> NOAA's National Environmental Satellite Data and Information Service staff and program representatives from funding agencies will identify the data sets that are likely to be of high utility and will require their principal investigators to submit these data and related information to the designated center.
>
> Funding agencies will apply this policy to their internal ocean data collection and research programs and to their contractors and grantees and will establish procedures to enforce this policy.
>
> A list of oceanographic data types and the centers designated to receive them are the following...:
>
> Data are to be submitted according to formats and via the media designated by the pertinent national data center.
>
> Principal investigators and ship-operating institutions are also responsible for meeting all legal requirements for submission of data and research results, which are imposed by foreign governments as a condition of that government's granting research clearances…

The full policy is available at: http://www.nsf.gov/pubs/stis1994/nsf94126/nsf94126.html

Division of Behavioral and Cognitive Sciences

NSF's Division of Behavioral and Cognitive Sciences (BCS), within the Directorate for Social, Behavioral and Economic Sciences, has a data policy that recognizes the diversity of types of data handled by the division. Excerpts from this policy follow:

> BCS supports a wide range of disciplines. The nature of the data, the way they are collected, analyzed, and stored, and the pace at which this reasonably occurs varies widely. There are different storage facilities and different access requirements for, e.g., archaeological data, specimens from physical anthropology, large-scale survey data, oral interviews with scientists and other subjects, data generated by experimental research, and field records of tribal ceremonies. Where appropriate and possible, grantees from all fields will develop and submit specific plans to share materials collected with NSF support. These plans should cover how and where these materials will be stored, at reasonable cost, and how access will be provided to other researchers, generally at their cost.

This policy explicitly recognizes that many complexities arise across the range of data collection supported by BCS programs, and that unusual circumstances may require modifications or even full exemptions. For example, human subjects protection requires removing identifiers, which may be prohibitively expensive or render the data meaningless in research that relies heavily on extensive in-depth interviews. Intellectual property rights may be at risk in some forms of data collection. The policy is intended to be flexible enough to accommodate the variety of scientific enterprises that constitute BCS programs. No comprehensive set of rules is possible, but the procedures indicated below are designed to provide guidance for broad categories of data collection.

Experimental Research
In experimental research, individuals, be they people, animals, or objects, are subjected to preplanned conditions and their responses tabulated in some fashion. Investigators should plan to make these tabulated data available to other investigators requesting them. In addition, complete information on how an experiment was conducted and any unusual stimulus materials should be made available, so that failures to replicate will not turn out to depend on one scientist's incomplete understanding of another's procedure.

Mathematical and Computer Models
Often in the course of conducting research, investigators develop mathematical and computer models, either as an innovative aid in the analysis of data or as a theoretical statement about the processes involved in generating some classes of data. Investigators should plan to make these models available to others wanting to apply them to other data sets or experimental situations…

Object Based Research
Some research supported by BCS is based on objects such as archaeological specimens or fossil remains. In these instances data consist of the objects themselves, contextual information such as geological sections and finally quantitative and qualitative descriptions of the materials. Because these physical objects rarely become the property of the investigator but belong to a host nation or cultural group, scientists often do not control access to them. This situation is further complicated by the fact that description of materials often must proceed slowly and may take several years to complete. However, it is still incumbent upon the investigator to make primary and contextual information available as rapidly as possible to permit other scientists to examine them and draw their own conclusions.

> Qualitative Information
> The kinds of qualitative information collected in research projects supported by BCS can range from microfilms and other copies of very old documents to oral interviews and video tapes about historical events in science or about contemporary technological controversies. They can consist of ethnographic or linguistic field notes or recordings or transcriptions, or hand written records of open-ended interviews. Investigators should consider whether and how they can develop special arrangements to keep or store these materials so that others can use them. If it is appropriate for other researchers to have access to them, the investigators should specify a time at which they will be made generally available, in an appropriate form and at a reasonable cost.
>
> Quantitative Social and Economic Data Sets
> For appropriate data sets, researchers should be prepared to place their data in fully cleaned and documented form in a data archive or library within one year after the expiration of an award. Before an award is made, investigators will be asked to specify in writing where they plan to deposit their data set(s)…

The full policy is available at http://www.nsf.gov/sbe/bcs/common/archive.htm

Other Agency and Interagency Data Policies

U.S. GLOBAL CHANGE RESEARCH PROGRAM

The interagency U.S. Global Change Research Program has a high level data policy that provides guidelines for more specific policies by participating agencies. Excerpts follow.

> The U.S. Global Change Research Program requires an early and continuing commitment to the establishment, maintenance, validation, description, accessibility, and distribution of high-quality, long-term data sets. Full and open sharing of the full suite of global data sets for all global change researchers is a fundamental objective.
>
> Preservation of all data needed for long-term global change research is required. For each and every global change data parameter, there should be at least one explicitly designated archive. Procedures and criteria for setting priorities for data acquisition, retention, and purging should be developed by participating agencies, both nationally and internationally. A clearinghouse process should be established to prevent the purging and loss of important data sets.

> Data archives must include easily accessible information about the data holdings, including quality assessments, supporting ancillary information, and guidance and aids for locating and obtaining the data.
>
> National and international standards should be used to the greatest extent possible for media and for processing and communication of global data sets.
>
> Data should be provided at the lowest possible cost to global change researchers in the interest of full and open access to data. This cost should, as a first principle, be no more than the marginal cost of filling a specific user request. Agencies should act to streamline administrative arrangements for exchanging data among researchers.
>
> For those programs in which selected principal investigators have initial periods of exclusive data use, data should be made openly available as soon as they become widely useful. In each case the funding agency should explicitly define the duration of any exclusive use period.

There are more details at http://www.globalchange.gov/policies/diwg/diwg-guidelines.html

NOAA COASTAL OCEAN PROGRAM (COP) DATA POLICY

Many of the programs and agencies involved in observational earth science data have data policies that are generally similar. The National Oceanic and Atmospheric Administration's Coastal Ocean Program is one example. Excerpts from its policies include:

> The COP Data Policy promotes: (1) full and open sharing of data and other products of COP-sponsored research by all COP researchers; (2) entitling the investigator who collects the data to the fundamental benefits of the collected data set, derived models, etc.; (3) selection of methods and equipment to ensure sufficient accuracy and precision to meet the project requirements for inter-comparisons and syntheses; (4) preservation of all data collected under COP sponsorship, including derived models, in an easily accessible archive with transfer ultimately to a permanent archive at a National Data Center…
>
> COP encourages the no-cost, open, voluntary and ethical exchange of data or other COP-related information among investigators. Publication of descriptive or interpretive results immediately and directly from the data is the privilege and responsibility of the investigators who collect the data. Prior to submission to a permanent data archive at a National Data Center, publication or presentation of any data derived by a co-participating investigator requires the permission of the scientist originating the data. Any scientist making substantial use of a data set should anticipate that

the data collectors will be co-authors of published results. Originating investigators may not unreasonably impede use or publication of archived data, models, or model application.

Methods and equipment used to take measurements and collect samples must be of sufficient accuracy and precision to yield data with quality adequate to meet the objectives of the COP field projects, associated modeling efforts, and larger-scale synthesis…

A data archive system will be established by each COP-sponsored project within six (6) months of the project start date for temporary repository of the data prior to their submittal to a permanent archive. The data archive system must facilitate the exchange of data and insure the long-term existence of the data set. The COP Project Manager (or a designated project Data Manager) will ensure the following data archive system conditions are met:

- data integrity and appropriate metadata are maintained;
- all users are provided access in a timely manner;
- and the data are transferred to a designated National Data Center (e.g., National Oceanographic Data Center) within two (2) years from the time of initial observations.

The submitted data will include the actual measurements and supporting descriptive information (i.e., metadata) sufficient to permit its effective use by researchers not familiar with the original project or the particular instrument making the measurements. The NOAA/Federal Geographic Data Committee Metadata Standard Format shall be used to describe the data.

This policy also encourages the project archive to include selected models, and model products or results. Measurements which do not involve manual analysis should be submitted to the project archive within six (6) months. All measurements, including metadata, should be submitted to a National Data Center for permanent archive.

Unclassified and/or unrestricted environmental data and information produced, sponsored, collected, or obtained by NOAA/COP are public property. It is NOAA policy to make environmental data and information available under NOAA's stewardship based on exchange, loan, cost of dissemination, or at no cost in the interest of full and open access to data.

The full policy is available at http://www.cop.noaa.gov/../Grants/datapolicy.PDF Other examples of earth-science policies that are generally similar in scope and terms are NASA's Global Change program (available at http://www.globalchange.gov/policies/agency/nasa.html and the Office of Naval Research's Ocean,

Atmosphere, and Space Science and Technology Department, available at http://www.onr.navy.mil/../02/docs/tcpsod.pdf. These policies are quite specific about what data and metadata must be provided, the timing of providing this data, and the data centers in which the data needs to be archived.

NATIONAL INSTITUTES OF HEALTH

The National Institutes of Health (NIH) has a relatively recent (2003) data sharing policy. This applies NIH-wide, but currently applies only to large grants. These policies apply to data sharing, but do not address long-term archiving. Excerpts from the policy include:

> Data should be made as widely and freely available as possible while safeguarding the privacy of participants, and protecting confidential and proprietary data. To facilitate data sharing, investigators submitting a research application requesting $500,000 or more of direct costs in any single year to NIH on or after October 1, 2003 are expected to include a plan for sharing final research data for research purposes, or state why data sharing is not possible.
>
> The NIH policy on data sharing applies:
> - To the sharing of final research data for research purposes.
> - To basic research, clinical studies, surveys, and other types of research supported by NIH. It applies to research that involves human subjects and laboratory research that does not involve human subjects. It is especially important to share unique data that cannot be readily replicated.
> - To applicants seeking $500,000 or more in direct costs in any year of the proposed project period through grants, cooperative agreements, or contracts.
> - To research applications submitted beginning October 1, 2003.
>
> Final research data are recorded factual material commonly accepted in the scientific community as necessary to document, support, and validate research findings. This does not mean summary statistics or tables; rather, it means the data on which summary statistics and tables are based… For most studies, final research data will be a computerized dataset.
>
> Given the breadth and variety of science that NIH supports, neither the precise content for the data documentation, nor the formatting, presentation, or transport mode for data is stipulated.
>
> … if an application describes a data-sharing plan, NIH expects that plan to be enacted. In the final progress report, if not sooner, the grantee should note what steps have been taken with respect to the data-sharing plan. In the case of noncompliance (depending on its severity and duration) NIH

can take various actions to protect the Federal Government's interests. In some instances, for example, NIH may make data sharing an explicit term and condition of subsequent awards.

Grantees should note that, under the NIH Grants Policy Statement, they are required to keep the data for 3 years following closeout of a grant or contract agreement… the grantee institution may have additional policies and procedures regarding the custody, distribution, and required retention period for data produced under research awards.

…NIH expects the timely release and sharing of data to be no later than the acceptance for publication of the main findings from the final dataset.

NIH recognizes that the investigators who collected the data have a legitimate interest in benefiting from their investment of time and effort. NIH continues to expect that the initial investigators may benefit from first and continuing use but not from prolonged exclusive use.

The rights and privacy of human subjects who participate in NIH-sponsored research must be protected at all times. It is the responsibility of the investigators, their Institutional Review Board (IRB), and their institution to protect the rights of subjects and the confidentiality of the data. Investigators may use different methods to reduce the risk of subject identification…

If research participants are promised that their data will not be shared with other researchers, the application should explain the reasons for such promises. Such promises should not be made routinely and without adequate justification.

For the most part, it is not appropriate for the initial investigator to place limits on the research questions or methods other investigators might pursue with the data. It is also not appropriate for the investigator who produced the data to require coauthorship as a condition for sharing the data.

…under the Small Business Act, SBIR grantees may withhold their data for 4 years after the end of the award. Issues related to proprietary data also can arise when cofunding is provided by the private sector (e.g., the pharmaceutical or biotechnology industries) with corresponding constraints on public disclosure. NIH recognizes the need to protect patentable and other proprietary data. Any restrictions on data sharing due to cofunding arrangements should be discussed in the data-sharing plan section of an application and will be considered by program staff. There are many ways to share data:

- Under the auspices of the PI
- Data archive
- Data enclave
- Mixed mode sharing.

Investigators will need to determine which method of data sharing is best for their particular dataset.

Regardless of the mechanism used to share data, each dataset will require documentation. (Some fields refer to data documentation by other terms, such as metadata or codebooks). The precise content of documentation will vary by scientific area, study design, the type of data collected, and characteristics of the dataset.

It is appropriate for scientific authors to acknowledge the source of data upon which their manuscript is based. Many investigators include this information in the methods and/or reference sections of their manuscripts.

NIH recognizes that it takes time and money to prepare data for sharing. Thus, applicants can request funds for data sharing and archiving in their grant application.

The full policy and implementation guidance is available at http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

Publications

In addition to government agencies, some publications have policies that affect the sharing and archiving of data.

SCIENCE

Science's policy is as follows:

Any reasonable request for materials, methods, or data necessary to verify the conclusions of the experiments reported must be honored.

Before publication, large data sets, including protein or DNA sequences and crystallographic coordinates, must be deposited in an approved database and an accession number provided for inclusion in the published paper, under the database deposition policy outlined below.

Database Deposition Policy
Science supports the efforts of databases that aggregate published data

for the use of the scientific community. Therefore, before publication, large data sets (including microarray data, protein or DNA sequences, and atomic coordinates or electron microscopy maps for macromolecular structures) must be deposited in an approved database and an accession number provided for inclusion in the published paper.

Macromolecular structure data. Atomic coordinates and structure factor files from x-ray structural studies or an ensemble of atomic coordinates from NMR structural studies must be deposited and released at the time of publication. Three-dimensional maps derived by electron microscopy and coordinate data derived from these maps must also be deposited. Approved databases are the Worldwide Protein Data Bank [through the Research Collaboratory for Structural Bioinformatics, Macromolecular Structure Database (MSD EMBL-EBI), or Protein Data Bank Japan], BioMag Res Bank, and Electron Microscopy Data Bank (MSD-EBI).

DNA and protein sequences. Approved databases are GenBank or other members of the International Nucleotide Sequence Database Collaboration (EMBL or DDBJ) and SWISS-PROT.

Microarray data. Data should be presented in MIAME-compliant standard format. Approved databases are Gene Expression Omnibus and ArrayExpress.

Large data sets with no appropriate approved repository must be housed as supporting online material at Science, or when this is not possible, on the author's Web site, provided a copy of the data is held in escrow at Science to ensure availability to readers.

For more information, see the Science Web site, http://www.sciencemag.org/feature/contribinfo/prep/gen_info.shtml#datadep

Nature has generally similar policies, available at http://www.nature.com/nature/submit/policies/index.html#6

AMERICAN GEOPHYSICAL UNION

The American Geophysical Union (AGU) has an extensive set of policies that govern (1) citations of publicly available data sets in regular AGU journal papers; (2) long-term access to small supporting data sets and graphics files that are published concurrently with, and are an electronic component of, some AGU journal papers; and (3) a special class of data and analysis papers that are offered in some AGU journals. Excerpts from these policies are as follows:

Citing Data in Regular AGU Journal Papers

1. Data sets cited in AGU publications must meet the same type of standards for public access and long-term availability as are applied to citations to the scientific literature. Thus data cited in AGU publications must be permanently archived in a data center or centers that meet the following conditions:
a) are open to scientists throughout the world.
b) are committed to archiving data sets indefinitely.
c) provide services at reasonable costs.

The World and National data centers meet these criteria. Other data centers, though chartered for specific lengths of time, may also be acceptable as an archive for this material if there is a commitment to migrating data to a permanent archive when the center ceases operation. Citing data sets available through these alternative centers is subject to approval by AGU.

2. Data sets that are available only from the author, through miscellaneous public network services, or academic, government or commercial institutions not chartered specifically for archiving data, may not be cited in AGU publications. This type of data set availability is judged to be equivalent to material in the gray literature. If such data sets are essential to the paper then authors should treat their mention just as they would a personal communication. These mentions will appear in the body of the paper but not in the reference list.

3. To assist scientists in accessing the data sets, authors are encouraged to include a brief data section in their papers. This section should contain the key information needed to obtain the data set being cited.

4. Data sets that meet the requirements stated in paragraph 1 above can be included in the reference list of an article in an AGU publication. The format for the reference will be specified in AGU's guide for contributors. The following elements must be included in the reference: author(s), title of data set, access number or code, data center, location including city, state, and country, and date.

Data Papers

1. Editors are free to establish a category of articles that are primarily designed to discuss the acquisition, preparation, and use of key data sets. The requirements for the substance of these articles and their lengths will be determined by the editor.

2. Data sets discussed in data papers published in AGU books and journals must be publicly available and accessible to the scientific community indefinitely. Authors of such papers are required to deposit their data sets in a data center that meets the criteria discussed above. In the event that an appropriate data center cannot be found by the author, AGU will take an active role in recommending the acceptance of the data by a suitable data center. AGU will provide temporary storage services, for a fee, and will facilitate the migration of the data sets to an approved center as soon as practical. (Also see section below on AGU's role in archiving data.)

3. Data sets that are the basis of data papers are subject to review. A sample of these data sufficient for the review process must be supplied with the submission of the paper. The reviewer is expected to comment on the data as if they were an integral part of the paper and on their usability.

4. Data sets for data papers must include a descriptive section that provides the user with key information about the collection, preparation and use of the data set. (This section is sometimes called the "metadata.") The format and content of this section will be specified in AGU's guide to contributors.

5. At the time of submission, authors must supply complete information about the archiving of the data sets. To avoid possible delays in the publication of the data paper, authors should consult with the data center(s) before submitting the paper to AGU. If the data sets have been archived before the paper is submitted, information on accessing them must be supplied to the reviewers.

6. The data sets will be listed in AGU's electronic index to publications (EASI). The citation in the index will include sufficient information for locating the data set.

Characteristics of Data Archive to be Maintained by AGU

1. Permanent archive: AGU makes a commitment to maintain and provide long-term access to the data sets.

2. Platform independent: The format of such data sets and graphics files shall be platform-neutral to allow the widest possible availability.

3. Future portability: Formats for archiving data and graphics files must be in a generic, preferably non-proprietary format consistent with conversion to future open standards if necessary.

4. Ease of management: Files shall not require significant pre-processing or reformatting for administrators in order to archive the data.

5. Usability: Compression techniques used for data sets should be available on multiple platforms, such as zip utility.

6. Flexibility: The guidelines and their recommended standards should be sufficiently flexible to allow for future incorporation of technology advances, and to allow for future user input gained from practical experience.

AGU's Role in Archiving Data

It is AGU's intent to ensure the continuity of archived data sets by providing long-term access to small supporting data sets and graphic files that are an electronic component of and other supplemental materials that are published concurrently with AGU journal articles; entering into agreements with data centers to acquire archived data sets should the center no longer offer this service; providing temporary storage if needed for these archived data sets until a new storage center is found; and maintaining a catalog of data papers which provides the current location of data sets.

1. AGU does not expect to archive data sets subject to this policy, except on a for-fee basis and for sets of a small size. In general AGU expects data to be deposited with and maintained by facilities that are specifically chartered for that purpose. AGU will work with these facilities as described below.

2. AGU will work with data centers to help advertise their services and to help inform authors about the formats and standards established by the data centers. This information will be provided in order to assist authors in finding an approved archive for their data sets.

3. AGU will take an active role in helping to expand the scope of data centers if authors have been turned down because the subject of the data sets does not fit the charter of existing data centers.

4. It is not AGU's intention to serve as an archive for large data sets that should be housed in data centers. Nor do we expect to take on the responsibilities of handling such data sets even temporarily unless they are an electronic component of a regular AGU journal paper.

5. It is AGU's intent to ensure the continuity of archiving of data sets in the data papers. Thus, AGU will attempt to enter into agreements with data centers to acquire archived data sets should the center decide to cease

storing them. AGU will provide temporary storage services while another approved center is found. To meet the continuity objective, AGU will maintain a catalog of data papers and the location of current storage.

6. AGU maintains a deposit service for supplementary material of different types in order to provide long-term access to small supporting data sets and graphics files that are published concurrently with, and are an electronic component of, some AGU journal articles. Procedures related to this service are discussed in "Guidelines for AGU Electronic Supplemental Data Set Archive."

These policies are available at http://www.agu.org/pubs/data_policy.html.

ANALYSIS OF DATA POLICIES

The examples of data policies provided here illustrate that there is a wide range in the scope, specificity, and terms of data policies within NSF, across Federal agencies, and across scientific communities. Some observations about these policies are as follows.

- Overall NSF policy is quite general, and does not address requirements for archiving (or sunsetting) data, requirements for metadata, or enforcement of policy.
- Some NSF programs have detailed data policies; others do not.
- Policies vary considerably in whether or not they require archiving of data or just sharing.
- Data policies are well established and stable for observational earth science data. This may arise in part because of the existence of a well-established system of world data centers that provide archives for data.
- Data policies are newer and evolving in the life sciences. Publication policies have an important influence on data practices in these fields. NIH policy is a recent addition to this field.
- Human subjects provisions and proprietary data concerns are major elements of data policies in the life and social sciences.

[Blank Page]