University of Rochester
Guidance on Developing Research Data Sharing Plans

This guidance has been developed to assist University of Rochester (UR) researchers and investigators in developing research data sharing plans when required or recommended in sponsored funding applications. Guidance on sharing unique research resources (e.g., model organisms) developed by NIH funds can be found at http://www.rochester.edu/ORPA/policies/organism_research.htm .

Definition(s) of Research Data: For the purposes of developing sharing plans, NIH defines research data as *"the recorded factual material commonly accepted in the scientific community as necessary to validate research findings. Final research data do not include laboratory notebooks, partial data sets, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects, such as gels or laboratory specimens."* While NSF does not specifically define research data, NSF policy has long advocated that investigators are expected to share with other researchers, *"at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants."* Other research sponsors, such as private foundations, may have other definitions for research data.

General Recommendations for Sharing of Research Data

- The type of plan that an investigator will develop will ultimately be discipline and project specific, and should be tailored to each individual situation.

- Data sharing for a small laboratory-based project may be best accomplished by making raw data available in publications (e.g., via appendices) or under the auspices of the investigator. For larger studies, providing data in publications is probably not the best vehicle for sharing. Therefore, other means for sharing must be made available such as through data archives, web sites or enclaves.

- Applicants may request funds for data sharing and archiving in grant applications.

- Data sharing plans should take into account proprietary information of the University and third parties, protection of human subjects information and Protected Health Information (PHI) and other situations where data sharing may not be appropriate or allowed. Where research is co-sponsored, the results may contain data that are proprietary to the sponsor that the UR is required to keep confidential. If data is required to support patent applications for intellectual property protection, delays in providing access may be required to allow exercise of UR's rights. Investigators who are involved in human subject research should provide for methods to protect the rights and privacy of the human subjects. Plans may be tailored to provide for restricted access, data-sharing agreements,

delays or other measures as appropriate to balance needs and rights of the UR with the data sharing mission.

Specific Sponsor Requirements:

NIH has required a data sharing plan with all applications exceeding $500,000 in direct costs in any single year (or if it is a special requirement in the program announcement) since October 2003. Effective January 11, 2011, NSF requires applicants to submit data management/sharing plans in all proposals (or to assert the absence of the need for such plans). Other major differences between the agency requirements are summarized below.

- NIH: Data sharing plans should be provided in the form of a brief paragraph (longer if necessary) immediately following the Research Plan section. Reviewers will not factor the proposed data sharing plan into the determination of scientific merit or priority score. Program staff will be responsible for overseeing the data sharing policy and for assessing the appropriateness and adequacy of the proposed data sharing plan.

- NSF: In describing the plan (of not more than two pages), NSF suggests a plan could include: the types of data, samples, physical collections, etc., that will be produced in the course of the project; any standards to be used for data and metadata format and content; policies for access and sharing including intellectual property provisions; provisions for re-use, re-distribution, and the production of derivatives; and plans for archiving data and for preservation of access to them. The data management should be described in the Special Information and Supplementary Documentation section. The data management plan will be reviewed as part of the intellectual merit or broader impacts of the proposal or both, as appropriate for the scientific community of relevance. Please note that NSF Directorates may have additional guidance and/or requirements that can be found at http://www.nsf.gov/bfa/dias/policy/dmp.jsp.

UR Resources for Investigators

UR Research (https://urresearch.rochester.edu) is a web-based University resource that may assist investigators with data sharing requirements. UR Research is an institutional repository system based on a relational database, with collaborative authoring and versioning capabilities. While not appropriate for extremely large datasets such as those produced by physics, astronomy, or similar departments, UR Research is a viable option for datasets in the 200-300 MB range. UR Research provides:

- *Support for researchers and collaborators:* UR Research provides a password protected private workspace that allows University of Rochester users to upload and store up to 2 gigabytes of files of any type. UR Research has been able to handle file uploads of over two hundred megabytes. All files uploaded have a checksum calculated upon upload for file integrity verification. UR Research allows for the storage and retrieval of multiple versions of any given file.

Registered users may share files with collaborators, either within or outside of the UR, for co-authoring purposes. Collaborators must create an account in UR Research and always log in to access materials being shared with them. A publication area is also provided and can be used to disseminate material to larger groups. The publication area allows the content to be restricted to a given set of users when necessary.
This progression from a work-in-progress file uploaded to the private workspace, to the sharing of the file with co-authors or peers for feedback, to the final publication of the work, parallels the data acquisition, analysis, and publication process.

- *Persistent URLs:* When an author publishes a file or files into the UR Research public repository, the record is automatically assigned a persistent URL based on the handle system. This address is guaranteed to always return users to that original work, regardless of changes in the software or hardware being used to run the system. We have already demonstrated the truth of this address guarantee by our successful migration from the DSpace to the IR+ software platform: all existing records carried over their handle address. Service and accessibility continued uninterrupted by any "404 Not Found" messages.
- *Encryption and security:* the system is run over a secure connection and uses a signed certificate. The current certificate is supplied by Verisign and is encrypted at 128 AES; on 1/15/2012 this certificate will be replaced by one from Entrust, encrypted at 256 AES.
- *Data preservation*: the UR Research system is backed up nightly. Thirty days of backups are maintained locally on disk; after 30 days the disks are written to tape, and tapes are rotated so there are always 12 months' worth kept off-site, in a fire and water proof rated safe.
- *Server environment*: the server holding the UR Research data is located in the University of Rochester River Campus Libraries' server room, which is temperature and humidity HVAC controlled. The server room is equipped with an alarm system that is wired to central university support for immediate notification of changes in prescribed temperature or humidity levels. The Libraries' server room is a secure, purpose-built room, accessible by keypad only by the libraries' systems administrators.

Sample data sharing language for projects utilizing UR Research (to be used in conjunction with the project-specific text such as appears in Example 1 of the Attachment):

Data from this research will be made available to the public in the University of Rochester's institutional repository, UR Research, at https://urresearch.rochester.edu. UR Research is fully searchable in its own right, and is also regularly indexed by Google, so that its contents are accessible from the results of Google searches.

Materials in UR Research are automatically assigned a persistent URL based on the handle system. This address is guaranteed to always return users to that original work, regardless of changes in the software or hardware being used to run the system. This ensures long-term service and accessibility uninterrupted by "404 Not Found" messages.

The UR Research system provides the following security and preservation features:

- *Encryption and security:* the system is run over a secure connection and uses a signed certificate. The current certificate is supplied by Verisign and is encrypted at 128 AES; on 1/15/2012 this certificate will be replaced by one from Entrust, encrypted at 256 AES.
- *Data preservation*: the UR Research system is backed up nightly. Thirty days of backups are maintained locally on disk; after 30 days the disks are written to tape, and tapes are rotated so there are always 12 months' worth kept off-site, in a fire and water proof rated safe.
- *Server environment*: the server holding the UR Research data is located in the University of Rochester River Campus Libraries' server room, which is temperature and humidity HVAC controlled. The server room is equipped with an alarm system that is wired to central university support for immediate notification of changes in prescribed temperature or humidity levels. The Libraries' server room is a secure, purpose-built room, accessible by keypad only by the libraries' systems administrators.

Attached are several other examples of data sharing plans that may be appropriate for certain research projects. Note that these have been provided as examples only, and the Office(s) of Technology Transfer are available to assist with specific cases. (Medical Center Office of Technology Transfer (784-8850) or the University Office of Technology Transfer (275-3998))

Revised February 2011
Attachment (Examples of Data Sharing Plans)

h:/letters/gjl/UR Guidance on Data Sharing Plans

Attachment
Excerpts from Data Sharing Plans

As noted in the NIH Policy, *"the precise content of the data-sharing plan will vary, depending on the data being collected and how the investigator is planning to share the data. Applicants who are planning to share data may wish to describe briefly the expected schedule for data sharing, the format of the final dataset, the documentation to be provided, whether or not any analytic tools also will be provided, whether or not a data-sharing agreement will be required and, if so, a brief description of such an agreement (including the criteria for deciding who can receive the data and whether or not any conditions will be placed on their use), and the mode of data sharing (e.g., under their own auspices by mailing a disk or posting data on their institutional or personal website, through a data archive or enclave). Investigators choosing to share under their own auspices may wish to enter into a data-sharing agreement."*

Please note the following examples are illustrative, and may not be appropriate or suitable for data sharing depending on project and/or discipline. While some of these plans were used specifically for NIH applications, they can be tailored for any research sponsor.

1) ***Example that adheres to the NSF requirements:***

Data Management Plan

The PIs are committed to sharing the primary data produced in the course of the proposed research, and to disseminating research results in a timely fashion.

Expected data

Research results will be disseminated to the broader research community through publications at major conferences and journals. To ensure repeatability and verifiability of the experiments, the PIs will release the source code for all developed simulation models along with the primary data used to generate the plots and tables included in the publications. Detailed simulation parameters and settings used to generate the results will be released in a publicly accessible document immediately after publication.

Period of data retention

All simulation results, developed simulation tools, and course materials will be released shortly after publication, and kept publicly accessible for a period of three years after the proposal termination date.

Data formats and dissemination

Published papers will be made available in PDF format, whereas the primary data used for generating the results reported in the papers will be tabulated and made accessible in raw text format. The PIs will also release a detailed description of the experimental setup to ensure the repeatability of their experiments: simulation parameters, software settings, benchmarks and input set sizes will all be summarized in a PDF document and made publicly available. Importantly, immediately following publication, the PIs will release the source code for the simulation infrastructure used to generate the results, in compressed ZIP file format.

Data storage and preservation of access

All released material will be kept available on a project website that will be maintained by the PI and hosted on the computer science department's web servers. The availability and preservation of the data will be assured through weekly scheduled backups of the project web site.

2) ***Example from the NIH Policy Web Site (but could be used for other research sponsors):***

The proposed research will include data from approximately 500 subjects being screened for three bacterial sexually transmitted diseases (STDs) at an inner city STD clinic. The final dataset will include self-reported demographic and behavioral data from interviews with the subjects and laboratory data from urine specimens provided. Because the STDs being studied are reportable diseases, we will be collecting identifying information. Even though the final dataset will be stripped of identifiers prior to release for sharing, we believe that there remains the possibility of deductive disclosure of subjects with unusual characteristics. Thus, we will make the data and associated documentation available to users only under a data-sharing agreement that provides for: (1) a commitment to using the data only for research purposes and not to identify any individual participant; (2) a commitment to securing the data using appropriate computer technology; and (3) a commitment to destroying or returning the data after analyses are completed.

3) ***Example for microarray data sharing*****:**

Microarray data will be made publicly available by depositing it in the Gene Expression Omnibus (GEO) maintained by the National Center for Biotechnology Information. This web resource is free to all users. GEO curators ensure that the metadata provided with the submission are compliant with MIAME standards, then assign an accession number that will be included in any publications referring to the data. Submission to GEO also requires inclusion of both the processed data for individual arrays and the microarray image files (e.g. CEL format for Affymetrix arrays, TIF format for non-proprietary arrays) so that

scientists who download the data can analyze the experiment with alternative algorithms.

4) ***Example for a small study with relatively simple data :***
(Note: this plan addresses "resource" sharing, including both data and reagents)

Sharing of data and reagents generated by this project will be carried out in several different ways. Our plan includes:

Presentation at national scientific meetings: We expect to have one to three presentations at national meetings, such as Keystone meetings on HIV/AIDS vaccines and immunology. These meetings include presentations of new information on a variety topics related to HIV/AIDS research. It is expected that the investigators from this project will be active participants in these meetings and share the data with scientists interested in HIV vaccine development.

Publications: Publication in Open Access journals, in which the scientist retains copyright on the published material, will greatly expand general access to our findings and allow us to post articles to our web server as they appear in print. For this reason, we will consider such journals for our own publications, as well as journals such as the *Journal of Virology*, which provide free-online access to all published content with a 6-month lag period following publication.

Reagents: Immediately following publication of experiments that describe novel reagents or materials (virus vectors), we will make all such clones and constructs available upon request to all researchers who request them, after completion of a Biological Materials Transfer Agreement. For most non-profit institutions, when the transfer involves a biological material, the UR Office of Technology Transfer (OTT) uses a simple AUTM implementing letter to document acceptance of the terms of the Uniform Biological Material Transfer Agreement (UBMTA). In those instances where the transfer involves a non-profit institution that is not a signatory to the UBMTA, the UR's standard Biological Materials Transfer Agreement or (if no intellectual property is involved) the AUTM Simple Letter Agreement is utilized.

See: http://www.urmc.rochester.edu/techtransfer/faculty/Faculty_MTA.cfm

5) ***Example for a large study with relatively complex data, including proprietary data:***

In accordance with NIH policy for research with budgets over $500,000 in direct cost per year, we have mechanisms for data sharing in place, following the guidelines set forth in:
http://grants2.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm .

The proposed research will generate preclinical datasets, but not clinical data – thus obviating confidentiality/HIPPA considerations. The major issues with the preclinical data will center around (i) publication/copyright-related considerations (since many journals view web-based dissemination of research findings as being equivalent to publication, thereby precluding acceptance of manuscripts that incorporate those data) and (ii) intellectual property/patent-related issues, particularly as they relate to materials provided via Project X (COMPANY) or experiments conducted under the terms of specific Material Transfer Agreements (MTAs). These considerations will, for example, prevent us from making datasets publicly available until after patent filing or manuscript publication (in the case of publication-sensitive "general" datasets) or until an acceptable confidentiality or data-sharing agreement can be worked out (in the case of studies that utilized reagents or materials that were covered by a MTA or other intellectual property-related agreement).

Therefore, we will make our "general" datasets available within 3 months of publication OR within 12 months following completion of a particular series of investigations, if publication is not being pursued and/or is unlikely (e.g., for negative results). Final datasets will be made available within 3-6 months of the termination of the grant award period and any associated no-cost extension period. Materials of this kind will be deposited into UR Research, where they will be assigned a permanent URL and electronically archived, for general access. Publications will reference the relevant URL (which will be allocated prior to publication, but not populated with datasets until after publication has occurred). As noted elsewhere, we will strive whenever possible to publish in Open Access journals or journals which provide free, general access with some reasonable delay period (e.g., 6 months).

In the case of datasets that involve results obtained using proprietary materials obtained under a MTA or other intellectual property agreement, access to datasets will require negotiation with the UR OTT, as well as with the originating parent company. Given the potential complexity of the issues involved, it is difficult to provide a "one size fits all" type of plan. As noted elsewhere, certain proprietary data, particularly data pertaining to Project X (COMPANY) may not be available for data-sharing because of potential patent/intellectual property issues (http://www.nih.gov/od/ott/200po6.htm). In other cases, however, we would anticipate making data available to outside parties under a data-sharing and confidentiality agreement that would provide for: (1) a commitment to using the data only for research purposes and not for commercial purposes; (2) a commitment to maintaining the data in an encrypted and secure manner, using appropriate computer technology; and (3) a commitment to destroying or returning the data after analyses are completed.

Finally, all data sets will be made available only in a password-protected and encrypted form, in generally available software formats that permit strong encryption - such as Adobe PDF. Original datasets will, of course, be maintained

behind the electronic firewalls of the various participating institutions, and original files and records will be subject to electronic audit trials, with data recovery and backup provisions (and attendant electronic security measures).

6) Sample language to address proprietary data:

This research has been co-funded by and is being performed in collaboration with Company under a Sponsored Research Agreement. Under said Agreement, Company has provided certain data that it considers proprietary and that UR is obligated to hold confidential. In addition, it is expected that the research may result in inventions which Company or UR may wish to protect by applying for intellectual property protection.

In order to ensure its ability to meet its contractual obligations and to exercise its intellectual property rights, UR will require data requests to be submitted in writing. Data requests which require access to Company data will require written permission from Company. Company may require the requestor to enter into a data sharing agreement restricting access and use of data. For requests for data which will form the basis of patent applications, the UR may delay providing access for up to 60 days in order to ensure intellectual property rights are secured.

7) ***Example from the UR's Clinical Translational Science Institute Competitive Renewal Proposal:***

Data sharing at the UR-CTSI is investigator-specific. Several methods of data sharing may be used, including publication in peer-reviewed journals, documentation of research in public archives, and dissemination of actual datasets from specific research projects through electronic means. The nature of the data and the restrictions that may apply to it will guide investigators in their choice of data sharing methods. The UR-CTSI encourages investigators to share data and will ensure that appropriate methods are developed and utilized to achieve data sharing in all appropriate situations. Data sharing must adhere to all provisions of HIPAA. The rights and privacy of people who participate in research must be protected at all times. Data for sharing should be free from any identifiers that would link the results to any individual research participants.

The University of Rochester makes available the "UR Research" digital repository as its main mechanism for sharing research data. UR Research is a service of the University of Rochester Libraries. The UR Research system was built using IR+, an open source repository software application developed at the UR and available through Google Code. The web address for UR Research is: https://urresearch.rochester.edu .

Data shared through UR Research can take almost any form. It will be appropriate to share certain data, for instance research protocols, in a text format, such as MS Word or PDF. Other researchers' use of such data is straightforward, and thus no

additional documentation and agreement are required. It may be necessary to utilize more sophisticated means to share other data. In such cases, the investigator will make available the actual datasets generated from research, in a commonly-used format such as a SAS® dataset. Since the dataset by itself is likely to be of little value without the knowledge of the study details and the context under which the research operated, the actual datasets will be associated with a related publication, research protocol or other documentation of the original research. UR Research will be the repository for such data. Some research data may be in multimedia form (such as recordings of focus group sessions). UR Research is fully capable of making these data easily available in digital format.

For investigators that desire a more controlled method for sharing their data, the Biomedical Informatics Key Function will implement web service applications as a solution to support automatic data sharing, integration, and analyses. For this purpose, the datasets need to be in specific formats that can be automatically interpreted by computer systems for further processing and analyses. Therefore, data sharing through this mechanism requires agreements among the parties involved in the data sharing activities. An example of such a format for automatic data sharing is XML, which can be used to wrap up both the metadata about the research (including the publication and documentation) and the actual datasets. The structure of the data represented in the XML format is defined by its schema, which must be agreed among all participating parties in order to correctly interpret the data. These common schemas for data sharing need to be standardized to serve a large research community. In addition to the actual datasets and the technical standards, we will also share the software tools (for example, the system for data integration and the execution engine of process knowledge interpretation) developed from this project with collaborators, such that they can be used in together with the data. It is important to note that sharing of actual datasets will also need to address other non-technical issues, such as regulatory requirements and intellectual property rights, which will also be included in the data sharing agreements. We will work with the CTSA Informatics Key Function Committee and other Consortium Committees to develop mechanisms to address these issues to ensure the effective dissemination of data and technology. In all instances, the UR-CTSI will insist that investigators benefiting from its resources adhere to the final NIH policy on data sharing (http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html).