PARADIM

Data Management Plan

V 1.0 4/13/2017

Prof. Darrell Schlom
Director

Lynn Rathbun, Ph.D.
Assistant Platform Director

Table of contents

1 Purpose

This plan describes the Data Management practices that PARADIM will implement with respect to samples, codes, and data generated by its users and staff. As detailed below, "data" may mean actual samples, simulation results, raw run data for crystal or thin film growth or a variety of structural or spectral characterization data, as well as polished published data and in all cases includes the metadata that is necessary to interpret (and repeat) the data acquisition conditions.

Data Management serves three related purposes:

> **Preservation of Data Assets:** Assure that all "data" associated with PARADIM projects is preserved. This applies to both proprietary and non-proprietary data assets. Data sets should include all information necessary to repeat a given experiment.
>
> **Protection of Privacy of Data Assets:** Assure that PARADIM users control access to their data sets for a period of time consistent with this policy. This period of time should be sufficient to analyze and publish the results of the research.
>
> **Dissemination of Data Assets:** Assure that non-proprietary data and results are made available in a timely manner to the larger community for verification and reuse.

The PARADIM data management plan is developed to serve these three purposes in a balanced fashion. The PARADIM data management policy applies to all data generated within PARADIM facilities.

It is the intention that PARADIM staff will have access to sufficient data and protocols to repeat **ANY** experiment done in PARADIM as part of a non-proprietary project, and that, for non-proprietary projects, those data and protocols also be made available to the user community in a timely fashion after publication or project abandonment.

2 Scope

While some overlap may exist, the PARADIM data management requirements are separate from the data policies of granting agencies, universities, and journal publishers. Because a research project may involve data taken at other facilities in addition to PARADIM facilities, it may be that the portions of the data associated with a particular research project or journal publication reside in different places. Likewise, it is possible that copies of a single set of data may reside in multiple places. PARADIM will only manage and be responsible for data assets generated in PARADIM-controlled facilities. In any case, all data assets developed by PARADIM staff, students, or users on PARADIM equipment will be managed in accordance with this policy. This policy covers what data will be saved, how it will be saved, who is responsible for saving it, who can access it, and if and how it will be released to the larger user community.

For the purposes of this policy, "data" includes physical samples, codes, simulation results, sample preparation data, growth data, and characterization data.

3 Applicability

3.1 Academic and Government Users

The procedures in this policy apply to all data generated by 1) academic users, and 2) Government users in PARADIM facilities. Such data is expected to be made public, eventually, and thus must, in most cases as described below, be preserved by PARADIM.

3.2 Industrial Users

Procedures and policies herein regarding PARADIM data storage and release explicitly do not apply to industrial users. Data generated by industrial users may be considered confidential by the users; the ability to retain confidentiality of data is a condition of use associated with the payment of full user fees. PARADIM will thus not, in general, store industrial users' confidential data longer than necessary to assure equipment safety and operation. Industrial users are expected to take any confidential data with them after facility use. PARADIM will archive industrial users data upon request but takes no responsibility for confidentiality beyond normal care.

4 Data Types

4.1 Physical Samples

4.1.1 Staff Development Samples:

Scope: PARADIM will retain excess "milestone" samples grown by staff in the course of their tool development and tool characterization activities. It is not the intention to save all samples or to make special samples for archival purposes, only to save excess pieces of those **significant** samples which might be useful for publication and documentation purposes.

Responsibility: This archive will be maintained by PARADIM Staff. A simple log will be kept, with links to the run data and characterization data.

Access: Archival samples are for internal use only. These samples will not be distributed outside of PARADIM.

Example: If staff grow the first ever sample of a novel material, excess pieces of that sample will be saved for later use. On the other hand, all the failed attempts leading up to that will, in general, not be saved. Neither will samples of routine materials grown during equipment characterization and evaluation be saved. PARADIM staff will determine which samples are significant enough to save.

4.1.2 Standard Samples

As recipes are developed and new materials discovered, PARADIM will make available certain standard samples grown via established recipes. This will only be possible once the original run data has become public (see below). These samples will be grown and characterized by the staff and made available to the user community through a simple "Sample Only" proposal process. These standard samples will most often be grown by PARADIM interns (e.g. REU and Masters of Engineering students) using the optimized recipes that were developed with PARDIM facilities by the discoverers of these materials. PARADIM interns will also characterize the samples they supply to interested users using the same methods that the discoverers utilized, with superimposed

analysis data so user can see how the samples they are provided with compare to the original samples. A list of these standard samples and their characteristics will be made available on the PARADIM web site. These will be grown on demand, and generally not stockpiled.

4.1.3 Internal Research Program Samples:

Scope: Samples generated by PARADIM's internal research program will be handled in a manner similar to staff generated samples. Excess significant milestone samples will be saved for later internal use as necessary.

Responsibility: This archive will be maintained by the PARADIM research group, with a log made available to PARADIM management periodically.

Access: Archival samples are for internal use only. These samples will not be distributed outside of PARADIM.

4.1.4 User Samples

Scope: PARADIM will not manage user samples, except for short term storage during actual lab use. PARADIM will maintain no archival user samples. This includes samples generated within the In-house research program.

Responsibility: The PI/User retains ownership of all physical samples and will remove them from PARADIM at the end of an experimental series. Responsibility for long term management, if any, of these samples resides with the user.

Access: As determined by user/PI.

4.2 PARADIM Developed Codes and Algorithms

4.2.1 Use of Software developed BY PARADIM supported students/staff

Scope: Software developed/modified by PARADIM staff will be made available to users for on-site execution, subject to license terms as appropriate.

Responsibility: PARADIM computational scientists and students/faculty are responsible for properly documenting and making available PARADIM developed software. Documentation shall include user documentation and well documented code.

Access: PARADIM developed software resources (original or modification to existing codes and algorithms) will be implemented on PARADIM computers and made available for use on PARADIM computation resources, with appropriate documentation. A standard Lab Use Proposal will be required to use codes on PARADIM computational resources.

4.2.2 Download of PARADIM Software for Use Outside of PARADIM

Scope: In addition, source code, executables, and documentation for PARADIM developed software resources will be available for download for use on other (non-PARADIM) computation resources.

Responsibility: PARADIM staff and the associated PARADIM faculty members will be responsible for making properly vetted and documented codes available for broader use.

Access: Access will be via the PARADIM.org web site via simple registration process (not a reviewed proposal).

Restrictions: Software will be released under an appropriate creative commons license or similar.

Timeframe: Software will be released within 1 year of development of version or upon publication.

4.2.3 Software Developed by Users on PARADIM Resources

Scope: During the course of a PARADIM project on PARADIM computational resources, users themselves may develop new or modify existing software resources, e.g., new psuedopotentials, new VDW functionals, etc. These should be made available to the user community as any other PARADIM developed work product.

Responsibility: PARADIM users and PIs will be responsible for making properly vetted and documented codes available for broader use. PARADIM users and PIs will be responsible for notifying PARADIM management when such software is developed.

Access: Access will be via web site via simple registration process (not a reviewed proposal).

Restrictions: Software will be release under an appropriate creative commons license or similar.

Timeframe: Software will be released within 1 year of development of version or upon publication.

4.3 Experimental Data (including Outputs of Simulations)

4.3.1 Classes of Experimental Data

Archival run data: Archival run data consists of **all** experimental data (and associated meta-data) generated within PARADIM facilities. This is unfiltered data. While this data will be kept, it is in most cases of limited general use. It will, however, be made available subject to the conditions described below.

Published/publishable data: After considerable time and considerable analysis and manipulation, a very small subset of run data turns into publishable data. This is the data that backs up graphs in publications, for example, or the final vetted recipe for growth or processing of a new material.

Metadata: All information necessary to identify and repeat the experimental conditions, such as instrument, operational parameters (voltage, flow, temperature, etc.), units of measure, experimental protocols, material identification, etc.

Protocols: Standard Protocols are certain recipes for routine processes that are repeated often, e.g., sample cleaning, material preparation, etc. Their use insures repeatability and simplifies the experimental description. Standard protocols will generally be developed by staff and made available to users.

Materials Properties: Ultimately complex run data gets distilled to quantified physical materials properties data, e.g., band gap, crystal structure, lattice constant, conductivity, melting point, etc.

4.3.2 Data Formats

To the maximum extent possible, data will be stored in non-proprietary, archival friendly formats. If at all possible it should not be tied to a particular instrument vendor's software, nor should it be tied to particular general purpose software package (Photoshop, PowerPoint, Excel, SigmaPlot, Origin, etc.). PARADIM will seek to acquire file format specifications for all instrument specific/embedded software at the time of instrument acquisition to minimize proprietary format issues.

Types of data and formats: The following is a non-exhaustive list of preferred file formats (at this time)

Text: (e.g., descriptions of experimental protocols) Text

Tabular (numerical) data: (e.g., spectral data, diffraction data, time/temperature growth recipes, pressure data, flow data, etc.) text, comma separated values or similar robust format; Proprietary format only as a last resort, and in that case a text file specification of the file format will be stored as well.

Images: (e.g., micrographs, images of graphs) JPG, TIFF, eps; Proprietary format only as a last resort

Scanned paper: (lab notebooks/log books); pdf

Video: (e.g., video of growth process video of diffraction data, etc.) MP4

Metadata: (all materials information (elemental, structure, etc.), protocol identification, and instrument parameters) Text with XML or JSON markup

Proprietary Instrument Data: PARADIM acknowledges that often considerable additional functionality is contained in the proprietary software and its associated file formats (e.g., manipulation of 3D AFM images). As a last resort, or when there is considerable added display/analysis value, data may be stored in proprietary format accessible only by the instrument vendor's software. In such cases, if possible, parallel copies in a more archival format should also be stored.

PARADIM acknowledges that preferred and accessible file formats evolve rapidly. Of necessity, these format recommendations will thus evolve and be updated periodically.

4.3.3 Digital Data Preservation Process

Scope: The majority of PARADIM data will be generated in digital format, i.e., computer data files.

File Names: Each data file will be given a unique file name, incorporating at a minimum PARADIM project number, date, and instrument information. File names will be generated manually or automatically according to a pre-defined schema.

Metadata markup: Accompanying each data file will be a metadata file describing the experimental materials and experimental conditions. Since the metadata is the handle for the search functions, different archives required different formats for this metadata. Metadata will be in XML or JSON format, depending on the archive.

Preservation: PARADIM staff will assure that all data and metadata files are properly named, tagged, and uploaded to archival storage. This includes markup of the necessary Metadata. Frequency will vary by instrument, at the discretion of staff, but shall be at least monthly or at the end of each project usage. In many instances, instrument specific scripts will accomplish this task automatically.

Resources: Data will be archived on the platforms described in section 5.

4.3.4 Analog Data Preservation Process

Scope: Some information will not originate in digital form. In particular, experimental notes will be contained in written form in laboratory notebooks. These notes, in both the notebooks of staff

and users, are a critical part of the data archive. In other cases, simple older instruments may not generate digital data (e.g., a simple oven or sample coater).

Capture method: Relevant sections of laboratory notebooks/log books/run sheets will be scanned and archived.

File names and Markup: Scanned files will be named and marked up in a manner similar to and compatible with that of digital data

Users Responsibility: At the end of a project run, or at least monthly, staff and users will work to preserve lab notebooks relevant to PARADIM activities; Users may redact any information reasonably considered confidential.

PARADIM's responsibility: Staff will archive their laboratory notebooks at least monthly.

Resources: Data will be archived on the platforms described in section 5.

4.3.5 Data from Affiliated Facilities

PARADIM offers access to hundreds of instruments in affiliated laboratories (CNF, CCMR, MCPF). These instruments are not under PARADIM control. Some of these data sets (e.g. TEM real time pixel array detector data) are truly massive, measured in terabytes. Data taken on instruments in affiliated facilities will not be managed or archived by PARADIM. It is the user's responsibility to work with the relevant equipment managers to obtain archival copies of data. It is the users responsibility to make such data available upon publication according to the requirements of his/her university, funding agency, and/or journal publisher. PARADIM may, at its discretion, include such data in the archived data set.

4.3.6 Data from Non-PARADIM, Non-Affiliated Facilities

Portions of the research on PARADIM projects will occur outside of PARADIM related facilities, e.g., in faculty laboratories at other universities or in faculty laboratories at Cornell University or Johns Hopkins University. PARADIM-supported students are required to abide by the spirit of this data management policy, regardless of where the data is taken. PARADIM, however, has no control over data taken elsewhere by PARADIM users. Management of this data is the responsibility of the user group, consistent with the data management requirements of their institution, the other facility and/or funding sources. Preservation of such data may be required by their granting agency or institution, but is outside the scope of this policy.

5 Data Accessibility/Availability by Project Type

5.1 Levels of Protection

Private: PARADIM data originates in the PRIVATE state. In the private state, only the user(s) (and PARADIM/IT staff) have access to the data. The user/PI will control who has access.

Restricted: A user may grant others access to formerly private data at any time, e.g., for collaborators. The person with restricted access, however, is not the owner of the data and cannot control who else has access to it.

Public: It is the intention that all non-confidential PARADIM data will become public at some point. Public data is accessible without restriction. The time line for data becoming public is varies by project type and status and is discussed below.

PARADIM staff and IT staff will have access to all data, but must respect the confidentiality of users' data, as well as any potential conflicts of interest.

5.2 Data –All projects that are not Industrial/Proprietary

Definition and scope: This section covers US Academic and US government (no charge users) as well as foreign academic, foreign government, and non-profit (cost recovery charge) users. Access to PARADIM comes with the explicit expectation that results will be published and shared with the larger user community in a timely manner. That being said, PARADIM recognizes that it takes time to analyze data and publish results. PARADIM obviously wants to protect the users' ongoing research and his/her intellectual property, whether publication or patent, and will be flexible in data release timelines. Here, Data means all the classes of data described in Section 3.

Release timeline: In most cases, data must be made public upon 1) publication or 2) project abandonment. Projects are considered ACTIVE during the period of active PARADIM use; they are considered ONGOING during the period after ACTIVE use, but before publication. Typically data analysis and supplementary characterization would occur during this ONGOING status period. Unless otherwise arranged, a project will be considered abandoned 12 months after the last active PARADIM facility use. PARADIM management will contact the PI/user after 12 months to ascertain if a project is truly abandoned or still in off-site analysis (i.e., status ONGOING). In cases where this timeline is deemed insufficient for the user's ongoing research needs, PARADIM will negotiate a mutually agreeable timeline. Eventually, however, all PARADIM data for subsidized use projects must be made public.

Release process: Upon publication or project abandonment, PARADIM management will contact the user/PI to discuss and set a timeline/conditions for data release. Data will not be released without attempted communication with the user/PI. Although concurrence is desired, PARADIM may unilaterally release the data if necessary.

5.3 Data –Industrial/Proprietary Projects

Definition and scope: The following apply to Industrial users of PARADIM Resources. This includes all users who pay full user fees.

Requirements: PARADIM run data will be archived in a protected file system with status PRIVATE. Only the user/PI, PARADIM staff and IT staff will have access to the data. The user/PI MAY make the data restricted or public at his/her choice, but is not required to. At the end of a project, the user may request that all copies be removed from primary storage.

Industrial users are under no obligation to publish; Industrial users are under no obligation to make data publicly available. For industrial/proprietary projects, the user is the PRIMARY owner and protector of his/her data.

PARADIM's responsibility: PARADIM staff will treat industrial user's confidential data with the same level of care and in the same manner as they would treat their own confidential data. PARADIM staff will not make industrial user's data available to others. PARADIM will seek to honor confidential user requests for primary data removal at the end of a project. This will not, however, affect backup tapes, for example.

5.4 Data –PARADIM Staff and PARADIM Supported Students, Faculty, and Post-docs

Definition and scope: This section covers data generated by PARADIM staff on PARADIM facilities in the course of their own research or instrument development/enhancement.

Requirements: Data management requirements for PARADIM staff and PARADIM supported students, postdocs, and faculty are the same as for other PARADIM users. All data must be archived and made publicly available. Students and staff have the joint responsibility of complying with the spirit of this policy. Staff will archive data generated on PARADIM instruments; students are responsible for archiving all other data.

Release timeline: Same as other academic users (section 4.3)

Release process: Same as other academic users (section 4.3)

6 Data Storage and Retrieval Infrastructure

PARADIM will take a multi-pronged approach to data management infrastructure. Different resources serve the different types and stages of data differently. Significant issues involve 1) privacy and access control, 2) markup, 3) and searchability and retrieveability. The use of multiple data management resources gives some additional security in case of failure of one of the partners and gives PARADIM some flexibility in implementing search and retrieval strategies. Gateways to these resources as well as other user friendly resources will be provided via the PARADIM website.

6.1 The Johns Hopkins University

PARADIM has enlisted the data management experts at Johns Hopkins University (JHU) within the Institute for Data Intensive Engineering and Science (IDIES) for storage of all PARADIM data generated under this policy. IDIES facilitates data storage for a wide range of big data users (e.g., from the Hubble Space Telescope and the Sloan Digital Sky Survey). They literally and figuratively store astronomical amounts of data. More recently, they have been transforming to store and analyze data from a broader range of scientific domains, including through NSF's Advanced Cyber Division (Award #1261715). They have committed to storing PARADIM data at no charge, and to help prototype online interactive data visualization and analysis tools.

6.2 Citrine Informatics

PARADIM will explore partnership with Citrine Informatics http://www.citrine.io/ for data management infrastructure. Citrine is a materials information company heavily involved in the Materials Genome Project. In Citrine's business model, they store massive amounts of materials data from multiple programs and institutions free of charge. And they make this information available to the research community free of charge. Their revenue comes from the use of their

proprietary machine learning algorithm against this database to make specific materials predictions for large corporate clients. From their point of view, the more data they have to work with the better their algorithm can serve their paying clients. To this end they provide free data infrastructure to help researchers meet data management requirements. Citrine has partnered with over 2000 institutions to provide this free data management platform.

At this time, Citrine does not offer private or restricted archives. All data is public. This archive would thus be, at the present time, most useful for final published data and derived materials properties data rather than raw data storage.

The public data archive is called the Citrination platform and it is accessible at http://citrination.com/

6.3 NIST Materials Genome Archives

NIST has invested significant resources in the development of resources and standards for the Materials Genome Project. These resources are being made available to the broader materials research community. PARADIM has attended a NIST sponsored workshop sponsored by NIST where results were presented and community input solicited. At that time, the tools did not seem to be at an appropriate stage of development for PARADIM use. PARADIM will monitor the development of these resources and may participate in the future. Like the Citrine archive, this repository is likely more appropriate for refined data rather than raw data.

6.4 PARADIM Maintained Archives

6.4.1 Simplified Archive of Vetted Material Recipes

As detailed above, PARADIM will archive all data, including raw run data. Since it is research, much of this data will relate to failed experiments or non-useful materials. While these archives will be searchable, this, however, runs the risk of burying useful refined data amongst mountains of less useful data. To simplify access to the most significant accomplishments, PARADIM will maintain on its website a simple list/database of significant, vetted materials recipes. This would be expected to be dozens of vetted (generally published) growth recipes rather than the thousands of data files in the main archive. These recipes will have links to the backup data directories at Citrine and JHU.

6.4.2 Publication Archive

PARADIM will maintain a simple database archive of links to publication citations resulting from PARADIM research projects, with links to the articles as available. As part of each citation, there will be streamlined links to the data associated with that publication, either at its location within the PARADIM archives or at journal or institutional archives.

6.4.3 Standard Protocols

Standard PARADIM protocols developed by PARADIM staff (and/or students and users) will be posted on the PARADIM web site. Their use and citation will help assure repeatability and simplify experimental descriptions.

7 Responsibility

Ultimate responsibility within PARADIM for implementation of the Data Management plan resets with the Assistant Director of User Programs. Responsibility for implementation, however, rests with the staff responsible for each instrument as well as with the users, project PIs, and supported students who generate the data. Compliance will be monitored annually.

----end----