Appendix V Data Management Plan

Data produced by XSEDE fall into two categories: primary data products and secondary data products. Broadly speaking, primary data products include software and records, while secondary data products include documents, materials, and documentation derived from project operations and primary data products. Both data types are discussed in detail later in this section. We do not address the issue of user data stored on XSEDE resources in this document; while XSEDE will provide facilities to assist in the execution of those user-defined data management plans, and it is expected that XSEDE allocated resources will be used for data management by a number of projects and users. This discussion of user-level data management plans is outside the scope of this document. The measures documented in this data management plan provide for the basic distribution of user and usage information to the research community, and include but do not detail plans to develop more robust, all-encompassing repository capabilities to handle secondary data products in particular. As XSEDE itself evolves, it is expected that the data products of interest, the sources of such data products, and the potential uses of these data products, will all continue to increase and evolve. Further planning and efforts regarding long-term preservation and access for both primary and secondary data products of XSEDE will be undertaken in concert with NSF-funded data infrastructure providers. The long-term goal for XSEDE data management planning and execution will be to minimize the proliferation of data sources and to provide simple and widely available access, both to document the usefulness of XSEDE infrastructure and to enable research on the use of XSEDE.

V.1 Primary Data Products

Primary data products of XSEDE will include: records of all users requesting allocations on XSEDE resources; records of the allocations granted and the usage of those allocations; records of resource usage, including specific users, job names and characteristics, including software used (if identifiable), and the allocated projects associated with each job; records of resources and services registered using the XSEDE information services, including in some cases availability information for specific resources and services; records of historical monitoring results; and of particular importance, user needs and requirements data gathered in the course of XSEDE development, including user surveys. Since user data may include personal information, access to some of this data must be restricted. Finally, there are software artifacts created by XSEDE staff, which may be either alterations to existing code or XSEDEspecific code for example code created for the XSEDE user portal. XSEDE recognizes the potential value of these data sets for current and future research and is committed to preserving and providing this data as appropriate. As most primary data products resulting from the allocations and accounting framework are also critical to the normal functioning of XSEDE, all recording of resource and usage information is currently stored in replicated databases, with at least two full copies of all data maintained at all times. Full backups are performed on a weekly basis, ensuring the preservation of both the data and the dependent service. These procedures will be continued into XSEDE2. Further, XSEDE plans to build on the existing software infrastructure and user database, while preserving all existing
data from XSEDE, the preceding TeraGrid effort, and prior NSF programs when available. The XSEDE infrastructure itself includes facilities to support preservation of data, such as metadata, replication, and other data management infrastructure. This infrastructure will be used within the project itself to provide a high-reliability storage layer for this primary data. Providing access to the primary usage data produced by XSEDE is dependent upon implementing appropriate access controls that ensure privacy for personal information while also making non-personal information widely available. For example, previous efforts have provided summary reporting on allocations and resource usage for public consumption. The XD Metrics Service (previously known as The XD Technology Audit Service for XSEDE) has provided new tools for examining usage and other information on Service Provider operations, using simple web interfaces for querying past usage data. Future efforts in XSEDE will focus on cooperation with the XD Metrics Service to expand the range of data available via these web interfaces, and to provide tools enabling the use of XSEDE primary data in research settings. A crucial component of the XSEDE engineering and development process involves the gathering of user needs and requirements information, largely through user surveys. These surveys will be performed by Indiana University under the guidance of an Institutional Review Board (IRB). Anonymized, de-identified versions of the raw user survey data will be made available, subject to IRB approval. Derived data on user needs, comprising the inputs to the ongoing systems engineering and management process, will be made openly accessible (subject to privacy and IRB constraints) as it is generated based on user surveys and other requirements-gathering processes. This approach will ensure the transparency of XSEDE engineering practice, as well as providing an important window into user perceptions of cyberinfrastructure. Software products generated by XSEDE will be sufficiently diverse to require a range of data management strategies. The most basic form of software product developed by XSEDE staff will be improvements or alterations to existing open source toolkits. In these situations, patches will be submitted to the software developers for maintenance by the open source development projects. For purpose-built software developed by XSEDE for XSEDE staff or users, such as web portal code, software deployment scripts, and other XSEDE-specific software products, the existing XSEDE software repository will be maintained and integrated into other XSEDE web interfaces for data and software. Software developed by XSEDE with potential for reuse will be made available via open repositories for others to download. Other products, for instance code developed for the XSEDE user portal, is unlikely to have significant potential for reuse and as such may not be made available in source form. If interest is shown in such software products, however, these decisions may be revisited. In all respects, XSEDE will attempt to provide open access to software code and other data generated by the project.

V.2 Secondary Data Products and Related Projects
In addition to the primary documentation of user activity, XSEDE creates a number of secondary data products with potential value. These include documentation of project activities such as quarterly and annual reports; ECSS work plans; publications written by project staff describing XSEDE activities or infrastructure; meeting agendas, minutes, and decisions; planning documents; source code developed for user interfaces; training materials developed for live, webcast and self-paced tutorials; materials developed by faculty for public dissemination and re-use in other undergraduate and graduate classrooms; and reports of advisory meetings. This information will continue to be available via the User Portal including publicly accessible project wiki pages. The self-paced tutorials are accessible via the Cornell Virtual Workshop website housed at Cornell University, and the CI-Tutor website hosted by NCSA. Long-term data preservation is handled through the XSEDE digital object repository based on IDEALS, utilizing handles with DOIs and persistent identifiers. Finally, documents and data will be jointly
created by the XSEDE project team and XD Metrics Service. Since most of these latter data are
generated explicitly for outside entities, in most cases XSEDE cannot be viewed as the provider of the
data in question. However, we recognize the potential value of these secondary data products and the
potential for integration and use of such data in novel ways. For these reasons, the XSEDE Program
Office will continue interactions with the XD Metrics Service on the development of an integrated
repository for these diverse data products. This may initially be as simple as a common, web-accessible
location for deposit and retrieval of data. As with the primary data products, providing access to this
repository to external researchers will be dependent upon privacy and confidentiality concerns, but the
goal will be to provide the maximum degree of access possible to all primary and secondary data. These
repositories, web services, and diverse data products will be discoverable by and accessible to the XD
community.