Data Sharing Plan for Matthew Turk GBMF DDD Grant

Summary: All data generated by this project will be made available in both raw and processed forms under CC0 or PDDL, and, except where constrained by external (i.e., “upstream”) copyleft licensing, all software developed will be made available in source form under permissive, BSD-style licenses. Source will be made publicly available on BitBucket.org and possibly GitHub.com, with additional backups possible. Data will be stored first on publicly accessible resources such as NSF resources and university resources, as well as backed up either in the Amazon Cloud or on local workstations. Every effort will be made to preserve integrity and accessibility of data.

Project Overview: The project will develop techniques for analyzing and visualizing data, particularly targeting data from physical science domains other than astronomy. Additionally, it will involve developing methods for sharing this data in an actionable way. Individuals employed by this project will be encouraged to pursue their own scientific interests in addition to further developing the project, as this will likely result in retaining domain-specific focus and encouraging cross-pollination of ideas. Data generated by non-Grantee employees will be strongly encouraged to be made available as described below, utilizing the methods developed within the scope of the project. Software will necessarily be developed to support this.

Data Description:

During the course of this project, we anticipate there being four primary sources of data. The first source will be that of data contributed by community members for the purposes of developing new techniques, visualizations, and as testbeds for data sharing and for continuous integration processes. This data, being supplied by community members, will be subject to whatever restrictions placed on it by the community members.

The second classification of data will be data ingested from external sources that are made public by some other agency. This data will be utilized and shared in accordance with the policy with which it is affiliated by the external agency.

The third type of data is raw data generated during the course of the project. As an example, the Grantee (Matt Turk) may conduct simulations of star formation or galaxy formation, or utilize other tools to generate non-astrophysical data. This data will be affiliated with considerable metadata (enough to generate the data ab initio). Others funded by this grant will also likely generate data.

The final type of data is generated, or processed, data, where some raw data has been manipulated (either interactively or in batch) to produce a different data product. The metadata affiliated with this will be a superset of the initial metadata, including the mechanism by which it was processed or manipulated.

Software will be produced with this project, and will have affiliated ephemera such as mailing list discussions, code review discussions, and artifacts of the design process. All code review and design discussions that are conducted electronically (i.e., not through in-person interactions) will be stored in a format recording the time, state of the code, and the various involved persons. Software will be versioned (likely in the distributed version control system mercurial) and affiliated with metadata recording individual, time, data, and commit message. Data file formats will necessarily be many; one aspect of this project is to develop a standard ingestion method, which will operate on non-uniform file formats. Data produced will include raw, processed, and processing tools.

Data Management:

Data will be stored primarily on systems either publicly funded (i.e., NSF or DOE) or provided by this grant. This grant will go towards funding both local storage and cloud-based storage, hosted by preference in an academic provider but also potentially in a commercial provider such as Amazon AWS, Rackspace or Google Cloud Storage. A primary aspect of this grant is to develop methods for sharing data in an actionable system. This will be the preferred method of storing data, inside a system where individuals are able to act upon it. This will include soliciting data for submission, providing persistent “handles” for data, and working in collaboration with libraries and cyberinfrastructure providers to ensure this data remains usable and accessible to individuals.

Over the course of the five year grant, the storage system will be maintained by funded individuals. Prior to the termination of the grant, we will work with the Gordon and Betty Moore Foundation to identify sustainability patterns for the storage and software systems developed. If no such patterns can be identified, we will make a best effort to return the data to individuals in a manner that is appropriate and usable for them.

Data Sharing:

When made available, data will be provided under either CC0 or a Public Domain Declaration license. This will allow for private, public, and commercial reuse of the data. We will make data available as early as possible, which will likely be pre-publication in most instances and immediately following publication when required by journals. Software will be developed in the open without access controls on repositories, and will be made available under the three-clause BSD license.

We anticipate that the Grantee and funded collaborators will participate in the open source ecosystem, and as such we will attempt to contribute patches upstream to other projects whenever possible. In some instances this may require copyleft licensing, but no proprietary software will be generated. External projects to which patches may be contributed that are copyleft include but are not limited to ownCloud, Mercurial, Git, and Kallithea.

A principal component of this project is to develop easy methods for acting on data. We will endeavor to make these methods are available to all individuals using the data and software produced by this project, including remote computation near data. This will be and aggressively open approach to data and software sharing -- not only will data be made available, but tools for using that data and even the computing infrastructure for storing and operating on that data will be provided.