SECTION J – DATA MANAGEMENT PLAN

1 Products of the research

The research described in this proposal will result in several types of data. The bulk of the data produced will be raw and processed simulation data (typically in time series), with the latter including individual fields extracted from the simulations, halo catalogs and merger trees, and synthetic observations. Additional types of data produced will include scientific software and a website describing the data products and software. This data will be stored exclusively in digital formats.

2 Data Formats

For simplicity, we break the discussion of data format down by the data product:

Raw simulation data will be stored in large binary files using the HDF5 data format. This is a commonly used scientific data format that has strong community support. The structure of the simulation datasets can be somewhat complex, but is explained by in-file metadata and an accompanying text file that describes important information about the grid patches. These data outputs can be read by a variety of publicly available, open source tools such as the YT data analysis tool, as well as the widely used Paraview and VisIt visualization tools. If necessary, the YT tool can be used as an intermediary to produce files that are more easily accessible to other researchers.

Processed simulation data, including most analysis products, will be stored in binary HDF5 files with easily understandable file structures. Images resulting from data processing will be stored in standard file formats such as jpeg, tiff, or png, and any mock observational data created will be stored using the FITS file format. All files, when appropriate, will have accompanying metadata to explain the contents of the file, including the nature of it, units, and related information. All processed data and analysis outputs will be created by a suite of publicly available tools.

Scientific software: Both Enzo and Enzo-E are publicly available. Any additions made to either code as a part of this project will be contributed back to the relevant projects.

Website: The website associated with this project will be ASCII plaintext using html, php, javascript, and so on.

3 Access to Data and Data Sharing Practices and Policies

Simulation data will be of sufficient size (hundreds of terabytes) that it will be difficult to make this available via a website, and wildly impractical to download via http. As such, we will make it available to other scientists through the National Data Service and hosted by archives at the national supercomputing centers (see part 5 of this document, below) with no restrictions on its use. Instructions on its access will be given via the project website, and also in README files in the top-level directories on the NDS site. This will be done by the end of the grant period, although if possible we will make data available upon acceptance of the relevant peer-reviewed papers.

Processed simulation data and analysis products will also be of large size (hundreds of gigabytes to terabytes), and it will be impractical to make it all available via a website. The most bulky data will be made available via the National Data Service (see directly above, and Part 5), and other products will be made available via a website. This website will allow users to download and use the data with no restrictions on its use. This will be done by the end of the grant period, although if possible we will make data available upon acceptance of the relevant peer-reviewed papers.

Source code will be contributed back to the Enzo or YT software projects (as appropriate) by the completion of the grant period (and, more likely, by the acceptance of the peer-reviewed papers using the relevant source code). This software will then be available without restriction to any user who wishes to download, use, or modify it.
The **website** will be available on the Internet. All materials, including movies, images, and text, will be licensed for re-use with appropriate credit given using a license such as the GNU Copyleft license.

Given the nature of the work being done as a part of this project and our collaboration's stance regarding open source scientific software and reproducible research, we have no significant privacy, confidentiality, or intellectual property requirements that pertain to this research.

4 Policies for Re-Use, Re-Distribution and Production of Derivatives

All data created as a result of this project – including raw simulation data, processed data, data analysis products, and source code – will be freely available and usable by the public and by other researchers. The only condition we will **require** is citation of our publications upon use of our data (raw, processed, or analysis products). We will further **request, but not require,** that users of our software contribute any changes, improvements, or additions that they make back to the appropriate open-source software project. Data or images that are copyrighted (by a journal, for example) will be marked as such by watermark or text header in the file.

5 Archiving of Data

Due to the type and quantity of the simulation data involved in this project, data archiving is a significant concern. Raw and processed simulation data and data analysis products will be stored in the mass storage facilities at the national supercomputing facilities (primarily the National Center for Supercomputing Applications and the Texas Advanced Computing Center) and accessible via the National Data Service. This data will be stored and replicated as long as the PI's research group continues to be allocated computing time at these institutes (which will be at least for the duration of this grant, and likely significantly longer). However, data archiving in perpetuity cannot be guaranteed at present, due to the impermanent nature of the data tapes used for storage.

Source code will be stored on github primarily. While it is unclear what the future holds for any individual provider, it seems likely that the Enzo and YT projects will continue long into the future, and this source code will propagate, along with these projects, to whatever Internet home the projects find in the future.