Policy on Acceptable Simulation Data Archiving

Simulation data is any data which is produced through numerical calculations which are based on a model. The model typically has inputs that are based on observed data, but also may have inputs based on physical theory. Simulation data may be archived in the PDS provided it has demonstrated scientific relevance. The scientific relevance of simulation data is established if the data can be compared to observed data and is discussed and referenced as part of a peer reviewed publication. The resolution of the archived data must be commensurate with the resolution used to determine the publishable results. In addition, the model inputs must be part of the provenance that is preserved along with the simulation data.

Simulation data may be archived prior to the submission of a paper to a peer reviewed journal so that an identifier can be assigned to the data in order to allow the data to be referenced in the paper. Simulation data does not need to be reproducible, for example by running the simulation, however the numerical methods used to generate the data must be described. Simulation data should be physically reasonable and should not contradict actual observations or fundamental tenants (e.g., cosmic abundances of elements).

Adopted by PDS Management Council X MMM XXXX

Addendum

Information relevant to the interpretation and application of the policy.

Context

This policy is to be viewed within the context of the overall PDS4 system which includes all other policies and requirements. When simulation data is archived it must adhere to established specifications and procedures, is subject to peer review, and is preserved by PDS on the same time frame as observed data.

Reproducibility

Reproducibility for simulation data is the ability to generate an identical set of data using the same methods that were used to generate the archived data. This may require software and computational capabilities which are not readily available (i.e., super computers). Also, as with observational data from missions, in some cases it can be very costly to acquire simulation data in terms of time and resources. In either case the numerical methods used to generate the data must be described in sufficient detail that the results could be reproduced should someone have enough time and appropriate resources.

Definitions

Derived: Information (data) that is an extension or modification of other information.

Model (aka mathematical model): A physical concept enabling the prediction of the behavior of the system from a set of parameters and initial conditions.

Numerical methods: Methods used to find numerical approximations to the solutions of ordinary differential equations (ODEs).

Observation: An act that results in the estimation of the value of a feature property, and involves application of a specified procedure, such as a sensor, instrument, algorithm or process chain.

Provenance: {ISO} The information that documents the history of the [digital object]. This information tells the origin or source …, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. [provenance] adds to the evidence to support Authenticity.

Simulation: The imitation of the operation of a real-world process or system over time using a model of a real-life or hypothetical situation.

Simulation data: Information which was automatically created from a computer process, application, or other machine without the intervention of a human.

Note: The model represents the system itself, whereas the simulation represents the operation of the system over time.

Examples

Simulation Data: The organization of theoretical magnet field vectors in a defined space over a period of time.

Observed Data: The measurement of the magnetic field vector by a sensor.

Observed Data: The results of applying a calibration algorithm to a set of raw observed data.

Derived Data: The results of analyzing or transforming any other type of observed data.

Scenarios

To illustrate which data from simulations or model results would be archived here are a couple of scenarios:

Published Simulation Results

A researcher runs a simulation for the magnetosphere around Jupiter. The resolution of the simulation is 1000x1000x2000. The simulation is run for a total of 40 hours (2,400 minutes). The simulation produces a set of data for multiple parameters (magnetic field vector, electric field vector, pressure, temperature) for every minute of the simulation. For all 8 parameters approximately 32TB of data is produced. Only the "interesting", inner part (500x500x1000) of the simulation is stored which is approximately 4GB per step. For all 2,400 steps the total amount of data stored locally is 9.6TB. This is "condensed" by selecting a step every 10 minutes resulting in 960GB of data ready for analysis. Upon analysis only 10 steps reveal the features relevant to the research, resulting in a 40GB data set. A paper is written and submitted to a journal. The 40GB data set is described and archived. All other data is "discarded".

Note: While this may meet the journal requirements that data used in a paper be archived, it could also be beneficial to archive the entire set of "condensed" data so that the full simulation results can be evaluated. The extent of the data to be archived could be determined during archive planning or at the peer review.

Multiple Simulation Runs

The researcher in "Published Simulation Results" scenario tries multiple configurations of the simulation, none of which produce publishable results. None of the data is eligible to be archive.