Towards the Incorporation of Cosmological Simulation Data into the Virtual Observatory

Laurie D Shaw, Nicholas A Walton and Jeremiah P Ostriker
Institute of Astronomy, University of Cambridge

Abstract
We present here our proposal for a Simulation data model defining the structure and metadata required to describe a simulated dataset. This model is an adaptation of the Observation data model, adjusted to account for the differences between observed and simulated data. We discuss the differences between Observation and Simulation and outline the current work in progress in developing this model.

1. Introduction
With modern computer resources and grid technology we are now able to perform N-body dark matter simulations on a cosmological scale – box sizes on the order of thousands of Megaparsecs containing several billion particles whilst incorporating algorithms that allow good spatial resolution (e.g. Bode and Ostriker, 2003). Alone, these simulations are important research tools much effort has been put in simulating dark matter clusters in an attempt to find a universal density profile, statistical correlations between halos or in following the tidal stripping of dwarf halos as they fall into larger structures.

This paper describes initial work we are undertaking to provide Virtual Observatory (VO) access to a large N-body simulation dataset, involving the use of the VOTable interchange standard (see http://www.us-vo.org/VOTable/). The ultimate aim is for users to be able to extract data from simulation archives, run their own analysis tools, compare simulated data directly to observed using tools that the VO offers and eventually even perform simulated observations of simulated data. We describe here one of the earliest steps towards the achievement of this vision - an attempt to define a simulation data model in imitation of the data models already suggested for observed data.

2. IVOA Observation Data Model
A comprehensive data model named 'Observation' for observational data is currently being defined. This model attempts to identify the different aspects that fully describe either a single observation of the sky, or a dataset derived from a number of observations. It therefore represents a description of all the metadata that may be required by both data discovery and retrieval services and data analysis applications. An example of the typical categories that make up a complete description of an observation is displayed in Figure 1 (taken from the current IVOA Data Modelling 'observations' draft at http://www.ivoa.net/internal/IVOA/IvoaDataModel/obs.v0.2.pdf).

Figure 1 demonstrates that an observation can essentially be broken down into three main categories - Observation Data, Characterisation and Provenance. Observation Data describes the units and dimension of the data. It inherits from the Quantity data model (currently in development) that assigns the units and metadata to either single or arrays of values. Characterisation describes how the data can be used - it can itself be broken down into Coverage (within what limits the data is valid) and Resolution and Precision (different aspects of how accurately we are able to measure any single value). Provenance describes how the data was generated. This includes the telescope/instrument configurations, calibrations, the data reduction pipelines and the target itself.

3. Simulation Data Model
We have made a first attempt to define a data model for simulation data (named 'Simulation') within the framework outlined by the Observation model (see Figure 2). We found that the three main sub-categories - Simulation Data, Characterisation and Provenance are still applicable. However, for simulation data it is the Provenance object, rather than Characterisation that contains the real descriptive content of the model. We now describe below each of the three main parts of
the Simulation model, noting the similarities or
differences to their equivalents in Observation.

![Figure 1 - The general model for Observation. See text for description](image)

### 3.1 Simulation Data

This object remains essentially the same as in
the Observation model; a subclass of the
Quantity object, which is used to contain the
main data output of the simulation. However,
for simulated data there is a much wider range
of quantities that could be stored. In
Observation at least one quantity in the data
must be an observable; this is not the case in
Simulation. The metadata structure - the set of
Universal Content Descriptors (UCD’s) - used to
describe each quantity must be enlarged to
incorporate data clearly labelled as being
‘theoretically derived’. It must be flexible
enough to be able to describe the many different
quantities that can be measured from a
simulation, yet accurate enough to allow their
identification by a general query service. We are
currently working on how this can be done.

### 3.2 Characterisation

Although simulation data is fundamentally
different to observed data (we can know
everything about a simulated object), the
Characterisation outlined in Observation is in
many ways still applicable to the equivalent in
Simulation. Although simulated data is
normally not subject to enforced data gaps or
exposure times, concepts such as Coverage,
Resolution and even Sensitivity are still
relevant. Rather than image resolution,
Resolution in the Simulation context refers to
mass resolution (the mass of the particles),
temporal resolution (time step) and spatial
resolution (grid size, or the particle-particle
interaction length, i.e. what is the distance at
which two particles will be able to resolve each
other as individual entities). These in turn will
determine the minimum size of objects that you
can resolve (thus providing the analogy to
sensitivity in observation) and distances to
which you can resolve over.

In Simulation, Coverage plays much less of a
role than in Observation. Quantities describing
the magnitude of the volume simulated and the
time period being simulated provide bounds to
the output data (the maximum distance between
any two halos is such and such), but there need
not be a well defined limit to particle velocity
for example. Likewise, a particular point in the
simulation is not of any significance in a
simulation - it is the relative position of objects
or locations that are normally of interest. This in
many ways simplifies Location as only the
coordinate system and the scale distance (and
time) need to be identified.

### 3.3 Provenance

As stated above, the Provenance object contains
most of the information describing the
simulation. This is because, unlike during an
observation, most of the effort in acquiring the
data is not through measurement but through the
execution of numerical routines, thus creating
the data set. The Provenance object is defined
as ‘the description of how the dataset was
created’ which for a simulation we are able to
describe entirely.

Provenance can be broken down into the
Theory, Computation and Parameters. Theory
describes the underlying fundamental physics
upon which the simulation is based. For
example, in a dark matter n-body simulation the
dominant effect that governs the evolution of
the simulation is gravity. In an experiment of
this type it will probably only be necessary to
use the Newtonian approximation of gravity
without having to account for general
relativistic effects. This is the kind of
information that would be included in the
Theory object - what processes have been
accounted for and which have been ignored?

Computation describes the technique used to
evaluate the physics described in Theory
through the execution of numeric routines. The
main components of Computation are the
organised sequence of algorithms that compute
the various stages of the simulation. The
algorithms are often chosen to provide a balance
between the time taken to complete the
simulation, the numerical accuracy and resolution, the complexity and requirements of the physics and so on, based on the hardware and software resources available. Often the sequence of algorithms will involve the 'main' simulation followed by a number of analysis routines.

![Diagram of Simulation Model](image)

**Figure 2 - The general model for Simulation**

From a metadata perspective it is anticipated that the Theory and Computation objects will consist of references or links to relevant papers and (in the case of Computation) a reference to the code itself (this could be secondary function of the web services that provide the astronomy community access to the simulation tools).

The third component of the Provenance is Parameters. The input parameters not only define the physical context of the simulation, but also the resolution and detail. If the algorithms are analogous to a mathematical function, the parameters are the values of the input variables. They are identified here as being separate to the Theory and Computation as they are normally the only elements that change between different runs of the simulation. Parameters can be broken into Physical and Technical sub-classes. In any cosmological simulation there are at least five physical parameters that must be defined (baryonic and total matter densities, initial density perturbations, etc). The technical parameters such as 'box size' (representing the volume of space being modelled) and the total number of particles will define the accuracy and resolution of the simulation and in some ways, the amount of processing power required. In an alternative view, the physical parameters could be seen as an input to Theory and the Technical parameters as an input to Algorithm. However, it seems sensible to separate the variable and static elements of a simulation. Parameters will be contained by the Quantity object. Although UCDs will probably already exist for the physical parameters, a new category will need to be created for the technical parameters. Work is currently in progress attempting to define the requirements of this new category.

### 4. Conclusion

We have outlined our proposal for a simulation data model. It was found that the basic structure of the model could be based upon the Observation data model. However, unlike the latter, the bulk of the metadata is found in the Provenance object (metadata describing how the data was created) rather than the Characterisation object (metadata describing how it can be used). This data model is being agreed and worked out at the international level through the International Virtual Observatory Alliance (Theory Interest Group and the Data Model Working Group). The work described in this paper is being inputted to that forum (of which the authors are active members).

### 5. References


VOTable: [http://www.us-vo.org/VOTable/](http://www.us-vo.org/VOTable/)

Observation Data Model: [http://www.ivoa.net/internal/IVOA/IvoaDataModel/obs.v0.2.pdf](http://www.ivoa.net/internal/IVOA/IvoaDataModel/obs.v0.2.pdf)

Quantity Data Model: [http://www.ivoa.net/twiki/bin/view/IVOA/IVOADMQuantityWP](http://www.ivoa.net/twiki/bin/view/IVOA/IVOADMQuantityWP)

Universal Content Descriptors (UCD’s): [http://www.ivoa.net/twiki/bin/view/IVOA/IvoaUCD](http://www.ivoa.net/twiki/bin/view/IVOA/IvoaUCD)

Theory Interest Group: [http://www.ivoa.net/twiki/bin/view/IVOA/IvoaTheory](http://www.ivoa.net/twiki/bin/view/IVOA/IvoaTheory)