International

Virtual

Observatory

Alliance

Data Model for Simulation

Version 0.2

IVOA Working Draft

2005 May 23

This version:

0.2 http://www.ivoa.net/internal/IVOA/IvoaTheory/SimulationDM.doc

Latest version:

0.2 http://www.ivoa.net/internal/IVOA/IvoaTheory/SimulationDM.doc

Previous version(s):

0.1 http://www.ivoa.net/internal/IVOA/IvoaTheory/SimulationDM.pdf

Author(s):

Laurie Shaw
Nicholas Walton
IVOA Theory Interest Group

Please send comments to: mailto:theory@ivoa.net

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.

Status of This Document

This is a Working Draft. The first release of this document was 2004 August.

This is an IVOA Working Draft for review by IVOA members and other interested parties. It is a draft document and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use IVOA Working Drafts as reference materials or to cite them as other than "work in progress".

A list of current IVOA Recommendations and other technical documents can be found at http://www.ivoa.net/Documents/.

Acknowledgements

This work is based upon the original Simulation data model proposal along with ongoing discussions within the IVOA Theory Interest Group.

Contents

1 Introduction

This paper describes initial work we are undertaking to provide Virtual Observatory (VO) access to simulated datasets. 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 simulations. 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 model already proposed for observed data. In doing this, we take the Observation data model as our starting point, modifying it where necessary. It is hoped that this will ensure that Simulation has a similar overall structure to Observation, differing only in the detail. There are two purposes to this approach. Firstly, it is hoped that a similarity between data models will aid the process of comparing simulated and observed datasets. Secondly, it maintains the possibility of defining an overall data model for astronomical data, real or synthetic.

2 Simulations in the Virtual Observatory

2.1 Astronomical Simulations

A simulation is a means of approximating a state, or a successive series of states of a (normally complex) system governed by an adopted set of physical laws and constraints. When a physical process that we observe in the universe is too complicated to model using a purely mathematical approach, astronomers resort to applying the physical effects that we believe to be influential in an iterative manner, evolving the system discretely from one state to the next. The final results are compared to what we observe, when possible, in the hope that they bare some resemblance. Hence, simulations provide another medium through which we can compare theory and observation.

Simulations are frequently used today in all areas of astronomy; from the birth, evolution and death of stars and planetary systems, to the formation of galaxies

in which they reside and the formation of large scale structures, dark matter halos, in which galaxies themselves are thought to form. There is much variety in the processes being investigated and the underlying physics that govern them. Simulations of star formation require nuclear physics; simulations of strong lensing must solve the equations of general relativity. Cosmological Nbody simulations must find ways to apply simple Newtonian physics to a vast number of interacting bodies. Many different approaches have been chosen to tackle each problem, often employing very different algorithms in which to solve and evolve the complex physics involved. There is clearly a huge amount of information that must be recorded in order to fully describe a simulation and its results, not all of which can be quantified in numerical terms. In order for simulated data to be included in the Virtual Observatory, we must first clearly identify all the different components that describe a simulated dataset. This is the purpose of defining an abstract data model for simulations.

2.2 Applications of simulated data

It is also useful to consider who might be interested in simulated data. The most obvious group is theorists who want to compare their results with those of other investigators. This group must be able to discover and compare the results of simulations published through the Virtual Observatory that are similar to their own. To achieve this goal, we must be able to find a means of describing simulated data so that astronomers, through query services, are able to discover and identify the data that is relevant.

Secondly, observers want to be able to compare theory, i.e. simulations, with observations. This is more difficult as simulated data sets do not always contain quantities that are observable, or results that can be directly compared with observations. For simulations of a large number of objects, one can compare the results with observations through statistical means, comparing correlation functions, number counts in a volume limited sample and so forth. In general, VO enabled web-services are currently being developed that will aid comparisons between observed and simulated data. One of the most exciting examples of this are tools that perform simulated observations of simulations. By mimicking the process of observing an astronomical object or a portion of the sky through a telescope (possibly incorporating various instrumental effects) we can attempt to calculate what would be seen if we were performing a real observation of the object or system in question. This produces 'observable' quantities or maps that can be directly compared to real observations. For example, one may recast the results of successive time-steps of a cosmological Nbody dark matter simulation into a lightcone (see Figures 1 & 2). By 'shooting' rays through the lightcone, S-Z and X-ray maps can be produced that can then be compared directly to those obtained from observations.

It is envisaged that services such as this will be developed so that it is possible to compare 'simulated observations' of different simulations (of the same phenomena) as well as to observed data. To enable astronomers to apply such services to their own results, we need to define a common data format for simulations, not only for the processed and analysed results, but also to enable access and extraction of the raw simulation data.

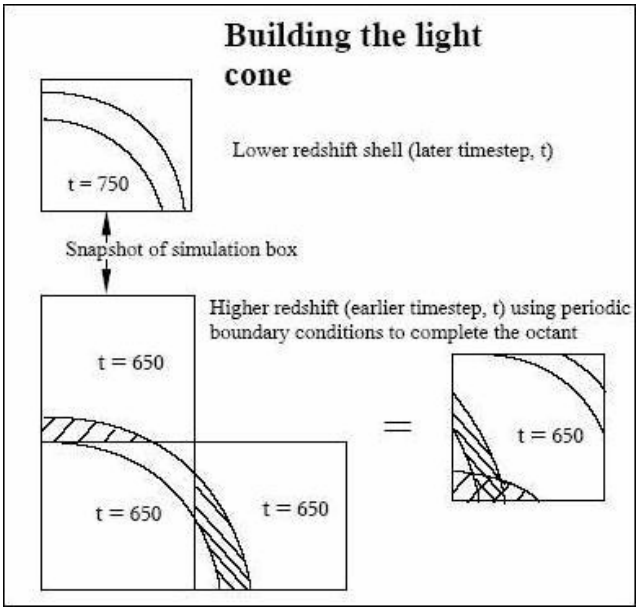

Figure 1 - The construction of a lightcone from succesive timesteps of a cosmological nbody simulation. Organising the results in this way enables theorists to perform 'observations' of the simulation, e.g. S-Z, temperature or weak lensing maps, which can then be compared to 'real' observations.

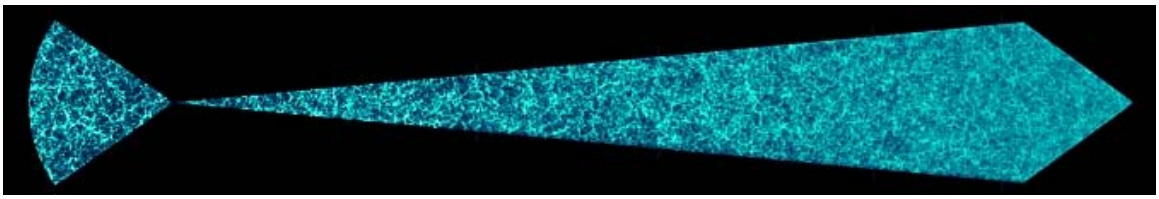

Figure 2 - Projected image of a slice through a lightcone constructed from a billion particle Nbody dark matter simulation. The gradual formation of structures is apparent as redshift decreases (towards the vertex of the lightcone).

Overall, it is important that we consider how simulations will be compared - both with each other and with observed data - by analyzing various use cases, such as the SZ mapping project outlined above. This will result in a set of basic requirements for the Data Access Layer (DAL), Astronomy Data Query Language (ADQL) and Registry working groups that must be met so that simulated datasets are interoperable in the virtual observatory. In order to be of general use, data

analysis tools must be able to query, retrieve and interpret the format and metadata of both simulated and observed data sets.

3 IVOA Observation Data Model

A comprehensive data model named 'Observation' for observational data is currently being defined [4]. 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 3 (taken from the current IVOA Data Modelling 'observations' draft [4]).

Figure 3 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) which assigns the units and metadata to either single or arrays of values. Characterisation describes how the data can be used. It can 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.

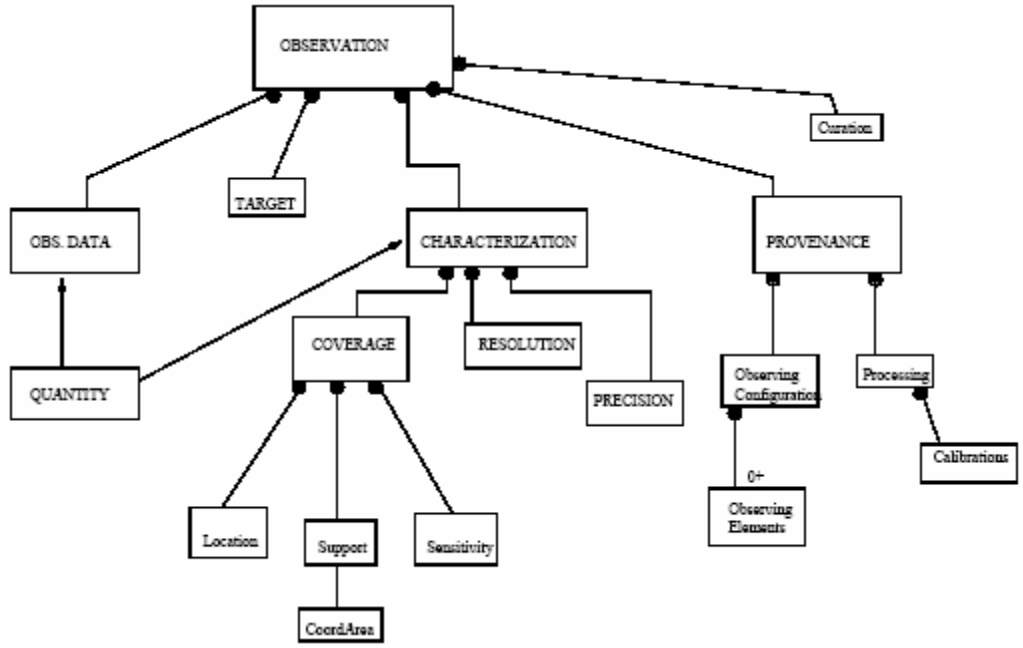

Figure 3 - The general model for Observation. See text for description.

4 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 4). In line with our aim of maintaining common aspects of both Simulation and Observation, we have found that the three main sub-categories - Simulation Data, Characterisation and Provenance are still applicable. We now describe below each of the three main parts of the Simulation model, noting the similarities or differences to their counterparts in Observation.

4.1 Simulation Data

This object remains essentially the same as in the Observation model - a subclass of the Quantity object [5], used to contain the main data output of the simulation. However, for simulated data there is potentially a much wider range of quantities to 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 [6] - used to describe each quantity must be enlarged to incorporate data clearly labeled 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 without creating a large quantity of highly specific definitions. We are currently working on how this can be done.

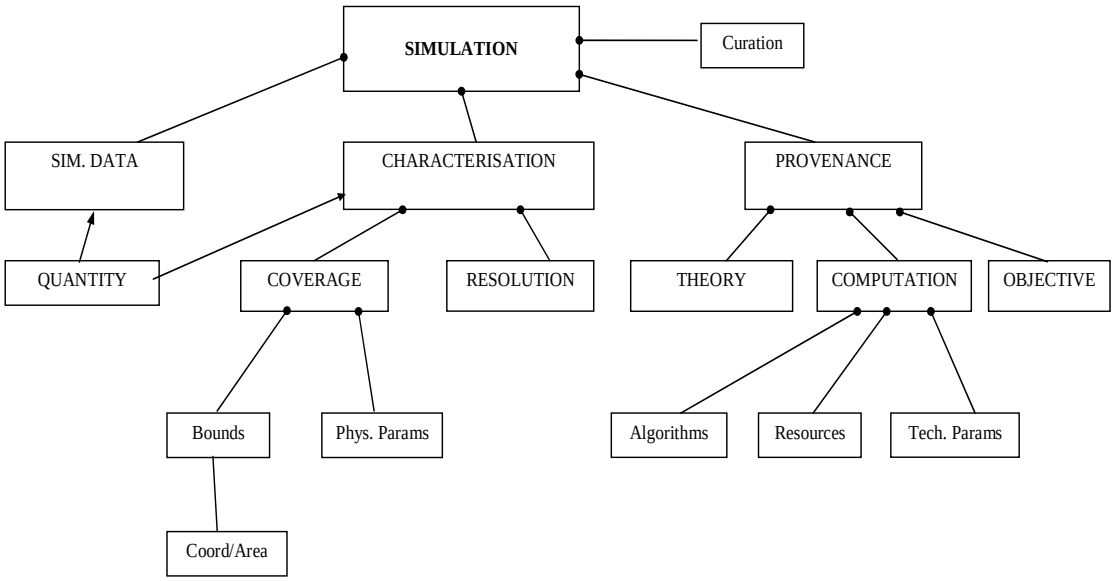

Figure 4 - The proposed model for Simulation.

4.2 Characterisation

Characterisation caries a substantial majority of the detail in the Observation data model. It describes not only the ranges over which each measured quantity is valid, but also how precise and how accurate they are. Essentially, it represents both the scope and limitations of the observed data and is therefore fundamental to the success of data query and discovery services.

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. Even though simulated data is normally not subject to enforced data gaps or exposure times, concepts such as Coverage and Precision are still relevant. Simulations are not sensitive to measuring effects, in the same manner in which observations are. However, they are sensitive to the choice of physical parameters, i.e. Ω_λ, σ_8, etc, that qualify the scientific validity of the results and the combinations of the raw technical parameters which determine the bounds and resolution of the results. We have approached the development of the Characterisation in the Simulation data model so that it attempts to provide analogues to that of Observation, i.e. summarise what and how well something is being simulated, whilst not delving too deeply into the Provenance of the data (how it was created). It describes the actual results themselves, independent of where they came from.

4.2.1 Coverage

In Simulation, Coverege contains less detail than in Observation. It is completely controlled by the initial technical and physical parameters of the simulation and is not subject to external or uncontrollable influences. It is strongly tied to the Provenance of the data. However, simulated data will frequently have undergone several stages of post-processing by the time it is published in a publicly accessible archive. Although a theorist could immediately make available the raw output of a large Nbody simulation, it is far more likely (and probably more useful) to first perform some kind of reduction of the data, e.g. identify objects that have formed in the simulation box and calculate their physical properties, or recast success snapshots of the simulation into a lightcone (see Figures 1 &2). Hence, the Coverage of the actual objects described by the data can be far more complicated than just specifying input parameters. They are determined by a very complex combination of the physical and technical parameters and the theoretical approximations in both the initial conditions and the adopted algorithms used to evolve and analyse the simulation. We have identified two components to Coverage: Bounds and PhysicalParameters.

4.2.2 Bounds

Bounds is identical to its equivalent in Observation, which describes the minimum and maximum values of each quantity. Clearly simulation data is not subject to data-gaps or variations in the response of observing instrumentation, hence we do not need to also define an analogue to Support, which describes in detail the domain in which observed data exists. Bounds provides more of a rough guide as to the region of parameter space in which the simulation probes. It may include the mass range of dark matter halos that have been identified in a cosmological Nbody simulation (or substructures within), or the redshift range of data arranged into a lightcone. Likewise, box size or volume would also be represented. Alternatively, it could include the wavelength range of synthetic spectra or the time period of a simulation of a variable luminosity source (black hole-neutron star pair, core collapse of a massive star, etc). Bounds will consist of a set of CoordArea objects with corresponding CoordSys objects to define the axes of the Coverage.

4.2.3 PhysicalParameters

The second component of Coverage is PhyscialParameters. The purpose of this object is to place in a physical or theoretical context the meaning/relevance of the simulated data and what it describes. The most obvious example is the adopted cosmology, represented by the baryonic and dark matter densities, the Hubble constant (at z=0) and the power spectrum of the initial density perturbations. Any input parameter to the simulation that has real physical meaning, i.e. in principle could be measure observationally, is included in PhysicalParameters. They are vitally important for interpreting – and placing into context - the simulation data. For many types of simulation, the results are meaningless without them.

4.2.4 Resolution

Whereas Coverage defines the scope of the data, Resolution in a Simulation context describes the limits in its accuracy and applicability. It should not be confused with Resolution in Observation which describes the 'smearing' of our knowledge about the data. Rather Resolution describes the limit at which the approximations of the simulation (as simulations are *always* approximations of the physical processes being analysed) begin to break down due to numerical artifacts. Spatial resolution will be limited by the finite grid size or the particle softening length (the distance at which two particles in a simulation will be able to resolve each other as individual entities). Temporal resolution will be limited by the time-step or the characteristic evolutionary timescale of an object. Mass resolution will be influenced by the particle masses. For example, there has been much recent interest in the inner density profile of dark matter halos. However, in order to probe the innermost regions, one must be sure that the results have

converged at that length scale. This is the type of quantity that will be described by Resolution. Many studies of simulations will include a section on numerical convergence to qualify the results they present.

However, as before, Resolution is not merely determined by the input technical parameters in Provenance. Numerical artifacts can also be limited by carefully written algorithms and are frequently dependant on the physical properties of the object being simulated (e.g. the characteristic timescale of an object may be dependant on its mass/size). The actual achieved resolution is frequently very difficult to predict from the input parameters alone - normally it must be determined from the final results, or by repeating the simulation at higher parametric resolutions (e.g. smaller mesh size). Therefore, the limits of the data are often more fundamentally linked with the data itself and what it represents than the parameterised components of the Provenance.

4.3 Provenance

Provenance contains information describing exactly how the simulation was performed. 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. By the information in Provenance, one should be able to completely and independently recreate the technical aspects of the simulation. Provenance can be broken down into Theory, Computation and Objective.

4.3.1 Theory

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?

4.3.2 Computation

Computation contains the technical information regarding the simulation. It describes the both the technique used to evaluate the physics described in Theory, the sequence of algorithms, their input variables or technical parameters and the hardware resources.

4.3.3 Algorithm

The components of Algorithm 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. For example, Figure 4 demonstrates the main steps towards the creation of a 'mock universe' - a catalogue of dark matter halos halos (identified in large scale dark matter Nbody simulations) populated with galaxies taken from observational surveys. Stage 1 represents the bulk of the simulation, the Nbody Tree Particle Mesh code (TPM, [3]) that evolves the dark matter particles from their distribution in the early universe to the present day. The stages that immediately follow are all 'reduction' algorithms that extract the useful information from the raw output of Stage 1. For example, the purpose of Stages 3 4 are to identify structures (halos) in the particle data and then to fit a density profile to each of them. This is roughly analogous to the data reduction performed on the raw data from observations. The details of the physics that goes on at each stage are unimportant here; the figure is just a good example of the sequence of algorithms that may be involved in such a simulation.

4.3.4 Technical Parameters

The second component of Computation is TechnicalParameters. If the algorithms are analogous to a mathematical function, the parameters are the values of the input variables. Parameters such as 'box size' (representing the volume of space being modeled), the total number of particles and the softening length, for example, will partly define the accuracy and resolution of the simulation and the amount of processing power required. They also partly determine the *purpose* of the simulation - a large box-size will be selected to simulate many objects at low resolution and a small box size to simulate a few objects in greater detail. However, the physical consequences of the technical parameter choice are summarised in Characterisation, which details the actual resolution and scope of the results. In Provenance we state the pure parameters so that independent groups are able to exactly reproduce the initial configuration of the simulation. The only exception to this are the physical parameters, or Constants, which are included in the Characterisation although - they would of course be required in order to perform an identical simulation.

4.3.5 Resources

The final aspect of Computation are the Resources. Included here are the details of the supercomputer/grid or cluster employed, the number of processors and the amount of memory used. Also included is the programming environment of the code (C++, FORTRAN, MPI, etc) and the time taken to complete the simulation.

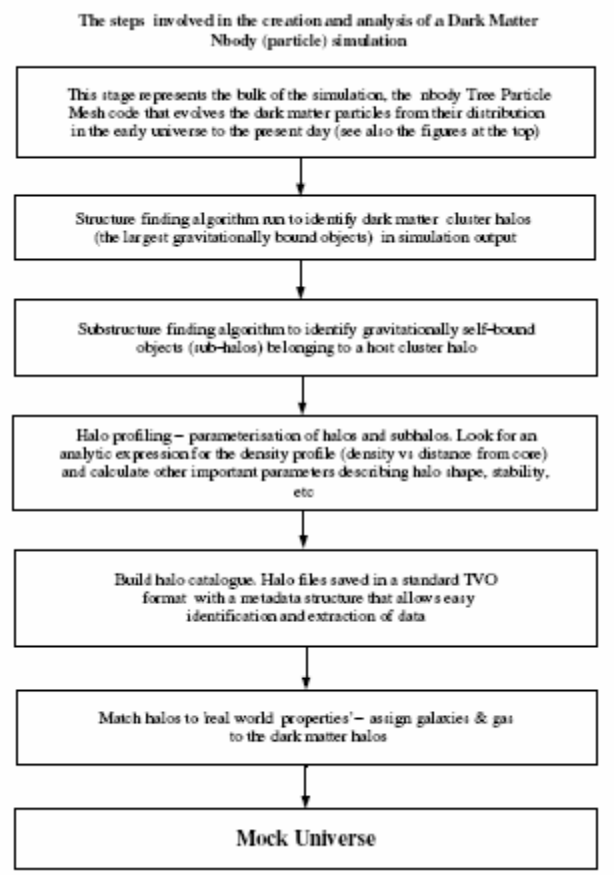

Figure 5 - Flowchart outlining the main steps towards the creation of a mock universe, from the raw TPM simulation (see [3]) to assigning gas and galaxies to the halos and sub-halos.

4.3.6 Objective

Objective is the simulation analogue to Target in Observation, which describes the thing (source, field) that the observation was made to study. It describes the aim of the simulation process: to produce the spectra of a stellar system, to model the formations of planets around a star, to analyse the formation of structure in the universe. We state *simulation process* as different simulation datasets may be derived from the same initial simulation, having undergone different analysis procedures. The final snapshot of a simulation box may be accessible through the VO independent of other datasets that contain more refined results. An astronomer may analyse a simulation with a different purpose to that which the simulation was performed in the first place. The content of Objective for the new results will be correspondingly more specific than the data that was used to create them.

From a metadata perspective it is anticipated that the Theory, Algorithm, Objective and Resources objects will consist in part 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). Although UCD's will probably already exist for some of the physical parameters listed in Characterisation, 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.

5 Model

- We have defined a data model for simulated data, using the Observation data model as a template.
- As in Observation, Simulation has three main parts – Simulated Data, Characterisation and Provenance. Loosely speaking, the three parts are metadata saying what the data is, metadata describing how to use the data in its current form, and metadata describing how the data was generated.
- Observation Data is a placeholder for the Quantity class (see the Quantity document, in work). It describes the axes and dimensions of the data.
- Characterisation consists of Coverage and Resolution. The axes of Characterisation are instances of Quantity. They denote the different parameters constraining the data.
    - Coverage describes the area of the Characterisation parameter space that the simulation occupies. It is compose of Bounds and PhysicalParameters.
    - Bounds is a set of CoordArea objects that define the range of values in which the simulation data exists (and is valid) in parameter space
    - PhysicalParameters consists of a set of physical constants that represent 'the Universe' in which the simulation inhabits and therefore represents to a certain extent the quantifiable physics (in a numerical sense) behind the simulation.
    - Resolution describes the scale at which is believed the simulation results begin to become significantly influenced by errors due to numerical effects. It is therefore not analogous to its namesake in Observation.
- Provenance describes how the data was created. It consists of Theory, Computation and Objective
    - Theory represents a description of the underlying physical laws in the simulation. It is expected to consist of a reference to a publication or resource describing the simulation.
    - Computation describes the technical aspect of the simulation. It consists of Algorithms, TechnicalParameters and Resources
    - Algorithm describes the sequence of numerical techniques used to evolve the simulation from one state to the next. It is expected that this also will contain a reference to a published paper or resource

- TechnicalParameters are quantities representing the 'inputs' to the algorithms, such as 'number of particles', 'softening length', etc
- Resources describe the specifications of the hardware on which the simulation was performed
- Objective describes the overall purpose of the simulation

References

[1] R. Hanisch, *Resource Metadata for the Virtual Observatory* , http://www.ivoa.net/Documents/latest/RM.html
[2] R. Hanisch, M. Dolensky, M. Leoni, *Document Standards Management: Guidelines and Procedure* , http://www.ivoa.net/Documents/latest/DocStdProc.html
[3] Bode, P., & Ostriker, J.P. 2003, ApJS, 145, 1
[4] Observation Data Model: http://www.ivoa.net/internal/IVOA/IvoaDataModel/obs.v0.2.pdf
[5] Quantity Data Model: http://www.ivoa.net/twiki/bin/view/IVOA/IVOADMQuantityWP
[6] Universal Content Descriptors: http://www.ivoa.net/twiki/bin/view/IVOA/IvoaUCD
[7] VO Theory Interest Group: http://www.ivoa.net/twiki/bin/view/IVOA/IvoaTheory