Mem. S.A.It. Vol. 80, 588

Memorie della

The Simulation Data Access Protocol

C. Gheller

CINECA – Via Magnanelli 6/3 I-40033 Casalecchio di Reno (Bo), Italy
e-mail: c.gheller@cineca.it

Abstract. SimDAP defines a protocol for retrieving data coming from numerical simulations from a variety of data repositories through a uniform interface. The interface is meant to be reasonably simple to implement by service providers. Data are selected by a proper search procedure. Once data of interest is identified specific quantities can be selected and sub-samples can be extracted and downloaded.

1. Introduction

In many fields of science, numerical simulations represent a valuable tool for verifying theories, analysing the development and the evolution of physical processes and for testing models. Present day computing systems represent virtual laboratories that can be used for virtual experiments. In some cases, e.g. in astrophysics applications, simulations are the only way to conduct experiments. The result of a numerical simulation, the so-called output data, is often exploited by a restricted number of experts only, usually the research group that ran the simulation. However, in most of the cases, the data could be of benefit to a much wider audience. Different researchers could study the data for different purposes, similar or potentially completely different from those the original group, who produced it, had in mind. They could carry out independent checks of scientific conclusions based on a given simulation or by comparing data to similar results, obtained with different numerical instruments. In order to offer effective access to a large number of numerical simulations, the Theory Interest Group (TIG http://www.ivoa.net/cgi-bin/twiki/bin/view/IVOA/IvoaTheory/) of IVOA is undertaking several core activities for laying the foundations for the development of a distributed digital archive of theoretical data, above which services and tools for its fast, intuitive and homogeneous access can be implemented.

In this process, the TIG adopted the standards and recommendations developed by IVOA (http://www.ivoa.net/), extending and adapting them when necessary. The proposed solutions are developed such that they are extensible, flexible, and adaptable to a wide range of astrophysical applications and they can be adopted by most of the research community.

The TIG group has presently focused on the definition of a general simulation data model (*SimDB*, Lemson et al. (2008)) and a first data access protocol, designed to provide the basic functionality over which sophisticated services can be built. This is the Simulation Data Access Protocol: *SimDAP*, Gheller et al. (2008).

Send offprint requests to: C. Gheller

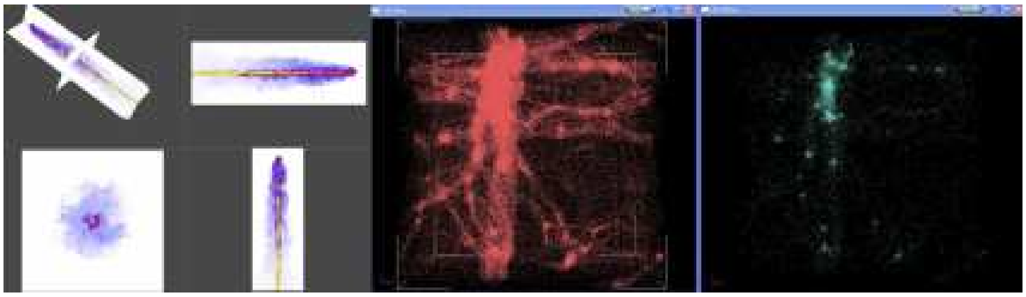

Fig. 1. Examples of preview services. Orthogonal projection of the simulated box (left). Decimated points set from an N-Body simulation (right).

2. SimDAP overview

The Simulation Data Access Protocol, hereafter SimDAP, defines a standard to access numerical simulation outputs (theoretical data), from a variety of astrophysical simulation repositories. The objective is that of defining a standard which allows humans and machines to access, use and process data, by means of different (distributed) services and tools.

SimDAP deals with datasets that can always be represented as (large/huge) tables in which raws identify a simulated element (a mesh cell, a particle, a pixel...) and colums represent the associated physical parameters (the 3D spatial coordinates, the velocity, the temperature...). Datasets can represent different timesteps (so, evolutionary configurations) of the the same simulated system. In the rest of this paper, we will refer to such datasets as *snapshots* of a numerical application. The snapshots and their data can be described and can be searched by means of the SimDM theoretical data model (Lemson et al, 2008).

Once identified, the simplest access mode to the datasets of interest is the download of the associated data file, possibly created adopting a standard file format, like HDF5 (`http://hdf.ncsa.uiuc.edu/HDF5/`) or FITS (`http://fits.gsfc.nasa.gov/`). However, in general, data is so large that its direct dowload is unfeasible. The SimDAP protocol describes a standard interface to access services which allows the user to reduce the data volume to move over the network (e.g. focus on a proper subsample of the data), permitting its download. The protocol defines also the interface to preview services which allow the user to choose between different datasets and to set the parameters to properly reduce the data volume.

The SimDAP protocol is designed primarily as a ”data on demand” service, with dataset created on the fly by the service given the position and size of the desired output dataset as specified by the client. This is not a simple task for various reasons. First, simulations data adopts specific units and coordinate systems, which depend on the nature of the problem, the characteristics of the algorithms and their implementation. Furthermore, simulation outputs can be represented by a wide variety of completely different data objects. For example, the output can consist in a set of particles in a given volume, where each particle has its physical position and a set of associated scalar and vector quantities, like velocity, mass density, temperature etc. On the other hand, mesh based simulations describe their data as discrete fields defined on a regular or adaptive mesh. The SimDAP protocol has the goal of providing a uniform description of the selection service trying keep it simple and, at the same time, to include as many different kind of simulations and data as possible.

3. SimDAP services

Search and exploration of available data archives and collections is part of the SimDB protocol, presented in Lemson et al. (2008). The result of the search and exploration phase consists in a set of parameters (metadata)

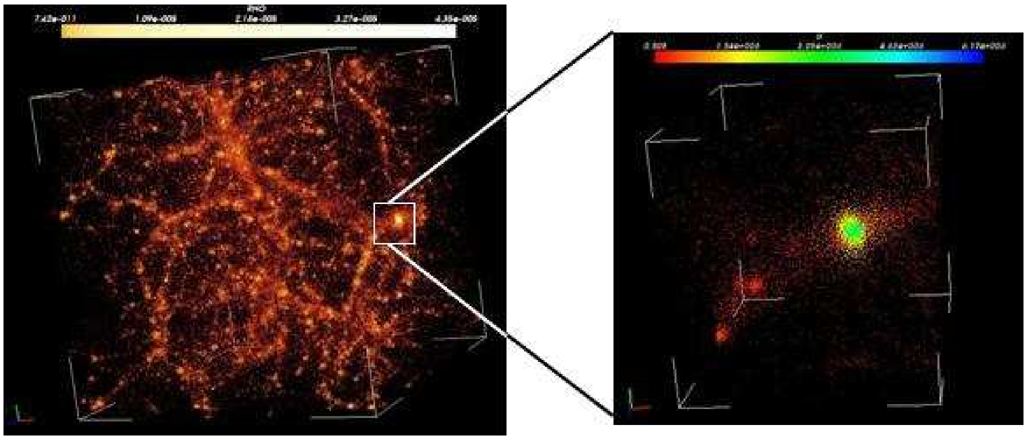

Fig. 2. Selection and extraction of a sub volume from a snapshot of a cosmological N-Body simulation.

which describe each dataset and an access reference to each specific snapshot, which provides an unambiguous link to the data source. SimDAP uses the access reference for the following services:

- data download
- data preview
- data cutout
- custom

3.1. Download

The search and exploration phase as well as other possibly available data services, find/produce one or more data products, which can be stored either in files or in databases, identified by their access references. The SimDAP download service allows the user to retrieve the data selecting only the fields (e.g. calculated physical quantities) of interest, which are listed as a result of the previous operations. The data is delivered according to the Theoretical Data File Format (TDFF), whose specification is in development (Gheller et al. (2008)).

The SimDAP download service proceeds as follows:
1. the user select the fields of interest (possibly, all available fields);
2. data are extracted and stored in one or more binary files according to the TDFF standard
3. An associated VOTable, following the TDFF standard, is created. The VOTable describes the content of the files according to the SimDB data model and contains their access references.
4. XML and binary files can be dowloaded by the user by means of any appropriate, available protocol (http, ftp, grid-ftp...).

3.2. Preview

In order to let the user explore the datasets and decide which snapshots he is interested in, the service provides tools to preview the available datasets. Such functionality must be available if the cutout service is supported (in order to select a sub-region specifying its size and position, see section 3.3)

This service could require the availability of a ”simplified” (but meaningful) version of the data, namely a thumbnail, easy to download and handle. Either, the service could allow the user to have a pre-defined view of one or more snapshots. The preview and the thumbnails features depend on the data and their implementation is up to the data provider. E.g. they could be represented by:

- projections of the computational box in the three coordinate directions (images),
- a random or decimated sample of the dataset (in particular for point like data),

• a reduced resolution realization of the dataset (e.g. averages over neighboring cells of a computational mesh)

• a ”clever” selection of regions according to specific criteria (e.g. ”overdense” regions) implemented by proper algorithms

• ...

3.3. Cutout

The SimDAP service can support a rectangular cutout of the data in a generic N-Dimensional (N-Dim) parameters space, in order to let the user focus on a region of interest, extracting the corresponding data and downloading the resulting file, strongly reducing the data movement. The service consists in extracting all the simulated elements for which some parameters have values in a given range. Notice that no assumption is made on the dimensionality of the problem (selection can be done on any number of parameters) or on the nature of the parameters (no restrictions to the parameters adopted in the selection operation). However, it can be convenient to consider as a favoured case, a 3D geometric selection, in which data are extracted according to its position in the 3D space. This means that spatial coordinates are used as cut-out parameters. This case is particularly simple and intuitive. Furthermore, it is common to a large number of applications. The cut-out region and the data will be selected according to the following parameters:

• position, as a N-uple of selection parameters which define a position in the phase space (e.g. the center of a the 3D geometric box)

• size, as the extension of the selected region in the N-Dim phase space (sides of the 3D geometric box)

• fields, a subset of the available physical quantities, result of the theoretical calculation, stored in a snapshot (e.g. the temperature or the velocity on a Cartesian Mesh)

The cutout region can be selected exploiting the preview tools introduced in section 3.2. The result of the cutout operation are stored in files according to the TDFF standard.

4. Conclusions

Nowadays the scientific community is witnessing an unprecedented growth in the quality and quantity of data coming from simulations and real-world experiments. This is due to dramatically improved computational power of modern computer systems and resolution of new imaging modalities; often datasets are measured in hundreds (or even millions) of gigabytes. To be useful, this output must be made easily accessible by researchers worldwide, at research laboratories, universities and other scientific institutions. Thus, the need to create and deploy new tools that allow data producers to publish their data and that allow data consumers to access that data flexibly, effectively and reliably is constantly growing.

The development of standards like SimDB, to describe theoretical data and search it across the network, and SimDAP, to select and retrieve data of interest, are contributing to effectively share and access data produced by numerical simulations, giving scientists new opportunities for accessing unlimited data resources. This will create a brand new approach to research, with unprecedented opportunities of scientific investigation, which would not be possible without a similar breakthrough instrument. Based on the common standards and above the basic data services, new tools will be built, to continuously improve the offer of the data infrasctructure. An example is given by the VisIVO server initiative by the INAF Astrophysical Observatory of Catania (`http://itvo.oact.inaf.it/visivoserver/`), which has the goal to provide advanced visualization services to data owners who adopts the standard proposed by the IVOA.

Acknowledgements. I wish to thank Gerard Lemson, Rick Wagner, Ugo Becciani, Patrizia Manzato, Herv Wozniak for discussions and feedbacks on the topic.

References

Lemson, G., Bourges, L., Manzato, P., Wagner R., Wozniak, in preparation

Gheller, C., Lemson, G., Wagner R. 2008, in preparation