Anexo 3a

The Virtual Observatory - A New Era for Astronomy

Albert Bruch (4)(7), Haroldo F. Campos Velho (2)(7), Alex C. Carcioffi (3)(7), Hugo V. Capelato (1)(7),Reinaldo R. de Carvalho(1)(7), Roberto Cid Fernandes(5)(7), Iranderly Fernandes (4)(7), Claudia B. Medeiros(6), William Schoenell (5)(7)

Executive Summary

There is a new paradigm in astronomy, the Virtual Observatory (VO). We trace it from the early developments only a decade ago to its current state. In a few years, astronomy will accumulate an unprecedented amount of data, on the order of 100 Pb, and growing at rate of 2-4 PB/year. This is an astonishing five orders of magnitude higher than it was in 2000. The VO is a response to the astronomical community's demands for improved and homogenized access to these data, combined with the tools to manipulate and exploit them. It is a complex enterprise with a decentralized, web centric nature, implying that astronomers need to rethink the old way of conducting their scientific programs. Most projects related to the VO started in the late 90's, and today an international effort is coordinated by the International Virtual Observatory Alliance (IVOA). The combination of growing data volumes and data complexity, coupled with computational and algorithmic advances, has made the VO a necessity. In Brazil, the National Institute for Science & Technology (INCT-Astrophysics) recently created by the Ministry of Science & Technology (MCT) is taking the lead in the VO development (BRAVO - BRAzilian Virtual Observatory). BRAVO is concentrating the effort on three distinct aspects of the VO development: 1) Database Development and Basic Infrastructure; 2) Data Grid & Processing Grid; and 3) Data mining. This white paper focuses some key VO-enabling technologies in Brazil, ranging from the hardware and software infrastructure needed, to databases and computational algorithms. A common theme among all these developments is the dire need for computational resources (CPUs, storage and network), software, and the expertise to design, install, and bring to life these complex systems. The international nature of astronomy implies that everyone can benefit, and everyone should contribute to this enterprise. VO's are a prime example of so-called eScience initiatives, which are basically multidisciplinary research efforts to develop computational tools and technologies that allow scientists to do faster, better or different research. We provide a basic and certainly not exhaustive outline of the components of the VO, and describe the specific contributions that the Brazilian astronomical and computer science communities have made and will be making to this effort. Our growing partnerships in large telescopes and unfettered access to large public datasets demand that we develop our own tools and expertise to leverage these investments and strengthen our scientific output. Finally, we describe the necessary next steps in terms of hardware, software, and personnel to advance BRAVO from an incipient program to a fully functioning project.

(1) Divisão de Astrofísica (DAS) - INPE
(2) Laboratorio Associado de Computação e Matemática Aplicada (LAC) - INPE
(3) Departamento de Astronomia - IAG-USP
(4) Laboratório Nacional de Astronomia
(5) Departamento de Física - UFSC
(6) Instituto de Computação - UNICAMP

(7) Comitê BRAVO

1 Introduction

For more than two decades, the international astronomical community has witnessed an exponentially growing capacity for accumulating astronomical data. Today, information is gathered in large surveys from the ground and from space, covering virtually the entire electromagnetic spectrum, from X-rays through the ultraviolet, optical, infrared, and beyond. Individual projects yield complementary data through specific, targeted scientific programs. Much of these data are made available to the community through public servers, usually in several different formats, and distributed at many institutions. Data, metadata, interfaces, and user accessibility needs are heterogeneous, with varying quality since each project typically curates its own data, presents it in a custom database, and even data formats in astronomy are instrument dependent with little effort made to unify them. In fact similar questions pervade the majority of areas of scientific research, as thoroughly discussed in [1]. Handling the data deluge and scientific modeling are two of five Grand Challenges in Computer Science for the decade proposed by the Brazilian Computer Society [2]. Astronomy is now an enormously data-rich science, and currently produces terabytes of raw data per day, with a few petabytes already in various archives. Both the data volume and data rate are increasing exponentially, with a doubling time of ~ 1.5 years. Even more important is the growth of data complexity (expressed, e.g., as the dimensionality of the parameter space spanned by the measurements of the detected sources) and heterogeneity. These data are now being federated in a global data grid under the umbrella of the Virtual Observatory (VO) paradigm. Computational hardware requirements are just one small part of the issues that arise when dealing with vast datasets. Processing takes a lot of time, so once completed, it is of paramount importance that querying and retrieving data be done quickly. This requires investment in not only database software, but the astronomical and computational expertise to design and implement efficient and scientifically useful data models and data structures supporting these models. This information, once structured in such a database, needs to be retrieved efficiently. This demands, on one side high-speed internet connections to which most research centers in Brazil do not have access; on the other side, it requires research on the design and development of a new generation of data mining algorithms, for which we must associate with computer scientists. For these reasons, our top priorities include implementing grid computing to enable the processing of massive datasets; creating a dedicated network for astronomy to enable access to the resulting data, and establishing a network of cooperative research between computer scientists and astronomers to develop these tools to produce cutting-edge science.

2 The Status of Information and Communication Technologies (ICT) in Brazilian Astrophysics

The new era of large data sets and the co-requisite data processing needs led to the recognition two years ago that we must modernize the tools for astrophysics in Brazil. In addition to the large photometric and spectroscopic surveys being carried out in both hemispheres, Brazil has committed significant resources to new facilities (including SOAR, Gemini, BDA, etc). As a result, we have access to extraordinary amounts of data in all portions of the electromagnetic spectrum, but without standard techniques for storage, retrieval, distribution, processing or analysis. Thus, the underlying concept of BRAVO is to federate these resources, using a common framework, standard interfaces, computational infrastructure, and analysis tools. In fact, in our conception, BRAVO should take on charge the planing of the VO activity in Brazil, orchestrating and coordinating the specific developments of different VO branches located at various universities and institutes. All of these developments are embraced within the concept of Information and Communication Technologies, encompassing all means for processing and communicating information. ICT is often used to describe digital technologies including

methods for communication, transmission techniques, communications equipment, and techniques for storing and processing information. The term has gained popularity partially due to the convergence of information technology (IT) and telecom technology. A major enterprise to develop the Brazilian VO is to understand our current hardware/software and personnel resources and their ability to meet our needs both today and in the future. The partners in this project comprise the totality of Brazilian astronomical institutes. We may conceive that the central repository of knowledge about computational hardware, software and personnel to be organized within a network as a Processing Grid in Astronomy (see also Section 5). Some concrete propositions on that have been given in [3], based on a census they conducted on the capabilities of institutes comprising the INCT-A showing the extreme deficiency of the current hardware, software and network infrastructure in Brazil. Moreover, an often overlooked (and underfunded) aspect of any computational project is the need for personnel with expertise in all aspects of the program. We cannot expect a computer scientist with experience in other scientific domains (e.g., bioinformatics) or even on commercial database applications to understand and implement astronomical databases without new training. Similarly, we would not expect an astronomer to develop efficient computational algorithms for, say, clustering analysis or pattern matching, without learning about recent advances in such applications.

3 Database Development and Basic Infrastructure

The current “gold standard” of databases in astronomy is the Sloan Digital Sky Survey. The imaging catalog has almost half a billion objects, in five filters, with nearly 500 columns of data on each object. While the volume of this single table (many Tb) is itself daunting, the SDSS database has nearly 100 unique tables, with an additional 50 views offering easy access to scientifically useful subsets of specific tables. The complexity of this database required years of consideration and cooperation between top researchers in astronomy and computer science, to design a workable schema, decide on which columns to generate indices to speed queries, understand how to load and update tables with new data, and how to provide public access. Just writing a portion of the table documentation was a full time job for a postdoctoral researcher for almost two years. Beyond the nearly 20Tb of catalog data, SDSS also allows users to access a comparable volume of images.

While the SDSS is quite complicated for an astronomical database, it pales in comparison to those from next-generation projects. Upcoming surveys such as Pan-STaRRS and LSST will yield a comparable amount of data - every time they survey the sky. These projects will create a new “SDSS” every few months. Not only do they produce multi-filter imaging, which must be processed, cataloged, stored, and distributed, they will also produce time series. Every object detected in one 3 image must be matched to its corresponding detection in all earlier images of that same area.

Optimal methods for differencing images must be developed to look for astronomical sources that vary or move. An entire pipeline is necessary to take moving objects, find them at different locations in images taken at different times, associate them, and generate orbits. Light curves for both stationary and moving objects must be created. All of this must be done almost instantaneously, because rare, one-time events such as supernovae must be found and notifications for follow-up observations disseminated before they fade. This means processing one gigapixel image every minute. The resulting database is correspondingly more difficult to model and populate. A “static” sky database must be created with everything detected, and updated as repeated observations allow for the creation of ever deeper images. Variable and moving objects must have all of their detections stored so that light curves and orbits can be derived. This scenario offers countless research challenges not only to astronomers but also to database experts.

4 Proposed Actions

In the past two years we have gained important experience and knowledge of VO development. New collaborations were established in all aspects of our planned investment. Within this context the priority must be to devise a roadmap for the near future to coherently invest in hardware software and, most importantly, *peopleware* that can meet our needs. BRAVO aims to create this synergy and contribute in strategic areas of the global VO. Below we list the main strategic points of this enterprise. We emphasize that besides developing a national VO structure, the actions that will be proposed below have all the same underlying principle: capacitation and enablement.

4.1 Investment on human resources

Its clearly essential for developing such a broad aimed project, to have a team of CS experts – for instance on database modeling and implementation, in programming, grid configuration and quality monitoring. It would be essential to create mechanisms capable of assuring such an expertise for all the research groups engaged in VO developing and VO science.

4.2 Establishing the VO culture within the Brazilian community

The Virtual Observatory is an emerging framework to harness the power of the Information technology for the astronomy of the next decade. Moreover, VO is an inherently multidisciplinary activity, involving not only Astronomy but also Computational Sciences, Information Technologies, Statistics and Applied Mathematics. It is therefore of paramount importance that we begin to promote this new kind of scientific culture - the VO culture - within our scientific community. Various kinds of actions may be envisaged:
- promoting National/International VO Schools, in which scientists would be trained not only on the VO techniques themselves, but, most important, to the VO way of doing science.
- promoting International Symposia and/or Conference with a focus on VO Science .
- To constitute a working group to study the pathways to bring VO education for under-graduate students
- Promoting small schools on basic computational principles and programming for undergraduate students to foster cooperation between astronomers, computer scientists and other scientists, needed to promote the VO culture through multidisciplinary research.

4.3 Investment on infrastructure

In order to succeed, a national VO structure should be established in terms of hardware. Ideally we should envisage such a hardware as a high performance distributed grid with high capacity storage, capable to be easily and quickly accessed using a high throughput network from any of the Brazilian research institute.
In contrast, the IT census demonstrates the level of insufficiency of the hardware, and more especially that of the network being used by a considerable portion of the scientific community in Brazil, particularly for astronomers. While it was possible to picture out an immediate solution for the question of hardware (see below), it is not clear how to handle the problem of the poor network actually linking the research institutes. It seems likely that the community should press RNP (“Rede Nacional de Pesquisa”) asking for better performing facilities.

4.3.1 Developing a Brazilian grid for Astronomy

Establishing such a grid is a task involving both hardware and software and most of all, man power and expertise. We propose as a first step on establishing the proper Brazilian grid for astronomy, to take advantage of the facilities already put in place by LNCC (Brazilian National Lab for Scientific Computing), offering a national grid for high performance computing and storage. Notice that this action strongly depends on man power which we do not have at present. Particularly, it requires expertise on computer networks, web-based algorithms and parallel

programming. This kind of expertise is at a premium all over the world, and falls back to action 4.1.

4.3.2 Network Infrastructure

The results of the IT census point out the level of insufficiency of the network used by the scientific community in Brazil, especially for astronomers. High speed and secure network connections are of paramount importance not only for simple tasks in our daily work but also for establishing a national grid processing facility. In our view the community should immediately initiate a discussion with people from RNP (Brazilian National Research and Education Network) in view of the development of a plan aiming to increase the accessibility of the astronomical community in Brazil to a higher level. This is one of the main points of this project - to conduct a study of the current situation and move to a modern network infrastructure.

4.4 Investment on New Technologies on data processing and management

- Astro-Wise (AW) as a national environment for data reduction and analysis.

Critical issues in VO development include the large amount of data and how it is to be processed - transformed from raw images to reduced data suitable for further analysis. We must be able to process raw images either on an individual basis or in a pipeline. Also, as more and more sophisticated algorithms for object detection, star-galaxy separation, photometric redshift estimates, morphological analysis and more flourish, there is an increasing pressure for having flexible pipelines capable to absorb new technologies as for reprocessing old, high quality data. The AW system [4,5] provides a general environment for image processing on large datasets, a problem that is inherent in the VO concept. The implementation of AW federated over the different institutes may represent a major step towards providing the community with an environment for *large* amounts of data processing. It is important to notice that AW is the only general facility being planned at BRAVO. AW is currently extant only in Europe in compliance with IVOA standards.

This would be the first AW node in South America.

REFERENCES

[1] Hey, Tansley & Tolle (eds), 2009,*The Fourth Paradigm: Data-Intensive Scientific Discovery,*
http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf.
[2] A.C.P.F. Carvalho et al. *Grand Challenges in Computer Science Research in Brazil - 2006-2016*. Available from http://www.sbc.org.br, 30 pages.
[3] R.R. de Carvalho, R. Gal, H.C. Velho, H.V. Capelato, F. La Barbera E.C. Vasconcellos, R. Rocha, J.L. Kohl-Moreira, P.A.A. Lopes, M. Soares-Santos, 2010, JCIS (in press)
[4] http://www.astro-wise.org
[5] Valentijn et al, 2007, http://xxx.lanl.gov/abs/astro-ph/0702189