New Horizons in Time-Domain Astronomy
Proceedings IAU Symposium No. 285, 2011
R.E.M. Griffin, R.J. Hanisch & R. Seaman, eds.

doi:10.1017/S1743921312000683

Data Management, Infrastructure and Archiving for Time-Domain Astronomy

David Schade

Canadian Astronomy Data Centre, NRC Canada, Victoria, BC, V9E 2E7, Canada
email: David.Schade@nrc-cnrc.gc.ca

Abstract. The workshop on Data Management issues for Time-Domain Astronomy was conceived as a forward-looking discussion of the primary issues that need to be addressed for science in the time domain. The very broad diversity of the science areas presented in the main Symposium made it clear that most of the general issues for astronomy data management—for example, large data volumes, the need for timely processing and network performance—would be pertinent in the time domain. In addition, there might be other tight time constraints on data processing when the output was required to trigger rapid follow-up observations, while science based on very long time-baselines might require careful consideration of long-term data preservation and availability issues. But broadly speaking, data management challenges in the time domain are not at variance to any significant degree with those for astronomy or data-intensive research in general. The workshop framed and debated a number of questions: What is the biggest challenge faced by future projects? How do grid and cloud computing figure in data management plans? Is the Virtual Observatory important to future projects? How are the issues of data life cycle being addressed?

Keywords. Data analysis, astronomical databases, surveys, standards, sociology of astronomy

1. Introduction

The purpose of this workshop was to invite both a general scientific audience and a few representatives of some of the major astronomy projects of the near future to discuss what the future will look like for astronomy data management. The present-day environment features fast networks, cloud computing, cheap mass storage, and the Virtual Observatory. Governments have funded major computing infrastructure to support research, and the private sector can deliver on-demand scalability for a price. Major data centres have grown in sophistication and power. This rapidly-evolving technology landscape offers a range of options to observational astronomers. How have they reacted? What choices are being made for projects of all sizes that are near enough that data management plans have been developed in some level of detail?

There were people present who identified themselves as representatives of some of the major coming projects: LSST, Gaia, Pan-STARRS and others. There was representation from the international Virtual Observatory (VO) community and from several data centres. This report was based on an audio recording of the workshop; it simply summarises the main areas of discussion, the most salient points raised, and the general consensus on an issue, if one presented itself. Some of the statements about individual projects may not be consistent with the "official" project position since they may not have been made by project representatives.

2. What is the Biggest Infrastructure or Data Management Challenge?

The first area of discussion focussed on identifying the challenge which this group considered the biggest data management or infrastructure problem that they faced. Was it processing or storage capacity? Was it I/O challenges or networking—either internally or between the data collection and the users? Was it the cost of infrastructure?

Some projects, including GAIA, reported that the most serious obstacle had not yet been clearly identified. One smaller project reported that some of the obstacles were rooted in funding rather than technical problems. It is a question of both money and human resources. Does each project need to build its own data centre to house its collection, or can shared facilities be used to reduce costs? It was reported by an LSST team member that one of the biggest challenges is to establish the bandwidth to get the data from Chile to the United States. Large investments were planned to guarantee that the necessary bandwidth would be available, and will involve private priority access to network capacity—a challenge that is far more onerous than obtaining the processing capacity. There was some discussion of whether it was better to do major processing near the data in Chile, with the goal of reducing the quantity that needs to flow through long-distance fibres. It was stated that the baseline plan was to use major computing power (e.g., NCSA) that would be available only in the USA.

It was pointed out that there are at least two distinct areas where network capacity might be a challenge. The first is moving data from the observatory where they are produced to the processing and storage facility. Then a second problem might be the effective transfer of (possibly) Terabyte-sized datasets from the storage facility into the hands of other end users. It was possible, if one had extra processing power, to use that processing power to reduce storage requirements through compression. Of course, many major data centres and projects use compression as a routine part of their data management plans.

A participant familiar with VISTA made the point that they were starting to find that the catalogues which they produce are larger than the image data from which they were derived. This surprising situation can easily occur if many parameters are measured from image data for a single object. VISTA was experiencing a number of bottlenecks in data transfer between Garching, Cambridge and Edinburgh, and they were forced to use shipment of hard drives part of the time. Some participants experienced in the shipment of hard drives remarked that they found that such shipment was a very unsatisfactory means of data transfer, and was neither cheaper nor faster than the internet in the long run. Others pointed out that internet transfer might be best from some locations, for example La Palma to the United Kingdom, but that international transfers from countries like South Africa were simply not feasible using the internet. We clearly live in a world that is not as well connected as we like or need. In fact, it is normally the case that network performance is substantially below what might be expected. Some participants had experienced research internet connectivity that was nominally 1 gigabit per second between two locations but which in fact performed at only a small percentage of that figure.

In summary, moving data was frequently identified as a more serious challenge than data storage, and storage was seen as a greater challenge than access to sufficient processing power.

https://doi.org/10.1017/S1743921312000683 Published online by Cambridge University Press

3. Grid and Cloud Processing

In Canada, the USA and Europe, governments have made major investments in computational infrastructure whose purpose is to support research. Facilities such as Compute Canada, TeraGrid, and now the Extreme Science and Engineering Discovery Environment in the USA and the European Grid Initiative are operated and managed to solve large-scale computing problems, including ones similar to those faced by astronomy and by its major survey projects in particular. In addition, there now exists a number of commercial clouds, such as Amazon Elastic Compute Cloud and Microsoft Windows Azure, whose ability to deliver on-demand processing and storage services might be attractive to large projects or to small teams as components of their data management systems.

Are any of the major projects trying to exploit the situation by using either government-funded or commercial facilities rather than building their own project-specific infrastructure? It was remarked by several participants that the commercial clouds are too expensive. Fully-costed estimates of storage, networking (I/O) and processing requirements indicate clearly that this is not a cost-effective way to approach the major problems of observational astronomy's data management. There were no opinions contrary to this viewpoint. A few participants mentioned that their projects had used commercial clouds in an experimental mode to evaluate performance and cost, but no one reported that those clouds were now in their long-term production data systems.

What about the research or academic clouds that are funded by governments as shared facilities? Are those resources being exploited? The history of high-performance computing had developed without data-intensive computing as part of its mandate—a situation that is only now beginning to change. According to one speaker, LSST has used TeraGrid for some of the data challenges and to do preliminary evaluations of their processing problems. However, the components of the official LSST data management plan does not include any public facilities, nor are there plans to utilize major shared facilities in the future, either public or commercial. It was considered that a custom-designed processing and storage system for LSST was the only way to make a system that was efficient enough to handle the scale of the problem.

The Canadian community is involved in a project that is a collaboration of universities and national astronomy data management infrastructure. Its specific aim is to bring the national high-performance research computing (HPC) infrastructure (Compute Canada) into service for observational astronomy research. Like other HPC organizations Compute Canada has not been deeply involved in data-intensive research. A great deal of progress has been made, and a petabyte of storage has been allocated as well as access to significant computing power. This is the first project where production-level cloud computing has been done on Compute Canada systems. Both Compute Canada and the user community is slowly, and somewhat painfully, getting used to operating in a new way. The vast publicly-funded research infrastructure that is available to the research community makes this approach potentially extremely important and valuable, but astronomy teams do not yet easily find ways to interface to this system.

One participant described the experience of users of the EVLA, where the data rates may produce several terabytes of data for a small-to-medium sized programme, and could take weeks to process. Where should the processing be done? It is not easy to *ftp* the datasets back to one's home institution for processing, so is the solution to use some grid local to the observatory or the US TeraGrid? It may be important to realize that NRAO itself does not have sufficient facilities locally to store and process that quantity of data, leaving astronomers to scrounge for disk space and processing capacity.

https://doi.org/10.1017/S1743921312000683 Published online by Cambridge University Press

To summarize this part of the discussion, both major and even moderate-sized, projects feel that they need to create custom-designed data management infrastructure for their projects. Although national research computing infrastructures may be able to provide sufficient processing power they have other shortcomings that prevent them from being used: they are not, in general, configured for data-intensive research; massive storage is not tightly coupled to fast internal networks; problems of efficient I/O persist, and the geographical layout of major shared facilities is, in general, not appropriate. Furthermore, there are issues of long-term commitment of resources to the projects.

None of the major projects has plans to use shared cloud or grid computing facilities, be they are academic or commercial.

4. Virtual Observatory and Time-Domain Astronomy

Most, if not all, major projects include in their data management plan statements such as: "Our data services will be compliant with the Virtual Observatory". But what do statements like that really mean? Do they mean that the application of the products of the VO will help their projects achieve their science goals? Do they mean that the projects will help the astronomy community reach the goal espoused by the VO, namely that all major sources of data be capable of a scientifically useful level of interoperability?

A number of the participants at the workshop are involved with the International Virtual Observatory Alliance (IVOA), which coordinates the work of the national VO groups. The IVOA considers it essential that major new projects participate in the development of VO standards and that they apply those standards in their own data systems. If not, it cuts the legs off the VO and defeats the main thrust of VO work, which is to create a unified data management system for astronomy, at least from the user's perspective. The research users are one potential beneficiary of the work of the IVOA. But what do survey teams or major projects gain from the use of IVOA protocols?

Is there really substantial buy-in to the use of VO standards? What do we mean by "substantial buy-in"? One answer is that organizing a project's observational metadata in a VO-standard way as close to the moment of observation as possible would deliver the maximum benefit to the projects' data management plan, in the sense that downstream from that point any VO standard could be incorporated into the design of that plan. A good example might be the Hubble Legacy Archive, where the Simple Image Access Protocol (SIAP—an IVOA standard) was built into the core of the system, so its data products were ready instantly for VO access. That is the way we would like to go but clearly there may be other reasons, such as efficiency, for structuring data in a specific way that differs from the IVOA standards. The IVOA does not provide a complete, end-to-end set of protocols that can be used as the basis for a project data management system.

Major projects would be motivated to adopt IVOA standards as parts of their systems if it were clear to them that it would lead to increased efficiency or effectiveness or reduced cost. Is it possible to make that case? There are examples, like the Hubble Legacy Archive, that demonstrate value for users. There is the example of the Canadian Astronomy Data Centre (CADC), where IVOA standards have been used in a variety of roles in the internal design of their data management system. In that example it was clear that using those standards saved effort in designing some parts of the system, although implementation was not always perfectly straightforward. It was also reported that an agreement was being developed between LSST and the VAO (Virtual Astrophysical Observatory, the VO organization of the USA) to define what those two organizations can do for one another to ensure VO access to LSST data products.

Are VO standards easy to implement? It was remarked that a presentation to a LIGO (Laser Interferometer Gravitational-Wave Observatory) meeting elicited the question: "Where can I find the software to implement VO standards?" It was agreed that there exists some software but not really enough to implement major pieces of the overall VO set of protocols. This is an area that clearly needs improvement if projects are to be able to derive benefits from the VO.

According to one participant a search of the IVOA documentation failed to turn up a clear explanation as to how one should proceed with applying IVOA standards and protocols. The VO documentation was desribed as "opaque" and difficult to understand. Several other participants echoed and endorsed that view. It was remarked that the IVOA had been in existence for a decade and had, over that time, successfully developed its own culture and jargon; it was referred to as a "maze of standards and documents" that was difficult for astronomers to navigate. There was a perception that there are no clear directions from the VO to the data providers on how to structure their data and systems. These points constitute serious criticism of the accessibility of IVOA products.

Proponents of the IVOA and of the VO in general suggested that it should be an IVOA priority to provide data managers with server-side and client-side libraries of software that support implementation of standards and protocols. In fact, there do exist some repositories of such software (e.g., `code.google.com/p/opencadc/`). A more complete set of tested and documented software would be a great step forward for the IVOA.

The adoption of standard data models was discussed. It was argued that the most fundamental benefit that could be delivered by the IVOA would be a clear and usable data model that met the needs of the vast majority of projects. It was pointed out that standard VO data models are not intended to cover every last eventuality and treat every piece of metadata, since specific observatories and instruments will always employ some set of unique characteristics. But there is a large subset of the attributes of an observation that are common across most datasets (an obvious one being the pointing on the sky), and where there are common attributes then we should adopt a common language to describe them. If we could develop a data model that standardized an extensive set of common attributes, then most observatories could adopt them as their starting point for their metadata systems.

In summary, there was a feeling that the VO has not made its products as accessible as they need to be if they are to be adopted by the developers of project data management systems. The main goal of research projects is to achieve their science goals; anything that assists might be adopted, but additional effort cannot be expended on the VO unless it is an assistance rather than an annoyance. *Those immediate considerations of reaching project science goals will always trump downstream considerations such as preserving data for the benefit of the community.*

5. Long-Term Preservation and Data Life Cycle

Many established data centres are philosophically bound to the idea that astronomical data have long-term value and therefore need to be preserved, but in reality they have no mandate or funding from external parties to support them in the pursuit of long-term data curation. The strong commitment that NASA has made in the past toward good data management and long-term archiving now appears to be suffering serious erosion. There was a worry that NASA, as an organization, was backing off from its understanding of the value of well-calibrated and well-archived data products. As a research community, we need to be vigilant in order to keep data preservation on the agenda.

https://doi.org/10.1017/S1743921312000683 Published online by Cambridge University Press

In the past decade or two the cost of maintaining archives of astronomical data from earlier generations of instruments has been negligible because each new generation has produced orders-of-magnitude more data than had the previous generation. In consequence, we have not had to face seriously one of the important questions related to data life cycle: when are we justified in discarding data? Remarks were made that every survey that has ever been done has required re-processing at least several times, so losing raw data would have been scientifically catastrophic. However, things seem to be changing in that the data volumes produced by some projects greatly exceed storage and network capacity. Hard decisions must therefore be made at some point in the foreseeable future.

It is anticipated that facilities like the SKA will drive us toward different models for our data management systems, and different approaches to data life cycle. Careful cost-benefit analyses, driven by science, will need to be done. None of the projects represented at this workshop expressed the intention of discarding data at some point in the data life cycle. It was widely agreed that, so far, little has changed in our ability to process data optimally the first time. Data therefore need to be kept.

Keeping data over a long term implies not only that the data exist but that they be accessible; otherwise their value is minimal. It implies continued investment in archiving far into the future. What about archiving analysis and processing software? This has been discussed for many years but has rarely been implemented.

Most of the projects represented at this workshop were focussed clearly on achieving their science goals. Considerations of data life cycle and long-term preservation are, justifiably, secondary and will be considered seriously only at a later time.

6. Conclusions

- Most projects intend to develop their own custom-designed infrastructure to support their data management systems.
- VO is not delivering major benefits to the data management plans for most projects.
- Commercial clouds and government-funded academic research clouds are not part of the planned infrastructure for any of the projects represented at this workshop. Commercial clouds are too expensive and academic clouds are generally still not designed with data intensive research as a high priority.
- Processing power is not a major challenge.
- Storage can be costly but is not a major technical challenge.
- The most challenging activity is moving data between the components of a data management system, both internally (local networks and I/O) and externally (long internet transfers from observatory to data management site, and from thence to users).

In a nutshell, the people attending the workshop and the projects which they were representing are committed to the principle of building project-specific data management systems from the ground up. Alternatives such as exploring the use of grid or cloud computing, the VO, or using expertise and facilities in data centres and existing projects and sharing facilities are not (yet) being considered.

Acknowledgements

I would like to thank all the participants at this workshop for a lively and interesting discussion, and particularly JJ Kavelaars for helping conduct the workshop.