

Integrating The MACHO Data-set with the Virtual Observatory.

Jonathan Smillie

Supercomputer Facility
Australian National University

Jon.Smillie@anu.edu.au

Roberta Allsman

National Optical Astronomy Observatories
Tucson, Arizona

robyn@noao.edu

Stuart Hungerford

Supercomputer Facility
Australian National University

Stuart.Hungerford@anu.edu.au

Jason Ozolins

Supercomputer Facility
Australian National University

Jason.Ozolins@anu.edu.au

Abstract

A Virtual Observatory (VO) provides enhanced distributed access to online astronomical data-sets, and frequently incorporates data processing and visualisation facilities. Almost all critical aspects of data-grid research and technology are encountered when developing a Virtual Observatory. Issues include interoperability of data-sets and processing software, metadata format standardisation, and resolution of ontological issues. Members of the ANU Supercomputer Facility (ANUSF) are undertaking research and implementation projects in several key VO areas. In this paper we discuss the data-sets and astronomy projects with which ANUSF is involved, and outline the VO projects we are undertaking based on these resources.

Keywords: Data-grids, Virtual Observatory

Copyright © 2005, Australian Computer Society, Inc. This paper appeared at the Australasian Workshop on Grid Computing and e-Research, Newcastle, Australia. Conferences in Research and Practice in Information Technology, Vol. 44. Editors, Paul Coddington and Andrew Wendelborn. Reproduction for academic, not-for profit purposes permitted provided this text is included.

1. Introduction

With the widespread adoption of CCD camera technology in recent decades, astronomy has become one of the most prolific generators of digitised data in any

field of scientific research. Large astronomical survey projects routinely generate data-sets on the order of tens of terabytes in size, and worldwide astronomical data volumes are doubling every year. The presence of these rich online data resources has been a key motivator behind the rapidly expanding interest in the Virtual Observatory (VO). The VO movement aims to extract maximum value from online astronomical data-sets by exploiting the leverage provided by emerging grid technologies, data-mining tools, and advanced visualisation techniques.

The Australian National University (ANU) (ANU 2004) is well positioned to contribute to Australian VO efforts. The Australian Partnership for Advanced Computing (APAC) National Facility (APACNF/NF) (APACNF 2004), which is operated by the ANU Supercomputer Facility (ANUSF) (ANUSF 2004), hosts the Massive Compact Halo Objects (MACHO) project (MACHO 2004) data-set. By applying APAC resources, GrangeNet project (GrangeNet2004) connectivity, and the expertise of the ANUSF Grid group (ANUSF 2004) personnel in collaboration with VO projects around the world, the ANU will be a key contributor to the realisation of an Australian VO.

2. Virtual Observatories around the world

2.1 International VO Projects

Currently, international VO projects include the European Astrophysical Virtual Observatory (AVO) (AVO 2004), the UK AstroGrid project (AstroGrid 2004), and the US National Virtual Observatory (NVO) (NVO 2004). These projects cooperate via the International Virtual Observatory Alliance (IVOA) (IVOA 2004).

International projects are characterised by plans to develop comprehensive national VO infrastructures via centralised efforts allocated millions of dollars of funding and employing from tens to hundreds of dedicated staff. Umbrella organisations such as the IVOA are coordinating substantial efforts addressing issues such as standards and interoperability.

2.2 The Virtual Observatory in Australia

Virtual Observatory implementation in Australia is being cooperatively undertaken by university and research institutions under the banner of the Australian Virtual Observatory (Aus-VO) (Aus-VO 2004). Several institutions have formed a partnership as Australian e-Astronomy (AeA) (AeA 2004) and have been successful in obtaining funding support for their VO activities. Aus-VO is also an active participant in the IVOA.

The model of operation adopted in Australia is the focusing of efforts around key data resources, such as the MACHO data-set, CSIRO's Australia Telescope Compact Array (ATCA) (ATCA 2004), and the upcoming Stromlo Southern Sky Survey (4S) (4S 2004), to which value is added by implementing VO compliant interfaces and processing tools around the individual data repository. Thus the Australian VO will be built in a bottom up fashion, via the cooperative efforts of individual institutions and research projects.

3. Data Grid issues facing Virtual Observatories

Many key data-grid issues are encountered when developing Virtual Observatory facilities. Of primary importance are issues of interoperability and standardisation: standardisation of data and metadata formats and standard data exchange protocols. Maintenance of data provenance information and resolution of ontological issues are also central concerns.

Much progress towards these necessary standards has been made within Astronomy, due to a long history of cooperation and collaboration within the discipline. The Flexible Image Transport System (FITS) (FITS 2004) has been a de-facto data storage and exchange format within Astronomy for many years. Metadata issues are being addressed by the emerging VOTable format (VOTable 2004), an XML based standard for exchange of tabular astronomical data. Ontological resolution is being facilitated by systems such as the Uniform Content Descriptors (UCD) (UCD 2004) developed by the Strasbourg Astronomical Data Centre (CDS) (CDS 2004). With significant progress already made towards

resolution of fundamental data-grid issues, the Astronomical community is well placed to realise many Virtual Observatory objectives.

4. The ANU Supercomputer Facility Grid Group

4.1 ANUSF Project Objectives

The ANU Supercomputer Facility hosts a Grid group which has been tasked with investigating and developing a range of advanced networking, data, visualisation and grid activities and infrastructures. This group is enabled through funding and resources contributed by bodies including APAC and the GrangeNet organisation. A key area of this group's activity is the investigation and development of production data-grid installations. To this end, Australian Virtual Observatory efforts have been identified as an area of significant interest and importance.

4.2 ANUSF VO Collaborations and Resources

The primary Astronomy resource available to the ANUSF Grid group is the MACHO data-set, hosted by the ANUSF Mass Data Storage System (MDSS) (MDSS 2004). This data-set comprises approximately ten terabytes of southern sky image data, collected over ten years, along with reduced time-series light-curve data for some twelve million southern sky stars. The objective in collecting this data was discovery of dark matter objects through detection of gravitational microlensing events, however the nature of this data makes it an excellent resource for discovery and study of many variable star phenomena. This data-set is now released for public access, and is highly valued and intensely utilised by astronomers worldwide.

The ANUSF Grid group is working on several projects intended to enhance the availability and accessibility of this data via application of grid and related technologies. By VO enabling this data-set, it will become a cornerstone of the Australian VO infrastructure, and an internationally valued VO resource.

Along with work involving the MACHO data-set, the ANUSF Grid group is participating in a variety of national VO initiatives and collaborations. Experience gained in these activities will be advantageous in upcoming VO projects, such as servicing the data needs of the Stromlo Southern Sky Survey (4S) project.

5. Current ANUSF VO Activities

5.1 MACHO Metadata Conversion

The MACHO project data was originally collected via the dedicated computer system which controlled the MACHO project telescope, the "Great Melbourne" fifty inch (50") telescope at Mt Stromlo Observatory. Calibrated image data was stored in FITS format, and photometry data was written as binary time series light-curves. MACHO metadata was stored in a custom database and accessed with utilities developed as part of the MACHO data collection software suite.

As part of ANUSF project efforts to make the MACHO data available via VO compliant web interfaces, the entire MACHO metadata database has been ingested and converted to VOTable compliant XML (XML 2004). Of particular concern in this process was the clarification of semantic issues within the metadata. These issues were resolved through consultation with MACHO project astronomers, and UCD format has been used to express the results. This XML repository is intended to be the legacy form of the MACHO metadata. The intention now is to develop thin wrappers or registries which will present this XML in VOTable format for the benefit of VO compliant search and analysis tools.

A sample MACHO image metadata record in VOTable format is presented in the Appendix to this paper.

5.2 MACHO Data Web Interface

Prior to development of the VOTable format, and the emergence of Aus-VO, a comprehensive, searchable web interface and delivery system was developed for the MACHO data-set by ANUSF staff (Allsman 2001) (MACHO WI 2004). This system uses a Z39.50 (Z39.50 2004) metadata database constructed from ingested original MACHO metadata. This data can be searched via an integrated search facility, and data corresponding to search results is delivered via FTP (FTP 2004) to the client browser.

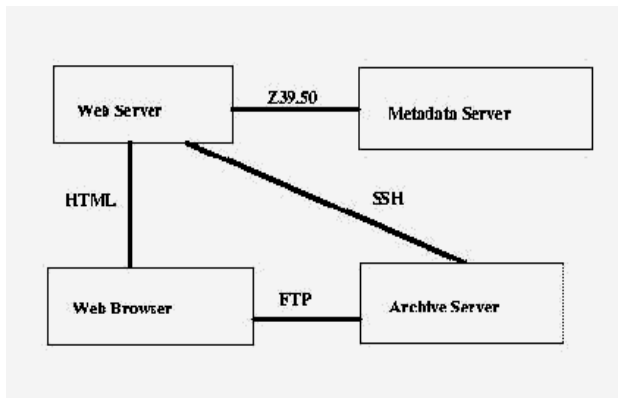


Figure 1 MACHO Web Interface structure

The interface to this system is implemented within an Apache HTTP server (Apache 2004) running on the ANUSF MDSS. Queries originating from the client browser are translated into appropriate syntax and submitted to a Zebra (Zebra 2004) server which provides the interface to the Z39.50 metadata. Query results are then used to formulate an FTP request to the FTP server running on the ANUSF MDSS, enabling archive data to be returned to the client browser. The interface is used on a regular and frequent basis by astronomers worldwide to access and retrieve MACHO images and light-curve data. It is available at:

<http://wwwmacho.anu.edu.au/Data/MachoData.html>

MACHO Project: Variable Star Catalog Retrieval

Submit Reset Startups Help Error Recovery

Click options below to expose or remove input fields.

Classification none

Variability Index

Sample: Units [HMS/DMS/arcmin] Equinox [J2000]
RA [5:1:15.2] Dec [-69:25:59.5] Search Radius [1]

Location by_RA_and_Dec

Sample: Units [radians] Equinox [J2000]
RA [1.31447] Dec [-.21183] Search Radius [.0003]

Sample: Units [degrees] Equinox [J2000]
RA [75] Dec [-69] Search Radius [.1]

Units HMS/DMS/arcmin Equinox J2000

RA Dec Search Radius

Average Magnitude none

Average Amplitude none

Average Period none

Review and edit search query n1 Display at most 20 records per page.

Figure 2 MACHO Web Interface example

5.3 Conversion of MACHO Image Data to a World Coordinate System

A key issue in VO implementation is data format interoperability and standardisation. An emerging requirement for astronomical data VO compliance is adherence to a World Coordinate System (WCS) (WCS 2004). Observations which are calibrated to a WCS have had their precise location on the sky mapped to a canonical position based on the locations of standard stars. Thus the region of the sky captured by any image can be unambiguously determined and inaccuracies in telescope pointing or image location metadata can be accounted for.

Conversion to WCS format involves the automated identification and location of key WCS reference stars in an image. This processing enables a precise WCS calibrated position to be assigned to the observation. In the case of images encoded in the widely used FITS format, such as the MACHO images, this WCS information is added to the FITS file header.

Staff at ANUSF have converted the entire MACHO image data-set to WCS format. This process required automated processing of each MACHO image FITS file and the writing into the image header of calibrated WCS locators. As the MACHO image archive contains approximately one hundred thousand images, all of which comprise approximately seven terabytes of data, a sophisticated pipeline processing system was constructed in order to complete this task in a reliable and expedient manner.

The foundation of this pipeline is a conversion utility developed by astronomers at Mt Stromlo Observatory using the Perl (Perl 2004) scripting language, the IRAF (IRAF 2004) image processing tool, and the US Naval Observatory UCAC2 star catalogue (UCAC2 2004). This utility is specific to MACHO image files, and calculates the necessary WCS information which is inserted into the image's FITS header. Coupled with this utility is Python

(Python 2004) code which queries the MACHO XML metadata repository and updates the metadata contained in the FITS header as required to ensure consistency. This Python code uses the PyFITS library (PyFITS 2004) to manipulate the FITS headers. The final processing step for each image is the packaging of the processed image in a multi-extension FITS (MEF) (MEF 2004) file, via a Python utility utilising the IRAF tool.

This per-image processing is wrapped in a Perl and shell script infrastructure which enabled the entire MACHO dataset to be processed on the APAC NF Linux (Linux 2004) cluster, via the PBS (PBS 2004) scheduling system, within wall-time on the order of one calendar month. Processing time was largely determined by the rate at which data could be streamed to and from the MDSS and distributed to processing nodes. The large amount of data transferred relative to compute time in this pipeline meant that the pipeline required considerable tailoring to the compute-oriented batch environment of the NF Linux cluster, in which MDSS I/O must be done in separate batch jobs, and compute jobs are meant to be a relatively heavyweight unit of work scheduling.

This is an example and instructive prototype of automated processing facilities that may compliment the data dissemination aspects of any comprehensive VO. Lessons and techniques learned in this task will be adapted and applied to the data processing objectives which will accompany subsequent ANU VO initiatives.

In particular, as large data-sets become grid-enabled, computation resources may need to be made more demand-driven, and more tightly coupled to mass data storage, in order to suit a wider range of access methods than have traditionally been offered. As an example from the VO space, the Simple Image Access Protocol (SIAP) (SIAP 2004) allow clients to request on-the-fly transformations on existing observation data. A balance must be struck between the twin goals of returning timely results to clients and making efficient use of processing hardware.

5.4 Stromlo Southern Sky Survey

The Stromlo Southern Sky Survey (4S) was scheduled to commence data acquisition in mid-2003 using the Mt Stromlo "Great Melbourne" fifty inch telescope, however the destruction of this telescope in the Canberra bushfire of the eighteenth of January 2003 (Mt Stromlo 2004) has resulted in a delay whilst an alternative instrument is prepared. This project will map the entire southern sky in six optical wavelengths using the Sloan Digital Sky Survey (SDSS) (SDSS 2004) filter set. It will run for approximately five years and generate approximately twenty-five terabytes of image data, cataloguing 10^9 objects in total. This data will include the necessary photometric and astrometric calibration to make it a valuable and fundamental VO resource.

The intention is to capitalise on the lessons learned and systems developed as part of the VO-enabling of the MACHO project. 4S data will be streamed to the ANUSF MDSS from where it will be made immediately available. Web interfaces and data delivery systems will be built

based on the existing MACHO project interfaces and VOTable demonstration installations. ANUSF staff will contribute to all aspects of 4S project data archiving, exploiting experience gained with the MACHO project and associated VO activities.

6. Conclusion and Future Directions

The Australian National University is well positioned to remain a central participant in Australian Virtual Observatory activities. The leveraging of grid technologies to the highly valued MACHO data set, and participation in a range of VO demonstrators and test-beds, will enable the ANUSF group to make an active contribution to this field, whilst developing and enhancing the expertise necessary to take the Australian VO into the future. The MACHO data-set will remain a heavily utilised resource into the foreseeable future, and will be a cornerstone resource of the Australian Virtual Observatory. Once the Stromlo Southern Sky Survey comes online, with the help of ANUSF expertise the resulting data-set will be a foundation of the Australian Virtual Observatory, and a priceless test environment for emerging VO and data grid technologies in this country.

7. Acknowledgements

The ANUSF Grid group is supported by funding and resources provided by the Australian Partnership for Advanced Computing (APAC), the GrangeNet project, and the Australian National University.

The code which performs WCS calculations for each MACHO image was supplied by Dr Brian Schmidt of Mt Stromlo observatory, and his ongoing assistance in this work is gratefully acknowledged.

8. Appendix

Sample VOTable format MACHO image metadata record:

```
<?xml version="1.0"?>
```

```
<!DOCTYPE VOTABLE SYSTEM "http://us-vo.org/xml/VOTable.dtd">
```

```
<VOTABLE version="1.0">
```

```
<DEFINITIONS>
```

```
<COOSYS ID="J2000" equinox="2000." epoch="2000." system="eq_FK5"/>
```

```
</DEFINITIONS>
```

```
<RESOURCE ID="18001" type="meta">
```

```
<PARAM name="obs_number" ucd="ID_NUMBER" value="18001"/>
```

```
<PARAM name="obs_date" ucd="TIME_MISC" unit="string" value="May 15 1994 14:04:16 GMT"/>
```

```
<PARAM name="obs_date_julian" ucd="TIME_DATE" unit="number" value="49487.6"/>
```

```
<PARAM name="obs_time" ucd="TIME_MISC" unit="number"
```

```
value="19940515"/>
```

```
<PARAM name="field_id" ucd="ID_FIELD" unit="number" value="108"/>
```

```
<PARAM name="domain_id" ucd="ID_IDENTIFIER" unit="string" value="Bulge1"/>
```

```
<PARAM name="RA" ucd="POS_EQ_RA_MAIN" unit="h:m:s" value="18:1:20.7"/>
```

```
<PARAM name="DEC" ucd="POS_EQ_DEC_MAIN" unit="d:m:s" value="-28:17:38"/>
```

```
<PARAM name="tel_RA" ucd="POS_EQ_RA_MAIN" unit="h:m:s" value="18:1:20.4"/>
```

```
<PARAM name="tel_DEC" ucd="POS_EQ_DEC_MAIN" unit="d:m:s" value="-28:17:39"/>
```

```
<PARAM name="seeing" ucd="INST_SEEING" unit="arcsec" value="3.82714"/>
```

<PARAM name="image_av_sky_value" </VOTABLE>
ucd="INST_SKY_LEVEL"
unit="ct"
value="2168"/>

<PARAM name="airmass"
ucd="PHOT_ATM_AIRMASS"
unit="number"
value="1.18024"/>

<PARAM name="exp_time"
ucd="TIME_EXPTIME"
unit="s"
value="150"/>

<PARAM name="refraction"
ucd="PHYS_REFRACT_INDX"
unit="number"
value="58.2"/>

<PARAM name="parallactic_angle"
ucd="POS_PAR_ANGLE"
unit="deg"
value="4.32529"/>

<PARAM name="focus"
ucd="INST_CALIB_PARAM"
unit="number"
value="-0.26"/>

<LINK ID="18001">
Image/O_18/Obs_18001.tar
</LINK>

</RESOURCE>

9. References

- 4S: The Stromlo Southern Sky Survey. http://msowww.anu.edu.au/news/archive/2003/01_jan/ Accessed 10 Sep 2004.
- AeA: Australian eAstronomy. ARC LIEF grant proposal 2004.
- Allsman, R. (2001): "Simplifying the Web User's Interface to Massive Data Sets" Eighteenth IEEE Symposium on Mass Storage Systems in cooperation with the Ninth NASA Goddard Conference on Mass Storage Systems and Technologies April 17-20, 2001, Hyatt Regency Islandia, San Diego pp 175-190. <http://storageconference.org/2001/2001CD/14/allsma.pdf>
- ANU: The Australian National University. <http://www.anu.edu.au/> Accessed 10 Sep 2004.
- ANUSF: The Australian National University Supercomputer Facility. <http://anusf.anu.edu.au/> Accessed 10 Sep 2004.
- APAC NF: The Australian Partnership for Advanced Computing National Facility. <http://nf.apac.edu.au/> Accessed 10 Sep 2004.
- Apache: The Apache HTTP Server project. <http://httpd.apache.org/> Accessed 10 Sep 2004.
- AstroGrid: The AstroGrid project. <http://www.astrogrid.org/> Accessed 10 Sep 2004.
- ATCA: The Australia Telescope Compact Array. <http://www.narrabri.atnf.csiro.au/> Accessed 10 Sep 2004.
- Aus-VO: The Australian Virtual Observatory. <http://www.aus-vo.org/> Accessed 10 Sep 2004.
- AVO: The Astrophysical Virtual Observatory. <http://www.euro-vo.org/> Accessed 10 Sep 2004.
- CDS: Strasbourg Astronomical Data Centre. <http://cdsweb.u-strasbg.fr/> Accessed 10 Sep 2004.
- FITS: Flexible Image Transport System. http://archive.stsci.edu/fits/fits_standard/ Accessed 10 Sep 2004.
- FTP: File Transfer Protocol. <http://www.w3.org/protocols/rfc959/> Accessed 10 Sep 2004.
- GrangeNet: The GrangeNet project. <http://www.grangenet.net/> Accessed 10 Sep 2004.
- IRAF: Image Reduction and Analysis Facility. <http://iraf.noao.edu/> Accessed 10 Sep 2004.
- IVOA: The International Virtual Observatory Alliance. <http://www.ivoa.net/> Accessed 10 Sep 2004.
- Linux: Linux operating system. <http://www.linux.org/> Accessed 10 Sep 2004.
- MACHO: The Massive Compact halo Objects Project. <http://www.macho.anu.edu.au/> Accessed 10 Sep 2004.
- MACHO WI: MACHO data web interface. <http://www.macho.anu.edu.au/Data/MachoData.html> Accessed 10 Sep 2004
- MDSS: ANUSF Mass Data Storage System <http://anusf.anu.edu.au/MDSS/> Accessed 10 Sep 2004.
- MEF: Multi-extension FITS format. <http://fits.gsfc.nasa.gov/> Accessed 10 Sep 2004.
- Mt Stromlo: <http://msowww.anu.edu.au/fire/> Accessed 10 Sep 2004.
- NVO: The National Virtual Observatory. <http://www.us-vo.org/> Accessed 10 Sep 2004.
- PBS: The Portable Batch System. <http://nf.apac.edu.au/facilities/userguide/> Accessed 10 Sep 2004.
- Perl: Perl. <http://www.perl.com/> Accessed 10 Sep 2004.
- PyFITS: PyFITS Python/FITS tools. http://www.stsci.edu/resources/software_hardware/pyfits Accessed 10 Sep 2004.
- Python: Python. <http://www.python.org/> Accessed 10 Sep 2004.
- SDSS: Sloan Digital Sky Survey. <http://www.sdss.org/> Accessed 10 Sep 2004.
- SIAP: Simple Image Access Protocol Specification, Version 1.0 [IVOA WG Working Draft] <http://www.ivoa.net/Documents/WD/SIA/sia-20040524.html>
- UCAC2: United States Naval Observatory CCD Astrograph Catalog. <http://ad.usno.navy.mil/ucac/> Accessed 10 Sep 2004.
- UCD: Uniform Content Descriptors. <http://cdsweb.u-strasbg.fr/doc/UCD.htm> Accessed 10 Sep 2004.
- VOTable: VO-Table format version 1. <http://vizier.u-strasbg.fr/doc/VOTable/> Accessed 10 Sep 2004.
- WCS: World Coordinate System. <http://www.harvard.edu/software/wcstools/>
- XML: eXtensible Markup Language. <http://www.xml.org> Accessed 10 Sep 2004.

Z39.50: National Information Standards Organisation
Z39.50 Information Retrieval Protocol.
<http://www.niso.org/z39.50/z3950.html>
Accessed 10 Sep 2004.

Zebra: GNU Zebra. <http://www.zebra.org/> Accessed
10 Sep 2004.