The Astro-Wise System: A Federated Information Accumulator for Astronomy

by Edwin A.Valentijn and Gijs Verdoes Kleijn

The progress of astronomy is about to hit a wall in terms of the processing, mining and interpretation of huge datasets. The Astro-Wise consortium has designed and implemented a fully scalable and distributed information system to overcome this problem for wide-field imaging. The same principles can be applied to other sciences.

Much of modern research involves the accumulation of huge amounts of digitized data. The analysis of this data by distributed communities represents a significant challenge to project management and ICT implementation, and is relevant to fields as diverse as biology, physics, astronomy, economics and cultural heritage projects. Furthermore, the projects are often global efforts requiring collaborators in many places to share, validate and combine processed data and derived results. It is therefore necessary to develop more efficient data lineage, mining and analysis systems to allow researchers to search intelligently through previously unmanageable volumes of data.

The Astro-Wise consortium has developed an information system to meet these challenges for wide-field imaging in astronomy. The Astro-Wise consortium is a partnership between OmegaCEN-NOVA/Kapteyn Institute (Groningen, The Netherlands; coordinator), Osservatorio Astronomico di Capodimonte (Naples, Italy), Terapix at IAP (Paris, France), ESO, Universitäts-Sternwarte & Max-Planck Institut für Extraterrestrische Physik (Munich, Germany).

Large data projects in high-energy physics, space missions and astronomy typically push data through various platforms in an irreversible way (eg a TIER node setting). In such a situation, the end user has little or no influence on what happens upstream. This 'classical' paradigm is characterized by fixed 'releases' of homogeneous, well-documented data products. In contrast, the Astro-Wise system allows the end user to trace the data product, following all its dependencies up to the raw observational data and, if necessary, to re-derive the result with better calibration data and/or improved methods.

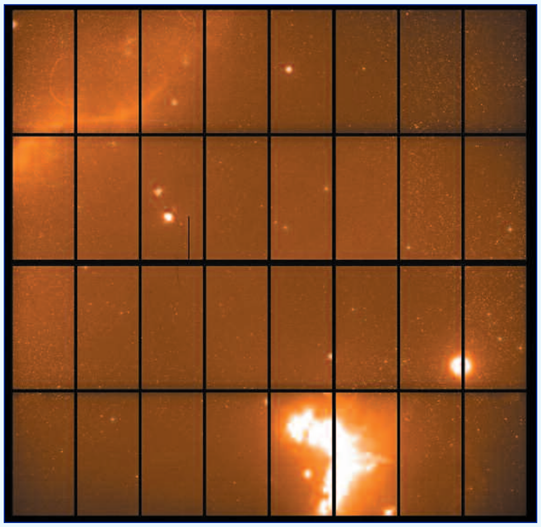
This improvement is achieved by:

- emphasis on project management; enforcing a global data acquisition and processing model, while retaining flexibility
- translating the data model to an object model, with full registration of all dependencies
- storing all I/O of the project in a single, distributed database, containing all metadata describing the bulk data (eg images) and derived results in catalogue form (eg lists of celestial sources).
- connecting to the database a federated file server that stores hundreds of Terabytes of bulk data
- an own compute-GRID which sends jobs (including clients) to single nodes or parallel clusters, which then request data from the distributed database.

The database with all metadata and catalogues provides the infrastructure to develop tools for a variety of purposes. These include rapid trend analysis of data, complex queries and fast hunting for 'needles in the haystack' of Terabyte-sized catalogues. Thus, the system provides the user with fully integrated, transparent access to all stages of the data processing and thereby allows the data to be reprocessed and the system to be improved and expanded.

For a given project/instrument, the system initially starts in a naive, 'quick look' mode, which gradually improves as various researchers add refined information to the system under the supervision of project leaders. Approved calibration modifications automatically become public, beyond the project boundaries. A mechanism for quality control is implemented which allows for changes due to one of:

- true physical changes of parameter values
- improvements in encoded methods, or
- improved insight in either of these.



A 256 Mega pixel test image of the OmegaCAM instrument, which consists of 32 eight Megapixel CCDs.

The core of the system exploits three properties in database environment. First, we apply the principle of inheritance using Object Oriented Programming (Python), where all Astro-Wise objects inherit key properties for database access, such as persistency of attributes. Second, the linking (associations or references) between instances of objects in the database is completely maintained, and for each bit of information, it is possible to trace those bits of information that were used to obtain it. Third, each step, and the inputs used for it, is kept within the system. The database grows constantly through the addition of new information or improvements made to existing information.

All system components are distributed over Europe, enabling research groups to collaborate on shared projects. Knowledge added by one group is immediately accessible by others via a Web portal, which includes data viewing, quality labelling and compute-services (see links). Currently, researchers use the Astro-Wise system with 10 Tbyte of astronomical images. Hundreds of Tbytes of data will start entering the system when the OmegaCAM panoramic camera starts operations in Chile. This camera is dedicated to various large surveys using the Astro-Wise system.

Astro-Wise coordinator OmegaCEN-NOVA is collaborating with the LOFAR consortium and CWI to explore usage of the Astro-Wise system for LOFAR, the next generation Low Frequency Array of radio telescopes, which is being built in the Netherlands and Germany. Astro-Wise can also be applied to other fields of science. The object-oriented use of the database allows for classes of objects dealing with arbitrary forms of digitized observational data. Scans of cultural heritage, DNA sequences, data from high-energy particle collisions or financial markets can be processed using similar principles to the images of the sky.

Links:
http://www.astro-wise.org
http://www.astro-wise.org/portal
http://www.astro.rug.nl/~omegacam
http://www.astro.rug.nl/~omegacen
http://www.lofar.nl

Please contact:
Edwin Valentijn, Astro-Wise Consortium
OmegaCEN-NOVA/Kapteyn Astronomical Institute, Groningen, The Netherlands
Tel: +31 50 3634011
E-mail: valentyn@astro.rug.nl

AstroGrid — Part of the European Virtual Observatory

by Peter M Allan

AstroGrid is the UK's implementation of the concept of a virtual observatory - being able to get at all the world's astronomical data directly from your desktop computer. It has been developed over the last 5 years with contributions from the Rutherford Appleton Laboratory and the universities of Edinburgh, Leicester, Cambridge, University College London (Mullard Space Science Laboratory), Manchester (Jodrell Bank Observatory), Queen's University Belfast, Bristol, Exeter, Portsmouth and Leeds.

The underlying concept is to provide data and computational services as grid services, such that a distributed data grid is built. The focus is on the provision of access to data, rather than on access to supercomputing power, although that can be one of the services offered. Within the UK, AstroGrid provides access to the large astronomical data resources held at the first six of the institutions listed above. If it did just this, AstroGrid would be useful, but not revolutionary. In fact, AstroGrid is an active participant in the International Virtual Observatory Alliance, a group dedicated to defining a truly international set of standards for grid-enabling access to astronomical data world wide. At the European level, AstroGrid is an active partner in Euro-VO, and leads the VO-Tech part of this Framework 6 project.

There are several virtual observatory projects around the world. AstroGrid has taken the route of deliberately deciding what astronomers really needed to make a major step forwards in their ability to analyse their data, and to build the infrastructure to do this. Some of the infrastructure has been a challenge to design and build, but we now have a system that can be used by astronomers in earnest. It is starting to be used to do real science.

The fundamental architecture behind AstroGrid consists of a set of web services with a workflow system that makes use of these services. In order to process data using AstroGrid, an astronomer builds a workflow and then executes it. They do not need to explicitly get the data from a data archive; the web services handle that. They do not need to be concerned with the storage of intermediate data; that is held in an area called MySpace. Depending on what the astronomer wants to be the result of some data search and processing, there may or may not be a requirement to have final data returned to the user's desktop. A typical simple workflow would be to get data from A and B, processes them at C and store the results at D. The actual location of A,B,C and D are not important to the user, only the result is.

As an example of the ease with which data can be obtained, as a test I recently tried to get some optical data on the quasar 3C273 (a famous object to astronomers). With a total of about ten clicks of the mouse and typing “3C273” (the system knows its position on the sky) I was offered data from the Hubble Telescope, which I chose to download from the archive in the USA to MySpace. I could then display the images on my computer. This is a fairly simple example of what is possible, but it demonstrates that a wealth of data are only a few mouse clicks away. This makes it much easier than in the past to