ASTRONOMICAL IMAGES AND DATA MINING IN THE INTERNATIONAL VIRTUAL OBSERVATORY CONTEXT*

FABIO PASIAN†

INAF – Osservatorio Astronomico di Trieste, Via G.B.Tiepolo 11, Trieste, I-34143, Italy

MASSIMO BRESCIA

INAF – Osservatorio Astronomico di Capodimonte, Salita Moiariello 16, Napoli, I-80131, Italy

GIUSEPPE LONGO

Università degli Studi "Federico II", Dipartimento di Scienze Fisiche, Via Cinthia, Napoli, I-80126, Italy

In the past ten years, the concept of Virtual Observatory (VObs) has increasingly gained importance in the domain of astrophysics, as a way of seamlessly accessing data in different wavelength domains stored in digital archives. There are many reasons why the VObs is useful for the development of science: to monitor time variability of phenomena, to compare phenomena in different bands, to increase return for investment (by fostering data re-use for scientific, educational and outreach purposes), to perform statistical analysis and mining on large quantities of data.

The International Virtual Observatory Alliance (IVOA) has paved the way for the VObs to become a really useful tool for the scientific community, by promoting standards, by defining data interoperability methods, by fostering the needed coordination among data providers.

But the VObs is more than just archives and standards: it is also infrastructure, basic software tools, advanced applications, evolution of methods and techniques, cross-fertilization with other communities. Discovering information in wide-field images and mining large archives are key items towards the use of the VObs as a tool for developing science.

Data mining, or knowledge discovery in databases, while being the main methodology to extract the scientific information contained in Massive Data Sets (MDS), needs to tackle crucial problems since it has to orchestrate complex challenges posed by transparent

* This work is partially supported by INAF through the VObs.it initiative and by the European Commission through the EuroVO-AIDA and EuroVO-ICE projects, and is endorsed by the International Virtual Observatory Alliance.

† Former Chair of the International Virtual Observatory Alliance (2008-2010).

access to different computing environments, scalability of algorithms, reusability of resources. To achieve a leap forward for the progress of astrophysics in the data avalanche era, the community needs to implement an infrastructure capable of performing data access, processing and mining in a distributed but integrated context.

1. Introduction: astronomical archives

Astronomy in recent years has expanded drastically its observing and modeling capabilities, increasing proportionally its demands for computational power and data access efficiency.

To understand completely the implications of this statement, one should remember that Astronomy is mostly an observational science, where any "experiment" is a snapshot of the fraction of the Universe being observed, and cannot be repeated as such. Astronomy measures $I(\lambda)$, light intensity as a function of wavelength (or frequency, or energy). But, since most phenomena are in fact variable, the measured intensity is $I(\lambda,t)$ which is convolved with the transfer function of the complex system (telescope, instrument, sky) actually involved in the observation. It is therefore easy to understand why every single observation needs to be kept, obviously creating a data preservation issue.

The increasing complexity of instrumentation and the need to optimize observing procedures (e.g. the meteorological conditions may be unacceptable for one type of observation while is suitable for others) lead to a change in concept on modern observatories. Classical observing (the scientist located at the telescope is given the whole night to perform his/her observations) is replaced by flexible scheduling and service observing, i.e. the observations and the basic data handling are performed by observatory staff. In this case, the scientist owning the observations extracts the data from disk storage, written in the FITS standard data format, which has been used by the community since 1977. Data might have been already partially processed (i.e. calibrated, with removal of the instrument signature).

The scenario described above has a pretty obvious extension, from data storages to full-fledged archives. Observatory policies have been defined so as to allow the data to become public typically after one year from the date of observation. Data taken for a specific scientific purpose can therefore be re-used for different purposes.

This approach, progressively emerged in the past fifteen-twenty years, has allowed the possibility of developing new science. Examples are monitoring the time variability of phenomena, comparing phenomena in different energy bands (the so-called multi-wavelength astronomy), performing statistical analysis or data mining on large quantities of data. From the technical point of view, the

possibility of re-processing the raw data when a better knowledge of instrumental effects is achieved.

An important effect of archives, and of the consequent re-use of data, is increasing the return for the investment: since a second of observation on modern facilities is worth roughly 1 US$, and the over-subscription factor is on the average a factor of 5, the possibility of re-using top-class data to do science, or for educational and outreach purposes has also important effects at the level of the society.

But archiving is an activity that is essential to cope with the data avalanche that is hitting the astrophysics community. As a matter of fact, the cumulative photon collecting area of telescopes world-wide has increased by a factor of 8 in the past 25 years, and the detector technology has allowed the production of panoramic chips which have multiplied the data flow. To give an example, the archive of the European Southern Observatory (ESO), the leading European institution, has grown in size a factor of 10 in three years. And of course, there are many dozens of ground- and space-based telescopes world-wide, covering the full electromagnetic spectrum, and equipped with instruments providing data which are telescope and band-dependent.

2. The Virtual Observatory

2.1. *Data and information services in Astronomy*

Besides data, many other sources of information for the Astrophysical community can be identified and can be summarized as follows:

- Telescopes (ground- and space-based, covering the full electromagnetic spectrum) are managed by Observatories;
- Instruments (which are telescope and band dependent) are managed by Observatories and/or Consortia;
- Data analysis software is usually instrument-dependent: it is managed by Observatories, Consortia and individual Researchers;
- Computing is supported by dedicated service Centers, or Organizations that can often be Observatories;
- Archives are managed by Observatories and/ or Agencies;
- Publications are managed by Journals; there is however a single point of search and access, the ADS system (which is the main source of bibliographic information for the whole of the Astronomical community);
- Data curation, i.e. management of metadata (i.e. the description of data), tables and catalogues) is performed by Data curators;

- Public Outreach is managed by Observatories and/or Agencies.

2.2. *The Virtual Observatory: definitions*

The Virtual Observatory (VObs) is an innovative, still evolving, system to take advantage of astronomical data explosion (e.g., use statistical identification to perform multi-wavelength, multi-parameter analysis, to allow astronomers to interrogate multiple data centers in a seamless and transparent way and to utilize at best astronomical data, to permit remote computing and data analysis and to foster *new science*.

The VObs aims at achieving the goal of having *all astronomical databases inside the scientist's PC* just as in the Web all documents are easily retrievable inside one's own PC. The situation depicted in the above section is pretty complex, and the desire has been felt to unify the sources of information within a federated framework.

Achieving the ambitious goal of allowing the different sources of information to interoperate is not trivial: as a matter of fact, the current situation shows both positive and negative aspects.

On the positive side, one can consider that observational data are freely available *on-line* in general 1 year after observation; both data and catalogues are normally stored in *astronomical archives*. Results are published in academic journals, all *available on-line* with *pointers* to data archives; as mentioned above, there is one *single entry point for journals* (ADS). The processing and analysis software *maintained* Observatories/Archives is made available *on-line*.

However, the different astronomical archives have *widely different access/search interfaces and standards/conventions*; and they mainly serve raw data. The widely specialized, *analysis software* for the various sub-branches is in general *complex*, thus yielding a *steep learning curve*, but this is needed, since multi-wavelength is now the norm to produce science. Publication-Archive links often point to *raw, unprocessed data* and are not complete. Furthermore, object *metadata are not homogeneously defined*.

2.3. *The International Virtual Observatory Alliance*

All of these considerations require that the various players speak the same language. It is therefore necessary that common standards and protocols are defined and adopted within the whole community. This issue is tackled within the International Virtual Observatory Alliance (IVOA), which coordinates 19 world-wide national or super-national VObs initiatives (shown in Figure 1).

Figure 1. The members of the International Virtual Observatory Alliance are national or super-national VObs projects and initiatives.

The IVOA mission is *"to facilitate the international coordination and collaboration necessary for the development and deployment of the tools, systems and organizational structures necessary to enable the international utilization of astronomical archives as an integrated and interoperating virtual observatory"*. The work is carried out by means of teleconferences, "TWiki" pages, and bi-annual meetings (the last one was carried out in Nara, Japan, in October 2010, the next will be in Naples, Italy, in May 2011).

The IVOA goals are the standardization of data, metadata and software, the definition of data interoperability methods, and the production of lists of available data and computing services (each provided by the individual VObs projects).

IVOA is an initiative that is based on the concept of collaborating and sharing. The absence of specific funding does not allow to force any decisions, which are basically taken unanimously. This practice is in principle good, since it builds consensus in the community. However, competition at the personal and/or project level brings as a consequence a slow convergence in the definition of standards.

From the organizational point of view, the IVOA structure is composed of an Executive Board which includes the representatives from all national VObs projects**,** a dozen of Working and Interest Groups coordinated by a Technical

Coordination Group. There is an estimated total of 300-400 individuals involved in IVOA development activities.

Among the various groups, it is worth mentioning the recent creation (2009) of a Data Mining and KDD Interest Group.

2.4. *Status of the Virtual Observatory*

The VObs is progressively evolving to a truly operational phase [1]. Its paradigm to utilize multiple archives of astronomical data in an interoperating, integrated and logically centralized way, allows to ‘‘observe a virtual sky’’ by position, wavelength and time. Not only data actually observed are included in this concept: theoretical and diagnostic are being included. VObs represents a new type of a scientific organization for the era of information abundance:

- it is inherently distributed, and web-centric;
- it is fundamentally based on a rapidly developing technology;
- it transcends the traditional boundaries between different wavelength regimes, agency domains;
- it has an unusually broad range of constituents and interfaces and
- it is inherently multidisciplinary.

The international VObs has opened a new frontier to astronomy. In fact, by making available at the click of a mouse an unprecedented wealth of data and by implementing common standards and procedures, the VObs allows a new generation of scientists to tackle complex problems, which were almost unthinkable only a decade ago. Astronomers may now access a ‘‘virtual’’ parameter space of increasing complexity (hundreds or thousands features measured per object) and size (billions of objects).

2.5. *Extending the Virtual Observatory concept*

In the last couple of years, the VObs received support at an International level: positive statements on the importance of the initiative were made by the Organization for Economic Co-operation and Development (OECD), by the European Strategy Forum on Research Infrastructures (ESFRI), and by ASTRONET, the European initiative aiming at enhancing the coordination and cooperation between national funding and research management organizations in Europe who are responsible for astronomical research. In the US, the VObs initiatives are supported by NASA and NSF. This recognition and support is important since, to make sense, the Virtual Observatory needs to be an international effort.

In this framework, an important comment was made in the 2006 ESFRI document, where the Virtual Observatory was indicated as the example to follow to create an international federation of libraries. As a matter of fact, the VObs concepts can be, and are being, re-used in different domains. Among the various interested communities, it is worth mentioning that the planetologists have their own International Planetary Data Alliance (IPDA), the Solar community has its own VObs and DataGrid, and the High-Energy Physics (HEP) community is interested in the VObs mechanisms.

3. Data Mining and the Virtual Observatory

As discussed in the previous section, up to now the VObs has mainly been a mechanism for data access, retrieval and manipulation. However, to obtain results, scientists need to deal with computational challenges related to the handling, processing and modeling of large quantities of data.

3.1. *Massive data processing in astrophysics*

Processing of huge quantities of data (large detectors, mosaics, images with high time resolution) is typical of the optical and solar communities. The amount of computations needed to process the data is impressive, but often ‘‘embarrassingly parallel’’ since based on local operators, with a coarse grained level of parallelism. In such cases, the ‘‘memory footprint’’ of the applications allows one to subdivide data in chunks, so as to fit the RAM available on the individual CPUs and to have each CPU to perform a single processing unit.

There are many software resources (legacy applications, libraries, etc.) widely used in past and current projects and experiments that form a wide range of tools, services and facilities used by key astronomical projects and environments, and by the individual astronomer.

Which kind of resources is necessary to tackle the processing and analysis of such large quantities of data in astrophysics? In most cases ‘‘distributed supercomputers’’, i.e. a local cluster of PCs such as a Beowulf machine, or a set of PCs distributed over the network, can be an effective solution to the problem. In this case, the Grid paradigm can be considered to be an important step forward in the provision of the computing power needed to tackle the new challenges.

As seen earlier, the concept of ‘‘distributed archives’’ is already familiar to the average astrophysicist, and a leap forward has been made by the VObs which was capable of organizing the data repositories to allow efficient, transparent and uniform access. In more than a sense, the VObs is an extension of the classical

Computational Grid; it fits perfectly the Data Grid concept, being based on storage and processing systems, and metadata and communications management services.

3.2. *Grid-based data mining: the DAME system*

As underlined above, most of the implementation efforts for the VObs has concerned the storage, standardization and interoperability of the data together with the computational infrastructures. In particular it has focused on the realization of the low level tools and on the definition of standards.

It is important to extend this fundamental target by integrating it in an infrastructure, joining service-oriented software and Grid resource-oriented hardware paradigm,1 including the implementation of advanced tools for MDS exploration, Soft Computing, Data Mining (DM) and Knowledge Discovery in Databases (KDD).

Moreover, DM services often run synchronously. This means basically that they execute jobs during a single HTTP transaction. This might be considered useful and simple, but it does not scale well when it is applied to long-run tasks. Typical long-running activities are the following:

- any archive query traversing a massive DB table;
- a data-mining job running from a batch (sequential) queue and
- a pipeline workflow with several computing-intensive steps, applied sequentially for many (and massive) data sets.

In any of these cases, the system is stressed if the activity lasts longer than a few minutes and becomes unreasonably fragile if it lasts longer than a few hours. With synchronous operations, all the entities in the chain of command (client, workflow engine, broker, processing services) must remain up for the duration of the activity. If any component goes down or stops then the context of the activity is lost and must be restarted.

To overcome this limitation, a new-generation DM system called DAME [2] has been devised. One of its main design strategies is to permit asynchronous access to the infrastructure tools, allowing running of activity jobs and processes outside the scope of any particular web-service operation and without depending on the user connection status.

The user, via client web applications, can asynchronously find out the state of the activity, has the possibility to keep track of his jobs by recovering related information (partial/complete results) without having the need to maintain open the communication socket. Moreover, the system is able to automatically

perform a sort of garbage collection for cleaning up resources, swap areas and temporary system tools used during the activity run phase.

Furthermore, as it will be discussed in what follows, the DAME design takes into account the fact that the average scientist cannot and/or does not want to become an expert also in Computer Science. In most cases he/she already possesses individual algorithms for data processing and analysis and has implemented private routines/pipelines to solve specific problems. These tools, however, often are not scalable to distributed computing environments. DAME aims at providing a user friendly web-based tool to encapsulate one's own algorithm/procedure into the package, automatically formatted to follow internal programming standards.

The natural computing environment for MDS processing is a distributed infrastructure (Grid/Cloud), but again, we need to define standards in the development of higher level interfaces, in order to:

- isolate end user (astronomer) from technical details of VObs and Grid/Cloud use and configuration;
- make it easier to combine existing services and resources into experiments.

Data mining is usually conceived as an application (deterministic/stochastic algorithm) to extract unknown information from noisy data. This is basically true but in some way it is very reductive with respect to the wide range covered by mining concept domains. More precisely, in DAME, data mining is intended as techniques of exploration on data, based on the combination between parameter space filtering, machine learning and soft computing techniques associated with a functional domain. The functional domain term arises from the conceptual taxonomy of research modes applicable on data [8,9,10].

Dimensional reduction, classification, regression, prediction, clustering, filtering are examples of functionalities belonging to the data mining conceptual domain, in which the various methods (models and algorithms) can be applied to explore data under a particular aspect, related to the associated functionality scope.

The analytical methods based partially on statistical random choices (crossover/mutation) and on knowledge experience acquired (supervised and/or unsupervised adaptive learning) could realistically achieve the discovery of hidden laws behind focused phenomena, often based on nature laws, therefore the simplest.

During the R&D phase of our project, aimed at defining and characterizing rules, targets, ontology, semantics and syntax standards, a functional breakdown structure was derived. It provides a taxonomy between possible data exploration

modes, made available by our infrastructure as data mining experiment typology (use case).

4. One step further: integration of the VObs, data processing and data mining facilities

To offer scientists a useful service, all of the components of the informatics infrastructure need to be thought as integrated, or at least fully interoperable. In other words, the various infrastructure components (applications, computing, data) should interact seamlessly exchanging information, and be based on a strong underlying network component.

The "big science" challenges in astrophysics call for an expansion of the computing infrastructures – and of network and data management infrastructures as well. The astronomical community is interested in using, and participating in defining, competitive computing infrastructures so to perform better research.

But there is furthermore the need to integrate network, data and computing infrastructures, or at least to let them interoperate. To fulfill this requirement applications, computing power, data repositories and databases holding metadata or catalogues should be accessed as a single utility. It needs to have a Virtual Observatory interface, which is more familiar to the community [7].

4.1. *VObs-Grid interoperability*

The documents of the ASTRONET initiative [3,4] have defined the Virtual Observatory as *"the e-infrastructure project in Europe"*. As a matter of fact, an e-infrastructure is a mechanism that, based on solid networking foundations, allows Computing, Data and Applications of various kinds (Data Reduction, Data Analysis, Theory, Numerical Simulations) to interoperate. Interoperability is also the key word of the Virtual Observatory.

Integration of the VObs with computational infrastructures, and in particular with Grid facilities, is of key importance. Effort in this direction has been carried out, also within EU-funded projects. In particular, specific IVOA standards have been developed on top of Grid middleware: a FITS driver has been built on top of *gridftp* [5]; the IVOA Single Sign-On standard has been implemented by means of the Grid Authentication & Authorization tools, VOSpace on top of the Data Management mechanisms, VObs Workflows using Job Management, VObs registries using the Grid Information Systems [6]. Furthermore, a "native" way of accessing databases from the Grid through a Query Element (similar in structure to the CE) was developed [12].

4.2. *Implementation in a distributed context*

One major issue to tackle is the capability of such an integrated e-environment to operate within a truly distributed environment. The key problem is that data processing and mining on distributed MDSs implies the distribution of the code as well.

As an example, let's consider the cross-check, or the merging, of science results obtained comparing two surveys (e.g. one based on ground-based and another one on space-borne instrumentation) covering a good fraction of the sky. In this case the VObs can help in identifying the location of the relevant data and in making the data themselves interoperable. But it would be unrealistic to move all of the data to the interested astronomer's home institution to perform the DM function. It is actually the data processing, DM or KDD codes that need to move, not the data.

The archive centers are bound to become something more complex in the future: complete full-fledged *service centers*, offering the possibility to provide processing capabilities. This is already happening e.g. for on-the-fly (re-) calibration of data with up-to-date parameters in case of observations [13], or for production of simulations on-demand [11].

But there is an additional step to perform: such centers should be willing to host *user-driven* processing applications. This is not difficult to achieve technically, given the good control of Grid and virtualization concepts and techniques that is becoming common practice, and the continuously decreasing cost of computing and storage hardware. DAME has proven the feasibility of the approach for DM applications. It is actually a "political" issue, tied to cost of resources and accounting.

5. Conclusions

Modern scientific data mainly consist of huge data sets gathered by a very large number of techniques and stored in much diversified and often incompatible data repositories. More in general, in the e-science environment, it is considered as a critical and urgent requirement to integrate services across distributed, heterogeneous, dynamic ''virtual organizations'' formed by different resources within a single enterprise.

In the last decade, Astronomy has become an immensely data-rich field due to the evolution of detectors (plates to digital to mosaics), telescopes and space instruments; new instrumentation is going to further enhance this growth. As a consequence, data archives have become an absolute must to cope with the data avalanche.

Data Mining techniques have also increased immensely their capabilities, and are now capable of extracting information from this enormous quantity of data, provided that appropriate processing resources are available. The DAME system is a good example of what can be achieved.

The effort to integrate and make the various data archives interoperable has been defined as the Virtual Observatory. The goal is to allow the users to feel all astronomical databases to be just "one click away". Achieving this goal would bring a degree of "democratization" in Astronomy, since first-class data are being made available to scientists in developing countries, to students of all levels, to amateurs. To make sense, the Virtual Observatory needs to be an international effort, which requires involvement at the *project* but also at the *data centre* level. This does not come for free, and work is required at the scientific and technical level to integrate one's own data archive in the VObs and make it interoperable with the others.

There is the need to expand network, data and computing infrastructures, and to integrate them, or at least let them interoperate thoroughly. This effort once again needs to be truly international, thus needing appropriate agreed-upon common policies. In this framework, the archive centers should become *service centers*, offering the possibility of hosting user-driven processing applications, in particular DM applications. In such a way astronomers will have the capability of using new powerful tools and techniques to improve their knowledge and understanding of astrophysical phenomena.

Acknowledgments

The Italian participation in EuroVO and IVOA is partially funded by INAF (National Institute of Astrophysics) through the VObs.it initiative. The DAME project, run jointly by the Department of Physics of the University "Federico II", INAF Astronomical Observatory of Napoli, and the California Institute of Technology, is financed through grants from the Italian Ministry of Foreign Affairs, the European FP7 EuroVO projects and by the USA-National Science Foundation.

References

1. R.J. Hanisch, in: ADASS XIX, *PASP* **434**, 65 (2010).
2. M. Brescia, G. Longo, G.S. Djorgovski, S. Cavuoti, R. D'Abrusco, C. Donalek, A. Di Guido, M. Fiore, M. Garofalo, O. Laurino, A. Mahabal, F. Manna, A. Nocella, G. d'Angelo and M. Paolillo, *arXiv*:1010.4843 (2010).

3. P.T. de Zeeuw and F.J. Molster (eds.), *"A Science Vision for European Astronomy"*, ISBN 978-3-923524-62-4 (2007).
4. M.F. Bode, M.J. Cruz and F.J. Molster (eds.), *"The ASTRONET Infrastructure Roadmap: A Strategic Plan for European Astronomy"*, ISBN: 978-3-923524-63-1 (2008).
5. G. Taffoni, A. Barisani, C. Vuerli, W. Pence and F. Pasian, in: ADASS XV, *PASP* **351**, 508 (2006).
6. G. Taffoni, C. Vuerli, F. Pasian and G. Rixon, in: ADASS XVII, *PASP* **394**, 289 (2008).
7. F. Pasian and G. Longo, in: ADASS XIX, *PASP* **434**, 357 (2010).
8. R. D'Abrusco, A. Staiano, G. Longo, M. Brescia, E. De Filippis, M. Paolillo and R. Tagliaferri, *ApJ* **663**, 752 (2007).
9. R. D'Abrusco, G. Longo, N.A. Walton, *MNRAS* **396**, 223 (2009).
10. M. Brescia, S. Cavuoti, G. d'Angelo, R. D'Abrusco, C. Donalek, N. Deniskina, O.Laurino, G. Longo, *Mem. SAIt Suppl.* **13**, 56 (2009).
11. F. Pasian, G. Taffoni, C. Vuerli, P. Manzato, F. Gasparo, S. Cassisi, A. Pietrinferni, M. Salaris, in: ADASS XVII, *PASP* **394**, 285 (2008).
12. L. Benacchio and F. Pasian (eds.), *"Grid-enabled astrophysics"*, Polimetrica International Scientific Publisher (2007).
13. F. Stoehr, D. Durand, J. Haase, A. Micol, in: ADASS XVIII, *PASP* **411**, 155 (2009).