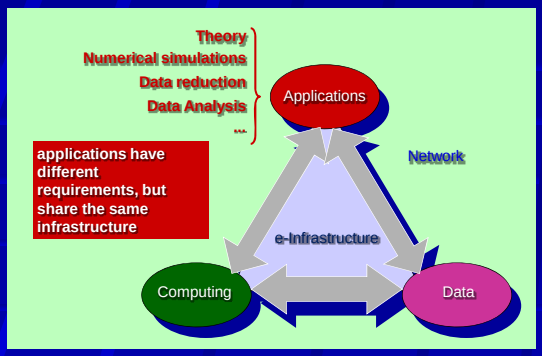
e-Infrastructures for Astronomy: an integrated view

F. Pasian ([1,2]), G. Longo([3])

([1]) INAF – Information System Unit, ([2]) INAF – Osservatorio Astronomico di Trieste, ([3]) Università di Napoli "Federico II"

As in the case of other disciplines, the capability of performing "Big Science" in astrophysics requires the availability of large facilities. Computational resources (e.g. HPC) are important, but are far from being enough for the community: as a matter of fact, the whole set of e-infrastructures (network, computing nodes, data repositories, applications) need to work in an interoperable way. This implies the development of common (or at least compatible) user interfaces to computing resources, transparent access to observations and numerical simulations through the Virtual Observatory, integrated data processing pipelines, data mining and semantic web applications. Achieving this interoperability goal is a must to build a real ``Knowledge Infrastructure" in the astrophysical domain. Also, the emergence of new professional profiles (e.g. the "astro-informatician") is necessary to allow defining and implementing properly this conceptual schema.



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.

DATA

In order to cope with the rapid increase in production and use of scientific information a high-quality and reliable repository infrastructure is essential. The emergence of "big data science" has a global dimension, as it reflects the increasing value of raw observational and experimental data not only in astronomy, but in virtually all fields of science.

The Virtual Observatory (VObs) has helped a lot in conveying the concept that data need to be shared to allow scientific advance in our field. But, besides raw and processed data, also publications, technical reports and white papers, software artifacts, numerical simulations, etc. all need to be taken in consideration. The VObs, within its IVOA 'standards' body, has recently made lots of progress in the direction of characterising 'scientific information' in a simple way. But this is confined to a limited set of data products and needs to be properly expanded.

Although astronomy is at the forefront of the digital data landscape, it is estimated that only a limited part of its research output is managed in digital repositories. Therefore, a new strategy for the management of scientific information and associated policies needs to be developed. This has to be done with the support of key research stakeholders as well as academic institutions and libraries.

But of course electronic access to data resources has costs involved (e.g. hardware, high-speed internet access, staff for maintenance), and most of these costs are left to the information providers.

Is developing domain-specific data repositories something still affordable in a period of economic recession? Or is it more cost-effective to develop a data infrastructure jointly with other disciplines, and implement on top of such a common infrastructure the domain-specific layer for data access (i.e. a VObs-enabled user interface)?

COMPUTING

To perform their research, astrophysicists heavily use a wide range of facilities: computing and storage capacities, archives plus, of course, the research networks. Astronomical institutions in all countries participate actively in the international initiatives in the field, and usually participate in defining national development policies together with the other disciplines.

Considering first the computing infrastructure, the facilities available to the scientists can be very different: from the laptop or desktop PC to the HPC center, passing through local clusters and ``the Grid". From the user perspective, ideally, all facilities should be seen homogeneously; in reality, they all tend to have different access modes. As a result, users find obstacles to their ideal exploitation.

In the medium-to-long-term it is important to reach complete interoperation of HPC centres: initially, Tier-2 (regional) facilities shall be integrated with the Grid, followed by full integration of local clusters (`Tier 3') and, as final goal, achieving inclusion of Tier-0/1 (national and international) within a common scheme. In any case, a necessary step is to achieve as soon as possible the full compatibility between HPC and Grid at least at the User Interface level.

The Grid concept should become far more widespread than currently. The old perception of the Grid re-using the idle cycles of a set of PC's (as was in the SETI@home experiment that originated the Grid concept) is nowadays completely misleading -- from the computing and storage viewpoints, the Grid has now the possibility of linking powerful clusters and even HPC hardware, provided it uses the proper middleware and high-throughput connections..

There is a further point. Up to now, the Grid has mainly delivered computing power, the main issue that its implementers, mostly involved in large High-Energy Physics experiments (e.g. at CERN in Europe), needed to solve. Accessing data, and databases using the Grid paradigm is a step still to be improved. Some activities are being carried out within EGEE. This is of course a further step in the direction of interoperability.

The "big science" challenges in astrophysics call for an expansion of the computing infrastructures -- and of network and data management infrastructures as well. The community is interested in using, and participating in defining, competitive computing infrastructures so to let its know-how in the field grow. But there is furthermore the desire to integrate the 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 our community.

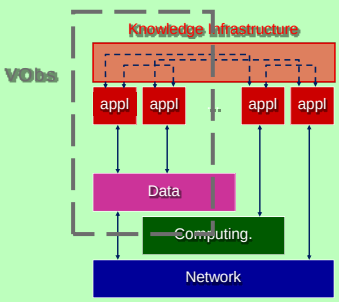
APPLICATIONS

Applications are the key of astronomical computer-aided research. Scientists develop codes of their own to perform personal research. To be integrated with other 'standard' data processing codes, such development has to embedded in a data processing environment.

The OPTICON (Optical-Infrared Co-ordination Network for Astronomy) proposal was funded by the EU FP7. It contains a network (N9.2) for discussing detailed designs and specifications for a future environment for data analysis in astronomy.

In a different paradigm, some applications are used as services provided by centres, and users can run them as black-boxes, without really caring about the computing resources they are run onto. This has been successfully applied for the production on-demand of numerical simulations (the BaSTI database), or for statistical analysis and data mining on large data sets (DAME).

A key issue for the complete exploitation of the potentiality of applications is the development of proper scientific gateways allowing seamless access to the codes.



The different layers of the integrated ICT infrastructure for astrophysics. Scientific applications may interact with the Data Infrastructure, or with the Computing Infrastructure, or directly with the Network Infrastructure (each interacting with each other). The interoperation of specific applications creates a Knowledge Infrastructure. The dashed box defines the area where the Virtual Observatory is expected to act.)

ASTRO-INFORMATICS

As it has been stressed in the report of the 2007 NSF workshop on data repositories "data-driven science is becoming a new scientific paradigm, ranking with theory, experimentation, and computational science". Astronomy has become a data-rich science , with data volumes rapidly growing from terabytes into tens (or hundreds) of petabytes. These data sets have posed scalability problems which are causing a strong shift in the way data are manipulated, reduced and understood. Problems which call for a set of know-hows in algorithms (scalability, parallel codes, etc), infrastructures (GRID, cloud computing, HPC) and access (web services and web applications) which do not belong to the traditional astronomical background. So far, these problems have been addressed in an informal way, leaving most of the required special expertise to e-Science, but have posed the foundations of an increasingly tight integration which will prove crucial in the near future when larger and larger data sets (both simulated and observed) will become available.

We believe that time is mature to promote (as it has already happened in other fields of research such as geo-informatics and bio-informatics) a major new discipline, which can be called Astro-informatics, to be integrated into Astronomy as a formal sub-discipline within agency funding plans, university research programs, graduate training, and undergraduate education.

In our understanding, Astro-informatics is an essential methodology for data-oriented astronomical research and the future of astronomy depends on it. Astro-informatics includes a set of naturally-related specialties: data modeling and description, data integration, information visualization, data mining and knowledge extraction, indexing techniques, classification taxonomies, ontology, etc.