

# A Common Data Analysis Environment

## Do we need one?

P. Grosbøl, P. Ballester and K. Banse

E-mail: pgrosbol@eso.org

European Southern Observatory  
Karl-Schwarzschild-Str. 2  
D-85748 Garching, Germany

### Abstract

Current astronomical data analysis systems were designed more than a decade ago and do not fully support or integrate many new technologies (e.g. web services and distributed computing). A common environment which provides a homogeneous interface to standard tools for analysis would increase scientific productivity. The top level requirements are outlined together with the explicit steps to define and implement them.

## 1 Introduction

This paper presents a short discussion of the needs for a common data analysis environment in the European astronomical community especially in view of the Virtual Observatory concept. Currently, several such environments are available but they were generally designed more than 10 years ago and do not offer adequate support for many important concepts (e.g. access to remote databases, distributed computing). Any one of them does not normally provide all the necessary tools to solve a given problem forcing researchers to spend time to learn several systems and migrate between them.

Whereas the creation of a single system may not be desirable, the definition of a common environment would improve efficiency by setting standards for interfaces between components used for analyzing scientific data. This would also enable different research groups to share their work and experience. Ideally, the efficiency of a research facility should be measured by considering the time it takes from a proposal is written to the results are published. Individual organizations can optimize their internal operation of their facilities but can only marginally influence the time used for the analysis of the data acquired. To improve the overall science performance, the establishment of an efficient environment for data analysis is of cardinal importance.

As the volume of data used by typical research projects increases, it becomes impractical to transfer and process the complete data at a single location. Many studies will also need to join data from different sources (e.g. satellite experiments or different wavelength regions) each with different processing requirements. This suggests that distributed access and processing of data will become increasingly important. Some of these issues are addressed by the Virtual Observatory and GRID studies but a common environment for data analysis which integrates these concepts will be an essential tool to increase scientific productivity of the individual researchers.

## 2 Basic Needs for Data Analysis

Many major astronomical facilities are starting to provide standard data reduction pipelines for their instruments in order to make it easier for users to start their analysis. These reduction pipelines are mostly limited to removal of instrument features from the data and to the calibration of them to physical units (e.g. flux, wavelength). The astronomical characteristics of the targets observed are normally not known to the standard reduction packages which therefore cannot extract more relevant information from the data (e.g. integrated intensities).

Even with access to reduced data, the astronomer is still left with the major part of the work, namely to extract useful quantities from the data and analyze them in an astronomical context. In a typical case, one would need to consider the following issues:

**Web Services:** Generic access to services offered on the Web is important to take full advantages of the ever growing number of such offers.

**Parallel and distributed computing:** Modern observing facilities can produce large quantities of data. With network access to cluster of machines (e.g. Beowulf systems), it is important that such resources can be used without a detailed knowledge of their internal functions.

In a scenario where data archives associated to major facilities also provide processing services to reduce raw data and extract quantities, it should be possible to specify the processing requirements in a generic fashion so that data from several different sources can be processed in a similar way. The GRID concept (see e.g. DataGRID and EuroGRID) will provide the technology to utilize distributed processing of large data sets, however, a data analysis environment must integrate this technology to make it feasible for scientists to fully exploit such resources.

**Extraction of quantities:** After calibrated maps have been obtained, the major task is to extract physically meaningful parameters.

**Propagation of errors:** Many new results are based on features detected at a marginal level in data sets. In order to quantify the significance of such claims, it is essential to have a firm trace of errors associated to individual values of the original raw data to uncertainties related to the final derived quantities.

**Statistical analysis:** Even with reliable errors known for a data set, one requires access to elaborate statistical tools to obtain safe estimates of the significance of the hypotheses being tested.

**Modeling:** For the physical interpretation of observational data, it is often necessary to compare them with theoretical models, either analytically or numerically, to understand the relative importance of different physical parameters. This suggests that it should be easy to integrate such theoretical models into the data analysis.

**Access to databases:** To compare a given data set with other similar ones it is often necessary to obtain data from remote databases.

**Interface to commercial packages:** It is neither possible nor reasonable that any single system can provide all features needed for the analysis of science data.

**Composite data types:** Most data have relations in the sense that they qualify other items (e.g. errors related to measured quantities, point spread function related to a spectrum). This can be a complex relation, but can often be mapped in a hierarchical structure.

### 3 Main objectives

A general issue is that, whereas technology evolves very fast, other parts like algorithms have a much slower rate of change. It is therefore important to have a clean separation of these two areas so that one can improve either part independently. It is essential that a wide community of users can use the environment for research of new algorithms as this is the best way to ensure a growing set of applications and increase the sharing of developments also between different branches of science. The close interaction with computer science and applied mathematics is crucial for the developments of efficient implementation of many numerical solutions to physical problems. Modularity is another key feature which would ensure that different groups could improve specific parts of the environment without the need to rebuild the full system.

The main objectives for a common Data Analysis Environment are summarized below:

- create open data analysis environment
- take full advantage of global communication (e.g. Internet)
- make simple usage simple
- encourage cross discipline development of algorithms
- be open for simple addition of new packages
- make access and sharing of raw and processed data simple
- design modular environment
- separate science algorithms and technology
- take advantage of new technologies (e.g. data mining and object orientation)
- support easy exchange of information with commercial packages
- support parallel and distributed computing

From the perspective of an individual researcher, ease of usage, availability of applications and stability of code are of prime importance. The ease with which one can develop new algorithms and interchange data is also an important issue for the full exploitation of data in hand.

For an organization considering to adopt the environment, ease of developments, reliability and maintainability are attributes to consider. The control of its configuration and options for including new requirements as response to new needs must also be taken into account.

### 4 New environment or upgrading an existing one

Most of the existing systems were designed over 10 years ago and therefore have not fully taken into account recent developments in information technology such as data mining, object orientation and fast Internet connections. They are, to a large extent, closed systems with a monolithic structure. Thus, it is difficult for non-expert users to add simple applications and integrate them with other packages.

Many of the systems have tried to respond to the new technical developments by providing hybrid interfaces to them. Although this may solve immediate needs, it does not resolve the basic problem of outdated design of the kernel parts of the system. As time goes on such additions tend to be more difficult to incorporate and maintain as compared

to a properly designed system. Thus, the modification or update of an existing system seems not to be an efficient way, especially since it would not solve the basic problem of a non-modular design.

Any general environment must at least provide a simple interface to legacy systems since a migration will take time and many important tasks will not be available immediately. Major archives and data centers may not see any advantage in changing an internal data processing system which is in stable operation. A full definition of the processing tasks available and their interfaces would be sufficient for a common environment to request operations to be executed at such centers and receiving the resulting data.

## 5 A scenario: How to proceed?

A common European environment for data processing and analysis can only be established if a broad consensus is reached in the community. Beside the technical definition, a critical issue to achieve is that the environment will be used by a sufficiently large fraction of the community as only this will yield the benefits of sharing and re-usage. The most important factors for getting a wide acceptance are:

- satisfy the main need for data analysis as outlined above,
- make it attractive and efficient to develop new applications,
- ensure stability and backwards compatibility, and
- provide adequate maintenance and support including documentation.

In order to achieve a consensus, one could take the following steps:

### 1. Preparatory meeting

This meeting should gather representatives from major European groups which could be interested in participating in the definition of a common Data Analysis Environment. The aim of this meeting would be to determine if there is a general interest in starting work on the definition of a common environment.

### 2. Requirement

A small working group should be formed with the mandate to make the draft of the requirements. A meeting to discuss the draft in detail should be arranged, followed by a formal review with participation of all parties.

### 3. High Level Design

The design should include a detailed list of the components required to create a baseline data analysis environment with their interface definition. An important consideration in the design will be compatibility with existing software in the free domain as adoption of such components could accelerate the implementation process.

### 4. Partition of tasks and responsibilities

The high level design has defined all components required for the environment. The implementation could follow two strategies, namely:

- (a) **Software Project approach:** A firm schedule for the deliveries is established and a group is created to monitor progress, make final acceptance and perform the integration of the different components.

- (b) **Open Source Internet approach:** The components are defined by the design document and groups interested in implementing a given part would advertise it and proceed. A board to coordinate the individual efforts should be established.

A substantial amount of data analysis software is available in the open source domain. It is important to take maximum advantage of this and try to reuse, adopt or adapt as much as possible as long as this does not threaten the homogeneity of the environment nor the requirements.

The configuration control of the environment must in all cases be well defined. Stability and reliability are essential for such a system to be adopted by the community.

## 6 Conclusion

It is clear that both individual astronomers and institutes would benefit from having a common data analysis environment as it would be easier to use and maintain than the current mixture of systems. However, the most important point would be an increase of collaboration, sharing of software and competition in development of new algorithms.

The main issues is less the definition and design than finding institutes which are willing to guarantee long term support for such an environment. Without that it would be very difficult to convince potential users that it would be worthwhile to make the personal effort in adopting a new environment.