Library and Information Services in Astronomy III
ASP Conference Series, Vol. 153, 1998
U. Grothkopf, H. Andernach, S. Stevens-Rayburn, and M. Gomez (eds.)

The VLT/HST Archive Research Facility

R. Albrecht[1,2], M. Albrecht, M. Dolensky[1,2], A. Micol[1,2], B. Pirenne[1], A. J. Wicenec

European Southern Observatory, Garching, Germany

Abstract. The European HST science data archive has been combined with the VLT science data archive. Beyond the capability to retrieve past observations the archive facility offers "added value products", such as recalibration using reference data and procedures which were not available at the time of the original observations.

At the same time the archive facilty will be optimised for knowledge discovery through data mining. During the data ingest procedure features and parameters which characterize the data will be extracted and made available in a rapid-access data base. Related information from other data bases and catalogues, and from the astronomical literature will be merged into the data base. In this manner previously unknown and unsuspected correlations can be discovered.

1. Introduction

The initial archive requirements for the HST science archive were derived from the need to store and retrieve the data taken by the Hubble Space Telescope (HST). However, in the near future the driver for the ESO Science Data Archive will be the Very Large Telescope (VLT): with four large Unit Telescopes, numerous smaller telescopes, large format detectors, and high sample rates, the data volume will be over 70 Gigabytes per day.

Given the rapid evolution of the World Wide Web the Archive is accessible to a growing community of users, who have become accustomed to rapid response and to the multitude of options and functions provided by large commercial services (e.g. the different search engines). In particular the Archive must evolve from a mere repository of data to a research facility which is capable of optimizing the scientific return from the telescopes. Therefore the development goals of the ESO archive have two main thrusts: the support of the VLT and of the HST; and the development of additional functionality to discover all information contained in the archive, either explicitly, or implied.

[1] Space Telescope European Coordinating Facility

[2] affiliated to the Astrophysics Division, Space Science Department, European Space Agency

2. Added Value Products: On-the fly Recalibration, WFPC Associations

New calibration reference data and new calibration procedures, based on improved operational experience allow improved calibration of archive data. Together with the Canadian Astronomical Data Center the ST-ECF pioneered the on-the-fly calibration concept: as the data are being retrieved from the archive they are calibrated using the best available calibration reference files and the best available calibration algorithms (Crabtree et al. 1997). This also allows to store ONLY the raw data. These are usually signed integers which compress well and fit on CD-ROMs or DVDs, keeping the data volume low and the hardware cost down. The ultimate goal is to perform both de-compression and re-calibration at the client site, using Web-compatible distributed software.

Associations are sets of logically linked observations, e.g. CR-split observations, long HST exposures interrupted by Earth occultations, or, ultimately, the Hubble Deep Field. The first goal is to identify suitable candidates and to perform the combination and re-calibration upon retrieval of the data from the Archive (Micol et al. 1998). It soon became evident that it was vital to assess the suitability of the data for re-combination on the basis of the jitter of the HST during the observations: only if the jitter is within the nominal limit of 7 milli arc seconds can the data be recombined.

A project to retrieve the jitter information for all WFPC2 observations was initiated and Java software to perform jitter analysis is available. It also turned out that the center of gravity of the "jitter ball" shifts between observations on a subpixel scale. This effect can be utilized to improve the resolution of the combined images using the "drizzling" technique developed for the combination of images of the Hubble Deep Field.

3. Data Mining

In view of the very large amounts of data generated by state-of-the-art observing facilities, the selection of data for a particular archive research project quickly becomes an unmanagable task. Even though the catalogue of observations gives a precise description of the conditions under which the observations were made, it does not contain any information about the scientific contents of the data. Hence, archive researchers first have to do a pre-selection of the possibly interesting data sets on the basis of the catalogue, then assess each observation by visually examining it (preview) and/or running an automated task to determine its suitability. Such procedures are currently used for archive research with the HST Science Archive. This is acceptable as long as the data volume is limited. However, already after the first year of Unit Telescope 1 operations, the VLT will be delivering data in quantities which do not make it feasible to follow the same procedures.

The ESO/CDS Data Mining Project aims at closing the gap and develop methods and techniques that will allow a thorough exploitation of the VLT Science Archive. The basic concept is not to have to ask for individual data sets, but instead to be able to ask for all information pertaining to a set of search criteria. In addition to parameters contained in the observation catalogue, the criteria should include parameters which pertain to the science content of the observations. This implies that science parameters have to be generated after the observations, either by performing science processing on the data, or by linking science-related information from other sources. The proper time is during the ingest of the data into the archive.

These parameters can then be correlated with other information. The concept is to create an environment that contains both extracted parametric information from the data plus references to existing data bases and catalogues. The environment then establishes a link between the raw data and the published knowledge with the immediate result of having the possibility to derive classification and other statistical samples.

Data mining as described above is a step in the process of Knowledge Discovery in Data Bases (KDD): application of specific algorithm(s) to produce a particular enumeration of patterns over the data base. Knowledge Discovery in Data Bases is the extraction of implicit, previously unknown, and potentially useful knowledge from data bases.

4. Determination of science-related parameters

The aim of parametrization must be the enumeration of statistically relevant and physically meaningful parameters. Examples are: integrated energy fluxes of objects, colors, morphology, distribution. This will lead to data archives which are organized by objects rather than by data sets (Albrecht et al. 1994)

A promising beginning are tools like SExtractor. This software package allows the extraction and parametrization of objects on large image frames (Bertin & Arnouts 1995). Electronic links will be used to collect parameters on the objects thus extracted from other sources, for instance data bases from other wavelength regions. These parameters will either be physically imported and added, or they will be attached to the objects through hyperlinks.

The most useful science processing consists of classification. The challenge of classification is to select the minimum number of classes such that objects in the same cluster are as similar as possible and objects in different classes are as dissimilar as possible. However, this has to be done in such a way that membership of an object in a particular class is meaningful in terms of the physical processes which are responsible for the condition of the object.

This is not always the case for traditional classification systems in astronomy: the binning criteria were determined by the characteristics of the detector and the physiology of the human classifier. This is also not necessarily the case for statistical approaches (clustering, pattern recognition, neural network), because no physics is involved in establishing the classes. The current emphasis is on automatic classification, and on the data base access mechanisms to mine terabyte-sized data bases.

5. The Archive Research Environment

It is evident that the optimum exploitation of the above concepts will require a special computational infrastructure (Albrecht et al. 1998). Given the large data volumes, the need to access heterogeneous data bases, and to execute different software packages we need an environment tailored to these requirements.

The goal is to offer the Science Archive as an additional instrument with the capability of feature extraction from raw data and a data mining environment on both data and extracted parameters.

Archive Research Programmes are either user-defined or standard processing chains that are applied to the raw data. Each of the processing steps is called a Reduction Block. Typically the first reduction block would be the recalibration of data according to a standard calibration pipeline. A reduction block consists of one or more processes which are treated by the system as 'black boxes', i.e., without any knowledge of its implementation. However, the reduction block interface does comply to a well-defined specification, which allows any reduction module to become part of the chain. In fact, this flexible architecture also allows the research programme to analyze different kinds of data from images and spectra to catalogues and tables of physical quantities. The output of an archive research programme will be derived parameters that are fed into an object-oriented data mining database (Objectivity).

Additional parameters for the data mining database will be fed in by a cross-correlation interface to the CDS and thus linking the ESO Science Archive to the published knowledge database of the CDS. The first step will be an interface to the parameters of catalogues of astronomical objects, the next step will also include means to extract information from published papers. This interface may be used from either side, CDS or ESO-SARE (Science Archive Research Environment) to fill the data mining database with relevant parameters for a KDD project. The implementation of this system will be a first step to realize an astronomical data mining approach as described above.

References

Albrecht, M.A., Angeloni, E., Brighton, A., Girvan, J., Sogni, F., Wicenec, A.J. & Ziaeepour, H., 1998, The VLT Science Archive System, Astronomical Data Analysis Software and Systems VII, R. Albrecht, R. N. Hook & H. A. Bushouse (eds.), ASP Conference Series, Vol. 145, 363

Albrecht, R., Albrecht, M.A., Adorf, H.M., Hook, R., Jenkner, H., Murtagh, F., Pirenne, B. & Rasmussen, B.F., 1994, Archival Research with the ESO Very Large Telescope. In: Proceedings of the Workshop on Astronomical Archives, Trieste, Albrecht, M.A. & Pasian, F. (eds.), ESO Workshop and Conference Proceedings Series No. 50, 133

Bertin, E., & Arnouts, S., 1995, A&AS, Vol. 117, 393

Crabtree, D., Durand, D., Hill, N., Gaudet, S. & Pirenne, B., 1997, Automatic recalibration – A new archive paradigm. ST-ECF Newsletter No. 24, 19

Micol, A., Pirenne, B. & Bristow, P., 1998, Constructing and Reducing Sets of HST Observations Using Accurate Spacecraft Pointing Information, Astronomical Data Analysis Software and Systems VII, R. Albrecht, R. N. Hook & H. A. Bushouse (eds.), ASP Conference Series, Vol. 145, 45