RevMexAA (Serie de Conferencias), **24**, 35–39 (2005)

THE GTC SCIENTIFIC DATA CENTRE

E. Solano[1]

RESUMEN

El desarrollo de un archivo científico que garantice el óptimo aprovechamiento de los datos producidos por GTC fue un aspecto claramente identificado ya desde las primeras etapas del proyecto. En este artículo se describe el diseño conceptual así como las principales características del futuro Archivo de Datos Científicos de GTC (GSA). La implementación, desarrollo y mantenimiento del sistema correrá a cargo del Laboratorio de Astrofísica Espacial y Física Fundamental (LAEFF).

ABSTRACT

Since the early stages of the GTC project, the need of a scientific archive was already identified as an important tool for the scientific exploitation of the data. In this work, the conceptual design and the main functionalities of the Scientific Data Archive of the Gran Telescopio Canarias (GSA) are described. The system will be developed, implemented and maintained at the Laboratorio de Astrofísica Espacial y Física Fundamental (LAEFF).

Key Words: **INSTRUMENTATION: MISCELLANEOUS — METHODS: DATA ANALYSIS**

1. INTRODUCTION

The final goal of a observational scientific mission is to obtain high quality data. In order to optimize the usage of these data it is necessary to build up an archive system that guarantees the long-term maintenance and easy access to the data.

The importance of Scientific Data Archives has been demonstrated by the space-based missions over last two decades. *IUE*, *HST*, *ISO*, *XMM*,..., are good examples of how important contributions to progress in Astronomy can be made using archived data. They also clearly show the difference between a safe data store and a real Scientific Data Centre in which the observations are pipeline processed, calibrated and catalogued in an uniform and homogeneous way and can be accessed by the community through easy-to-use data interfaces.

Although only in the last years the large ground-based telescopes have considered the development of Scientific Data Centres, there is no fundamental reason why a Science Data Archive from a ground-based facility should be more difficult to implement or less valuable than an archive from a space-based observatory. These differences between space and ground-based observatories have been motivated historically by the fact that a higher level of automation is required in space where real-time human intervention is much more difficult.

A well-designed and properly implemented Scientific Data Archive, as part of the capabilities of the GTC telescope, is a major contribution toward the full exploitation of its unique characteristics. The access to GTC data by the wider community after the proprietary period is over will ensure that the data are fully exploited and that maximum scientific value is returned to the astronomical community and, ultimately, the citizens that support GTC through taxes. The collective impact of the GTC archive will far exceed what could be produced by the programs of individual observers or teams. Three are, at least, the types of research projects that can be conducted with archived data. The first consists of cases where the data are used for an entirely different scientific project that they were obtained for. The second is the case where new, improved or more effective methods of analysis are considered. Finally, the third exploits the collective effect of the archive where a large dataset spanning a wide range in some important parameter is available to the researcher. Moreover, the linkage across archives and wavelengths regimes in the framework of the Virtual Observatory initiatives adds still more value to archive data.

This paper is structured as follows: the archive conceptual design is given in §2, the main functionalities are outlined in §3,4 and 5, the role of the GTC archive in the framework of the Virtual Observatory is explained in §6 whereas the conclusions are summarised in §7.

2. ARCHIVE CONCEPTUAL DESIGN

A data archive system typically comprises the data storage, data management and data retrieval

[1]Laboratorio de Astrofísica Espacial y Física Fundamental, Madrid, Spain.

through an interface. The process starts with the receipt of data from a processing pipeline and ends with a dataset transferred over the Internet to the user's computer.

Four different functional blocks can be identified:

- Data products: FITS files containing the data itself.
- Database: It contains the metadata of the data products. It will make possible to perform complex searches among the data products.
- Database ingestion: extraction of the metadata from the data products to store it in the database.
- Archive applications: it comprises all applications developed to access the archive (Web application, Virtual Observatory interfaces, etc).

The applications will not interact directly with the database and data products. Instead, a middle tier is added to manage the interaction between the applications and the archive. This provides the system with a better flexibility and independence of the Database Management System.

A general schema of the proposed system is shown in Figure 1.

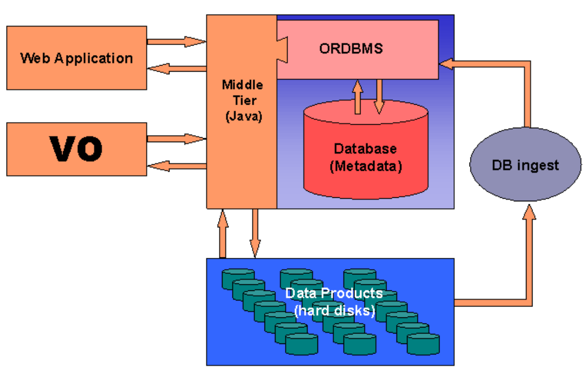

Fig. 1. Conceptual design of the GTC Scientific Data Centre.

3. INGESTION STORAGE AND MANAGEMENT OF GTC DATA

The primary task the GSA must fulfill is being a data repository; this is the first step that needs to be accomplished before any of the possible added-value features can be implemented. The present scenario assumes that the pipeline data processing and the shipping of the output products in FITS format to LAEFF via the File Transfer Protocol (ftp) system is performed by GTC staff at the telescope. A critical issue is, therefore, the accurate definition of the GTC-LAEFF data transfer system. If required, metadata information not present in the FITS headers would be included into separate files (e.g., ASCII tables). The possibility of performing the reduction process at the Scientific Data Centre is also contemplated and may be implemented in the future.

Once at LAEFF, all the metadata describing the data contained in the files need to be ingested into a database. A Java application will perform the following operations:

- Extract the metadata from the FITS headers and associated files and insert them into the database in an automated way.
- Move the FITS files into the data storage system. To ensure the data integrity, a well-defined backup and data safekeeping policy will be established. Also, the archive must respect the proprietary period assigned by GTC.

The mechanism should be dimensioned to be able to cope with the data flow expected from the Day-1 instrumentation ($\approx$ 5 GB/night) and further implementation of new instruments.

4. GTC DATA RETRIEVAL AND USER INTERFACE

The friendliness of the GSA user interface is extremely important for the archive itself to be effectively used by the community at large; a possible awkwardness in its use may turn away large fractions of potentially interested people. On the other hand, complex searches must be possible to provide the user with an efficient tool, which allow finding exactly what he/she is looking for.

The proposed user interface is HTML-based to allow simple use through the Web. The HTML pages will be dynamically generated using JSP/Java from the information present in the database and from the data products. Therefore, in the very moment the data obtained from the GTC have been ingested, they will be accessible through the user interface. A prototype of the user interface is given in Figure 2.

5. ADDITIONAL FUNCTIONALITIES

Apart from the core search engine, additional tools will be developed to allow an effective use of the archive according to the user needs. Some examples are:

- Visualization and analysis tools: A modern scientific archive cannot be a simply data provider but must implement analysis tools to facilitate the scientific use of the archive. In order to optimize the efficiency of the tools implemented at the Scientific Data Centre it is of fundamental importance a close feedback loop between the tools developers and the archive users to ensure that the tools match the scientific requirements of the astronomical community.
  - Plotting facilities: Fast visualization of the data as well as some basic operations (zooming, scaling, histogram computations,...) must be guaranteed. Previewed data must contain enough information to allow user to judge the data quality.
  - Data mining tools: This is an added value to a scientific archive of undoubted importance. Data mining, understood as the process of finding information buried in a mass of data, has been identified as a key activity within the Virtual Observatory project and from where the greatest scientific benefits are expected to flow. In this framework the user will remotely run sophisticated analysis software through click-and-play packages at the Scientific Data Centre without having to transfer vast amount of data to his/her desktop.
- Off-line data reprocessing: The ability to retrieve raw data for its further off-line reduction will be implemented. The associated information and files should be complete enough to allow users to repeat the same reduction steps as the pipeline does.
- HelpDesk: A multi-layer approach to cope from the most general to the most specific questions is foreseen. The system will contain a *Frequently Asked Question* section with the answers to the most common questions asked by the users.
- On-line Access to project documentation: A detailed description of the project, the archive and access system shall be given on-line and stored in a format allowing an easy browsing (e.g. PDF).

6. INTEROPERABILITY: THE VIRTUAL OBSERVATORY

Although astronomical archives constitute a basic tool for modern Astrophysics as revealed by their intensive usage, it is also true that the efficiency in the information retrieval is seriously limited by the lack of interoperability among them. Historically, archives have been built independently of each other and their remote interoperability is neither easy nor efficient. The query to multiple databases simultaneously is presently done but slowly and painfully by hand as the databases are spread all over the globe with a variety of formats, access levels and policies. In most cases the user must often have a detailed knowledge not only of the particular instrumentation but also of operational details and of the specific structure of the database.

The Virtual Observatory (VO) is an international project aiming at solving the problems that this lack of interoperability creates for multiwavelength astronomy. The project is designed to provide the international astronomical community with the data access, research tools and systems, research support, data interoperability standards, data-flow practices and data centre coordination necessary to enable the exploration of the huge amount of data resident in the international astronomical data archives. One of its main objectives is the creation of a federation of astronomical archives connecting all major astronomical data centres that, with the implementation of new technologies and standards, provides an easy and efficient access to the astronomical data.

Although VO is an emerging project still in its developing phase, it is considered both from the technical and scientific point of view as a basic requirement for the astrophysical research and the framework where to settle in the short-term the astronomical archive-related activities. LAEFF is leading the Spanish participation in the VO through the development of the Spanish Virtual Observatory (VOE) and its further inclusion in the Virtual Observatory federation. The development of the GTC Scientific Data Centre in this framework will allow a smooth integration in the VO structure by taking into account the compliance with the requirements and standards defined within the VO project. This integration will permit the GSA users to compare GTC data with other data sets, measurements or identifications available in other astronomical archives and thus giving rise to an enhancement in the number of the multi-wavelength research projects in which GTC can be involved.

7. CONCLUSIONS

The GSA shall provide long-term storage of scientific data from GTC, will implement visualization and analysis tools with special emphasis on advanced

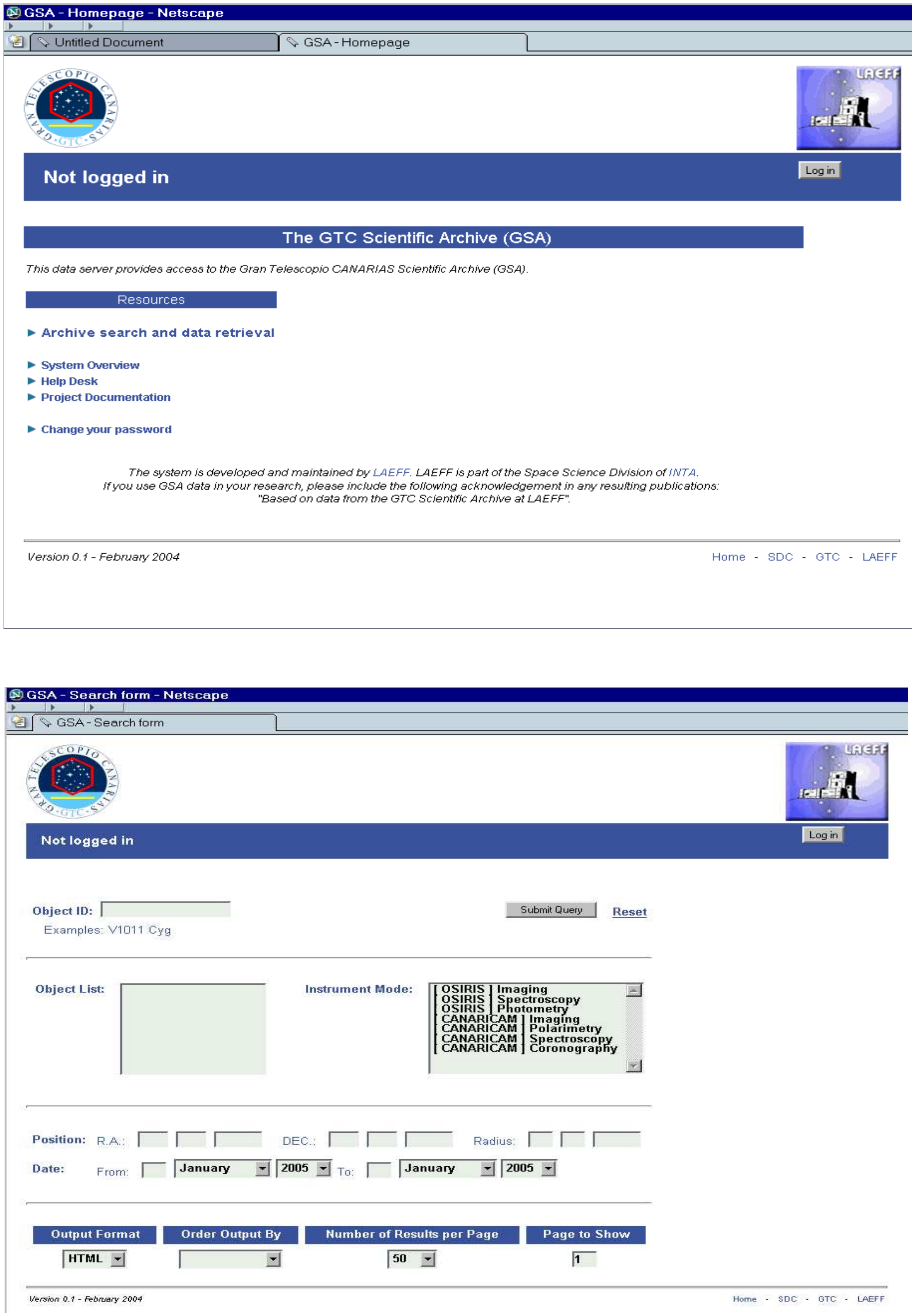

Fig. 2. Prototype for the User Interface of the GTC Scientific Data Centre.

data mining tools and will guarantee the interoperability with a number of existing data centres and archives, in particular with those fostering the idea of the Virtual Observatory. GSA will guarantee that the GTC legacy, i.e., the collection of excellent observations produced by its innovative instruments, will play an important role in the astrophysical research long after GTC itself ceases operations.

The expertise gained over the past years has demonstrated that LAEFF possesses, both at individual and Institute level, the know-how necessary to carry out the development of a Scientific Data Centre fulfilling all the GTC requirements for archiving, maintaining and distributing data. The GSA will be implemented at LAEFF as an extension to the existing Scientific Data Centre what will reduce the overall cost and benefit from the accumulated experience.

I would like to express my gratitude to the Group of Operations and Archives at LAEFF, in particular to Raúl Gutiérrez, for his valuable help in this work.

Enrique Solano : Laboratorio de Astrofísica Espacial y Física Fundamental. P.O. Box 50727, 28080 Madrid, Spain. (Enrique.Solano@laeff.esa.es).