Querator: an advanced multi-archive data mining tool

F. Pierfederici[a]

[a]European Southern Observatory, Karl-Schwarzschild-Str, 2, D-85748 Garching, Germany

ABSTRACT

In recent years, the operation of large telescopes with wide field detectors – such as the European Southern Observatory (ESO) Wide Field Imager (WFI) on the 2.2 meters telescope at La Silla, Chile – have dramatically increased the amount of astronomical data produced each year. The next survey telescopes, such as the ESO VST, will continue on this trend, producing extremely large datasets. Astronomy, therefore, has become an incredibly data rich field requiring new tools and new strategies to efficiently handle huge archives and fully exploit their scientific content.

At the Space Telescope European Coordinating Facility we are working on a new project, code named Querator*. Querator is an advanced multi-archive search engine built to address the needs of astronomers looking for multicolor imaging data across different astronomical data–centers.

Querator returns sets of images of a given astronomical object or search region. A set contains exposures in a number of different wave bands. The user constraints the number of desired wave bands by selecting from a set of instruments, filters or by specifying actual physical units.

As far as present–day data–centers are concerned, Querator points out the need for:
- an uniform and standard description of archival data and
- an uniform and standard description of how the data was acquired (i.e. instrument and observation characteristics).

Clearly, these pieces of information will constitute an intermediate layer between the data itself and the data mining tools operating on it. This layered structure is a prerequisite to real data–center inter–operability and, hence, to Virtual Observatories (VOs).

A detailed description of Querator's design, of the required data structures, of the problems encountered so far and of the proposed solutions will be given in the following pages. Throughout this paper we'll favor the term data–center over archive to stress the need to look at raw-pixels' archives and catalogues in an homogeneous way.

Keywords: Astronomy, Multi-Color photometry, Virtual Observatory, Data-Mining, Databases, Search Engines

1. INTRODUCTION

We are now witnessing a radical change in Astronomy and space science. While in the past astronomers typically studied in great detail only a small number of object on the sky, nowadays the numerous ongoing digital sky surveys enable the scientific community to carry out systematic studies of the whole sky, often in several filters and in different regions of the electromagnetic spectrum.

This is a definitive shift in the way astronomers look at the universe, which requires a new way of thinking about Astronomy and about the tools and methods to use. Indeed, Astronomy has now become a very data rich field, with the 'production' of several Gigabytes of astronomical data per day. The the next surveys are expected to substantially increase these figures, producing Pentabytes of data per year (Brunner et al., 2001).

The big challenge Astronomy has to face now is, therefore, to find a way to organize this enormous amount of data so that it can be quickly accessed and throughly exploited. This process is made difficult by the amount itself of the data involved, by the intrinsic heterogeneity of astronomical data (acquired with different instruments, in different filters, in different regions of the electromagnetic spectrum, in different epochs etc.) and by the peculiarities of how each data–center stores this information.

Further author information: E-mail: fpierfed@eso.org
*http://archive.eso.org/querator/

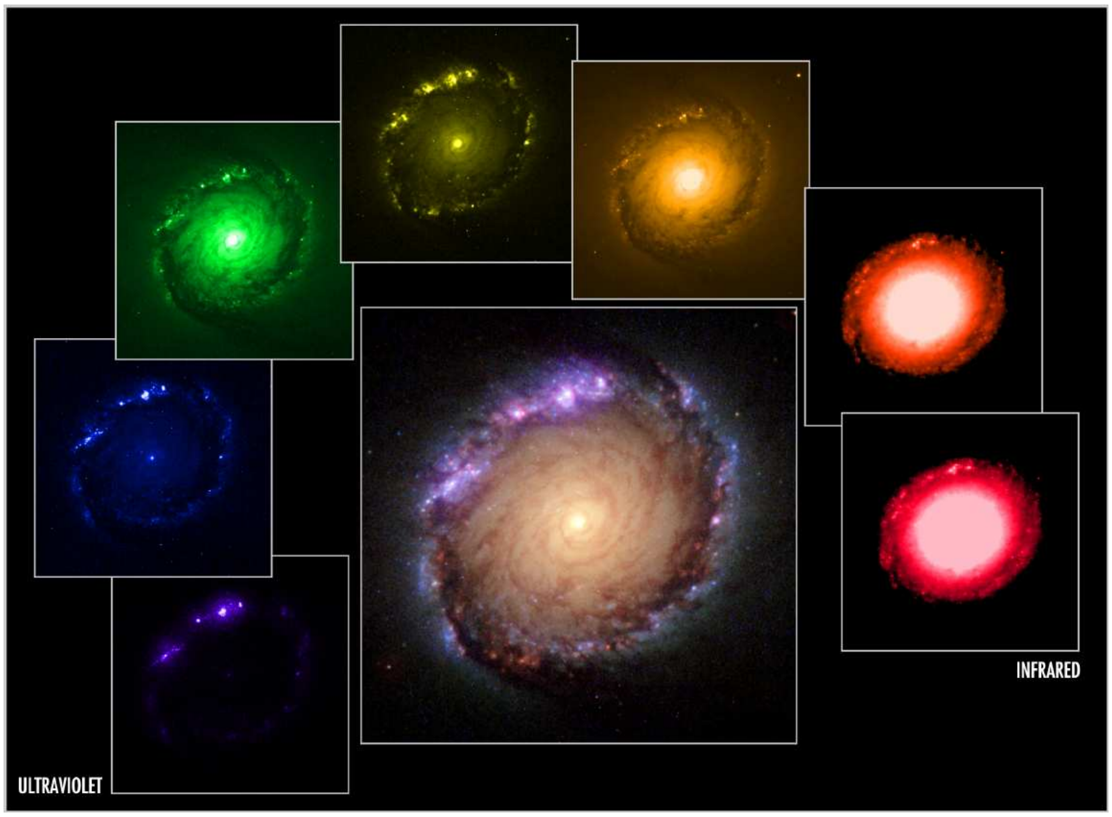

Figure 1. This composite picture shows how differently a galaxy can appear if observed in different filters. Image: NASA, ESA, Dan Maoz (Tel-Aviv University, Israel, and Columbia University, USA).

It's already available, both on the Internet and in specific software packages, a large number of tools to explore multiple data–centers at once, more or less transparently. Nevertheless, we felt the clear need for something more complex and complete.

We wanted, in essence, to address two separate issues. On one side, the necessity to overcome the current granularity of astronomical data–centers. We wanted astronomers to be able to explore all the available astronomical information as if they were connected to a single Virtual Observatory. Users, in other words, shouldn't even need to know that they are looking at data which is actually stored in different sites. We wanted them to be able to look at astronomical data as a whole. This approach will enable new ways of doing Astronomy. At the same time, statistical studies which are too difficult, demanding and time consuming to be carried out with the current technologies will be made more easy.

On the other hand, we realized that current tools let the users find information about a given astronomical object or a given region of the sky. We wanted to proceed in the opposite direction: in our case, the object names (and/or field names) were meant to be the results!

Astronomical data has several dimensions. These dimensions are defined by parameters describing the physical characteristics of the different objects (e.g. morphology, magnitudes, temperatures etc.) and how, when, with which instrument and under which conditions they was observed. Parameters describing the characteristics of the observation itself, such as its seeing or its limiting magnitude, define other dimensions as well.

With our search engine – code named Querator – users can now find objects of interest simply defining cuts in this multidimensional parameter space. This is made easy by the use of a simple web interface and by letting the users specify constraints on physical parameters in physical units rather than forcing them to learn an obscure query syntax.

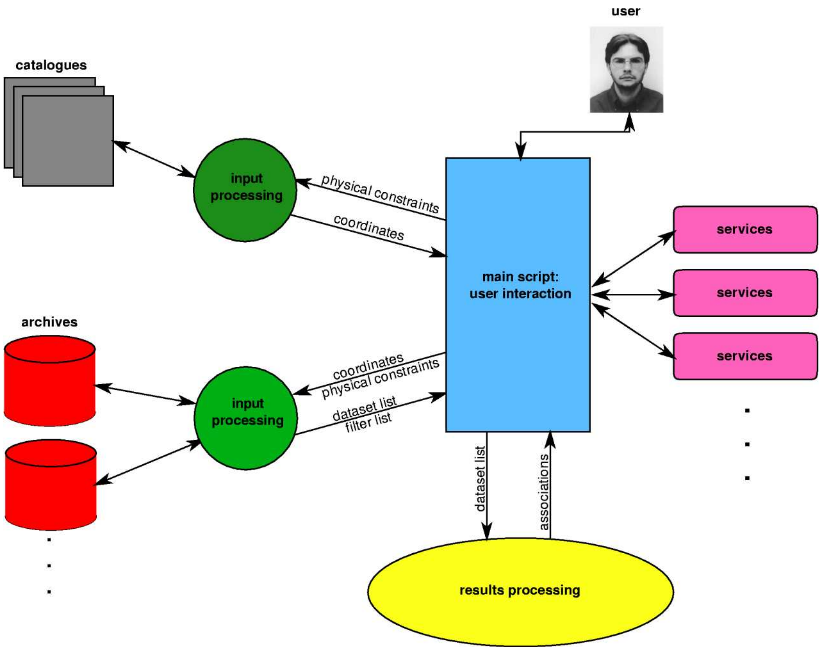

Figure 2. Querator's design.

As widely recognized (NVO Informal Steering Committee, 2001), data–center inter–operability is a key ingredient for the construction of the Virtual Observatories of the future. It is also a prerequisite for the development of the data mining and information visualization tools VOs are going to need – much like telescopes need sophisticated instrument to produce exiting results.

By data–center inter–operability, we also mean the design and implementation of a layer between the data itself and the tools which operate on it. The intermediate layer should provide an uniform and standard description of the data stored in archives and catalogues. This meta-data has to be uniform and standard in order for different tools to be able to perform the same actions on all different kind (and provenience) of data they 'know of'. The 'meta-data' layer should also describe the quality of the stored information (which has to be defined and measured homogeneously) as well as how the data was acquired.

Another important point is the need for a description of the evolution in time of instrument characteristics and performance. This is of particular importance if one needs to take into account, for instance, filter transmission curves to characterize archival data, as in our case.

Among all the different possible developments, we have chosen to work on astronomical multi-color imaging data. This choice has been driven by the awareness that there wasn't any way, using the available tools, to select only astronomical objects observed in a given number of different colors. Querator addresses this very need and has proven to be of extremely useful to astronomers interested, for instance, in multi-color photometry, in the determination of photometric redshifts or in studying the evolution of galaxy morphology with z.

2. QUERATOR

In the following subsection we'll describe in great detail the structure of Querator. Next, we'll describe what has been implemented so far. In the last subsection, we'll explore open issues and future developments.

2.1. DESIGN

Querator has been thought as a rather complex search engine. It is able to query different archives at once and to perform cross-correlations between archives and such catalogues as the Lyon-Meudon Extragalactic Database

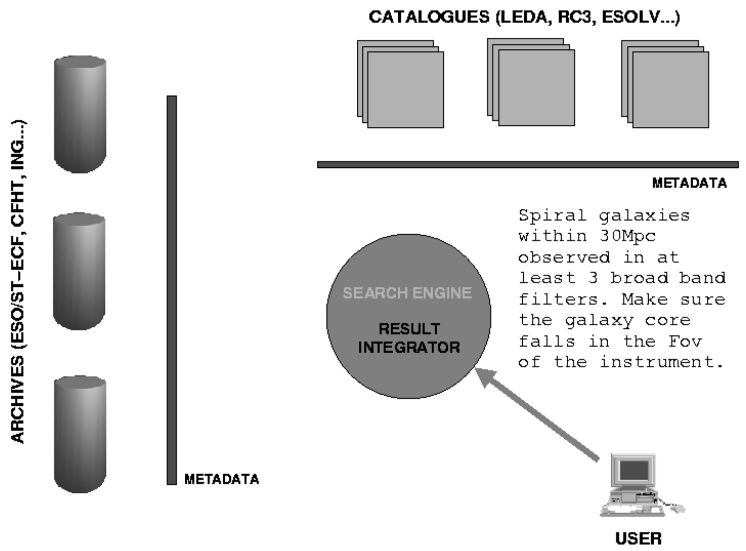

Figure 3. The user sends a request to Querator.

(LEDA[†]) or SIMBAD[‡]. It collects the results, integrates them and make sure that they satisfy all the constraints imposed by the user. Finally, users are presented with nicely formatted HTML pages listing the output of their queries. From there, they can retrieve the selected images or perform some further analysis.

This design is open to the inclusion of add-on programs able to offer post-query services once results have been displayed. Examples include queries to bibliographic services (e.g. the NASA Astrophysics Data System[§]) to retrieve the literature references on a particular object or advanced visualization tools.

Since our code spends the vast majority of time waiting for answers from the different data–centers, we realized that its execution speed was not a major issue. This leaves us with great freedom in the choice of the programming language; choice which has been driven by the availability of existing code libraries, more than anything else. Some of our scripts have therefore been developed in perl, some in python. A major rewrite of the software in python is underway. This has been suggested by the desire of a purely object oriented design which would offer a natural support for add-on applets providing services to the user and to the search engine itself.

Anyway, especially in this prototyping phase, a scripting language was preferred because of its quick develop–test–debug cycle which enabled us, in the beginning, to come up very quickly with a working prototype and, later on, to quickly implement new features.

Given the need to query different sources of information at the same time, we preferred a distributed design in which all the needed services are provided by small, highly specialized applications (which are being turned into python classes in the new object oriented design). In order to speed up the execution of the 'parent query' all the independent sub queries are run asynchronously in a multi threaded environment.

For ease of development and to provide users with an already familiar graphical user interface (GUI), we opted for a web based program. This solution does not require any special software installation on the user side, which is, by itself, an important bonus.

In the current design, there are three groups of applets involved in a typical query: a first group handles the data retrieval process. This group is divided in two subgroups: on one side there are applets building the list of celestial coordinates and/or regions on the sky the user is interested in. This job can be as simple as reading the input from

[†]http://leda.univ-lyon1.fr/
[‡]http://simbad.u-strasbg.fr/
[§]http://adswww.harvard.edu/

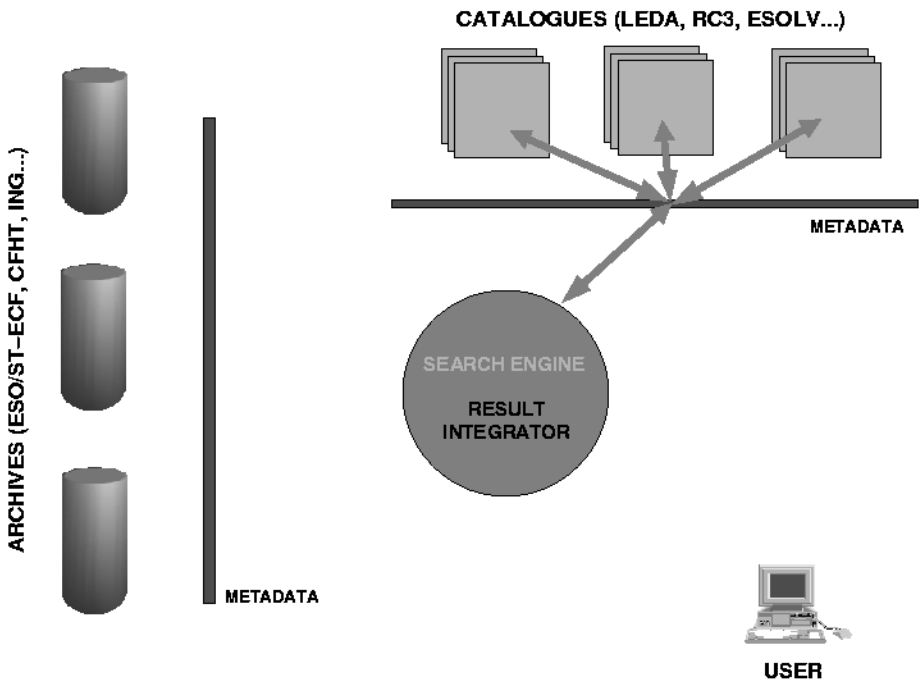

Figure 4. Coordinate lists are built.

HTML forms or as complex as making use of constraints on physical parameters to extract information from known catalogues. The applications retrieving the list of filters (on a per instrument basis) the user is interested in, also belong to this subgroup.

As far as filters are concerned, users are asked to input the wavelengths and/or ranges of wavelengths they are interested in. This information is then complemented by their choice of the minimum number of different filters and of filter type (e.g broad band or narrow band filters). The Filter transmission curves are stored in auxiliary database tables used to translate these constraint in filter lists on a per instrument basis.

The second subgroup of applications is composed by applets querying the different data–centers. They make use of the coordinate and filter lists produced by the previous subgroup. The querying of archives for images starts as soon as the first coordinates are received.

A second group of applets organizes and summarizes the collected data. This step is needed to make sure that every user constraint is satisfied. This could mean, for example, trying to associate the different pointing on the sky in clusters of observations and in making sure that the constraints on filters hold for each cluster.

The last group of applets is composed by applications performing post–query actions. These can address different needs of the user and can be added in subsequent stages as they are developed.

These three groups of applications are 'directed' by the script handling the interaction with the user. This scripts has, therefore, to be responsible for the creation of HTML pages to collect users input and to organize and present the results.

2.2. IMPLEMENTATION

Ideally, the access to different data–centers should go through an intermediate layer of meta-data information, as discussed in the Introduction. This layer would provide the required translation of our query syntax into the different 'dialects' spoken by the different data–centers.

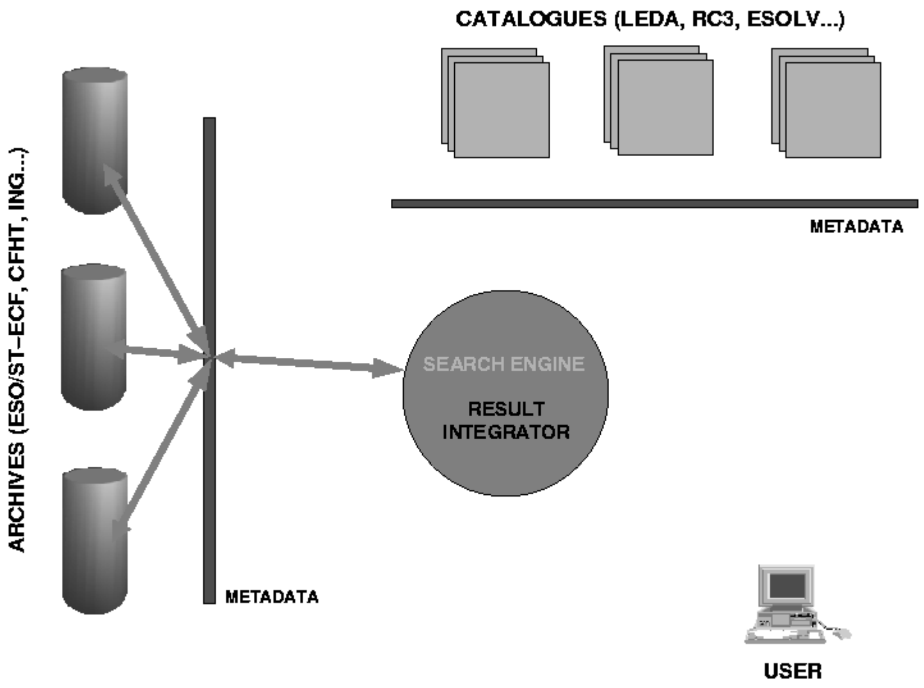

Figure 5. Filter lists are retrieved. Archives are queried.

In reality, at the moment, this is not the case as such a layered structure has yet to be implemented. Both in the United States and in Europe, the leading data–centers are jointly designing the structures and developing the technologies needed for archive (and catalogue) inter–operability.

For the time being, therefore, a number of ad–hoc solutions had to be developed in order for Querator to be able to effectively query and gather results from multiple data–centers. These ad–hoc solutions involve low level, direct access to the database servers used by the different data–centers.

This temporary solution works, in the sense that users are actually able to get the information they need. It is not a fully satisfactory solution, tough, mainly because of the large amount of work needed to support it. Another serious drawback of the current implementation is its requirement of further translation of the results coming from different sources into a common format and syntax. This step is particularly important for a nice and consistent user experience, but also for the internal work of Querator itself.

2.3. OPEN ISSUES AND FUTURE DEVELOPMENTS

From the discussion of the previous subsections, it is clear that, once the international Virtual Observatory will become a reality, Querator will have to be retrofitted in order to make efficient use of the meta–data descriptions provided by the archives and catalogues federated in the VO.

On the features side, we are very much interested in adding further capabilities to Querator in order to make the exploration and exploitation of astronomical data–centers more easy and more ‘deep’. This will involve the development of applications providing value added services to the users, the support to new instruments, new archives and new catalogues. In this phase, comments and suggestions coming from early adopters have been of invaluable help.

APPENDIX A. LIFE AND DEATH OF A TYPICAL QUERATOR SESSION

In this Appendix, we describe the life cycle of a typical Querator session.

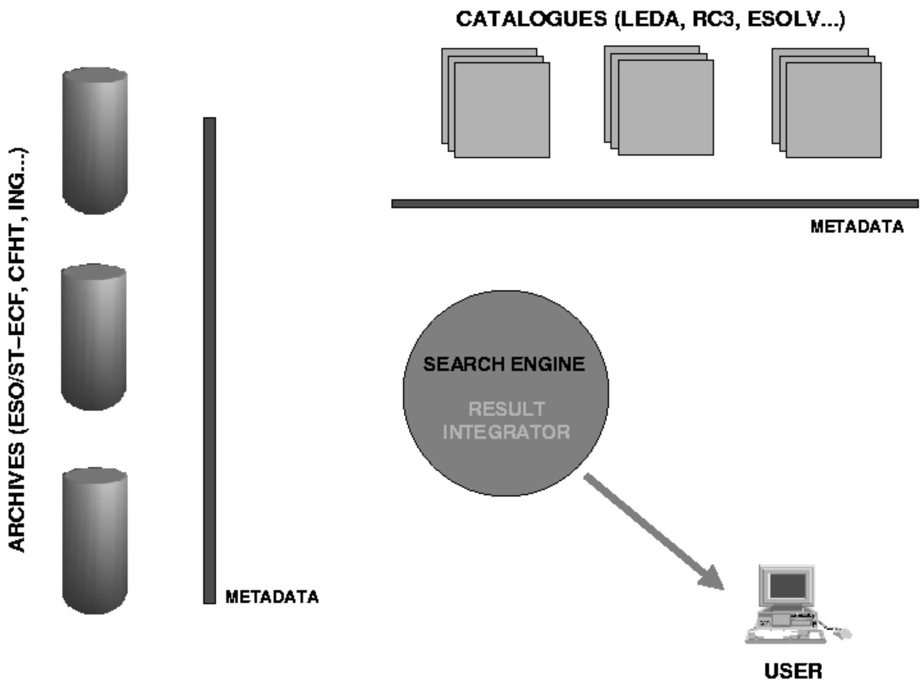

Figure 6. Results are integrated and displayed.

To start with, the user sends a request to Querator setting constraints on physical parameters. Querator will be able to translate the user's request in constraints to data–centers' specific parameters. This translation is expected to be made transparent and automatic, in the near future, by the meta–data layers provided by each data–center (figure 3).

In a second step, if needed, a number of catalogues are accessed in order to translate the relevant portions of user's query into coordinate lists (figure 4). Using the coordinate lists provided by the user of retrieved by catalogue cross-correlation, Querator is able to query the image archives for the available datasets. At the same time filter lists are retrieved from auxiliary database tables (figure 5).

Results are then analyzed to make sure each of the users' constraints is satisfied. At a later stage, small highly specialized application will provide additional services to the user. These could include, for instance, advanced visualization or statistical analysis tools (figure 6).

ACKNOWLEDGMENTS

The author is particularly grateful to S. Cristiani and to S. Smartt for their invaluable suggestions and their constant testing of Querator. B. Pirenne, M. Dolensky, A. Wicenec and A. Micol helped in understanding the ins and outs of the HST and ESO archives. L. Christensen has been of invaluable help in assessing Querator value in the fields of public outreach and education. M. Kornmesser provided much of the artwork used in posters and papers describing Querator.

REFERENCES

1. R. J. Brunner, S. G. Djorgovski, T. A. Prince, and A. S. Szalay, "Massive datasets in astronomy," *astro-ph/0106481* , 2001.

2. NVO Informal Steering Committee, 2001. Towards a national virtual observatory: Science Goals, Technical Challenges and Implementation Plan. In *Virtual Observatories of the Future*, R. Brunner, S. G. Djorgovski and A. S. Szalay editors, p. 353-372, Astronomical Society of the Pacific, San Francisco.