Presentation of VO-Paris Data Centre and feedback on data archive

Pierre Le Sidaner [(1)], Jean Guibert [(2)], Francoise Tajahmady [(2)], Albert Shih [(1)], Jonathan Normand [(1)], Jacques Vetois [(2)], Ivan Zolotukhin[(1)], Jean Aboudarham [(3)], William Thuillot [(4)], Franck Le Petit [(5)], Daniel Nieto[(1)]

[(1)] DIO-VO, UMS2201 CNRS, Observatoire de Paris
61 Av de l'Observatoire, 75014, Paris, France
EMail: Pierre.LeSidaner@obspm.fr

[(2)] GEPI, UMR8111, Observatoire de Paris
5, place Jules Janssen, 92195 Meudon, France
EMail: Jean.Guibert@obspm.fr

[(3)] LESIA, UMR8109, Observatoire de Paris
5, place Jules Janssen, 92195 Meudon, France
EMail: Jean.Aboudarham@obspm.fr

[(4)] IMCCE, UMR8102, Observatoire de Paris
61 Av de l'Observatoire, 75014, Paris, France
EMail: William.Thuillot@obspm.fr

[(3)] LUTH, UMR8109, Observatoire de Paris
5, place Jules Janssen, 92195 Meudon, France
EMail: Franck.LePetit@obspm.fr

ABSTRACT

VO Paris Data Centre is a collaboration between Observatoire de Paris, IAP, and IPSL to promote the Virtual Observatory and develop data centre activities. Our mission is to share the data centre structure as well as the knowledge and competence in associated technologies for the benefits of the astronomical research community. In this proceeding we present the different components of our archive and their key interoperability features in the Virtual Observatory framework.

The archive is implemented on top of the Active Circle storage virtualisation system that allows one to ensure security, integrity, and high availability of the data, at the same time remaining independent from the hardware evolution by using a specific software solution for data replication in the open format stored on hard drives or tape storage systems.

As a value-added service, we have built the infrastructure to use the computing cluster from the web-portal by running user tasks and codes as RESTful services.

Specific points related to the data distribution from our archives are: (1) the metadata are stored in a relational database (mainly PostgreSQL); (2) we use the standards developed by the International Virtual Observatory Alliance (IVOA) for the data description, output formats, and access protocols. We have developed a web-portal making heavy use of these technologies and accessible at http://voparis-srv.obspm.fr/portal/. Its success lead to our involvement into two European projects aimed at providing access to the data on: (1) planetology, and (2) atomic and molecular database.

We present our archive organization using Virtual Observatory interoperability standards ensuring the smooth data distribution to the large user community.

Keywords: astronomical data bases,: Virtual observatory tools.

INTRODUCTION

VO Paris Data Centre (VOPDC) is a collaboration between all the departments of Observatoire de Paris, Institut d'Astrophysique de Paris (IAP) and Institut Pierre Simon Laplace (IPSL) supported initially by Plan Pluri formation (PPF) and the VO France initiative. VO Paris was built to develop VO at Observatoire de Paris and to disseminate the knowledge, promoting data in the VO with compliant services. Then the structure grew and became also a data centre.

When we started to host data from legacy archives or living projects, the problem of data storage and data preservation became of a great importance. Observatoire de Paris is one of the oldest organizations in astronomy and it was established in 1667. But the problem of long time preservation of numerical data was not clearly defined outside the museum and library. Since its beginning VO Paris team participates in the "Cercle des compétences techniques du CNES" in "preservation et mise à disposition de l'information numérique". Later, the collaboration between the Virtual Observatory and the Open Archival Information System (OAIS) was set. The VO can solve some of the questions raised by the OAIS and the International Virtual Observatory Alliance (IVOA) interest group of data curation and preservation has started to take long time preservation into account.

We present below the feedback from the work started at VOPDC and the IVOA interest group devoted to standardization in data archives.

PRESENTATION OF VOPDC ARCHIVE

As VOPDC is a group formed by researchers coming from different backgrounds, the range of data and services in VOPDC is very broad, from classical 2D images to physical and molecular data, to spectra, tables, planetary profiles, etc.

The responsible body which decides whether to include some new project or organization into data centre is the VOPDC scientific council. In case of positive decision we discuss proper data formats and the complete description of the data model with the team of authors. It is necessary to gather all the metadata as well and to promote the data model in the IVOA. The engineering team then implements VO protocols to access the data and adds a project to the web portal [1] http://voparis-srv.obspm.fr/portal/ to provide a user-friendly access and to advertise VO Paris content. This standardization allows user to access the data not only with well-known client applications (Aladin, VOSpec, TOPCAT, etc), but also in a batch mode via script in his/her favourite programming language. VO protocols were also designed to allow data mining [2] [3], especially because data format is clearly defined and all the metadata are joined. Due to this VO allows direct cross correlation between data coming from different archives.

As the project grew, the amounts of data stored became the first technical challenge crucial for project's success. We have started with classical spectra and images. Due to the volume of digitized survey of the southern hemisphere we have organized the storage at Observatoire de Paris. In the beginning it was a basic classical storage organized as Direct-attached storage (DAS) on Linux platform. Then as the data volume increased, the archive was migrated to the virtual storage that abstracts from the physical devices because a software layer (virtualization) handles the access to the data. Using this technology one can easily change the physical storage transparently to the accessing clients. This technology also proposes a replication of storage as a simple way to avoid access failures.

We have finally chosen Active Circle, a software solution that aggregates heterogeneous devices into a single and scalable storage pool. This hardware-independent storage allows to use tape as well as disk array that differ by access time only. This solution was specially chosen because of the manpower it saves for the classical scenario of important data replication to the tape. Storage virtualization is the only scalable solution that limits the cost per terabyte. For example, Moore's law for storage devices describes hard disk capacity increase of 45% per year on average. The networking performance does not follow this rate, which effectively means that the time scale to move an archive becomes so large that it should be managed by a complete software solution. Hard disk replication is the only reasonable alternative to avoid the denial of services for a long time.

OAIS PRESCRIPTION AT VOPDC

OAIS is a well-known recommendation of an archive management in a standardized way. Following the detailed OAIS recommendation prevents one from missing any important aspects of data preservation and archive organization. The OAIS archive flows are well described in the common schema (Fig. 1).

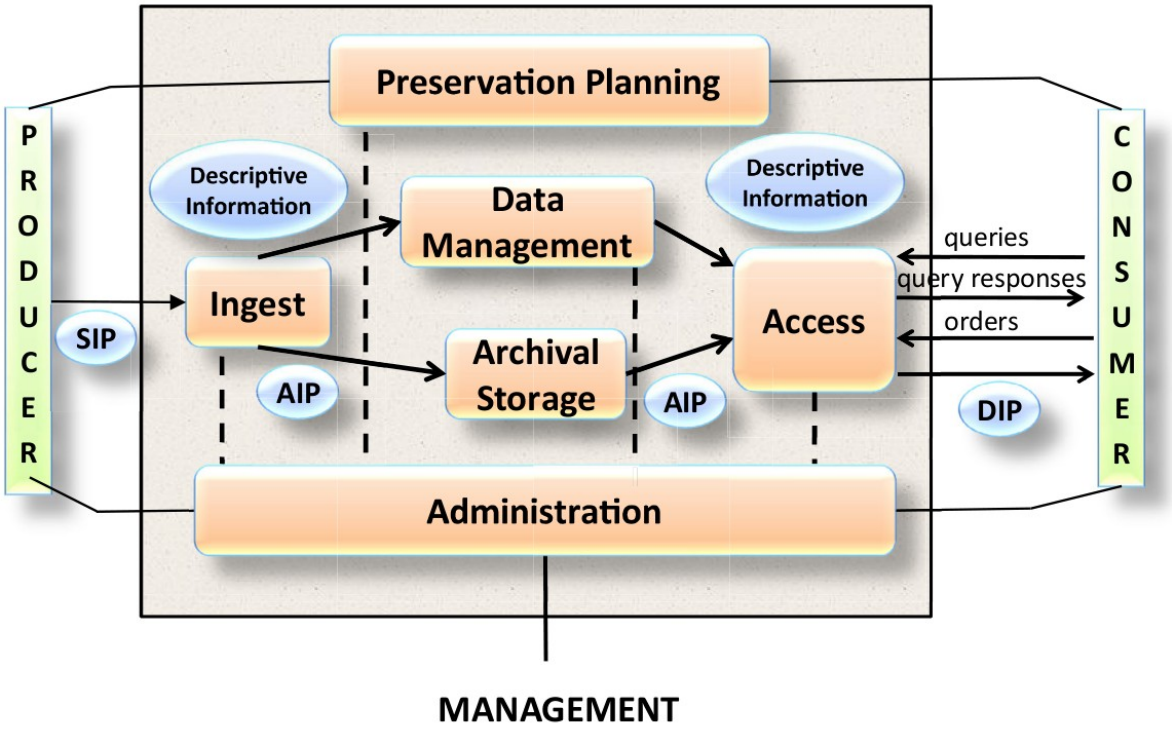

Figure 1: OAIS Functional Entities

taken from CCSDS document

In Fig. 1 the flow of data shows at first the ingestion of data defined by the Submission Information Package. In the VO case, things are simple because we deal with the science-ready package. But VOPDC generally works with ground-based data which are not fully described compared to astronomical ones and therefore ingestion procedure is split in two steps: first, the researcher or the small team, who have acquired the data, reduce them and produce science-ready data, second, the team of VOPDC takes charge of the archive.

From experience we faced with a collection of data with changing format due to evolution of the detector, version of the pipeline and so on. Then with the data comes the difficulty to collect the descriptive information package that is firstly needed to construct the VO service and also to build the OAIS Archival Information Package (AIP). There is always a gap between procedure recommendation and the reality and when an archive asks for the data description, data providers are always busy. The requested information comes slowly and in general getting information is usually an iterative time consuming process. This Submission Information Package

(SIP) is the key point of a long time preservation since it must contain the information for the user community to make all scientific usage of the data. It must take also into account the evolution of the user community language and usage patterns. Typically the neighbouring scientific communities do not yet deal with this types of data but likely will do that in the future.

We compare information description in OAIS and IVOA in five topics below.

- Reference Information

is defined in OAIS as “The information that is used as an identifier for the Content

Information

". The VO is built over an information system called registry that describes services and data collections in a set of records. One block of an information in each record is the IVOA identifier http://www.ivoa.net/Documents/latest/IDs.html constructed as an Uniform Resource Identifier (URI) that refers to organization, publisher, data service, collection. The IVOA identifier is constructed hierarchically as

ivo://<root>.<entity>.<organization>/<service>/<VO Protocol>

The Registry gives correspondence between this identifier and the URL to access the service. To access a data product the URI becomes ivo://<root>.<entity>.<organization>/<service>/<VO Protocol>#internal_identifier. This gives persistent identifier to a simple data that is referenced in the IVOA registry.

- Context Information

is defined in OAIS as "the information that documents the relationships of the Content

Information to its environment". VO registry gives only possibility to link IVOA relationship string to associated services like: mirror of, derived from, served by and so on (see http://www.ivoa.net/Documents/latest/RM.html).

But Context in the sense of OAIS includes also all the information related to the resource, e.g. references to papers describing the data, their acquisition and their processing. This information can also be included in the IVOA registry. There is a work in progress in the IVOA to construct a data model describing the history of a dataset by defining some kind of quality assessment characteristics like seeing, observing configuration of the telescope, data processing steps undertaken. IVOA provenance is mostly context information in OAIS.

- Provenance Information

is defined in OAIS as "the information that documents the history of the Content

Information"

. In VOPDC we try to keep the latest version of processed data with all the context information associated.

- Fixity Information

is defined in OAIS as "The information which documents the mechanisms that ensure that the

content Information object has not been altered

". In VOPDC we use the MD5 hash algorithm with md5sum function to compare existing file signature with the original one stored. This ensures the integrity of the file stored.

- Access Rights Information

is defined in OAIS as "The information that identifies the access restrictions pertaining

to the Content Information

". In VOPDC one of the initial rules was to deal only with public data. But if restriction is not the limitation there is a eternal discussion about the data visibility exposed by the provider. VO is a virtual observatory where all data available are accessed simultaneously in all data centres all over the world. A third party client like Aladin or Topcat usually does this. There is an historical work in IVOA clients to indicate the origin of a data to the user.

User Access to the Archive

Discovery of the archive: The IVOA registries harvest each other and contain the full list of registered VO services. By querying the registry user can obtain a full list of VO services that can contain the type of data requested.

All the access to the archive is fully VO-compliant: discovery of the archive, standardized protocol to find and access the data, data mining and cross correlation possibility, and data access directly from client applications.

The added value of the VO is that independently from the actual archive organization all operations are transparent to a final user. The VO standardization is a layer above the whole system that abstracts the access from specific service details. Moreover VO has developed tools that act as a graphical interface to access data. The VO interest is that all the tools are built to discover all available services and access all VO compliant archives.

IVOA METADATA AND OAIS

The question of metadata standardization is one of the good practices to start long-term preservation of data. The IVOA has adopted OAI-PMH metadata to support harvesting and to standardize metadata. There is great similarity between the PREMIS dictionary (http://www.loc.gov/standards/premis/) and the IVOA Registry field names.

The specifics of astronomy are that the vocabulary keeps a constant designation all over history. If for example supernova does not define a new star, it still describes the same object. This is not the case in human sciences where all vocabulary must be placed in its context. This simplifies the semantic work on metadata.

IVOA has developed a vocabulary and a grammar to describe physical quantities calling Unified Content Descriptor (UCD) and also works on Unit description. All this work is done by the semantic working group that keeps alive the dictionary and extends it to the new domains of the VO like planetary science.

IVOA also works on Data Models to describe observational data with Observation and Characterization data models. Some more generic things like Space Time and Coordinate are there to define coordinate and reference and also more specific one like Spectral Data Model. All these data models give vocabulary and context of application and also like UCD a metadata vocabulary to qualify data.

All the links between IVOA works and long time preservation constraint is discussed in an interest group called “Data curation and preservation”.

CONCLUSION

IVOA fulfil the request of a user community that wants to access in a standard way and format well described scientific data. OAIS tries to describe the proper way to organize and manage a long time preservation archive.

In VOPDC we have implemented the VO standards and now follow the VO evolution. Then started to build a data centre and an archive. We have tried to reuse the job done by IVOA at the same time taking an approach described in OAIS. The main part used from the IVOA in data description for preservation comes from Data Models, Registry and Semantics workgroups.

Acknowledgement: We want to thank the Observatoire de Paris for funding the project and also the VO France project for its support.

REFERENCES

[1] – P. Le Sidaner, J. Normand, A.Shih et al. SF2A- 2008 Proceedings of SF2A, p.91

[2] I. Zolotukhin., I. Chilingarian ., 2011 A&A 526, A84

[3] I. Chilingarian, V. Cayatte, Y. Revaz et al. 2009, Science 326, 1379