Historical lessons, inter-disciplinary comparison, and their application to the future evolution of the ESO Archive Facility and Archive Services

Paul Eglitis [(1)], Dieter Suchar [(1)]

[(1)] ESO

ESO, Karl-Schwarzschild Str. 2, 85748 Garching near Munich, Germany

EMail: paul.eglitis@eso.org

dieter.suchar@eso.org

ABSTRACT

The ESO/ST-ECF Science Archive Facility is a joint collaboration of the European Organisation for Astronomical Research in the Southern Hemisphere (ESO) and the Space Telescope - European Coordinating Facility (ST-ECF). The Archive Facility has developed in response to the evolution of observatories, instruments, the characteristics of raw data, processed data products and associated files as well as specific constraints pertinent to the data flow necessary to support scientific programs carried out on world-leading observatories such as the ESO Very Large Telescope. Since 1998 the Archive Facility has grown from a 0.5 TB capacity optical disk system into a state-of-the-art PETABYTE class archive based on spinning disk technology in a dedicated data centre designed to optimise technical maintainability. The historical development has been in response to the rapidly-changing and domain-specific requirements of astronomical data acquisition, archiving and distribution and this has led to a sophisticated and multi-layered set of services presented to the end user. The lessons learnt throughout the archive growth are now mature and can be applied to future development in shaping new storage architectures, homogenising end-user interfaces and providing value-added services. In addition, since the need and use of large-scale archive solutions is widespread throughout science and technology and in commercial enterprises, much can also be learnt from experiences in other disciplines. This paper will provide an overview of the current Archive Facility, describe lessons learnt from the historical development and inter-disciplinary comparison, and detail the plans for the next steps in the evolution of the Archive Facility.

Keywords: ESO Primary and Secondary Archive, Astronomy, Data Preservation, Long Term Archiving, Interoperability, Storage Technology, Lessons Learnt, Inter-disciplinary

INTRODUCTION

The ESO Science Archive was conceived in 1991 [1] with shelved-storage of optical platters and operator co-ordinated media-reading at a dedicated machine on request. With the advent of the ESO Very Large Telescope (VLT) in 1998, rapid development of the ESO Archive had to take place [2] and the archive systems subsequently evolved from a 2 CD jukeboxes configuration into a fully redundant facility with data managed on both spinning disk and near-line tape. A multi-petabyte capacity has been achieved through scalability by upgrading storage media to higher density formats. With a decade of successful operation now demonstrated, a review of the archive development in that period is important to support the next steps in the archive evolution. Reference is also given to similar systems from other scientific fields in an effort to also learn from interdisciplinary comparison.

ESO ARCHIVE SYSTEMS OVERVIEW

The ESO Science Archive was established to preserve and exploit observations acquired by the instrumentation at the La Silla and Paranal Astronomical Observatories in Chile, including the APEX

sub-millimeter telescope on Llano de Chajnantor where the Atacama Large Millimeter/sub-millimeter Array (ALMA) is also under construction. The data holdings at ESO are complemented by observations from the Hubble Space Telescope (HST) through the ESO collaboration with the Space Telescope European Coordinating Facility (ST-ECF). A further mission of the ESO Archive is to receive scientific data products released by Principal Investigators on the completion of research with ESO observations. Figure [1] provides a simplified schematic of the archive facility in the context of the data flow from the La Silla and Paranal Observatories taken as an example.

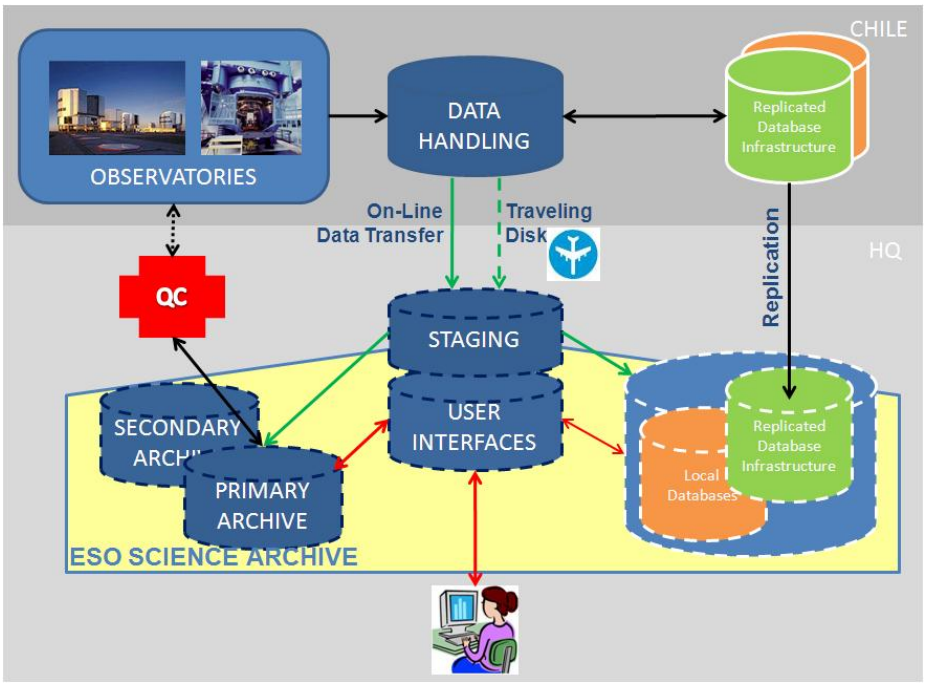

Figure. 1: Simplified schematic of the ESO Science Archive (yellow shaded region) and its context in the overall data flow, between observatories in Chile and the ESO Science Archive located at HQ in Garching, Germany.

Data originating from the ESO telescopes and their varied detectors comprise observations, calibration files and auxiliary information such as operations' logs. The end-to-end data flow includes processes taking place on-site at the observatory location and at ESO Head Quarters (HQ) in Garching. On-site data management includes the extraction of header information into local databases. The contents of the database are immediately replicated to HQ, providing the first ingestion of metadata to the ESO Archive. Historically the remoteness of the observing location and limitations in the availability of network bandwidth led to the full datasets always being shipped to HQ by initially optical disk and then hard-disks. In 2008 an on-line Scientific Data Transfer (SDT) system became available. Shipments of disks continue to maintain redundancy and manage the large data volumes produced by some of the instruments in operation and those expected in upcoming programs such as the survey telescopes VISTA (Visible and Infrared Survey Telescope for Astronomy) and VST (VLT Survey Telescope).

An important internal user of the archive are Quality Control Scientists (QC), who in the analysis of new data, metadata and calibration files are able to provide direct feedback to the observatories on the last set of measurements. The QC scientific team pipe-line process data and return advanced data products to the archive further increasing the richness of the stored volume.

On ingestion to the ESO Archive, data is written to a Primary Archive based on spinning disk technology and a copy is also stored in a Secondary Archive comprising a Sony PetaSite tape library. The data storage across all archive components is managed by the ESO-developed globally distributed file management system NGAS (Next Generation Archive System) [3] with internal reference to all holdings in a central NGAS database. Descriptive metadata for all archived entities linked to a unique NGAS file id is held in a set of Sybase Databases that support the web interfaces for user-queries and the systems for data retrieval; on-line through ftp or by writing to media with operator support. The NGAS system is able to track all disks either mounted in a machine or in transit in a Diplomatic Bag

between Chile and HQ. The management system intrinsic to NGAS is implemented as an HTTP server and includes many modern archive management software features such as periodical self-verification of archive holdings through checksum analysis, sleep mode functionality and features to allow the mass migration of data and support scalability as well as the ability to process bulk data within the archive itself.

LESSONS LEARNT

Storage Technology

The initial hardware configuration of the ESO Scientific Archive in 1998 consisted primarily of 2 CD jukeboxes hosting approximately 1000 CDs with a total capacity of 0.5 TB. In 2005, the entire ESO science data archive was based on a heterogeneous set of digital media and media management systems, including: off-line CDs and DVDs that must be manually mounted for reading, DVDs mounted in four high capacity jukeboxes and small RAID systems attached to Sun Solaris systems. The main ESO data holdings were stored in a Linux-based hard disk farm consisting of 100x200 GB PATA disks, 25x250 GB SATA disks, and 48x400 GB SATA disk. Over the past 10 years the data volume of the ESO Scientific Archive has been growing to 110 TB located on 432 SATA hard disks currently located on 18 NGAS systems. With the ESO archive facing the challenge of data volume growth to over 1 PB over the next 4 years, a Petabyte Class Archive [4] at ESO has been designed, including the Primary Archive, an online copy with fast access of the entire dataset; the Secondary Archive, a full second copy of the Primary Archive at a physical different location on SAIT tapes; a Compute Stack with CPU, memory and bandwidth on demand including a high performance storage, the accessibility of the scientific data through internal and external network and the media production for data distribution, leading to the configuration depicted in Figure [2].

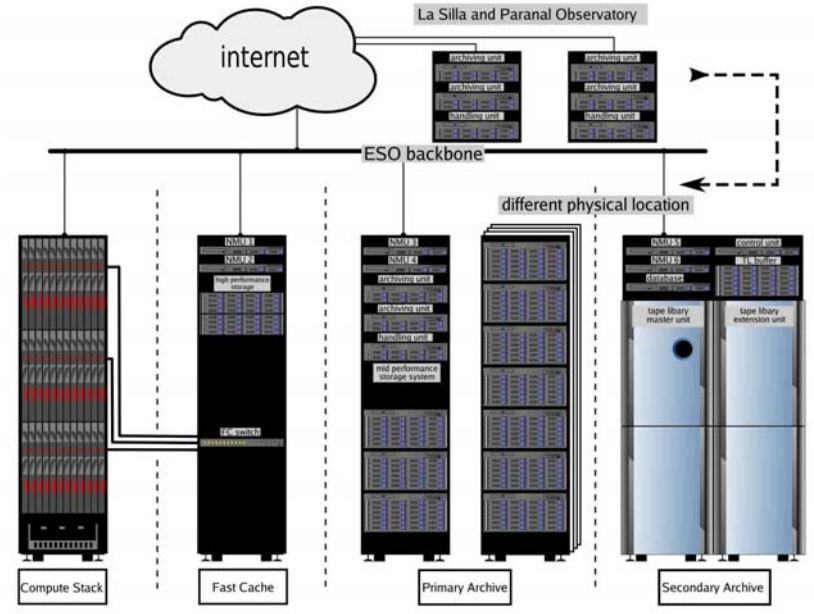

Figure. 2: The operational IT infrastructure for the ESO Archive.

In addition, a complex Data Centre environment had to be designed and constructed to host and to support the various components of the Petabyte Archive in a reliable and flexible manner with respect to safety and availability. Whereas the initial costs to set up a data centre are high, the expenditure has been recouped over time since upgrades of the storage technology are efficient and economic relying on the exchange of disks to a higher density. Figure [3] illustrates the change in storage technology (CD, DVD, and Disk), the volume of the unit media used (blue line) and the number of racks required to hold 100TB of non-compressed data (footprint: red line). The changes in storage from a CD jukebox to DVD jukebox, and then to disk arrays all had an additional cost overhead (timing of these changes marked by red asterisks). Once an archive solution on spinning disks was established upgrades involved changing

to higher density disks and were comparatively inexpensive (green asterisks) – i.e. always considering the cost as a fraction of the total cost of the archive at any time, rather than making absolute comparisons.

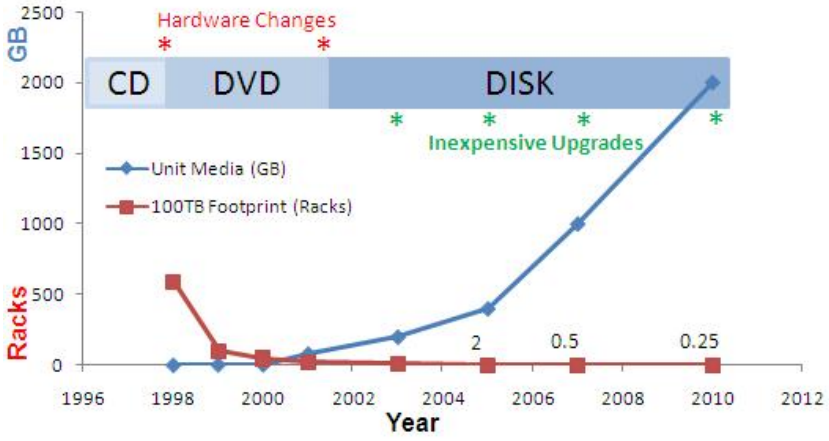

Figure. 3: The change in storage media, unit media size and the corresponding archive physical size represented as the number of racks required to host 100 TB of data.

Over the next decade, the volume of stored data will continue to increase. The current holdings comprise raw, processed and calibration data from the La Silla and Paranal Observatories (LSO, PO) with a data volume of 80 TB in 2008 arising principally from the instrumentation suite on the ESO Very Large Telescope (VLT).

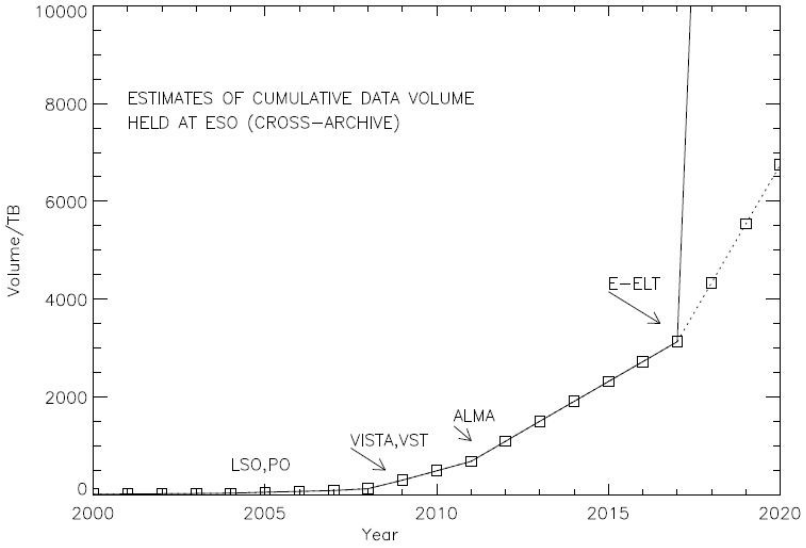

Figure. 4: This figure illustrates the change in the total data volume stored in the ESO Data Centre over the last decade and an estimate of the growth in the total volume over the next decade.

At the time of writing the survey telescopes VISTA and VST are due to come into operation during 2009 and 2010 leading to a ten-fold increase in data rate. In 2012 ALMA is expected to contribute a further 200 TB each year. The ALMA data will be stored in a dedicated archive, one of the ALMA Regional Centers (ARCs), hosted at ESO. Therefore, from 2010 to 2017, an increase of up to 400 TB/year is expected with the operation of both survey telescopes and ALMA. With the current data storage configuration this corresponds to the addition of approximately 1 rack per year given that 2 TB disk technology is available (i.e. in Figure [3] 0.25 racks hold 100 TB in 2010, note: this figure increases to 0.4 for 100 TB racks if all the data is stored in a RAID 5 configuration on disk). This illustrates the scalability of the current solution, but nonetheless a full cost benefit analysis and review of end user requirements is being undertaken to assess future options, including switching to different technology, e.g. tape or even outsourcing part of the data storage. Such cost benefit analysis must take into account

the effective cost of the change to- and design of- a new system, the subsequent operations' costs and also those associated with future migrations and upgrades as compared to the costs of maintaining the current solution, and cannot rely on comparing media unit prices.

The next leap in data volume is expected to coincide with the European Extremely Large Telescope (E-ELT) from 2018 at the earliest. The E-ELT is in the Detailed Design Phase and exact figures for the expected data volumes are uncertain. Based on preliminary analysis [5] possible instruments running on the telescope may deliver up to 46.4 TB/night when running for one hour in burst mode that leads to an increase in the total volume by over 17.4 Petabytes each year after 2017 or when the E-ELT becomes operational (this is indicated by the continuous line that leaves the graph boundary before 2018). If we assume that any particular instrument only operates 5% of the total operational time then the dotted line gives an estimate of the predicted data volume after 2017.

Petabyte scale archives are already in operation throughout particle physics research. At Fermilab the Enstore and dCache [6] data management system has been developed for a multi-petabyte data volume. Storage is on tape libraries and scalability is achieved by periodically switching to higher density tapes. The possibility of handling tens of petabytes per year is already anticipated for data from the Large Hadron Collider at CERN by the Cern Advanced STORage manger (CASTOR) [7] with storage based on tape silos with a NAS based disk cache. It is interesting to note that large scientific projects have often needed to develop in-house software solutions to meet the scale and special requirements of the data storage and management needs, e.g. Enstore at Fermilab, CASTOR at CERN, Google File System for Google [8], and EADS introduced SatSTORE as their own storage management layer in commercial HSM systems [9]. Typically, costs are mitigated with commodity hardware and most expenditure is on knowledge management with the use of industry standard databases.

Database and Web Services

The original metadata schema was developed in the mid 90s for the ESO New Technology Telescope (NTT) as a pilot for the VLT and then was updated over 1999-2000 in support of the first VLT operations. User access to data was granted by web query forms based on Perl-CGI exploited by an in-house toolkit; WDB [10]. The requirements on the system quickly outgrew the initial design and instrument specific database tables and instrument specific query forms were introduced during 2001 to 2003. The next evolutionary step was the introduction of Virtual Observatories (VO) services from 2004 onwards. The Virtual Observatory [11] is the set of e-Science projects that endeavours to enable astronomical research through the internet based on interoperable software applications and data sources. In moving towards compliance with VO standards and protocols additional internal database tables and software processes were introduced inside the ESO archive system. While the VO is leading to a global standardization of data sharing within astronomy the initial impact on the data archive was to produce additional complexity in the system architecture. The highly complex set of services and supporting infrastructures (databases and servers) is under complete revision. The first step has been to design a new keywords repository [12] (now in place) to contain all the required metadata to manage ESO observations. A further abstraction is under development to provide a unified architecture to support all interfaces to the ESO Archive. Benefit is also being taken of new software technology to generate reusable and more maintainable software components and provide services to end users more efficiently.

Domain Drivers and Interoperability Considerations

Reviewing the history of the Archive design described above it is apparent that the underlying software systems and supporting database structures have had to change rapidly in response to the need for end results, changes in telescope instrumentation and international projects such as the VO, with much system complexity arising as a by-product. The business objectives of an organization such as ESO are to provide the best possible data for scientific research. There is always the possibility to modify or just reconfigure the instrumentation and bring in new instrumentation. Thus, it is not unusual that any supporting data infrastructure must adapt as well and this leads to a continual system development overhead. Lessons have to be learnt as the observing facilities evolve and engineering changes are made

to simplify the supporting systems and make them more amenable to the changes that will occur, while also continuing to address new User Requirements.

In other disciplines, there are more constraints on the operations of a system and often more methodical system engineering is observed. For example, Earth Observation missions are typically driven by the need to make continuous observations for monitoring and observing changes. The data is used by scientists but also operational services, e.g. national meteorological services and commercial enterprises, that demand (and may pay for) a certain quality of service. Space missions are driven directly by scientific goals but the nature of the project requiring the launch and operation of a satellite need careful planning and standardization of processes to ensure the mission success that very often is a single opportunity. For a ground-based astronomical observatory, science is the main goal and the scientists and engineers developing the observation techniques are able to keep pushing the boundaries and data management systems, including archives, have to be flexible not to stifle the creativity of the scientist that will often have direct access to the data acquisition facility.

Conversely, other disciplines can learn lessons from the data management policies present in astronomy research. A very successful standardization achieved in the astronomical community has been the long term use, acceptance and standardization of the FITS (Flexible Image Transfer System [13]) standard as a native format for many years. The wide use of a single format is an effective method for the long term preservation of data with regards to maintaining access and use and also provides a basis for interoperability in that data can be easily handled at different data centers. The Virtual Observatory projects have also adopted a common approach to the exchange of data in defining the VOTABLE standard [14].

In other domains such as Earth Observation, the challenges for interoperability include the need to coordinate systems and standards across many different disciplines and where each discipline has naturally already adopted their own set of formats, protocols and best practices for data management. The GEO sponsored Global Earth Observation System of Systems [15, 16], addresses this problem by working to establish agreed interoperability arrangements [17] and compiling registries of all operating components and systems. Clearly, there are many fundamental issues connected to managing data, enhancing interoperability, exploiting and preserving information that are shared between different domains and lessons can be learnt in each field by studying the successes, pitfalls and main drivers of each discipline.

CONCLUSION

Lessons have been learnt in the development and operations of the ESO Science Archive with respect to the system design, storage policies and its role as a data provider in the modern context where internet applications strive to exploit data in a transparent manner irrespective of the location or nature of the source. The design of the archive (and supporting data centre) is found to be scalable and ready to address the future requirements of the next set of ESO facilities, but nevertheless systems are under review and in the process of re-design to improve services and affirm future decisions on storage technology selection.

REFERENCES

[1] - F. Ochsenbein: The ESO Archive Project, Databases and On-line Data in Astronomy, edited by Miguel A. Albrecht and Daniel Egret ISBN 0-7923-1247-3; 1991 Astrophysics & Space Science Library vol. 171, p. 107 (1991).

[2] - R. Albrecht ,M. Albrecht, M. Dolensky, A. Micol, B. Pirenne, and A. Wicenec: The VLT/HST Archive Facility. in: U. Grothkopf et al. (eds.), Library and Information Systems in Astronomy III, ASP Conference Series, Vol. 153, p. 261, (1998)

[3] - A. Wicenec, J. Knudstrup, ESO's Next Generation Archive System in Full Operation, The Messenger, 129, 27, (2007).

[4] - D. Suchar, J. S. Lockhart, A. Burrows: Operating a Petabyte Class Archive at ESO. Observatory Operations: Strategies, Processes, and Systems II, edited by Roger J. Brissenden, David R. Silva, Proc. of SPIE Vol. 7016, 70160N, 0277-786X/08/$18 doi: 10.1117/12.789263, (2008)

[5] - ESO Internal Document, E-ELT Programme estimated data flow rates. E-TRE-ESO-750-0282 Issue 1.1 (2008)

[6] - G. Oleynik, B. Alcorn, W. Baisley, J. Bakken, D. Berg, E. Berman, Chih-Hao Huang, T. Jones, R. D. Kennedy, A. Kulyavtsev, A. Moibenko, T. Perelmutov, D. Petravick, V. Podstavkov, G. Szmuksta, M. Zalokar: Fermilab's Multi-Petabyte Scalable Mass Storage System, DOI Bookmark: http://doi.ieeecomputersociety.org/10.1109/MSST.2005.1, (2005).

[7] - G. Lo Presti, O. Barring, A. Earl, R. M. Garcia Rioja, S. Ponce, G. Taurelli, D. Waldron, M. Coelho Dos Santos: CASTOR: A Distributed Storage Resource Facility for High Performance Data Processing at CERN. Mass Storage Systems and Technologies, Digital Object Identifier: 10.1109/MSST.2007.4367985 (2007).

[8] - S. Ghemawat, H. Gobioff, S.-T. Leung: The Google file system, December 2003 SOSP '03: Proceedings of the nineteenth ACM symposium on Operating systems principles (2003). Also published in: December 2003, SIGOPS Operating Systems Review, Volume 37 Issue 5 and at http://labs.google.com/papers/gfs-sosp2003.pdf

[9] - G. Petitjean, P. Marchadour: Long-term preservation and advanced access services to archived data:The approach of a system integrator, PV2002 Conference Proceedings, (2002).

[10] - B. F. Rasmussen , WDB---A Web Interface to Sybase Astronomical Data Analysis Software and Systems IV, ASP Conference Series, R.A. Shaw, H.E. Payne, and J.J.E. Hayes, eds., p. 72, Vol. 77, (1995).

[11] - P. J. Quinn: Virtual Observatories: the Future of Astronomical Information. Proceedings of Library and Information Services in Astronomy IV, P. 55-58, (2002).

[12] - M. Vuong, A. Brion, A. Dobrrzycki, J.-C. Malapert, C. Moins: Applications of the ESO metadata database. Proceedings of SPIE, the International Society for Optical Engineering ISSN 0277-786X CODEN PSISDG , Observatory operations : strategies, processes, and systems No2, Marseille , FRANCE (2008)

[13] - R. J. Hanisch, A. Farris, E. W. Greisen, W. D. Pence, B. M. Schlesinger, P. J. Teuben, R. W. Thompson, and A. Warnock: Definition of the Flexible Image Transport System(FITS). Astronomy & Astrophysics, 376:359–380, (2001). For the latest version, see http://fits.gsfc.nasa.gov/fits documentation.html.

[14] - VOTable Format Definition, Latest Reference: http://www.ivoa.net/Documents/latest/VOT.html

[15] - GEO, 2005a. Global Earth Observation System of Systems (GEOSS) 10-Year Implementation Plan, ESA Publications Division, the Netherlands, Bruce Battrick, ed. ISSN No.: 0250-1589, ISBN No.: 92-9092-495-0, (2005).

[16] - GEO, 2005b. Global Earth Observation System of Systems (GEOSS) 10-Year Implementation Plan Reference Document, ESA Publications Division, the Netherlands, Bruce Battrick, ed. ISSN No.: 0379-6566, ISBN No.: 92-9092-986-3, (2005).

[17] - S. J. S. Khalsa, P. E. Eglitis: How the GEO Standards and Interoperability Forum (SIF) Advances the Interoperability Goals of GEOSS, In Proceedings of the 33rd International Symposium on Remote Sensing of Environment—Sustaining the Millennium Development Goals, May 4-8 2009, Stresa, Italy, (2009).

Acknowledgement. The authors would like to thank Lowell Tacconi-Garman, Andreas Wicenec, John Lockhart, Adam Dobrzycki, Nathalie Fourniol, Nausicaa Delmotte, Alberto Micol, Fernando Comerón and Uta Grothkopf for many helpful discussions in the preparation of this paper.