Archive Functional and User Requirements

Authors:

Approved By: (institutes)

L.A. Antonelli INAF-OAR/ASDC

M.Capalbi INAF-OAR/ASDC

A. Carosi INAF-OAR/ASDC

F. Lucarelli INAF-OAR/ASDC

A. Di Paola, S.Gallozzi INAF-OAR

S. Lombardi, V. Testa INAF-OAR

History:

1.0	2012-10-26	First Version
2.0	2012-12-29	Second Version
2.0.3	2012-12-29	Inserted some comments by G. Lamanna
2.0.4	2013-01-14	Inserted comments by J. Schwarz. Further revision by F.L.
2.1.0	2013-01-30	Inserted comments by R. Smareglia and C.Knapic (INAF -OATs). Further revision by A.C.

Distribution: DAFA Group

Contents

1 Introduction

1.1 Scope of the document

The scope of the archive system lies on making available and accessible the high level data products of the CTA observatory to the international scientific community. This document describes the high level general functional and user requirements as well as the use cases for the CTA data archive.

1.2 Purpose of the document

The purpose of the document is to serve as a first fundamental tool in building a suitable and efficient archive software architecture.

2 List of acronyms

AUX	Auxiliary
CEIN	Computing and E-Infrastructure Work Package
CTA	Cherenkov Telescope Array
DAFA	Data Archiving, Format and Analysis Work Package
DAQ	Data Acquisition
DB	Database
DL	Data Level
HL	High Level
HK	Housekeeping
IACT	Imaging Air Cherenkov Telescope
IRF	Instrument Response Function
IVOA	International Virtual Observatory Alliance
LL	Low Level
MC	Monte Carlo
ML	Medium Level
MWL	Multi wave-length
RT	Real Time
TAC	Time Allocation Committee
TBC	To Be Confirmed
TBD	To Be Defined
VHE	Very High Energy

3 Glossary

According to [5] the following data level naming convention is used throughout the document.

DL0 files: RAW data files produced by the DAQ (possibly FITS formatted) and used as first input for the standard analysis and reconstruction pipelines.

DL1 files: Calibrated data plus image shape parameters.

DL2 files: Reconstructed shower information.

DL3 files: List of gamma-candidate events for science analysis.

DL4 files: High-level scientific products (sky maps, lightcurves, spectra).

In more details, the current CTA data flow foresee the steps summarized in the following table

	Data Level	**Short Name**	**Description**
LL	Level 0	RAW	The raw PMT tube and ancillary telescope data. *One file per camera.*
	Level 1A	TEL_ EVENTLIST	PMT calibrated data of Level 0. *One file per camera.*
	Level 1B	TEL_ EVENTLIST	Level 1A data with the corresponding image shape parameters. *One file per camera.*
ML	Level 2A	TEL_ EVENTLIST + EVENTLIST	It is the merging of all cameras Level 1B data including the reconstructed shower information. Multiple EVENTLISTS are foreseen: one per reconstruction method. *One file per observation.*
	Level 2B	REDUCED(TEL_EVENTLIST + EVENTLIST)	Reduced version of Level 2A data where pixel information is removed. *One file per observation.*
	Level 2C	EVENTLIST	Contains only shower-base information and does not include any single-camera entry. *One file per observation.*
HL	Level 3	REDUCED EVENTLIST	Standard high level data format for scientific analysis. Level 2C data are reduced according to a gamma/hadron separation cut. These files will contain a list of gamma-candidate events. *One file per observation.*
	Level 4	HL PRODUCT	High level binned data products as sky maps, spectra and lightcurves.
	Level 5	OBS PRODUCT	Observatory level files such as CTA survey sky maps, CTA source catalog.

4 General overview

The CTA observatory is expected to produce an amount of data ranging from 1 up-to 10 PB per year including both observation and calibration data. Furthermore, being CTA operated as an Observatory, the acquired data should remain available and easily accessible to the community for a long time, well after the project ends, so that many other scientists could reproduce the results and do new science with those data. The principal components of the CTA archive are 1) Data and Metadata; 2) the software used for managing the archive and the related database(s); 3) hardware (i.e. storage device, RAM and CPU); 4) services to access data.

The CTA Data Archive will be formed by two separate units:

- An **on-site archive**, which will contain raw data files, the engineering files (HK and AUX files), a set of pre-defined IRFs and the products of the On-site and Real-time Analysis, including preliminary science products;
- An **off-site archive**, containing the raw data, the Data Reduction products and the high-level science products as well as RT analysis results, engineering archive and Monte Carlo simulations.

5 General Archive Content

Each telescope of the CTA array(s) will produce data of different types: science data, calibration data, housekeeping. The array itself will produce ancillary data regarding the array configuration. Other ancillary data will be also collected from different subsystems providing all the information relevant to the observations (e.g. sky brightness, wind, humidity, etc.). Monte Carlo simulations for the reduction and calibration of the acquired data will be also massively produced and archived.

The complete archive of the CTA observatory will be composed by five main logical levels:

- **Raw Data archive.** It contains the definitive archive of the RAW data produced by the arrays and all data products generated from the reduction pipeline excluding DL3.
- **Science archive(s).** It includes reduced DL3 data and DL4 products (sky maps, light curves etc.) as well as RT analysis results.
- **Engineering archive.** It contains the archive of calibration data as well as the housekeeping and auxiliary information.
- **Monte Carlo archive.** It contains all the Monte Carlo events simulated for the different array configurations.
- **Final Products archive.** High-level scientific results to be compared to other products from other observatories at other wavelengths.

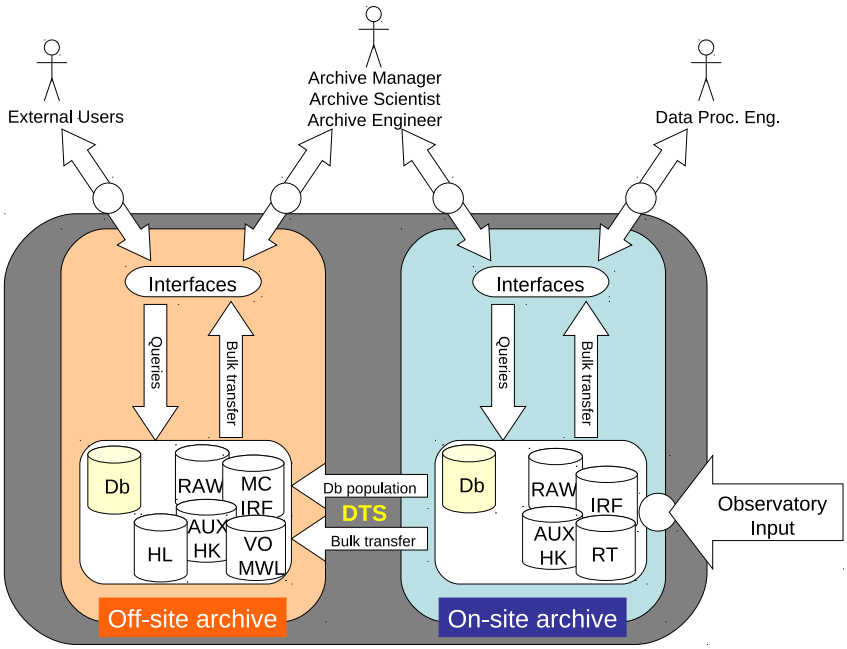

Figure 1: The CTA Data Archive logical structure and the foreseen accessibility for different users.

6 Archive users

Three different general users of the CTA Observatory have been identified [2]. In order to define the general Archive requirements, within the internal users, dedicated user profiles devoted to the archive management have been pointed out:

- **CTA Internal users:** include different user profiles as defined in [2]. Moreover, a dedicated Hardware Engineer profile must be foreseen in order efficiently monitor and check hardware performances. Users in this category have unlimited access to the CTA archive. The archive-related profiles are:
  - *Archive Manager:* responsible for configuration and archive integrity.
  - *Data Archive Engineer:* responsible for all raw (raw, calibration, engineering, MC...) data ingestion, backups and delivery of data to higher level archive.
  - *Data Processing Engineer:* supervise and operate the processing pipeline to calibrate the observation at each site (responsible for the quality of the data processing software).
  - *Archive Scientist:* responsible for quality and integrity of the archives and for the correct distribution of the science products (responsible of the data validation)

- **Guest Observer:** scientist that submits an observing proposal.
- **Archive User:** scientist that uses the science archive in a research project.

7 Use cases

Some of the following use cases for each of the CTA archive users are already described in [2]. We will mention them here again for sake of comprehension.

- **CTA Internal Users**
  - *CTA Archive Manager*
    1. The Archive Manager supervises the quality and integrity of the whole CTA data archive.
    2. The Archive Manager notifies anomalies and failures in the data archive to the Data Processing engineer.
    3. The Archive Manager monitors and manages the housekeeping and auxiliary information storage (TBC).
    4. The Archive Manager manages the remote access to the data in the different ways foreseen for the data distribution (web access, access to VO protocols, dedicated software applications etc.)
    5. The Archive Manager makes reference to scheduler and PI's policy and permission.
  - *CTA Archive Scientist*
    1. The Archive Scientist makes available the RT scientific products to the Guest Observers.
    2. The Archive Scientist makes the data available to the Guest Observers after the requested observation is completed.
    3. The Archive Scientist releases, according to the Data Access Policy, the data and the high-level products to the scientific community.
  - *CTA Archive Engineer*
    1. The Archive Engineer manages and populates the database of metadata.
    2. The Archive Engineer monitors and validates the generation of the Raw Data Archive.
    3. The Archive Engineer supervises the generation of the Raw Data Archive backups.
    4. The Archive Engineer monitors the transfer of raw data to the Data Reduction Unit.
    5. The Archive Engineer monitors the insertion of MC simulation, IRFs and reconstruction pipelines products into the archive.
    6. The Archive Engineer monitors the insertion of housekeeping and calibration data into the archive.
  - *CTA Data Processing Engineer*
    1. The Data Processing engineer produces all the intermediate and high-level data products to be stored in the archive.

2. The Data Processing engineer reprocesses raw data to generate a new version of science data products due to changes in the pipeline (e.g., due to updated calibration information.).

– *CTA calibration scientist*

1. The calibration scientist monitors the storage of the Monte Carlo simulations and Instrument Response Functions (IRFs) to be used for the data reduction.

– *CTA Array operator*

1. The Array Operator accesses the HK and AUX archive information to monitor array performances.

– *CTA Hardware Engineer*

1. The Hardware Engineer accesses the HK and AUX archive information to monitor performances and status of relevant hardware components.

- **CTA External Users**

– *Guest Observer/Archive user*

1. Guest Observer retrieves data according to the Data Delivery Policy accessing the archive through different interfaces.
2. Guest Observer accesses to RT data products in the dedicated archive area.
3. Archive User accesses the public archive data through different interfaces.

8 Requirements

In what follows, each requirement will be defined according to the following scheme:

- **ID:** Unique categorization of the requirement in the form of CATEGORY.NNN, where CATEGORY denotes the major requirement category and NNN is a unique number identifying the requirement. In our case, CATEGORY==AR (Archive Requirement).
- **Justification:** Description of the scope and constraints of this requirement.
- **Impact:** Relevance of the requirement according to performance, cost and priority.
- **Interfaces:** Cross-correlation of the requirement with other CTA Working Group. Most of the interfaces are related to the software pipelines defined in [3].
- **Validation:** Description of how and where the requirement is planned to be validated.

Whether not specified, we refer to CTA archive as the off-site archive.

8.1 Functional requirements

AR1.10 The CTA archive will allow the access and handling of scientific and calibration data by the different archive users (defined in §6) at different levels.

Justification Being an open observatory, CTA will provide, for the first time, open access to VHE data to the astronomical community. To this end, data storage and access to the archive must be as efficient and easy as possible.

Impact `performance:` high; `cost:` high.

Interfaces Data Acquisition Unit, Standard Reconstruction Pipeline, Realtime Reconstruction Pipeline

Validation

AR1.11 CTA on-site archive will be accessible only by the internal users.

Justification The CTA on-site archive will temporary store and buffer data of the last TBD days. External users, should not be allowed to access the on-site archive for safety and bandwidth occupancy reason. Moreover, following **AR1.50**, there is no special motivation for granting access to external users.

Impact `performance:` low; `cost:` low.

Interfaces

Validation

AR1.12 CTA archive will guarantee data storage and access in accordance with standard guide lines regarding the long-term preservation of scientific data.

Justification Data should maintained by the CTA Data Archive for a sufficiently long period and must be preserved for at least 10 years after decommissioning of the CTA instrument (motivated by **A-USER-0120** in [1]).

Impact `performance:` high; `cost:` high.

Interfaces Standard Reconstruction Pipeline, Realtime Reconstruction Pipeline

Validation

AR1.20 The off-site archive will permanently store all RAW data produced by the CTA Observatory.

Justification The RAW data should be maintained and available to all members of the CTA project (TBC) in order to run reduction and analysis pipeline at any time, if necessary (e.g, after new calibrations and/or new software releases).

Impact `performance:` high; `cost:` high.

Interfaces Standard Reconstruction Pipeline, Realtime Reconstruction Pipeline

Validation

AR1.21 The on-site archive will temporarily store all RAW data produced by the CTA Observatory.

Justification Data from the array will be produced by the DAQ Unit locally. For this reason, the on-site archive must guarantee the RAW data preservation until the definitive transfer to the off-site (permanent) archive.

Impact `performance:` high; `cost:` high.

Interfaces Realtime Reconstruction Pipeline

Validation

AR1.21.bis The on-site archive will temporary store all RAW data produced by the CTA Observatory as well as part of the engineering archive.

Justification Data from the array will be produced by the DAQ Unit locally. For this reason, the on-site archive must guarantee the RAW data preservation until the definitive transfer to the off-site (permanent) archive. Moreover, engineering data must be stored for health monitoring purpose. Not necessarily, all the engineering information will be transferred to the off-site archive

Impact `performance:` high; `cost:` high.

Interfaces Realtime Reconstruction Pipeline

Validation

AR1.22 The on-site archive data must be sent to the off-site archive on a TBD period.

Justification The on-site archive is not expected to have the same storage capacity as the off-site archive. Periodic transfer of data and data products must be implemented in order to guarantee enough disk space each observation night.

Impact `performance:` high; `cost:` high.

Interfaces Standard Reconstruction Pipeline, Realtime Reconstruction Pipeline

Validation

AR1.30 CTA archive will store all the intermediate data products obtained with the latest software version and calibrations.

Justification To keep track of the all analysis reduction steps, it is important to store all the intermediate products. *This possibility is strictly related to the available amount of storage and to the expected data amount foreseen by the DAQ and data reduction working group.*

Impact `performance:` high; `cost:` high.

Interfaces Standard Reconstruction Pipeline, DAQ and RECO working group, Realtime Reconstruction Pipeline

Validation

AR1.30.bis CTA archive will store intermediate data products from DL1b on in the off-site archive (TBD/TBC). Calibrated DL1a files will be stored and made available to CTA members to perform quality controls for a TBD period in the on-line archive (TBC).

Justification Due to the large amount of data produced by the CTA, it might be foreseen to temporarily archive intermediate products of the data reduction chain. *Depending on the expected ratio between data reduction time and available storage amount one of the two requirements AR1.30 and AR1.30.bis can be applicable.*

Impact `performance:` high; `cost:` high.

Interfaces Standard Reconstruction Pipeline, Realtime Reconstruction Pipeline

Validation

AR1.31 CTA archive will guarantee the possibility to re-reduce the data products generated by oldest software version and calibrations.

Justification To preserve storage capabilities, oldest data reduction products will be removed. However, by using proper metadata, there will be the possibility to recover and reproduce data with oldest calibrations and software versions.

Impact `performance:` medium; `cost:` medium.

Interfaces Standard Reconstruction Pipeline, Metadata working group

Validation

AR1.40 CTA archive will make available high-level data (DL3, DL4) for any archive users through web interface/database query on multiple parameters (TBD).

Justification The archiving, long-term preservation and distribution of scientific data is of crucial importance in order to maximize the scientific return of the observatory. To realize this goal, it is necessary to allow an easy and efficient data discovery and download and to make available to the users facilities for data visualization and analysis.

Impact `performance:` medium; `cost:` medium.

Interfaces Standard Reconstruction Pipeline

Validation

AR1.50 CTA archive will store and make available to the Guest Observers the Real Time high-level data products.

Justification Guest Observers may need a short time-scale check/monitor of their proposal-driven observations. This could be assured by RT pipelines that produce first high-level products on relatively short timescale. The access to the RT results will be granted to Guest Observers through a dedicated archive area.

Impact `performance:` medium; `cost:` high.

Interfaces Realtime Reconstruction Pipeline

Validation

AR1.60 CTA archive will store housekeeping and auxiliary data produced by the arrays. They must be available for any users able to run reconstruction analysis pipeline and monitoring tools.

Justification Housekeeping and auxiliary information are essential to perform scientific data analysis as well as to monitor system status and performance.

Impact `performance:` medium; `cost:` medium.

Interfaces Instrumental Response Generation Pipeline

Validation

AR1.70 CTA archive will allow an efficient management of queries and data retrieval in order to collect the necessary information about data generation as well as to allow data reproducibility.

Justification In order to provide an easy and efficient access to the archive, it is essential to define a set of metadata describing the archive content at all data levels and to organize them in a database [4].

Impact `performance:` medium; `cost:` medium.

Interfaces Metadata working group.

Validation

AR1.80 CTA archive will host all the Monte Carlo (MC) data produced by the MC Simulations Unit.

Justification MC data required by the reconstruction process will be produced by the MC Simulations Unit. MC data will be produced at the science data centers and transferred to both the on-site and off-site archive in a dedicated high-level archive tree (TBC).

Impact `performance:` medium; `cost:` medium.

Interfaces MC working group

Validation

AR1.81 CTA on-site archive will host all the IRFs needed by the RT analysis pipelines.

Justification In order to guarantee an efficient and quick data reduction by RT analysis, a set of predefined IRFs must be directly accessible by the RT reconstruction unit. Such an IRFs must therefore be sored locally in the on-site archive.

Impact `performance:` medium; `cost:` medium.

Interfaces MC working group

Validation

AR1.90 CTA archive will store and make available to the team members and Guest Observers the different produced Instrument Response Functions (IRFs).

Justification Motivated by **A-USER-0080** in [1].

Impact `performance:` medium; `cost:` medium.

Interfaces Instrument Response Function pipeline

Validation

AR1.100 CTA archive will be continuously and efficiently accessible worldwide (24/7).

Justification In order to maximize the scientific return of the CTA observations, the data must be accessible for visualization and download at any time in an efficient way. Due to the large expected amount of data, better than the standard mirror archives solution, this could be achievable with advanced archive structure architecture as distributed storage (*cloud paradigm*).

Impact `performance:` high; `cost:` high.

Interfaces

Validation

AR1.110 CTA archive will have backup storage.

Justification Due to the importance of all CTA data, it is important to prevent any type of data loss. For this reason, a full data backup must be organized. It will host a copy of the data to assure long term preservation.

Impact `performance:` high; `cost:` high.

Interfaces

Validation

AR1.120 CTA archive will provide software applications for source query, data file download and high-level scientific products visualization to the archive users.

Justification Due to the expected huge amount of CTA data and the importance of a multi-wavelength analysis, software tools must be robust, fast and runnable from any operative platform. DB queries shall be optimized for searches based on source coordinates, observation times and photon energies.

Impact `performance:` high; `cost:` medium.

Interfaces

Validation

8.2 Non functional requirements

AR2.10 The CTA data archive must be dimensioned according to the expected data amount.

Justification The expected amount of data per year is $\sim$ 5.5 PB/year [2]. Suitable storage devices, either hosted in mirror archives or cloud distributed, must be projected and developed according to this constrain.

Impact `performance:` high; `cost:` high.

Interfaces

Validation

AR2.20 Data Transfer System (DTS) must ensure transfer of data and data products (**AR1.22**) according to the adopted storage solution.

Justification DTS must rely on automatic procedures that check the transfer result, notify failures and preserve data integrity.

Impact `performance:` high; `cost:` high.

Interfaces

Validation

AR2.30 CTA archive will guarantee the availability of the data exclusively to specific users for a predefined proprietary period. After that, data will be released to the scientific community.

Justification Motivated by **A-USER-0100** in [1].

Impact `performance:` low; `cost:` low.

Interfaces

Validation

AR2.40 CTA archive will guarantee remote access to the public high-level data products using Virtual Observatory (VO) protocols.

Justification The International Virtual Observatory Alliance (IVOA) coordinates an international collaboration aimed to reach an agreement about standards and protocols to be used to locate, retrieve and analyze astronomical data within a common framework. It aims to provide uniform access to multiple archives distributed worldwide. The CTA archive should therefore be designed to be VO-compliant. This will guarantee an efficient and easy access to high-level data after the proprietary period.

Impact `performance:` medium; `cost:` medium.

Interfaces Metadata working group.

Validation

AR2.50 The high-level products metadata will be defined following the IVOA specifications.

Justification The International Virtual Observatory Alliance is working on data description with the aim to reach an interoperability among different archive resources based on the agreement on a common metadata schema. It is therefore important to work in agreement with the IVOA working groups. The goal is to define the necessary and sufficient metadata for the CTA archive to avoid metadata becoming as large as data.

Impact `performance:` medium; `cost:` medium.

Interfaces Metadata working group.

Validation

AR2.60 The CTA data archive must be secured from any malware and other external harmful software.

Justification malwares might pose severe threats to the integrity and preservation of the CTA Obs. Archive. Thus, it will be mandatory to secure the archive with the latest available protection.

Impact `performance:` medium/high; `cost:` medium/high.

Interfaces CEIN working group

Validation

AR2.70 Password protected access must be guarantee to the Guest Observers during the proprietary period.

Justification Guest Observers shall have exclusive access to the data deriving from their accepted proposals. That will be achieved through different interfaces (web-based, sftp, ...).

Impact `performance:` low; `cost:` low.

Interfaces

Validation

AR2.80 CTA archive will maintain the database of the accepted observational proposals as well as the scheduler functionalists.

Justification The CTA will operate as an open, proposal-driven observatory and it will guarantee the access and handling of the different proposals. Moreover, in order to assure the management and optimization of the available observational time, the archive will organize the scheduler algorithm.

Impact `performance:` medium; `cost:` medium.

Interfaces

Validation

References

[1] *User Requirements for CTA V1.0* W.Hofmann & J. Hinton

[2] *High-level user requirements for CTA data management V0.6.* J.D. Ponz & R. Walter

[3] *Reconstruction Requirement V1.0* K. Kosack & A. Djannati-Atai

[4] *Functional MetaData and Archive Model Requirements V1.0.0* C. Lavalley

[5] *CTA-data management PBS & WBS V1.3* G. Lamanna & C.Boisson