J. Astrophys. Astr. (2021) 42:43
https://doi.org/10.1007/s12036-021-09747-x

MISSION

Data processing, archival and dissemination pipeline for AstroSat: Challenges and strategies

C. BALAMURUGAN[1,*], SACHIN NARANG[1], U. PRIYANKA[2], NAVITA THAKKAR[3], HIMANSHU PANDEY[1], AMIT PUROHIT[1], A. L. SATHEESHA[2], P. GUNASEKHAR[1], T. P. SRINIVASAN[3], A. S. SHASTRY[2] and B. N. RAMAKRISHNA[1]

[1]ISRO Telemetry Tracking & Command Network (ISTRAC/ISRO), Bengaluru 560 058, India.
[2]UR Rao Satellite Centre, Airport Road, Bengaluru 560 017, India.
[3]Space Applications Centre, Ahmedabad 380 015, India.
*Corresponding Author. E-mail: balamurugan.c@istrac.gov.in

MS received 4 November 2020; accepted 7 February 2021

Abstract. The ground data processing, archival and dissemination of the science data for India's first multi wavelength astronomical spacecraft, AstroSat is carried out from Indian Space Science Data Centre (ISSDC). The proposals submitted by users worldwide are grouped into observations and are observed in multiple orbit segments. Large volume of data in various configurations from different payloads and the inherent diversity of data products resulting from observations had demanded intricate procedures in handling the data processing, archival and dissemination scenario at ISSDC. The paper describes the layered approach followed in implementing a software stack for establishing and operating a completely automated pipeline for data processing, archival and dissemination of astronomical data. The operations of the automated pipeline, challenges faced during the implementation and the strategies adapted to overcome them are also summarized.

Keywords. AstroSat—data pipeline—archive management—dissemination management—event management—automation—good time interval—bad time interval.

1. Introduction

AstroSat (Singh *et al.* 2014; Navalgund *et al.* 2017) is a proposal-driven multi-wavelength first Indian space observatory capable of simultaneous observation over a wide range of electromagnetic spectrum. The primary science objectives of the mission are met with 5 science payloads namely Ultraviolet imaging telescope (UVIT) (Tandon *et al.* 2017), Cadmium Zinc Telluride Imager (CZTI) (Rao *et al.* 2017), Soft X-ray Telescope (SXT) (Singh *et al.* 2017), Scanning Sky Monitor (SSM) (Ramadevi *et al.* 2017), and Large Area X-Ray Proportional Counter (LAXPC) (Agrawal *et al.* 2017). The astronomical community and researchers worldwide submit proposals for the astronomical object/event to be observed. The proposals submitted by users for different targets are grouped into observations. The data for a single observation would be recorded and received at the ground stations in several data dumps. Each dump session refers to the availability of the station for data collection from the spacecraft during a single orbit. Hence the data products generated are dump-orbit wise and when the data belonging to a complete observation are available on ground, observation wise products are generated. The data processing at ISSDC includes Level-0 and Level-1 processing. The Level-0 process carries out time correlation, segmentation, attitude filtering and orbit corrections. The Level-1 processing focuses on applying filters and corrections to properly interpret the data.

This article is part of the Special Issue on "AstroSat: Five Years in Orbit".

Published online: 11 June 2021

The data products of different payloads of AstroSat are archived and disseminated to the respective Payload Operation Centers (POCs) on regular basis immediately after generation at ISSDC. POCs qualify the data and generate higher level data products and send them back to ISSDC for archival and dissemination to different classes of users worldwide. The dissemination during lock-in and after lock-in period is handled in a flawless manner without any deviation to the defined data policies of AstroSat mission.

In Section 2, the overall pipeline for ground operations at ISSDC has been elaborated. The details of Level-0 and Level-1 processing are outlined in Section 3. Section 4 describes the specifics of archival and dissemination pipeline. Section 5 describes the layered approach adopted for automation of data processing, archival and dissemination. Section 6 describes how this pipeline caters to the observation cycle specific - proposal based data dissemination of AstroSat mission through Astrobrowse and Generalized Access and Dissemination Software (GADS).

2. AstroSat pipeline at ISSDC

The intrinsic diversity of data products, the distributed network and storage system over which the pipeline is established has imposed challenges in implementing and completely automating the activities of the payload data pipeline at ISSDC.

2.1 *Network architecture*

The architecture of the ISSDC consists of different layers namely, Mission layer (MIS), Archive layer (ARC), Exchange layer (EXH), External Network (EXT) layer. MIS layer interfaces with all the supporting ground stations of AstroSat mission and provides access to the payload and auxiliary data. In addition to interaction with the ground elements, it also provides support for quick look processing and display, data ingest and Level-0 and Level-1 processing. The ARC layer encompasses the ISSDC archives and higher levels of processing. The support provided in this layer includes ingest, media handling, data product generation, migration, integrity and security for data archives.

The EXH layer stages and controls exchange of data products between the internal layers (MIS, MOX, and ARC layers) and the EXT Network. Ingestion of higher level data products from POCs to the ISSDC archives and dissemination to POCs is carried out in this layer through the Ingest and Dissemination software. This layer is separated from other internal layers by high end security devices. Main functionality of EXT network is to host Web, Mail and File transfer services. EXT layer provides an entry point for users to access data and submit proposals through various web based applications. Figure 1 shows the set of software elements that form part of the pipeline at ISSDC for AstroSat.

2.2 *Pipeline elements*

The proposal processing pipeline has elements for proposal submission, evaluation, scheduling and command generation. The AstroSat Proposal Processing System (APPS) is the main software in the proposal processing pipeline. Proposers, Guest Observers and Instrument teams can submit the proposal through the proposal processing system. The proposals are submitted under various categories which include calibration, opportunity based observation and emergency requests to observe active targets and transient events. The proposals submitted are evaluated for their merit and are accepted or rejected.

Various other web applications are deployed along with the APPS software to facilitate the proposal submission activity. Astroviewer is a web based software utility used to generate the probable visibility periods of the celestial bodies which can be observed by AstroSat (Nagamani *et al.* 2010). AstroSat Exposure Time Calculator is a simulation tool to predict the exposure time required for the various payloads of AstroSat considering parameters like count rate and energy. The visibility periods from Astroviewer and Exposure time prediction are mandatory for proposal submission.

The Mission Control and Proposals database (MCAP) is the repository of all the data from the approved proposals. This is accessed for command generation of the payloads. The data received from the spacecraft are termed as raw data. (Pandiyan *et al.* 2017). The raw payload data collected at the ground station is ingested to ISSDC for further processing, archival and dissemination. The ingested data is subjected to Level-0/Level-1 processing.

The Level-1 processed data at ISSDC is disseminated to the Payload Operation Centres (POCs) for quality checks and higher level product generation.

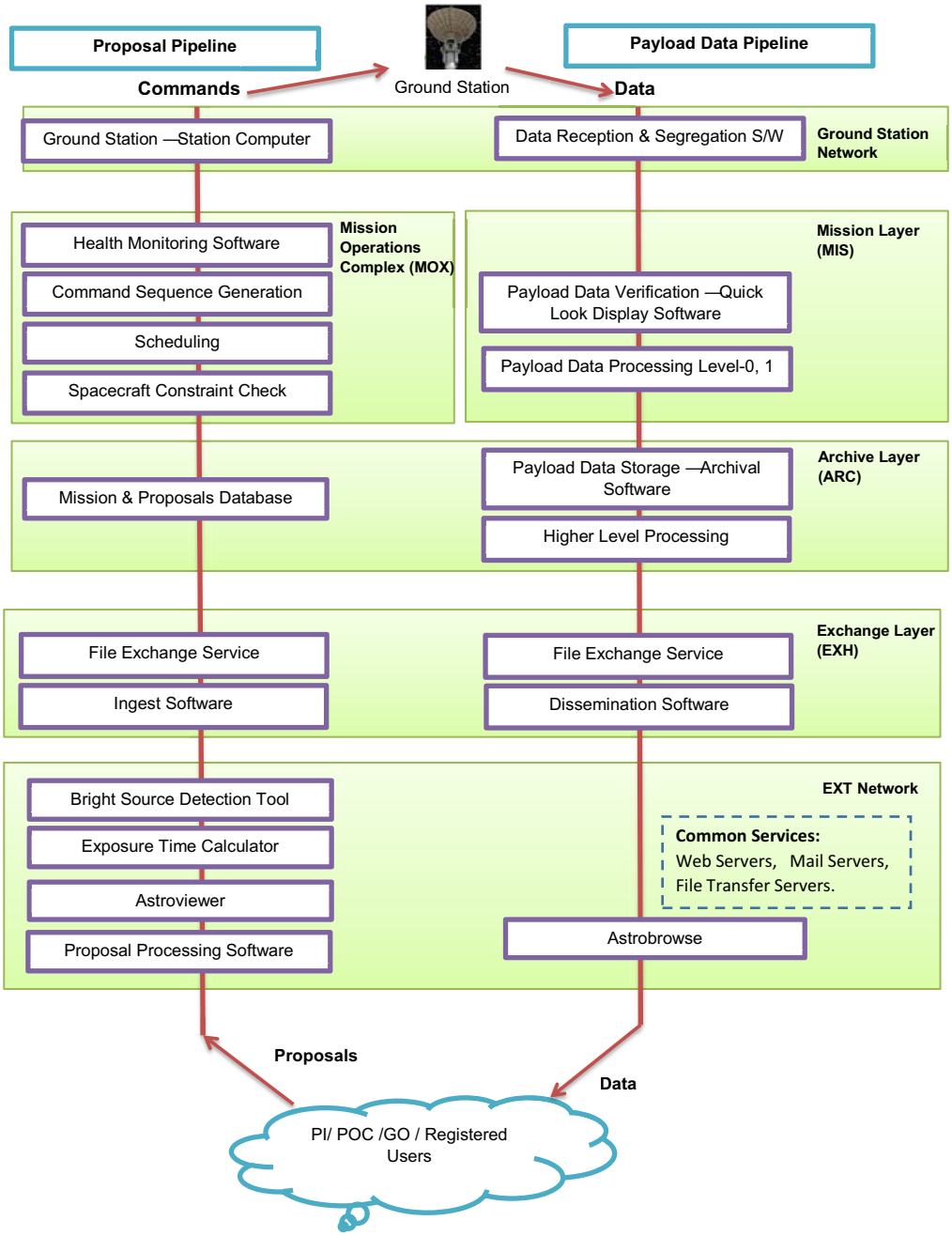

Figure 1. Proposal processing pipeline software stack.

POCs send back the higher level products for dissemination to Principal Investigators during the lock-in period and to all registered users after the lock-in period. The validated and processed data are also archived at ISSDC. The dissemination of data to the POCs is through Virtual Routing and Forwarding (VRF) over the public internet. The dissemination of qualified data to proposers is through the web based software Astrobrowse.

3. Data pipeline for Level-0 and Level-1 processing

Level-0 products: Level-0 products are generated from the raw data ingested by the Data Acquisition software. Level-0 products are segmented time-tagged science data in binary format with auxiliary information, for each observation-id. Level-0 products are input to Level-1 pipeline for generation of Level-1 products.

Level-1 products: Level-1 products are mode-segregated science data files along with Time Correlation Table data, Orbit, Attitude, MKF (Make Filter File), and GTI (Good Time Interval) and BTI (Bad Time Interval) data provided in Flexible Image Transport System (FITS) format for Astronomical data analysis. Level-1 Products are provided to POCs for generation of higher level products for public release.

AstroSat Level-0/Level-1 data pipeline is a complete automated system. Orbit-wise data products as well as observation-id wise product generation are done in automation without any manual intervention. Observation-id is made available in Mission and Proposals database (MCAP) by constraint check and scheduling software which is fetched by processing software to generate observation-id wise products.

Auxiliary data are processed to update Time database and SPICE database (SPICE stands for S = Spacecraft ephemeris, P = Planet, I = Instrument information, C = C-matrix, E = Events used for planning and interpreting scientific observations) for Orbit, Attitude and Time data. Payload chain generates time-tagged, formatted, mode-wise segmented data files along with Time Correlation Table data and Orbit Attitude files also in FITS format. These data files are bundled together and provided to the secure file transfer system software for archival at ISSDC.

At Level-1 processing, MKF consisting of House-Keeping (HK)/health parameters and Auxiliary data are generated. Level-1 software also provides the GTI and BTI data which takes a combination of parameters from each payload and their range to mark the data interval as good or bad. Figure 2 illustrates a simplified version of one of the payload (LAXPC) processing at Level-1 for AstroSat mission.

All the data received are Level-0 processed and kept in separate observation-id wise folders. Once the given observation is completed, a separate processing chain gets triggered to generate merged Level-0 product. This chain merges all the data available in the observation-id folder after removing duplicate records. It also triggers Level-1 processing chain and thus an observation-id merged product gets generated.

4. Data archival and dissemination pipeline at ISSDC

The complete proposal based operations of AstroSat takes place through different proposal cycles like Performance Validation (PV), Guaranteed Time (GT), Calibration (CAL), Announcement of Opportunity (AO) and Target of Opportunity (ToO).

The data products generated is in the FITS file format and is a bundled data set. The archival of these

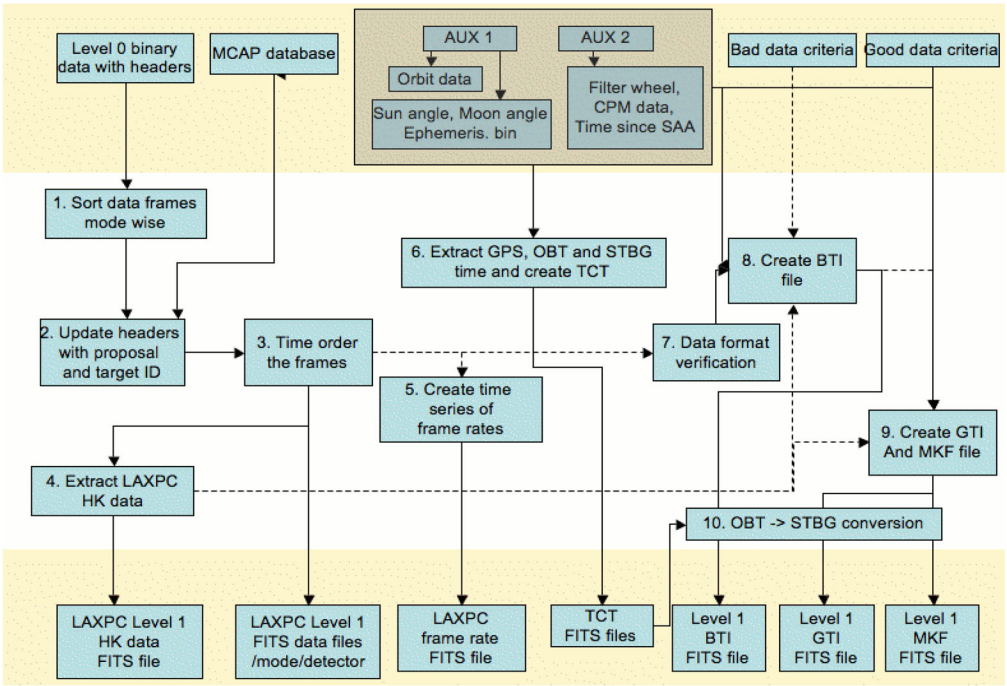

Figure 2. Typical Level-1 pipeline for LAXPC payload.

data products is done by the organizer software based on observation sequences along with the meta-information from the APPS. The dissemination policy and the lock-in period for the data observed and processed during various cycles is specific to every cycle. A lock-in defines the period within which the rights over the access to data are restricted only to the proposer of the data. The data dissemination strategies have been devised to handle such diversity.

The Level-1 data products of UVIT, SXT, CZT and LAXPC, Level-0 data products of SSM are generated at ISSDC and disseminated to the POC on regular basis immediately after generation.

POCs post back the merged and processed Level-1 data and Level-2 data along with the quality feedback. Observation-id wise Level-1 and Level-2 data products for CZTI, UVIT and LAXPC payloads and orbit wise products for SXT and LAXPC are received at ISSDC. Figure 3 shows the elaborate dissemination scenario at ISSDC.

The processed data along with the quality feedback are transferred to the scheduler software for compression and archival. The AstroSat Data Products Tracking Tool (ADAPT) software interfaces with the scheduler and configures the public release date for ingested data in the database. The Astrobrowse software accesses the compressed archive, lock-in information from the database and makes it available for download based on the data dissemination policy. The notification to users and POCs regarding the availability of data for download is done through the ADAPT software as shown in Fig. 4.

5. Strategies and challenges

5.1 *Information and data exchange between data pipeline entities*

The pipeline elements are distributed over multiple networks spanning different security domains across geographically separated locations. The web based applications APPS, Astroviewer, Exposure Time calculator, Bright Source Detection Tool are hosted in the external network of ISSDC. The Mission-proposals database, flight dynamics, scheduling and command generation software are located in the internal mission network layers at ISSDC and at Mission Operations Complex.

The network architecture at ISSDC and the distribution of pipeline entities across various network locations posed an inherent challenge in transferring the required information and data in a reliable, automated and secure manner between the pipeline entities.The interactions between the pipelines entities are through files. The layered approach as described below is used for transfer of files between entities and had ensured a prioritized, guaranteed transfer of necessary inputs for entities in the pipeline.

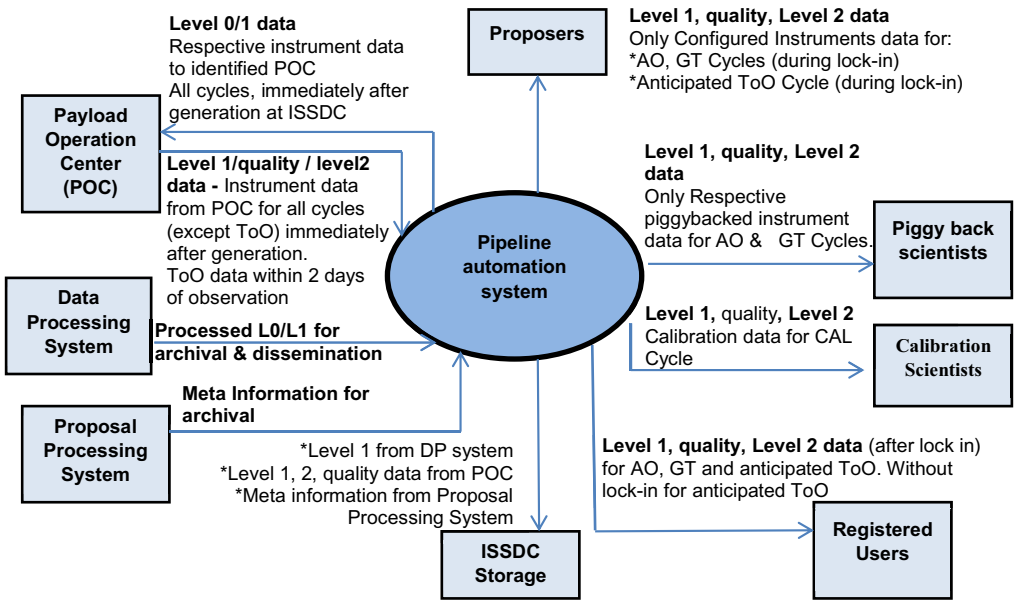

Figure 3. Astronomical data processing, archival and dissemination scenario.

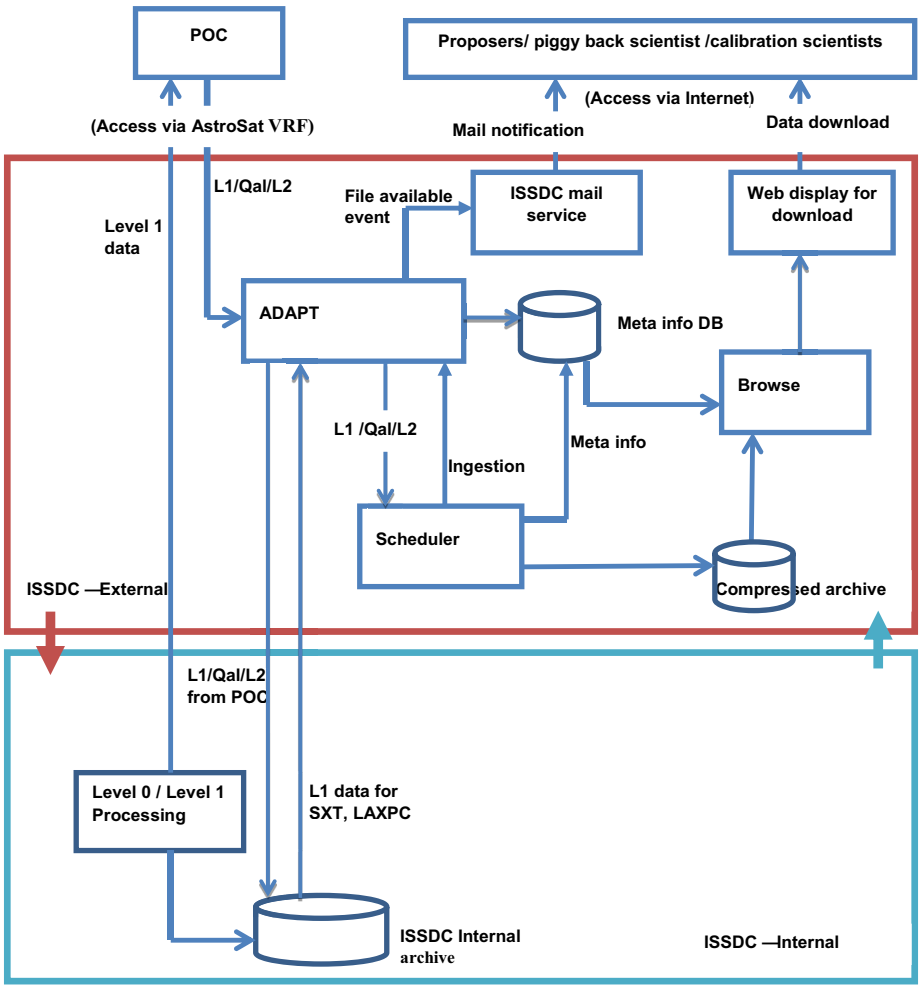

Figure 4. AstroSat archival and dissemination pipeline.

5.2 *Layered approach: Automation of astronomical data processing, archival and dissemination*

A layered approach to handle the Automation of processing, archival and dissemination of diverse categories of AstroSat data has been implemented at ISSDC, as shown in Fig. 5. This section describes the function of different layers.

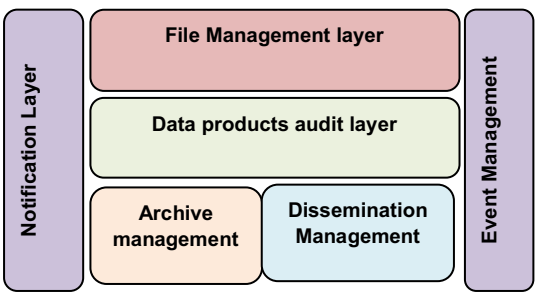

Figure 5. Layered model for astronomical data processing, archival and dissemination.

File management layer: This layer verifies the completeness and correctness of the files received from Level-0 and Level-1 processing software and the POCs based on the check sum file sent along with the data products. The software layer interfaces with the Level-0/Level-1 processing software ingestion of raw data received from different ground stations and collects the processed data for archival and dissemination to the POCs. This layer also handles the reception and ingestion of final products from POCs for archival at ISSDC and for dissemination to the proposers and public release.

Event management layer: Events are defined to track every activity in the pipeline, which include events for data ingestion, reception of incomplete/ incorrect data sets, unavailability of data sets, data product generation and successful ingestion of data products for archival. This layer tracks and transfers the events to the relevant modules for necessary action.

Notification layer: This layer receives the events related to the successful ingestion of data products to the archival software. Then it accesses the database for proposer related information like the email-id and user-id, generates file availability messages, and transfers them to the mail server for notifying the proposer. It receives the events related to incomplete/incorrect files received from the POC and sends notification to POC to repost the incorrect data sets. This layer interfaces with the Exception Notification System (ENS) to alert and send logs to the designers of data processing software upon encountering a software exception for quick resolution of issues.

Data products audit layer: This layer receives the events related to data products generated at ISSDC, data products posted from POC, their availability and dissemination status. It performs an audit to find the missing data products and keep track of the different versions of the data products. It periodically generates statistics on the data status at ISSDC and generates messages for notifying the same. It also receives the feedback from POCs regarding the duplicate data and data-sets which cannot be processed at POC and records them in the database.

Archive management layer: This layer receives the files for archival from Level-0/Level-1 and POCs, interfaces with the proposal processing software and generates meta information file for every observation like the co-ordinates of the celestial sources, source name, category, mode of observation, science values of the observation, quality of data, percentage of usable data etc. It organizes the data products in the storage based on the observation- id, proposal-id and instrument-id along with the meta-info file.

Dissemination management layer: This receives the events regarding the successful ingestion of data products and sets the public release dates for every data set in the database. If the ingestion event corresponds to a ToO data, then it sets the release date as the ingestion date. The Astrobrowse software reads the release date information from the database and presents the authorized view to the proposers and general users for downloading the data. A check sum file is generated along with the actual data to ensure the integrity of data during transmission. Older versions of the same data products are removed from the storage based on the data retention policies.

5.3 *Data processing at ISSDC*

At the data processing level, the main challenge has been that every payload in AstroSat can be considered a separate mission in itself considering the varying time correlation, payload data handling and observation-id wise data product generation requirements. The challenges in handling large volumes of data from different payloads have compelled the system to be more robust.

Observation-wise product generation has the complexity of knowing when the observation has ended as the order of dumping data is based on station availability and spacecraft constraints. Observation which have spanned for a number of days (7–10 days) has provided data volume in the order of $\sim$50 GB (especially UVIT payload) to be processed. This has called for the optimal use of compute and storage resource to process high volume data without any failure in reasonable timelines. Another challenge in designing the software was to develop highly reliable software system to successfully process 14 orbits of data per day. In spite of all the above, there have been backlogs due to late update of attitude information, any minor malfunction of a payload component necessitating reset etc. A dedicated processing infrastructure has been established at ISSDC to process the backlog data without affecting the regular operations.

Taking into consideration the redundancy in data acquisition at the ground stations, there are multiple streams of data available on ground for processing at the same time. Availability of datasets from multiple data ingestion system triggers the data merging software which merges the data reception at any of the ground station chains. These merged data are provided to the Level-0/1 data pipeline which takes care of the payload specific processing. A Filter File which is basically collection of required HK and Auxiliary data information is also generated.

Every payload has its own science data format, processing requirement and product definition. For e.g., UVIT data are organized as data blocks with header providing payload configuration information, while LAXPC and SXT data records have mode information in every payload line whereas CZTI payload processing requires to consider quadrant information along with mode and SSM payload processing requires event information decoding. Each payload has its own clock which needs to be decoded and correlated with Spacecraft Positioning System (SPS) time accurately. UVIT even has a separate

clock for each detector, with one of them set as master, which added further complexity to the system. At Level-1 processing, MKF which is basically collection of required HK and Auxiliary data are also generated. Along with this, there has been a requirement of generation of Good Time Interval (GTI) and Bad Time Interval (BTI) data which takes a combination of parameters from each payload. For BTI files, qualify flags are also provided to mark which parameters are at fault for the bad data indication. To take care of different payload parameters and data processing conditions, "expression evaluation" techniques have been employed to make the software configurable and adaptive. The data are processed in-memory to avoid multiple disk access and to provide better throughput.

5.4 *Target of Opportunity (ToO) data dissemination challenges and strategies*

Target of Opportunity (ToO) proposals require the immediate observation of an astronomical event. These proposals are subjected to fast-track review process which is different from normal procedure for Announcement of Opportunity (AO) proposals. Hence, as a policy, the dissemination of the Target of Opportunity data has a turnaround time of 7 days from the time of observation. This turnaround time includes any delays in processing from both ISSDC and POCs. Once the quality feedback, Level-1 and Level-2 products are available from POC, the ADAPT ingests the same to the archival software and sets the release date of the data products as per the ToO data dissemination policy. The Astrobrowse software configures the release of data as per the release date set by the dissemination management layer.

The challenge encountered in this scenario is that in cases where the turnaround time is not met and the data are not available for public download within 7 days; this would result in delay of analyzing critical celestial events, especially when there are observations carried out in co-ordination with other international observatories. To overcome this challenge, the ToO data immediately after generation at ISSDC are also made available for public download with a disclaimer on the quality of data. This is done using the Generalized Access and Dissemination System (GADS) software. GADS provides an authorized access to data at ISSDC through customized views, while maintaining transparency from various underlying system, storage and network complexities.

After the qualified data are available from POC, the data are made available to the public users through Astrobrowse and the unqualified ToO data from GADS are removed. This allows users to effectively download and use the data observed under the ToO category

6. Future work on pipeline operations

Data mining techniques based on resource usage heuristics and archived meta-information is planned to be implemented to bring about intelligent resource planning. Natural Language Processing (NLP) based agents for autonomous event generation, pipeline monitoring, pipeline configuration, alert notification and self-recovery from pipeline failures through backtracking is also envisaged.

7. Conclusions

The strategy adopted for handling the challenges in astronomical data processing, archival and policy based dissemination for AstroSat has proved to be effective and has provided a platform to handle future proposal based missions. The strategy has led ISSDC to envisage a unified intelligent framework for having access to download data against the serviced proposals.

Acknowledgements

The authors would like to thank and acknowledge the support provided by Indian Space Science Data Centre-ISTRAC, Space Application Centre (SAC), IUCAA, IIA, TIFR, RRI, Space Astronomy Group-URSC, PPAD – URSC, Spacecraft Operations team – ISTRAC and Space Science Programming Office, ISRO HQ for their extensive support in establishing and maintaining the AstroSat Processing, Archiving & Dissemination pipeline at ISSDC.

References

Agrawal P. C., Yadav J. S., Antia H. M. *et al*. 2017, J. Astrophys. Astr. 38, 30

Nagamani T., Bharadwaj N., Dakshayani B. P., Pandiyan R. 2010 AstroSat Software Tool to Aid Celestial Source

Viewing, 61st International Astronautical Congress, Prague, CZ, Paper IAC-10-A3.4.6
Navalgund K. H., Suryanarayana Sarma K., Gaurav P. K. *et al*. 2017, J. Astrophys. Astr. 38, 34
Pandiyan R., Subbarao S. V., Nagamani T. *et al*. 2017, J. Astrophys. Astr. 38, 35
Ramadevi M. C., Seetha S., Bhattacharya D. *et al*. 2017, Exp Astron 44, 11
Rao A. R., Bhattacharya D., Bhalerao V. B. *et al*. 2017, Curr. Sci. 113, 595
Singh K. P., Tandon S. N., Agrawal P. C. 2014, SPIE 9144, 1
Singh K. P., Stewart G. C., Westergaard N. J. *et al*. 2017, J. Astrophys. Astr. 38, 29
Tandon S. N., Hutchings J. B., Ghosh S. K. *et al*. 2017, J. Astrophys. Astr. 38, 28