

The data pre-processing method based on observation constraints for astronomical mission and its application

JI Zhen^{a*}, MA Fuli^a, LI Bing^b, YUQinsi^a, YANG Xiaoyan^b, TONG Jizhou^a

^a National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China

^b Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China

* Corresponding Author

Abstract

With the advancement of astronomical observation technology, obtaining high precise and more accurate data has become one of goals of space astronomical missions. Meanwhile, in astronomical mission scientists also propose complex observation plans which involve various observation priorities and observation modes to implement accurate observation. Therefore, the efficiency and integrity of ground data pre-processing has more stricter requirements. In response to these requirements, this article proposes a data processing method and designs a real-time data processing pipeline, which integrates the steps of occlusion state calculation, observation integrity evaluation and so on. The pipeline uses the method which processes the raw data based on data types and processing dependencies. Firstly, it performs time matching between engineering and scientific data to gather the data with the same observation identification observed in different time periods, and completes the spatial coordinate transformation between the detector and the celestial sphere. Then, it calculated important filtering parameters such as payload normal direction and the occlusion state through different algorithms. Third, based on the observation plan and preprocessing results, statistical analysis was conducted on data from different observation modes such as pointing observation, sky survey, and target of opportunity (ToO) observation. Statistical information mainly includes time accuracy calculation and target position deviation analysis. At the end of the pipeline, the integrity of the observation data is evaluated based on the observation priority, and standardized data product files are generated according to the format definition. The method and the pipeline have been successfully applied in the data preprocessing of Einstein Probe (EP) mission, and the results show that it can quickly process data while ensuring compliance with the constraints in the observation plan and the data product integrity for one observation. This provides new technical ideas for efficient preprocessing of scientific data for subsequent astronomical satellite missions.

Keywords: (maximum 6 keywords)

EP, Data preprocessing, Astronomy

1. Introduction

With the implementation of advanced telescopes and space missions, the volume and complexity of observation data have surged exponentially which brings the challenges for researchers. Time domain astronomy, as an important branch of astronomy, is listed as one of the five frontier nurturing discovery opportunities in the United States Ten Year Plan for Astronomy and Astrophysics (2012-2021) [1]. With the continuous advancement of space science exploration technology and significant improvement in information processing capabilities, the understanding of the universe is converting from static images to dynamic processes, making it possible to make rapid time-variant observations. This transformation not only provides new opportunities for exploring and understanding the universe and related physical laws, but also reveals physical information about radiation mechanisms and historical evolution through observing changes in radiation, such as transients and explosive events. Especially with the rise of gravitational wave astronomy [2], time-domain astronomy has entered a golden age of multi-band, multi-messenger joint observations.

Efficient data pre-processing is a pivotal step for the scientific research. The pre-processing of astronomical data is deeply depended on the specific observation requirements and constraints of every mission. There are some factors such as the observation priority, the characteristics of the instruments used, and the scientific goals of the mission introduce unique constraints on the way of data pre-processing. For instance, time-domain astronomy demands rapid and accurate data processing to capture the transient phenomena. The high timeliness requirement for time-domain astronomical research poses significant challenges to the real-time and accuracy of ground data preprocessing. The traditional pre-processing techniques are too generic, falling to solve the mission-specific nuances and account for the changes of observation conditions.

To address these challenges, this paper proposed a novel data pre-processing method that incorporates observation constraints into the pre-processing pipeline. The method aims to enhance the efficiency, accuracy and relevance. The

observation-based method has the ability to perform high-precision temporal and spatial position calculations, collaboratively process data from multiple communication channels, and interpret the data integrity of an observation task.

The remainder of this paper is organized as follows: Section 2 details the proposed method. Section 3 presents the practical implementation of the method. Section 4 presents the results of the application in EP mission and discussion of its effectiveness and potential impact. Section 5 concludes with the summary of the main content and potential future application of the method.

2. Key processing algorithm

In the procedure of pre-processing, it is necessary to distinguish scientific events data frames based on the information such as instruments working mode and working status. Meanwhile, the data observed in the special space areas such as the Earth occlusion (EOC) zone and the South Atlantic anomaly (SAA) zone. During the observation process, the priority of observation types varies based on the scientific objectives of the mission, but there are cases where low priority observations are interrupted by the higher priority observations, requiring integrity analysis of the interrupted observation data. The hierarchical collaborative processing algorithm and the data integrity analysis algorithm based on observation task organization will be respectively introduced in the following subsections.

2.1 Pre-processing algorithm

The hierarchical collaborative processing algorithm is an algorithm for identifying, separating, and extracting scientific events frames based on the parameters such as observation plans, physical events types, working status and working modes of instruments and the attitude states of satellite. The procedure of algorithm is shown in the Fig. 1.

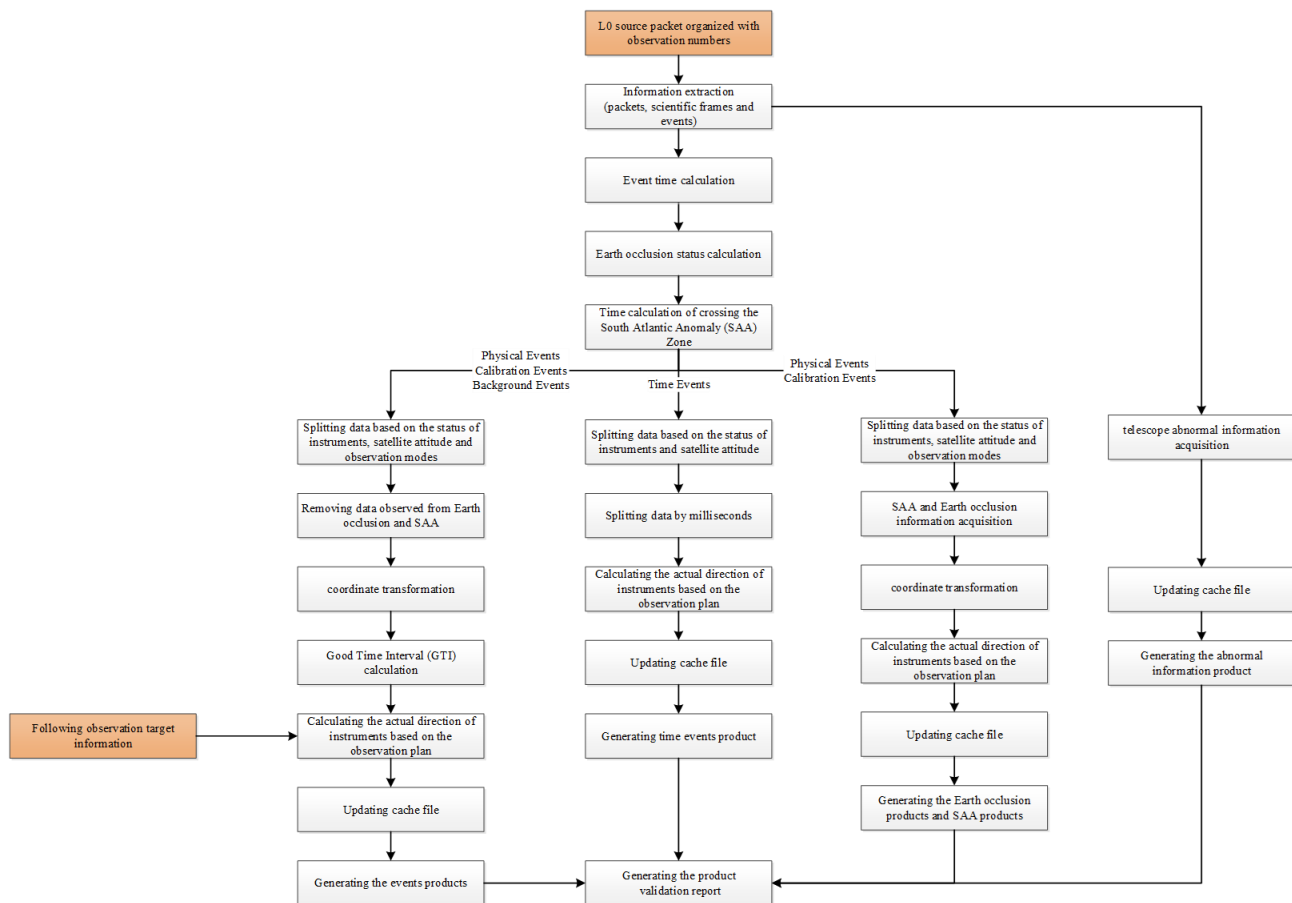


Fig. 1 The data pre-processing algorithm

In order to facilitate researchers to conduct scientific analysis based on observation tasks, it is necessary to first process the raw data, obtain and sort source packets in chronological order, at the end generate the level 0 data products organized by observation task numbers. The algorithm takes level 0 data products as input, and the specific processing steps are as follows.

The first step is to process the level 0 data products, including extracting scientific source packets, scientific data frames, and scientific events, obtaining abnormal data from the instruments and generating abnormal products. The second step is extracting the time events contained in each scientific frame according to the scientific data frame format definition. And then combined with the local time of scientific data frame, the time of each frame is calculated which is accurate to milliseconds. The third step is based on the engineering data, analysing the EOC and SAA status of each frame data, as well as the observation data generated in the abnormal zones. The fourth step is splitting the events data into data blocks according to the working modes, working status, the instrument type, and the attitude state of satellite. After removing the data generated in EOC and SAA zones, the coordinate transformation is performed. According to the scientific frame count continuity, the good time interval is calculated. At the end of step, the actual observation direction is recalculated based on the processed data blocks by obtaining target information from observation plans or ground follow-up observations. The fifth step is similar to the fourth step, splitting the time events data into data blocks according to the instrument type and the attitude state. Then calculating the actual observation direction by obtaining target information from observation plans or ground follow-up observations. The final step is calculating the actual observation direction of EOC and SAA data by obtaining target information from observation plans or ground follow-up observations. At the end of the processing, the accuracy of the time and observation direction is verified and the standardized data products are generated.

2.2 Observation data integrity analysis algorithm

There are two basic requirements for the integrity of data products organized by observation task. On the one hand, the data of current observation task is continuous with the data of the previous and subsequent observation tasks. There is no data loss between observation tasks. On the other hand, there were no data loss in the current observation task, namely, there were no missing data in the middle, beginning, or end of the data products of current observation task. But for the astronomical observation, there are different observation types with different priority, such as the ToO observation has priority over the sky survey observation. So, there are scenarios where the current observation is interrupted or restored based on the observation priority, and high priority observation task data is prioritized for transmission.

Therefore, if the ground system finds that the scientific data frames are discontinuous during data integrity analysis, there are two normal scenarios for data discontinuity: 1) high priority data is transmitted before low priority data, and low priority data appears discontinuous; 2) The observation task in progress is interrupted by a high priority observation task, and it is resumed after the high priority observation task is completed. However, it is also possible that data loss may occur due to transmission between satellite and ground, which poses a challenge for ground systems to analyse data integrity. The data integrity analysis algorithm based on observation priority, as shown in Fig. 2 and the steps of algorithm are described in the following.

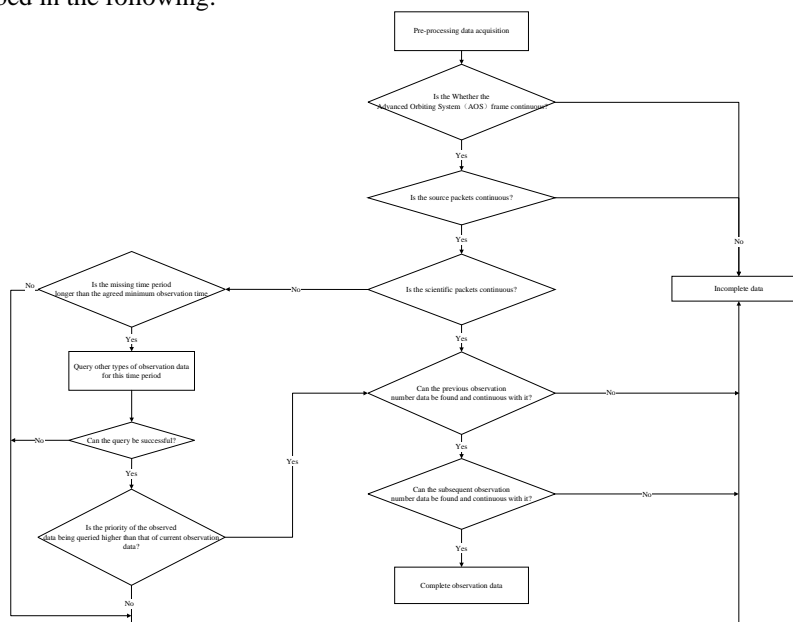


Fig. 2 The data integrity analysis algorithm

First step is analysing the transmission continuity. All data from the same observation task is obtained and based on the pre-processing report the continuity of the AOS frames and the source packets are evaluated. When the frame

count or source packet count is discontinuous, the data is incomplete and the analysis process is end. The second step is when the count of frame and source packet is continuous, the continuity of scientific data packets of source packets is evaluated. Retrieving scientific data packets and sorting them with time and packet count. When the scientific data packet is discontinuous, it is necessary to analyse whether there are other observation tasks with high priority, which is the fourth step. When the scientific packets are continuous, the third step is performed. The third step is analysing the continuous of adjacent observation tasks. Based on the data time, the nearest previous observation task is searched and the continuity of the time and scientific packet count is analysed. If the previous observation task data is continuous with the current observation data, the nearest subsequent observation task is searched. Then the continuity of the time and scientific packet count is analysed. If the data is continuous, the data of current observation task is complete. If the data is discontinuous, the data is incomplete and the analysis process is end. The fourth step is analysing the duration of the interruption. When the scientific packets are discontinuous, if the interval time is less than the designed minimum observation duration, it is incomplete observation data. If the interval time exceeds the minimum observation duration, there is a possibility that the observation data has not been transmitted or the observation has been interrupted. The fifth step is performed. The fifth step is analysing whether there is other observation data during the interruption time. According to the data time of observation task, retrieving data of all observation tasks in this time period, sorting by data time and analysing the data continuity segment by segment. If the data is continuous within the time range, the sixth step is performed. If other observation data within the interruption time cannot be obtained, or if the period is not continuous, the observation data is incomplete and the analysis is terminated. The sixth step is observation priority comparison. When obtaining other observation data within the interruption time, extracting the observation types from the sequency number and comparing it with the observation type of the current observation task. Then, whether it is a higher priority observation type is determined. If it is a same level or higher priority observation type, the procedure skip to the third step to complete the continuity evaluation with the previous and subsequent observation tasks. If it is a low priority type, it will be interpreted as incomplete data and the progress will end.

3. Application in EP mission

The pre-processing method has been applied in EP mission as an example. The EP mission is a small astronomical exploration aimed at time-domain astronomy. The EP mission is led by China, and cooperated with the French Space Agency and the European Space Agency. It conducts high-frequency time-domain sky surveys and transient source rapid follow-up observations of the soft X-ray to obtain the high-quality X ray flux and spectral data [3]. The EP mission has two scientific payloads, the Wide field X-ray Telescope (WXT) and the Follow up X-ray Telescope (FXT). The WXT payload mainly implements sky survey observations, triggers and locates X-ray sources, and triggers on-board autonomous follow-up observations for important transient sources. The FXT payload performs follow-up observations on the trigger source, high-precision fixed-point observations on opportunity targets, and high-sensitivity sky survey observations. The EP mission has designed four priorities, from high to low: the important ToO observation, follow-up observation, ordinary ToO observation and sky survey observation.

Based on the method presented in this paper, a pre-processing and distribution platform for EP mission was constructed, which utilizes the general framework for ground data pre-processing of space astronomical missions [3,4]. The details of pre-processing pipeline is shown in the Fig. 3.

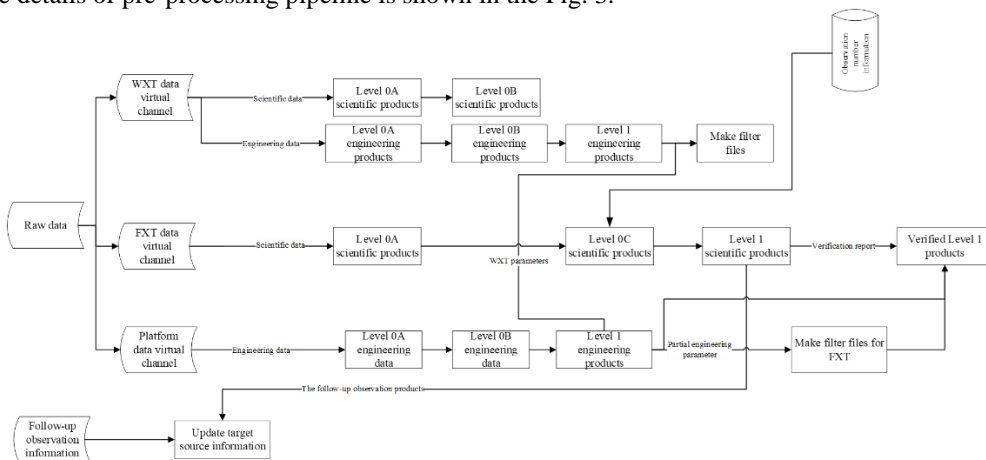


Fig. 3 The EP data pre-processing procedure

Each horizontal line in the figure is a separate type of data pre-processing procedure. The relationships between the different processes are established base on the data source and the data dependencies. In the real processing procedure,

the processing tasks state need to be evaluated based on the relationships between the processes. Only when all the data dependencies are ready, the pre-preprocessing task will be initiated.

4. Results and discussion

EP mission was successfully launched on January 9, 2024. There are 10 times of data transmission per day and for each transmission the raw data volume is about 5GB. The raw data includes satellite status data, engineering data and scientific data. Alarm data is generated irregularly based on triggering conditions. The data processing time of EP mission is shown in the Table 1.

Table 1. EP data processing performance

Data Type	Data Level	Data Volume	Processing Time
Scientific data	L0	1484.8MB	45 minutes
	L1	3686.4MB	69 minutes
Engineering data	L0	146MB	45 minutes
	L1	902MB	67 minutes

After each completion of data processing, the integrity analysis will be performed on data products organized by observation sequence. By verification of researchers, the accuracy of integrity analysis is close to 100%. The observations of EP mission may be interrupted according to the observation priority and then restored, and data of different priorities are arranged in chronological order onboard. For this situation, it is important for researchers to evaluate the integrity of one observation task. The data product integrity analysis algorithm mentioned in this paper can accurately interpret complete data according to the observation sequence. For instance, during the execution of the sky survey observation task with observation number 13600006198 (2024-05-17 21:44:03Z~2024-05-19 18:38:47Z), it was interrupted by the important ToO observation task with observation number 08500000099 (2024-05-18 15:00:01Z~2024-05-18 18:33:44Z). After the completion of observation task 08500000099, the payload returned to observation task 13600006198. The ground system received observation data from 13600006198, but the data for the time period of 2024-05-18 15:00:01Z~2024-05-18 18:33:44Z was missing. The data product analysis algorithm designed in this paper can accurately interpret that missing data is observation task 08500000099 data and evaluate the two observation task products as complete. Table 2 shows the results of the integrity evaluation.

Table 2. Product integrity evaluation results

Observation Number	Observation Type	Start Time	End Time	Scientific Data Integrity	Engineering Data Integrity
08500000099	Important ToO	2024-05-18 15:00:01Z	2024-05-18 18:33:44Z	Complete	Complete
13600006198	Sky survey	2024-05-17 21:44:03Z	2024-05-19 18:38:47Z	Complete	Complete
13600006197	Sky survey	2024-05-17 20:07:44Z	2024-05-17 21:44:03Z	Complete	Complete
08500000098	Important ToO	2024-05-17 13:35:01Z	2024-05-17 20:07:44Z	Complete	Complete

5. Conclusions

With the in-depth astronomical research, there is a higher requirement for real-time and efficient processing of observation data. Facing this trend, this article introduces a hierarchical data processing method based on observation priority and data dependency, as well as a data product integrity analysis algorithm based on the observation task. The method has been applied as a specialized processing pipeline in the EP mission. The operation results show that the method can process data efficiently and the processing time is acceptable. The accuracy of data product integrity assessment has been verified by scientific team in the subsequent processing. These results provide important support for scientists to carry out astronomical data processing and scientific analysis in EP mission. These results also provide valuable experience and technical references for the design of data processing systems for similar astronomical mission such as SVOM.

Acknowledgements

Thank you to the EP Science Centre and Instrument Centre for the contribution in the research and development of data pre-processing pipelines and products validation.

References

- [1] National Research Council, *New Worlds, New Horizons in Astronomy and Astrophysics*, The National Academies Press, Washington, 2010.
- [2] ABBOTT B P, ABBOTT R, ABBOTT T D, GW151226: observation of gravitational waves from a 22-solar-mass binary black hole coalescence, *Physical Review Letters*, 116(2016): 241103.
- [3] YUAN Weimin, ZHANG Chen, CHEN Yong, Einstein probe: exploring the ever-changing X-ray universe, *Scientia Sinica Physica, Mechanica & Astronomica*, 48(2018): 039502.
- [4] MA Fuli, CHEN Ling, LI Bing, Design and implementation of GECAM preprocessing pipeline, *Astronomical Research & Technology*, 19(2022): 274-282.
- [5] CHEN Yajie, WANG Feng, DEGN Hui, The application of elastic search in the massive astronomical data retrieval, *Acta Astronomica Sinica*, 57(2016): 241-251.