Space Mission Data Provenance Traceability

Yi Qu, Haitao Wu, TingLiu and Yue Zhao

Academy of Optical-Electronics, Chinese Academy of Sciences, Beijing, 100094, China

First, this paper proposes the concept of “space mission data provenance traceability” based on relevant concepts from other disciplines. Also the unique features of space data provenience traceability are presented. In particular, its significance to space mission interoperability is considered. Next, this paper analyzes some typical requirements for space mission data provenance traceability from different parties, including mission managers, data producers and data consumers, each of whom will benefit from space mission data provenance traceability. Then, this paper summarizes the design constraints of space mission data provenance traceability. As to the key elements for data provenance traceability, this paper refers to the “W7” model that has been accepted by many other disciplines, and analyzes its applicability to space mission data. Finally, this paper takes telemetry data as an example to identify the key problems of space mission data provenance traceability, including form of data annotations, frequency of data annotations, order and time sequence of data annotations, semantics collision avoidance of data annotations and last step of space mission data provenance tracing. The significance of this paper is to propose the concept of “space mission data provenance traceability”, and to analyze its significance to space mission interoperability, which enriches the content of space data management. This paper also discusses typical requirements on space data provenance traceability, summarizes the design constraints, and identifies key problems. It is expected to provide a theoretical reference to the establishment and implementation of industry standards, and to promote the popularization and application of future space mission data.

I. Introduction

Space missions are quite costly in money, effort and time, therefore their data are very precious and valuable. So much attention should be paid to space mission data management and application in order to promote more scientific research, to achieve ideal scientific objectives, and to accumulate more experiences for later space missions.

However, current state of space mission data management and application is not as satisfactory as imagined. For a long time, space mission data management and application are confined to mission participants and interested parties in mission lifecycle. Even in the background of network interconnection and data sharing, the speed and coverage of data transmission much greater than ever before, space mission data application is still far from universality, durability and efficiency. Some of the reason can be attributed to the absence of data authenticity proof and data quality evaluation.

Actually, when data are published after multiple calculations, extractions, modifications and reprints, it is often difficult to determine their origin, authenticity and quality, therefore their accuracy and integrity are also questioned, which dramatically reduces the value of data, and limits their applications too. This is especially true for space mission data applications which seek for data authenticity strictly.

Therefore, it is very significant to explore the provenance of space mission data, tracking their entire evolvement from generation, processing to publishing. Furthermore there are eager requirements to establish space mission data provenance standard.

In this paper, the concept of “space mission data provenance traceability” is proposed to provide data identification proof to space mission data applications. The concept focuses on the processing and transformations throughout data lifecycle, which may help to improve data quality, credibility, security, reliability and other aspects.
The remainder of the paper is organized as follows. Section 2 discusses the concept of “space mission data provenance traceability”, section 3 lists provenance traceability requirements from mission managers, data producers and data consumers, section 4 analyzes design constraints and key factors of space mission data provenance, section 5 identifies some key problems in implementation, and section 6 provides a brief summary.

II. Concept of Space Mission Data Provenance Traceability

A. Concept

At present, there is no official definition of space mission data traceability. In other disciplines, data provenance has not been defined uniformly yet, and different disciplines provide different explanations.

World Wide Web Consortium (W3C) defined provenance as “a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing” [1]. Buneman defined data provenance as “the process of tracing and recording the origins of data and its movement between databases” [2]. Lanter termed lineage of derived products as “information that describes materials and transformations applied to derive the data” [3] to describe provenance in GIS. In science workflow, Simmhan regarded provenance as “one kind of metadata, pertains to the derivation history of a data product starting from its original sources” [4]. Deelman considered provenance as “end-to-end process at all levels of abstraction, from the science domain level down to the system level” [5]. In archival information systems, provenance was defined as “information that documents the history of the content information. This information tells the origin or source of the content information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated” [6].

Although the expressions listed above are not identical, they imply common meanings, i.e. data provenance describes the history of data, including all information about status changes since their generation. Broadly speaking, data provenance can be seen as the history of data generation and evolution.

Referring to data provenance concept from other disciplines, this paper regards space mission data provenance as the whole process from their acquisition to publish. Subsequent analysis is based on this point.

B. Feature

First, the object of space mission data provenance tracing is heterogeneous data. Space mission data itself is a rich concept, including telemetry data, house-keeping data, science data, auxiliary data and other data. In their generation phase and transmission phase, these data are constrained by many factors, such as payload data acquisition mode, onboard data management mode, onboard data storage capacity, satellite-ground data transmission capacity and so on. In the preprocess phase and deep process phase, they are affected by different discipline backgrounds. These strict constraints and complicated processes lead to variable space mission data structures, also lead to tremendous challenges to space mission data provenance tracing.

Second, space mission data standards are not sufficient in practice. Most space mission systems and data systems are customized at present. Though some standards have been released, they cannot cover all aspects of space missions. Meanwhile, these standards cannot meet growing requirements. Besides, legacy system heritage also leads to the increasing of customized space mission data. Non-standard data and systems will bring difficulty to space data provenance tracing.

Third, there are huge differences in space mission data granularity. In those missions lacking TT&C resources, some device states are represented by only one bit of telemetry data. Whereas in some earth observation missions, the size of an image can exceed 1G bytes. The granularity difference between them can reach $10^{10}$ or even larger. In fact data granularity has a great impact on data provenance traceability. Fine-grained data tracing and coarse-grained data tracing will put forward dramatically different requirements on computing capacity and storage capacity.

Fourth, there are associations among different types of space mission data. Most space mission data are collected and transmitted in data frames or packets, and many of them affect each other. Final data product may be the fusion result of multiple types of data. This feature’s impact on data provenance tracing is that cross platform tracing is inevitable. In addition, most space missions are implemented by many agencies, and distributed systems are the mainstream of space mission systems, which strengthens the requirement on cross platform tracing further.

Because of the existence of these features, it is impossible to copy results from other disciplines directly. Therefore, it is necessary to establish special data tracing methodology and systems for space mission data.

C. Significance to Space Mission Interoperability
Current discussions about space operation interoperability are mainly concentrated on architecture design, software implementation and point-to-point test [7-9]. Few is done on overall system performance evaluation and fault diagnosis, even less is done on the risks and measures when function modules switch.

It is undeniable that in interoperable systems, all function modules will be tested and verified rigorously before putting into practice, also their processing flow will be carefully designed, and most of them will have log functions. However, “plug and play” means the increasing of function module option and replacement, resulting in the increasing uncertainty of data processing path. Besides, plugging during mission operating may lead to data loss and data processing exceptions, especially in multiple mission scenarios.

If there is neither detailed deployment description nor systematic methods to track the plugging behavior, mission systems can rely on nothing but traditional operating log that is rarely open to the public. Since most logs are local, mission systems can hardly exactly tell which function module implemented the data process, which function module triggered the exception and which plugging caused the variation of data quality or integrity.

What’s even worse is that data consumers are unable to investigate data origin and unable to know overall data quality variation. They may question the quality and credibility of all of the mission data just because of a small amount of abnormal data.

Therefore, it is valuable to take space mission data provenance traceability requirement into account in the establishment of interoperability standard.

III. Requirements on Space Mission Data Provenance Traceability

A. Requirements from Mission Managers

By data provenance tracing, space mission managers can get an overview of complete mission data process, from data downlinking to publishing. As we all know, space mission data process is very complex, and many agencies participate in it. Therefore, it is a great challenge to illustrate the complete and detailed process history of space mission data. If plugging occurs in long term operating process, it will increase the difficulty of exploring data evolution history undoubtedly. As a new method, data provenance tracing can provide direct proof to solve this problem.

By space mission data provenance tracing, mission managers can acquire system status in long term operation. In fact, during space mission life cycle, system may experience personnel flowing, equipment aging, software rebuilding and network upgrading. As a result, data process will also change in such a dynamic environment. If data provenance information is available, mission managers can grasp system configuration changes and make adjustment strategy in time.

By space mission data provenance tracing, mission managers can evaluate the accuracy, stability and reliability of every function module, and furthermore evaluate the quality of data it processed. Long term excellent provenance information can accumulate credibility for function modules, and accordingly function module credibility can prove data credibility. Provenance information can be published along with data as a kind of data quality evidence to enhance data consumers’ confidence.

By space mission data provenance tracing, mission managers can evaluate system ability of supporting multiple missions in parallel. As commercial space mission market is gradually forming, sharing modules in time-division mode among systems will be a convenient and economical choice. If statistics on space mission data provenance is made, it can evaluate the continuity of system workflow and the operating efficiency in multiple mission scenarios, and can accumulate experiences for rapid system construction and function module dynamic combination in future.

By space mission data provenance tracing, mission managers can enhance data security and system security to some extent. Data provenance information will increase the cost of malicious tampering, and reduce the cost of data authenticity verification, thus enhancing data credibility. If data are tampered or corrupted, both historical data provenance information and current data provenance information can help to detect attacked functional modules, improving system security.

B. Requirements from Data Producers

By space mission data provenance tracing, data producers can master the accuracy, real-time and other performances of every function module, which can help to improve system efficiency and data processing quality. Similarly, data producers can also fully grasp data filter and extraction process in virtue of data provenance information, helping to evaluate data integrity.

By space mission data provenance tracing, data producers can accelerate system fault location, and implement efficient fault diagnosis. In case of emergency, data producers can obtain the backup information of function modules from data provenance information, and recover the system in time.
By space mission data provenance tracing, data producers can acquire the availability and load of every function module in the mission system, and draw more reasonable planning and scheduling, which can promote system load balancing, and improve overall system efficiency.

C. Requirements from Data Consumers

By space mission data provenance tracing, data consumers can understand data content deeply and carefully, including data background, providers, producing time, process procedure and other related information, which may play an important role in evaluating data integrity, authenticity and science value. Based on these, data consumers can compare or integrate data from multiple missions, outputting more valuable results.

By fine-grained space mission data provenance tracing, data consumers can find out all influence factors in data evolution process, which can form a large data lineage with many branches. If deep analysis is implemented on the lineage, it may even develop into a data association mining pattern, and obtain novel findings.

In summary, by space mission data provenance tracing, mission managers and mission designers can build better systems, and provide better services. While data consumers can inspire deeper understandings, and make better applications. It will promote to form a splendid cycle from mission design, mission operation to mission application.

IV. Design Constraints and Key Factors of Space Mission Data Provenance Traceability

A. Design Constraints

Constraint of data size: space mission data acquisition rate is very high, and many space missions last for a long time, so the size of space mission data themselves is very large. If excessive provenance information is added to space mission data, it will bring serious challenges to network transmission capacity, system storage capacity and computing capacity. Therefore, space mission data provenance information should be designed carefully to seek the balance between information details and implementation environments.

Constraint of uniqueness: space mission data evolution path recorded by provenance information should be specific, without ambiguity or contradiction. The action, participating entity and occurrence time of each step should represent data process logic clearly, though it is difficult to implement in distributed systems. Furthermore, uniqueness requires systematic planning in mission design stage to avoid duplication definitions and conflict definitions in operating stage.

Constraint of verifiability: provenance verifiability and reproducibility is especially emphasized in science workflow in order to achieve the same outputs in subsequent repeated experiments. To space missions, complete reproducibility is very difficult, at least the reproduction of data acquisition onboard and data transmission between satellites and ground is costly. But verifiability should not be ignored. After all, one of the primary purposes of data provenance tracing is to enhance credibility, whereas provenance unable to be verified is not convincing.

Constraint of security: provenance information itself is data, so it may suffer from malicious attack or tampering. The situation is even worse to space missions due to their distributed feature and different security standard adopted by different agencies. Therefore, provenance information security should be considered as early as possible, and PDRR (Protection, Detection, Reaction, and Recovery) principle should be used as reference when necessary.

Constraint of standardization: standardization can promote the interoperability of provenance information, facilitating their storage, visualization and application in future. So this paper suggests making standards for space mission data provenance as soon as possible.

B. Key Factors

Many researchers from other disciplines have studied provenance information model. A representative model is the “W7” model proposed by Sudha, which suggests data provenance information should contain “W7”, what, where, who, when, which, why, and how. [10]

In this model, “What” means events that change data status, including creating, transmission, transforming, and even archiving. “Where” means the place of event occurring (in some cases, it is used as query condition). “Who” means persons or agencies involved in the event. “When” means the time of event occurring. “Which” means tools or software used in the event. “Why” means the reason of event occurring. “How” means actions triggering the event. “What” is the core of this model, and others are details and explanations centered on “what”. [10]

As a complete set of provenance information, these seven aspects are perfect. “what”, “who”, “when” and “where” are main purposes of data provenance tracing, and “which”, “how” and “why” can provide details critical to reproduce data evolution history. They can cover most space mission data provenance tracing requirements, so this paper adopts this model as space mission data provenance information model.
However, this model may bring pressure to system capacity in practice, so proper references or omissions when possible may be more feasible.

V. Key Problems Identified

As mentioned previously, there are many types of space mission data, including telemetry data, house-keeping data, science data, auxiliary data and so on. This paper takes telemetry data as an example to analyze space mission data provenance tracing. The main reason is telemetry data’s universality to space missions, since every space mission has its own telemetry data. What’s more, compared with other data types, telemetry data is more standardized in format, and less affected by discipline backgrounds. Fig.1 illustrates a simplified flow of telemetry data process.

Fig. 1 Simplified Flow of Telemetry Data Process

Traditionally data provenance tracing methods are often divided into methods based on annotations and methods based on query inversion. Space mission data provenance tracing, because of their unique features, such as heterogeneous data, non-standard data, persistent data, data granularity difference and data associations, needs to integrate both of the methods.

Methods based on annotation require data annotations to record data provenance information, and require data annotations to be distributed and transmitted along with data, so that data consumers can get provenance information.
from annotations directly. Their advantage is easy to implement, and is also easy to manage provenance information. Their deficiency is that they need extra storage to store data annotations, and they need special design to realize fine-grained tracing.

Methods based on query inversion are implemented by constructing query inversion functions. Their advantage is that they need neither special annotations nor extra storage. Their deficiency is that their implementations are complex, which need special inverse functions and corresponding verification functions (not all functions have reversibility).

According to space mission data features, methods based on annotations can be used in data streams and data files in/between GS, MCC and PCC, while methods based on query inversion can be used after space mission data enter databases, tracking their processes of migrating, merging and filtering in databases.

In the application of these methods, this paper identified some key problems as followed.

A. Form of Data Annotations

Data annotations can be expressed in concrete text or abstract code.

Text is easy to implement, and its readability is good, but its data size is huge. If annotations of all the 7 aspects mentioned above are expressed in text, annotation size will get much bigger than data size. Furthermore, if every function module adds its annotations to data, the size of data packet may exceed network limitation eventually and the data packet will encounter mandatory splitting, which increases network burden and error probability during transmission. In addition, compared with code, text is less secure, more easily tampered and corrupted.

If annotations are expressed in code, systematic design of encoding algorithm and decoding algorithm is needed. Algorithm complexity should be acceptable so that it cannot degrade system real-time performance. Algorithm portability and reliability are another challenge because of space missions’ distributed feature. The ideal state is to reach standardization level.

B. Frequency of Data Annotations

Current downlink rate of space mission data is getting higher and higher. Usually telemetry data downlink 2 frames per second, and science data may be faster. Adding complete annotations to each frame is neither feasible nor necessary.

Normally data processing flow and function module configuration are identical to all the data frames of the same type in the same downlink cycle. To these frames, provenance information about “who, which and why” are the same, therefore pointers may be used to refer to former frame’s provenance annotations, reducing frequency of data annotations.

However, it should be considered how to keep up with system status and how to add provenance annotations in time when the system configuration is changed or the system is abnormal. Furthermore, how to refer to former frame’s provenance annotations correctly in multiple mission systems is another thorny problem. In the deep processing of space mission data, data from multiple periods and multiple types may be fused, and annotations frequency may increase accordingly.

C. Order and Time Sequence of Data Annotations

As Fig. 2 illustrates, annotation appending position, i.e. annotation is appended in the head of data or in the rear of data, should be systematically standardized in design stage. Besides, data annotation inserting principle, i.e. the relationship between process step order and their annotation order is “first in first out” or “first in last out”, should also be systematically standardized in design stage.
The purpose of keeping annotation order is to enhance readability, while the vital problem is to keep the time sequence of these annotations. In real-time space mission systems, the time duration of many processing steps is very short. If the precision of time code is not enough, it may lead to the phenomenon that provenance annotations of multiple processing steps share the same time code, which is easy to cause system processing logic confusion.

In addition, time sequence maintaining also depends on the unification of system time benchmark (e.g. UTC time vs local time) and the performance of system time synchronization. If GS, MCC and GCC use different time benchmark, or the time difference among them is large, it can lead to asynchronous timestamps in provenance annotations, resulting in time logic error.

If data acquisition onboard is included in space mission data provenance tracking, time difference between satellite and ground should be provided to data consumers as a subsidiary item of provenance annotations, in order to avoid understanding ambiguity of data provenance annotations.

### D. Semantics Collision Avoidance of Data Annotations

Due to space missions’ distributed feature, formalized and standardized design of annotation contents in the whole system should be done in advance, in order to avoid duplication definitions and semantics conflicts among different agencies. If annotations are encoded, encoding resources and encoding algorithms may require pre-allocation in the whole system.

During long term mission operations, some function modules may be replaced and plugged. To these backup modules, their ability of provenance annotation semantic normalization and standardization should be taken into accounts when implementing, and their ability should be verified before putting into system.

### E. Last Step of Space Mission Data Provenance Tracing

The last step of fine-grained space mission data provenance traceability is to trace back to the data onboard acquisition process and downlink process. Thanks to the detailed design of space mission data management systems, we can get the most important and live information from data packets and data frames.

The packet version, APID, packet length and other information can be extracted from packet primary header, and packet time can be extracted from packet secondary header or data area (it doesn’t work to those packets without secondary header and time parameter in data area). The frame version, spacecraft ID, virtual channel ID and other information can be extracted from frame primary header.

![Telemetry Data Package Structure Illustration](image)
After decades of development, space missions are no longer limited to a small number of science missions. Space missions have been widely applied in our daily life, facing global commercialization and industrialization. Space mission data embody the value of space missions to a great degree. As space mission data consumers are expanding, the requirements of data quality and reliability are also growing. It is a critical problem to implement effective management to enhance the reliability and the value of space mission data. Data provenance tracing can be used as a technical means to meet these requirements.

This paper proposed the concept of space mission data provenance traceability, analyzed its significance to mission managers, data producers and data consumers, discussed the design constraints and key elements, identified some key problems, hoping to provide reference for the establishment of the relevant standards, and to promote the popularization and application of space mission data.

**Acknowledgments**

This work was supported by the National Natural Science Foundation of China (No. 91438207).

**References**


---

**Fig. 4 Telemetry Frame Structure Illustration**

<table>
<thead>
<tr>
<th>Version</th>
<th>Spacecraft ID</th>
<th>Virtual Channel ID</th>
<th>Operation Control Flag</th>
<th>Frame Identification</th>
<th>Frame Counter</th>
<th>Virtual Channel Counter</th>
<th>Frame Data Area Status</th>
<th>Frame Secondary Header (Optional)</th>
<th>Frame Data</th>
<th>Operation Control (Optional)</th>
<th>Error Control (Optional)</th>
</tr>
</thead>
</table>

VI. Summary

**Table:** Telemetry Frame Structure Illustration

![Telemetry Frame Structure Illustration](image-url)