Long-term archiving of Japanese space science data

Ken EBISAWA [(1)], Iku SHINOHARA[(1)], Kei-ichi MATSUZAKI[(1)], Yukio YAMAMOTO[(1)]

[(1)] ***Center for Space-Science Operation and Data Archive (C-SODA), ISAS/JAXA***
3-1-1 Yoshinodai, Sagamihara, Kanagawa, 252-5210, Japan

EMail: Ebisawa.Ken@jaxa.jp

ABSTRACT

We have started discussion how to ensure long-term archives for Japanese space science data. Here, we present an interim report of our discussion. We have first defined purposes and principles of the data archives, and then systematically reviewed current status of the archive developments of Japanese space science missions. We have identified common problems of data archive development in different science fields or satellites. Results of our discussion will apply not only to Japanese space science communities, but also to wide international science communities.

Keywords: data, processing, preserving, services, archive, space, science, satellite

INTRODUCTION

JAXA (Japan Aeronautics exploration Agency) is Japan's sole public space organization, which launches scientific satellites as well as practical satellites. These scientific satellites cover extensive fields, space astronomy (X-ray, infrared, radio), solar physics, solar-terrestrial physics (STP), and lunar and planetary science. Most of the space science data taken by JAXA's satellites are accumulated at ISAS (Institute of Space and Astronautical Science), which is a part of JAXA, and their high-level data products are publically available from DARTS (Data ARchives and Transmission System; http://darts.isas.jaxa.jp), operated by Center of Science-satellite Operation and Data Archive (C-SODA). At present, data products from about a dozen scientific satellites launched after late 1980's are archived in DARTS.

Today, as JAXA's satellites are becoming larger and more complex, a large amount of high quality science data is being produced. However, compared to the current status of satellite and instrument developments, status of the science data archives at JAXA may not necessarily be satisfactory. This is primarily because there are hardly clear principles and guidelines of science data archives at JAXA.

Under these circumstances, we have started discussion in order to ensure high quality science data archives for Japanese space science data. The discussion is primarily made in the steering committee of C-SODA, involving experts in different science communities. We aim to define purposes and principles of science data archives, to identify common problems involved in archive developments, and to establish practical guidelines and strategies to resolve these problems. We hope results of our discussion are valid not only Japanese space science community, but also international science communities. Thus, here we present an interim report of the summary of our discussion.

SCIENCE DATA ARCHIVES AND THEIR PURPOSES

All the scientific data taken by satellites should be properly processed to produce the high-level data products so that anybody can obtain them, via some data services, for scientific purposes for free of charge, and publish results. The system to enable this, an entity composed of the data processing, data preserving and data services, is called "data archive" here (see the Appendix). Purposes of archiving science data are the following:

1. **Ensure reproducibility and universality of the results obtained from data**

   Any scientific results should be reproducible and universal. If some results obtained from data by some methods cannot be reproduced by other parties, the results may not be trusted. When anybody can reproduce the same scientific results from data archives, the results will become universal. In particular, we may not observe exactly the same phenomena repeatedly in space science. Thus, in order to examine observed phenomena by plural scientists from different perspectives, data archiving is indispensable.

2. **Expand the lifetime of data**

   Lifetime of satellites is an order of years. On the other hand, lifetime of data is indefinite, as long as the data are archived. Thanks to the data archives, scientific results can be produced even long after the satellite missions are terminated. Also, since timescales of cosmic phenomena are often much longer than those of satellites, past satellites data are commonly used to study long-term phenomena. Namely, archiving science data will support scientific research in future.

3. **Expand the range where data are used**

   Scientific results which may be produced by the satellite project members are limited. By opening the data archive to the world-wide community, more scientific outcomes are expected.

4. **Contribute to international development of science**

   Developments of satellites are publically supported, thus their outcome should be returned to general public. Since products of the science satellites are scientific data, the return should be guaranteed by scientific results produced from the data. Thus, opening the science data archive to the worldwide science community is considered to be Japanese contribution to the international development of science.

PRINCIPLES OF DATA ARCHIVES

Currently, situations of science satellite data archives significantly vary in different scientific fields and/or satellites; sometimes, high-quality data archives are constructed and scientific results are being produced, but sometimes not. This is primarily because we do not have such basic principles and guidelines in data archive developments that should be followed in any stages of planning, designing and developing science satellites. In many cases, developments of science archive are made as “best-efforts”, so that their outcomes are not guaranteed.

In order to improve the situation and to maximally utilize the science satellite data by domestic and international communities, we propose the following principles regarding the use of science data. Both sides of the parties proposing and evaluating satellite missions should keep these principles in minds in any stages of the satellite projects:

1. **Principle of the data processing**: All the science data should be properly processed by applying instrumental calibration and algorithm to produce high-level data products, from which scientific results are derived only assuming common knowledge.
2. **Principle of the data preserving**: All the acquired science data should be permanently preserved under the usable conditions.
3. **Principle of the data service**: Data centers should, acknowledging the data providers, provide basic data services, for free of charge, toward a wide range of user communities, so that the data are used for a long period of time.

Some explanations for each principle follow:

Principe of the data processing

Raw telemetry data from satellites may not be interpreted as physical quantities as they are. Only after the telemetry data are decoded taking account of specific data formats and the instrumental characteristics are subsequently removed (i.e., calibration of data), the observed data are interpreted as physical values. It is not very easy to process data up to that level, but the satellite projects should be responsible for their own data processing. In this manner, scientists who do not have knowledge specific to the satellites can use the satellite data, as long as they have common scientific knowledge in the fields.

In general, satellite observation data, as opposed to ground-based observation data, are more suitable for standardized data processing ("pipeline processing"), because observational conditions tend to be more uniform or standardized. Such pipeline processing should be introduced to every science satellite, while data product levels of the pipeline processing are different for individual satellites. Each science satellite project should define what kind of data products are produced as the final products of the pipeline processing.

Executing pipeline processing is considered to be an engineering activity rather than scientific activity. In order to develop the pipeline processing system, both scientific and engineering knowledge are necessary. Thus, the processing system should be developed by the satellite project in cooperation with scientists, but once the system is developed, engineers or technicians may run the pipeline processing systematically. The pipeline processing may be run somewhere besides the space agency or satellite operation center. Each satellite project is responsible for clarifying the organizational structures of executing the pipeline processing to produce data products. Also, the space agency is responsible for clarifying the quality of the data products, as requested by users, regardless of where the pipeline processing is executed.

Principle of the data preserving

It is obvious that all the scientific discoveries have to be objective and reproducible. Scientific results from satellites have to be reproduced by third parties, but this is often difficult because celestial phenomena are not controllable. In fact, it occasionally happens that scientific "discoveries" made by instrumental teams using proprietary data are denied much later by another group based on higher quality observations. What is common behinds such false discoveries is that those original data and algorithms (or software) were not preserved nor released, so that independent examination of the data by third parties was impossible.

In order to avoid such circumstances, any data, software, and algorithms used to obtain new discoveries or new knowledge should be properly preserved so that third parties can reproduce the same results from the same data. This has a merit for data providers too, to ensure fidelity of the scientific results. Those preserved scientific data should be considered as "public properties", not as the objects to claim for intellectual property right. In the end, all the scientific data should be publicly released for free of charge.

On the other hand, often it is a strong motivation of progressing the projects for scientists to make a discovery using their own instruments. Therefore, naturally, the satellite project team may well have initial proprietary right to use data. Also, for engineering data, data are often not released to keep technical competitiveness. In such a case, the satellite project should clarify the reason and the expected merit of not releasing the data.

When satellite observations are carried out by open proposals, original proposers may have initial data right. The satellite project is responsible for clarifying the data policy and the open proposal system, such as allocation of observation time, data proprietary periods, dates of data release, etc.

Principle of the data service

Data providers create data products, and data centers should provide services to preserve and release data. Even if science data are properly processed and preserved, if they are not easy to use for outside users, scientific outcomes may not be expected. Thus, data centers should provide user-friendly services for free of charge. Developing such high quality services may not be straightforward, and should not be

considered as a secondary job for scientists. Developing data service should be defined and respected as an independent task of the satellite development or scientific research.

Also, in order to promote data usage, adequate user supports (such as a help desk) are necessary as well as data services. It should be reviewed if proper resources are invested into user supports in accordance with reviewing status of the data archive development.

When data are released, the data providers should be explicitly indicated. In this manner, responsibility of the data will be clarified, and the data provider is properly acknowledged by data users.

RELATIONS OF DATA PROCESSING, PRESERVING AND SERVICE

In the discussion of data archive, it is important to distinguish data processing, data preserving and data service, and their relations. In this document, we consider the data archive as an entity composed of data processing, data preserving and data service (see Appendix). Their relations are schematically indicated in Figure 1. Namely, data from the data providers are preserved via data service and data processing. Data users obtain the preserved data via data service (optionally after data processing).

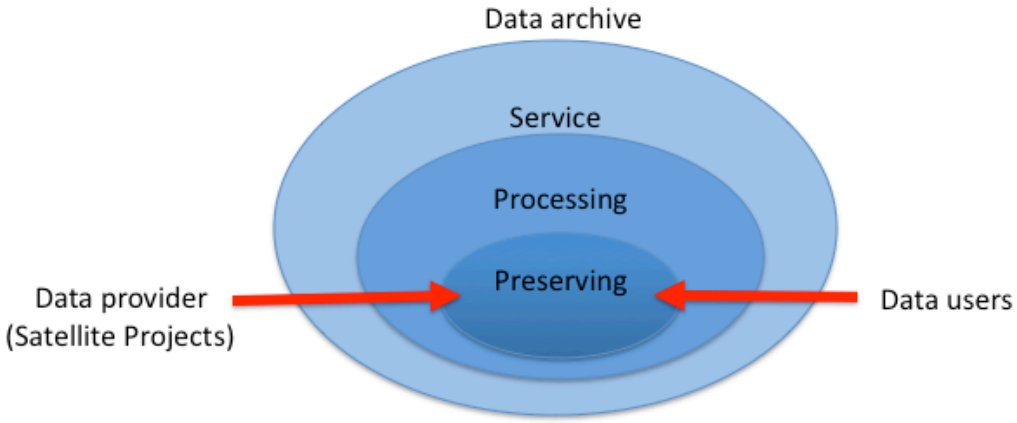

Figure 1: Relations of data processing, data preserving and data service

It is important to recognize that requirements for data preserving, date processing and data services are different. This also applies to the required skills for the people in charge of each process. For example, ones in charge of the data preserving require knowledge of the secure data storage, rather than scientific knowledge or engineering knowledge of the satellite. Ones in charge of the data processing need knowledge of the instruments and/or scientific background in the field. Data services may include, for example, functionalities of searching or browsing data, and their developments may require advanced knowledge of information technologies.

For data users, data archive may just look like a place where the data are found, and the users do not have to know presence of the three layers behind it. However, for archive developers, it is important to distinguish the three layers (processes), and clarify interfaces between them and the bodies in charge of each process.

COMMON PROBLEMS IN ARCHIVE DEVELOPMENT

In order to identify common problems in archive development for different satellites and scientific fields, we have thoroughly reviewed status of the current data release and usage for all the Japanese science satellites. We have evaluated the status for each satellite, instrument or data product, based on the following points of evaluation:

1. Data policy – Are data policies defined, preferably in the initial phase of the science project, to ensure producing, preserving and releasing data products, and constructing long-term data archives?
2. Complete data processing – Are all the acquired data properly processed to produce high-level data products?
3. Internal documents – Do internal documents exist which enable one to continue archive development and operation after the persons in charge have left?
4. Long-term data preservation – Are scientific high-level data preserved indefinitely in a usable form?
5. Data release – Are the high-level scientific data products publicly released?
6. Data services – Are any data services provided for general users to facilitate data use?
7. Data descriptive documents – Do such documents exist to describe data formats and their meanings that enable one to use the data without relying on particular software?

After evaluating each satellite, instrument and data product, we have extracted problems for each science field. After that, we have extracted common problems over different scientific fields. We thus identified the following six common problems:

1. Non-uniformity of the satellite projects subjective of the long-term data archiving:

   In the scientific fields with long histories and traditions, long-term data archives are relatively developed well compared to rather new scientific fields. For example, data from the instruments on-board International Space Station are, currently, not archived envisioning for long-term use (at lease in Japanese science community).

2. Ambiguity of data policies:

   It is almost always true that when data processing, preserving or services are not going smoothly, proper data policies are lacking. The opposite is also true, that long-term archives are developed well, as long as data policies are properly defined. It is desirable to define data policies in early stages of the satellite projects to describe how data are processed, preserved and released, envisioning the long-term archive after the mission is terminated.

3. Uncertainty of data usability:

   It often happens that apparently unusable or unnecessary data are archived, and resources are invested to maintain such systems. It is important to distinguish usable data and unusable data in the archives, and to specify how long unusable data is preserved.

4. Ambiguity of responsibilities among different organizations:

   It often happens that two or more institutes or organizations are involved in data processing, preserving and services. When data archives are developed outside of the space agency, often their long-term operability is not ensured. It may be necessary for the space agency to take over the archives when the outside parties are not able to maintain. In such a case, it is important to clarify division of responsibilities among different institutes. Even within the same space agency, different branches may be involved in the development of data archives, so it is important to clarify division of the responsibilities for each satellite project.

5. Missing documents to ensure transition from the short-term archive to the long-term archive:

   Short-term archives are developed by primarily satellite project members to satisfy missions' specific needs during the mission period. After the mission is terminated, long-term archives are to be maintained indefinitely by a data center for a wide range of user communities (see Appendix). Rolls and requirements of the short-term archive and the long-term archive are different, and it is important to ensure smooth transition from the former to the latter. However, it often happens that the transition does not go smoothly, primarily because of lack of the documents to describe short-term archives and make the transition possible. It is necessary to fully describe the short-term archive in design documents, so that the transition from the short-term archive to the long-term archive go smoothly.

6. Shortage of resources and/or non-systematic developments:

   Even when it is clear what to do in order to develop archives, it often happens that the development does not proceed smoothly due to lack of resources (human and/or budget). In the same science community, often operation and archive development of the current satellite go in parallel with development of the future satellite, and more resources tend to be put in the future project. Therefore, it is important to put aside resources for the development of data archives of the current project, independently of the future projects.

   Also, archive developments often necessarily depend on particular personnel or teams, and quality of the archive is dependent on their abilities. It is more desirable that archive development is carried out in a systematic manner, so that the developments do not dependent on abilities of particular personnel or teams. To that end, it is desirable to develop generic software, tools and documents that are usable in the archive developments in different scientific fields.

MANDATORY REVIEWS OF ARCHIVE DEVELOPMENTS

In order to solve the problems mentioned above, we propose to mandate reviews of the data archive development at each step of the satellite mission phases. We are going to fix, for example, when the archive reviews should be held, how the reviews should be chosen, and what should be reviewed at each review. In this manner, we are planning to make a clear guideline document for archive development, which each satellite project may follow. Details of the document are under-discussion, and we hope to report summary of our discussion in the next occasion.

CONCLUSION

In order to ensure long-term archiving of Japanese space science data, we first defined purposes and principles of the data archives, and then systematically reviewed current status of the archive developments of Japanese space science missions. We have identified common problems of data archive developments in different science fields or satellites. In order to solve these problems, we are going to propose mandatory archive development reviews, of which details are currently under discussion. We consider our discussion applies to not only Japanese space science communities, but also international science communities in general.

APPENDIX – GLOSSARY OF TERMINOLOGIES

In space science community, meanings of the terminologies regarding data archive developments are not necessarily clarified. Therefore, we define and clarify some of the terminologies used in our discussion.

- Calibration

  "Calibration of instruments" is the action to study instrumental characteristics and determine physical parameters (calibration data) to characterize instruments. "Calibration of data" is the action to apply the calibration data to the observed raw data, and determine physical parameters of the observed targets.

- Data products

  Products of a posteriori processing of data obtained by scientific observations, so that they become more suitable for scientific study.

- Data processing

  To determine and apply algorisms to produce data products, and related tasks. This includes not only converting data product levels (calibration of data), but also producing ancillary data, software, documents, constructing models, and calibration of instruments. What should be included in data processing depends on each satellite project.

- Data product levels

  Data processing consists of several steps, and outputs of each step are distinguished by data product levels. We follow the definition of data product levels in the PDS reference manual [1] from Level 1 to 5, which are, Raw data, Edited data, Calibrated data, Resampled data and Derived data. However, individual satellites may have different conventions of the data product levels. In such a case, it is desirable to clarify the mapping between the PDS convention and the local convention.

- Pipeline

  Automatic data processing carried out without human intervention. "Pipeline processing" will create data products automatically, using "pipeline scripts".

- Data provider

  The one who will provide the data products. In addition to the satellite projects and the instrument Principle Investigators, the data processing teams can also be data providers.

- Data service

  Functionalities to facilitate using and preserving data products, aimed for data users and data providers, respectively.

- Data archive

An entity of combination of the data processing, data preserving and data service, assuming to be publically available. Not only data products, but also such documents, software and other peripheral information that are necessary to use data products are included in the data archive. In general, after some proprietary period is over, anybody can obtain and use data products from data archives and publish scientific results.

- Data archiving

  To construct data archives by data processing, preserving the data products and providing data services.

- Short-term archive

  Data archives primarily developed and used by on-going satellite projects. Main users are those involved in the satellite projects or those in the same scientific communities.

- Long-term archive

  Data archives that are maintained indefinitely after satellite projects are finished, so that the ones who are not familiar with the original satellite projects can use the data products. Users in other science communities are also assumed.

REFERENCES

[1] – "Planetary Data System Stadard Reference", Jet Propulsion Laboratory, California Insittute of Technology, Pasadena, California

We acknowledge all the members of the C-SODA steering committee for fruitful discussion.

Ken EBISAWA, professor, Department of Space Science Data Analysis, and Center for Science-satellite Operation and Data Archive (C-SODA), ISAS/JAXA, Japan

Iku SHINOHARA, associate professor, Department of Space Science Data Analysis, and Center for Science-satellite Operation and Data Archive (C-SODA), ISAS/JAXA, Japan

Kei-ichi MATSUZAKI, associate professor, Department of Space Science Data Analysis, and Center for Science-satellite Operation and Data Archive (C-SODA), ISAS/JAXA, Japan

Yukio YAMAMOTO, assistant professor, Department of Space Science Data Analysis, and Center for Science-satellite Operation and Data Archive (C-SODA), ISAS/JAXA, Japan