A STRATEGY FOR THE LONG-TERM PRESERVATION OF SPACE MISSION DATA

Claude HUC
French Space Agency
CNES, CT/TI/PS/AP
18 avenue Edouard Belin
31055 Toulouse Cedex
France
HUC@CST.CNES.FR

Abstract

After explaining why the long-term preservation of data stemming from scientific space experiments is essential, this article analyses the main difficulties which at present obstruct access to this data. On the basis of this analysis, and taking into account the rapid advances within the technological context, a strategy is proposed to improve methods for ensuring the long-term preservation and availability of this data.

The data concerned

An enormous amount of data stemming from scientific space missions has been accumulated since the beginning of the Space Era. This data covers the study of the universe (the sun, the planets, astronomy, etc...) to the same extent as the observation of our terrestrial environment and its evolution (observation of the Earth, atmospheric study, oceanology, etc...)

The data collected is as highly diverse with respect to the way the information is organised as it is with respect to the nature of the observations themselves. Part of this data is available through directly readable media: graphs on paper, films, microfiches, etc... but most of it has been archived in a digital format. This article concerns itself uniquely with the latter form of organisation.

The quantity of information accumulated is enormous and amounts, on the international level, to millions of magnetic tapes, Compact Discs, cartridges, and other forms of digital storage.

Reasons for long-term data preservation

The need to preserve our historical and cultural heritage has long been acknowledged, as this heritage is such an integral part of our societies collective consciousness. On the contrary, a clear understanding of the need for the long-term conservation of data stemming from scientific space experiments is quite recent. During the 1970's, it was still thought that science would progress with each space mission. Thus a new mission, with more effective instruments, could be defined in terms of the previous results and would in turn provide much higher quality information than its predecessors. Experience has shown that this linear view of scientific progress does not match the reality. Moreover, while numerous physical and chemical experiments are entirely reproducible in a laboratory, this is no longer the case when the experimental objective is to analyse the evolution of our terrestrial and planetary environment over time.

The investigation and analysis of a space mission data can in practice last twenty years or more. Numerous observed phenomena are unique (solar events, comets, etc...). The origin of many phenomena has not yet been elucidated, thus requiring that relevant observations be conserved. Over the past 20 years, many scientific theories have evolved, allowing for the re-examination of phenomena from a fresh point of view. Certain phenomena, in particular those which concern our terrestrial environment, evolve very slowly, and observations spread over several decades are needed to understand them. This is the case for transformations of the global climate, the atmosphere and the ozone layer. Our time is witnessing the appearance of new branches of science, as well as the overlap of disciplines, which leads to the use of data in domains unforeseen at the outset.

In economic terms, space missions are expensive for national communities, and it is much cheaper to preserve the data than to launch a new mission.

Analysis of current difficulties in data preservation and access

The preservation of digital data can be summed-up by three simple requirements:

- The data should be physically intact,
- The necessary hardware and software for reading this data should be available,
- As scientific data is not simple to use, the information allowing for correct interpretation of this data should be available.

The loss of information linked to the deterioration of physical data storage media is today a significant potential risk, as the loses are generally irreversible. Luckily, only slight loses have been recorded up until now. During the 1970's and 1980's, the data was generally stored on digital magnetic tape (Compatible computer Tapes). For example, nearly 100,000 such tapes are being stored by the French Space Agency at the Toulouse Space Center. As for internationally available information, this data on magnetic tape has been transferred or is in the process of being transferred to other forms of physical media: magnetic cartridges, optical discs, etc...

The main current obstacle to access to old data is of another order: it is most often the case that each space experiment is a unique prototype which is different from all preceding experiments, and thus the nature and organization of the data collected changes with each experiment. Consequently, each experiment requires new software for the reading and analysis of its data. This has lead to the development of an enormous quantity of software tools, and these are far from being totally independent of the systems within which they operate. In consequence, the software used in the 1970's, on computers which are nowadays obsolete, no longer works, and the corresponding data can only be accessed after costly adaptation of the reading software.

Another important difficulty should be highlighted: in order to correctly process the data, it is generally necessary to know how the experiment was constructed and how it operated. Furthermore, it is necessary to know with no ambiguity how the data is organised, what each elementary piece of information signifies, which measurement unit was used, etc... In other words, it is not only a question of preserving the data, but also of preserving all the information required for interpreting and using this data. This information is usually called the metadata. The main difficulty is that this metadata must allow for the re-use of the data in the future, and that by researchers who will not have participated in the initial experiment. For the oldest data, certain parts of this metadata are lacking, making re-use of the data highly difficult, if not impossible.

Aware of these problems, the national and international space agencies have started to organise institutions or departments for salvaging and preserving this data heritage.

It is on the basis of this analysis and this awareness of current access difficulties that a new preservation strategy has been defined.

The spectacular development of technology

Over the past 20 years, a veritable explosion has occured in the diversity and, above all, in the capacity of digital data storage media. This storage capacity has been multiplied by more than 1000 for unitary media, and a sizeable additional increase beyond that provided by the organisation of media inside robots.

The information storage capacity of these media is measured most frequently in bytes. A byte allows for the storage of one character, for example a letter or digit. For large capacity media, terms typically used are megabytes (10^6 bytes), gigabytes (10^9 bytes) and terabytes (10^{12} bytes).

In 1972, the 800 bpi magnetic tape allowed for the storage of 18 megabytes. Then, between 1975 and 1980, this tape's capacity increased first to 36 and then to 120 megabytes. Since the 1980's, new high-capacity storage media have appeared, for example:

- optical techniques: the Compact Disc (600 megabytes),
- magnetic techniques: the DAT (1.3 gigabytes),
- video techniques: Ampex or Sony cartridges (up to 96 gigabytes).

In 1975, 30,000 magnetic tapes, or 150 cupboards, were necessary for the storage of a terabyte. Today, robot manufacturers boast 3 to 10 terabytes of storage per m^2.

It may be estimated that the cost of storage diminishes by no less than a factor of two every two years. Naturally, the progress concerns not only storage capacity, but also media lifetime, reading and writing rates, conformity to acknowledged industry standards, etc...

As far as long-distance access of archived data via information networks is concerned, the techniques required for access at a very high transmission rate have become available. The rate of data transfer in an information network is measured in bits per second. The bit is by definition the most elementary piece of information, and can take only two values: either 0 or 1. A byte constitutes of 8 bits.

- for the local-area or short-distance (100 km) networks, rates of 100 megabits per second are available with Ethernet and FDDI (Fiber Data Distributed Interface) networks,
- for long-distance networks, the X25 protocol currently used by France Telecom registers a rate of 2 megabits per second. Rates from 34 to 155 megabits per second are available with new technology called ATM (Asynchronous Transfer Mode) already in use in certain sectors.

Over the period of a few years, technology has advanced from thousands of bits per second to millions of bits per second. Today, 600 megabits per second is already being envisaged. **The consequences of this development in network capacity is that the place where the data is archived is no longer necessarily the place where it is exploited**; these two sites can in fact be a great distance from one another.

Similar advances have occured in computing power.

These technological advances present **obvious advantages:**

- conservation of digital data with no or negligible deterioration,
- ease of duplication and transport without deterioration,
- increased performances and lower costs.

However they also involve **serious inconveniences:**

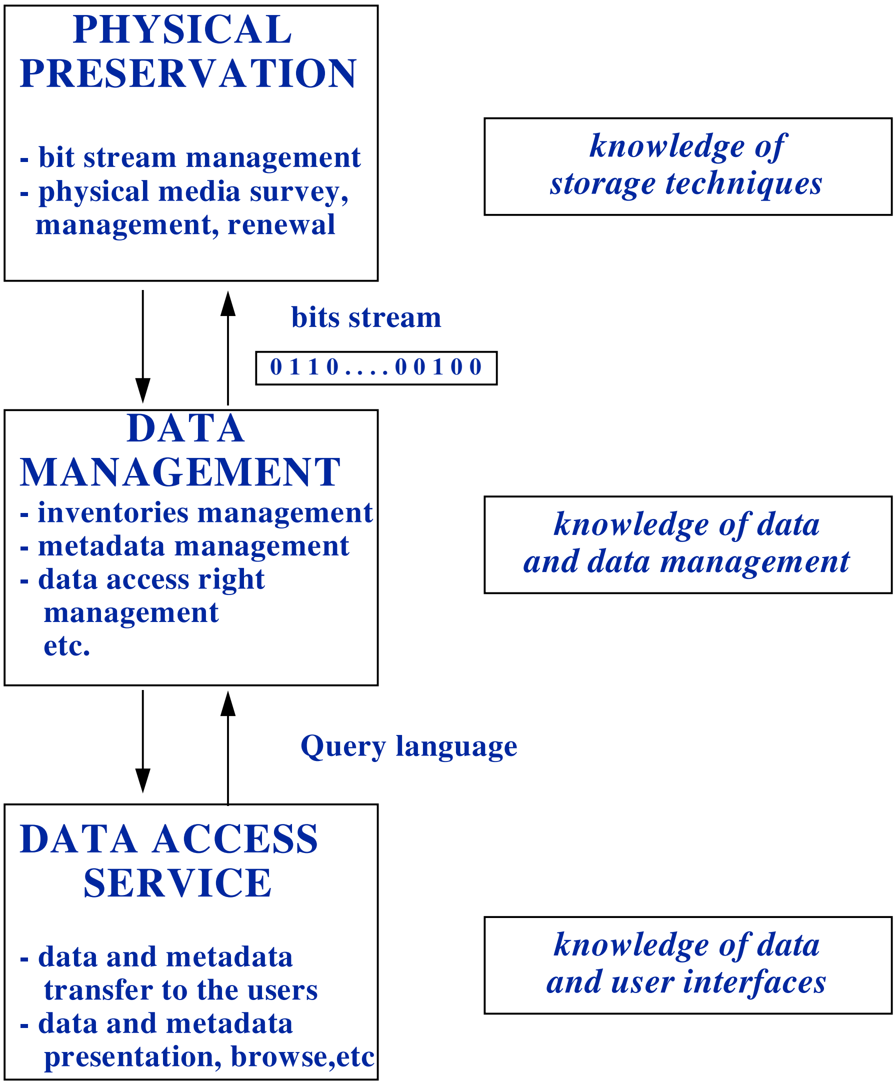
- the technology being used very quickly becomes obsolete,
- risks of data and information loss are considerable if adequate precautions are not taken,
- the users are completely dependent on the development of a market which obeys its own laws.

The main guidelines of the preservation strategy

In order to meet the necessity for long-term data preservation while remaining in control of technological advances, the preservation strategy should be based on the following principles :

- a system model built of successive layers,
- the application of certain standards.

A layered model : This model defines a system reference architecture allowing data preservation and access.



This model, defined by the French Space Agency, does not pretend to be unique. Nevertheless, it presents two major advantages:

- each layer communicates with the other layers by means of standardised interfaces. Each layer is thus autonomous as regards its own development. Development linked to technological progress is possible, but the related undesirable effects and chain reactions noted in the previous generations of systems have been minimised. A change in the physical storage medium will thus have no effect on data management. The use of the new data management system will likewise have no impact on the storage media.

- the second advantage resides in the fact that the human skills required for each layer are distinct: the layered structure greatly facilitates the organisation of these skills.

Consider, for example, the physical data preservation layer. Despite the sizeable progress achieved in quality and durability of storage media, an institution in charge of the preservation of a heritage cannot blindly trust the storage media qualities announced by the manufacturer. The body must introduce a technical policy concerning the surveillance and preventive renewal of these media. For this purpose, use must be made of methods, hardware and software to handle this surveillance and these renewals. The specific task in question is independent of the nature of the data to be preserved, be it text, sounds, images, scientific data, etc...

Likewise, data management, in logical terms, is similar to inventory management. This management consists of having at one's disposal a list of available 'data objects', and of knowing the characteristics of each data object without having to be aware of the type of medium on which the data is being stored. This activity is organised on the basis of knowledge and logical understanding of the data, which was not the case for the previous layer.

The last layer corresponds briefly to what the user (generally a laboratory researcher) can see and do from within his work environment. In particular, this layer concerns such functions as data and information presentation, data selection, and data requests.

The French Space Agency is currently testing this model. An archiving service which assures physical data preservation, the STAF (*Service d'Archivage et de Transfert de Fichiers*), has been created and put into operation. Space data servers for the national and international scientific communities are currently being developed.

The use of standard formats

It is frequently the case that when data is created with the help of a computer, this latter inserts additional information into the user's data, often called 'control words'. These control words are used by the 'system' which drives the computer in order to administer this data. These control words are invisible to the user, but are nevertheless present in the data. Furthermore, these control words are not standardised and are incompatible from one computer to another; they are the cause behind much of the difficulty in accessing old data. When creating this data, particular precaution should therefore be taken to choose options which do not make use of these control words. Archived data should consist uniquely of usable data. Likewise, the encoding used for the representation of numbers (integers, real numbers) should be standard encoding, and not the specific numerical representations of such and such a manufacturer.

The basic precautions indicated above, which consist of making the data independent of the computers, are however far from being enough. It has already been said that it is illusory to try and maintain in working order all of the specific software used to read and analyse the preserved datasets. The only possible solution would be to create so-called 'generic' software, i.e. software capable of accessing and reading all the data in a given scientific field, regardless of the data's origin. The condition for this solution to be workable is that the data in question must respect a certain number of organisational and structuring rules. In other words, this data must conform to standards that define not only the data organisation and structuring but also the data description. These standards have emerged little by little: note for example the SFDU (Standard Formatted Data Unit) defined by the Consultative Committee for Space Data Systems, and likewise the FITS (Flexible Image Transport System) acknowledged worldwide in the field of astronomy, or even the HDF (Hierarchical Data Format) used mainly in the terrestrial environmental sciences, etc... Certain of these standards are acknowledged by the international standardisation organisations, while others are *de facto* standards. These standards do not all have the same functions, and some are competing. Still, their emergence and undeniable success is a major asset when it comes to resolving the difficult question of long-term data preservation.

Global availability of space data through the International Directory Network (IDN)

It is impossible to end this article without having mentioned the International Directory Network.

In scientific experiments, users are researchers and laboratories interested in data analysis. Generally, each experiment embarked has been carried out under the supervision of a scientist (the principal investigator) who, along with his team, owns exclusive rights to use the data for several years. After this period, the data becomes available to the national or international scientific community. Taking into account the high cost of space experiments, it is in fact justifiable to market the data produced by these experiments whenever possible. This means making the data widely available, so that the greatest number of researchers can make use of it.

Moreover, taking into account the international character of most scientific missions, it follows that the community of users is worldwide and that these users are almost always found at a great distance from the place where the data is archived.

This is why, in response to the suggestions of the Committee on Earth Observation Satellites (CEOS), the IDN was created as a type of great international catalogue of available scientific data, both spatial and otherwise. The IDN is administered by the CEOS, and access facilities are provided jointly by the American, European and Japanese space agencies. It consists of a data base which itemizes the available data sets in the international community. This data base is freely accessible on a permanent basis through digital networks (Internet, etc...). The available information for each set of data consists mainly of its general scientific and technical characteristics (scientific discipline, parameters measured, measurement periods, instruments used, etc...), as well as the name and address of the 'data center' in whose possession the data can be found. This description of the data set is provided in English according to a standardised format. A specialised consultation system allows each user to consult the IDN in terms of his own criteria of interest.

The IDN therefore plays an essential role in providing access to data to the worldwide scientific community today. Furthermore, it paves the way towards a greater compatibility among current data access systems.

Conclusion

This article has concentrated on the technical aspects of data preservation and access. It should be emphasized that these technical aspects must be accompanied by the arrangements of human organisations, notably as regards the durability of those bodies in charge of preserving the data and of making it available.