Bull. Spec. Astrophys. Obs., 2000, 49, 128-140

Storage of observational data and establishment of archive systems

V.K. Kononov, V.E. Panchuk

Special Astrophysical Observatory of the Russian AS, Nizhnij Arkhyz 369167, Russia

Abstract. The role played by archive systems in the technological chain of obtaining astrophysical results is discussed. The most important factors that have an effect on performance of the present-day automated systems of experimental data storage are considered.

Key words: methods: data analysis – astronomical archives – methods: numerical

1. Introduction

Two basic processes that complement each other provide for functioning of SAO as an astronomical centre:

1. Doing a variety of astrophysical experiments (observations).
2. Subsequent processing and analysis of obtained results.

The two processes are directly bound with the professional activity of astronomers who, on the one hand plan and implement their observing programmes on the telescopes and, on the other hand, carry out subsequent processing of observational data.

The observations are supposed to use a set of different facilities and techniques which, in the aggregate, determine the technology of the observational process. Here belong the telescopes, their control systems, astronomical apparatus, receiving-and-measuring complexes, data acquisition systems, and possible modes of their joint operation. Results of observations are always presented in a certain form allowing the astronomers to perform their further transformation and analysis.

The processing of experimental data has a technology of its own, which is based on using computing means and consists in application of diversified programme systems to implement both the simplest procedures of reduction and complex procedures of interpretation. The processing of experimental material is essentially an iterative process and normally takes more time than the observations proper. In particular, this is due also to frequent involvement of additional results of observations for the same programme or other programmes that may have even been carried out at other telescopes.

Thus the technological chain "observation – processing" provides for existence of an intermediate link that will ensure accumulation and storage of experimental data. This link is an *observational data archive.*

In the general case the archive of observational data may be viewed as serving a double purpose:

- Any observation results in acquiring experimental data in a specific form determined by a suitable acquisition system. If, in addition to this, regular accumulation of the information to be stored for a long time is executed, the concept of astrophysical experiment can then be naturally widened: the standard process of archiving becomes its part. In this case, the archive, as an informational system, is one of the technological links of the system of automation of astrophysical experiment. Here one may regard the observations to be formally finished after the results have been archived or, at least, after their preliminary buffering for consistent communication to the archive system.
- Subsequent processing of data makes provision for access to any data previoulsy accumulated and being stored. This highlights that the archive is closely allied with a second major step — obtaining of astrophysical results. The archive here is the principal source of original data, being thus one of the technological links of the process of transformation of information.

So, until the final results are obtained, the activity of the astronomers is supported by the collection of means representing a combination of interrelated systems (Verkhodanov et al., 1996), a particular part being played by the chain

$$\text{Acquisition} \longrightarrow \text{Archiving} \longrightarrow \text{Processing},$$

which forms a basis of the present-day technology of working with experimental information at any astrophysical institution (Fig. 1).

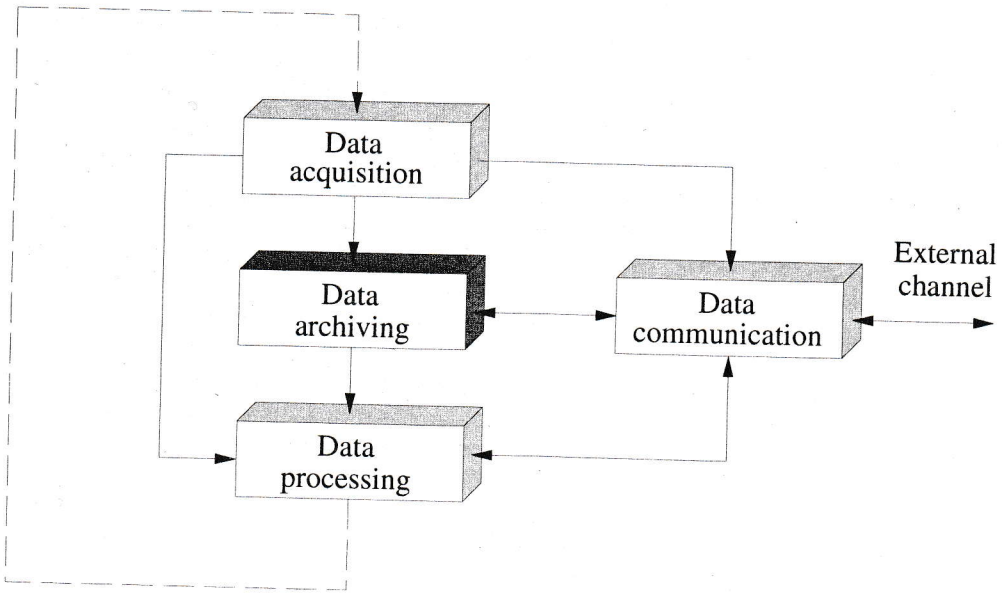

Figure 1: *A fragment of technological chain of obtaining astrophysical results.*

Experience in the establishment of different archive systems at SAO has shown that the most promising line in this area is a development of integrated programme complexes intended to be employed with a broad gamut of astronomical devices both in the optical and radio ranges (Kononov, 1996a; Kononov, Panchuk, 1999). It is precisely with the aid of such kind of systems that a comprehensive and effective solution to the problem of retaining unique experimental information can be provided. At the same time, the very process of developing systems like these depends on a number of circumstances. Among them the most important are:

- heterogeneity of observational data;
- dynamics of information streams;
- storage medium;
- data representation form;
- completeness of information;
- access to data archives;
- homogeneity of archive bases;
- standardization in different areas;
- authorization of access;
- organizational support of archives.

This paper analyses all the factors listed for taking them into account to optimize design of flexible archiving systems and establish eventually an integrated Bank of observational data.

2. Heterogeneity of data

There are two major sources of *data heterogeneity*. First, several instruments are operated at SAO: the radio telescope RATAN–600, the optical telescopes BTA, Zeiss–1000 and Zeiss–600 that provide for investigation different wavelength ranges. Second, when carrying out observing programmes on these instruments, a variety of astronomical devices are used, which, together with their acquisition systems, determine specific techniques of observing. For instance, the high-resolution echelle spectrometer LYNX at the Nasmyth–2 focus of BTA, the BTA prime focus photometer, the broad-band radiometers of the feeds No. 1, No. 5 and No. 6 of RATAN–600, the panoramic spectrum analyser of the feed No. 3 of RATAN–600.

The heterogeneity of data obtained with these facilities shows up, first of all in the diversity of logical data structures and their physical interpretation, identification, parameterization, and procedures of their processing. In all this, individual programmes and even specific observers may be additional factors introducing heterogeneity.

The very fact that experimental data are heterogeneous is a natural consequence of multiaspect astrophysical research at the observatory. This is why, the integrated system of archiving must take account of the particularities of observational data on the basis of some classification, for instance, by the *types* of data (Kononov, 1996b).

Each *type* is to be associated with a particular device, irrespective of whether the device is currently operative or its *superset* has ended (Kononov, 1995a). The main thing is that the system must locate the information of a certain type in a particular section of the observational data Bank, and each section acquires the properties of a homogeneous database and is interpreted as *local archive*. Any *type* is characterized by a set of parameters, including the identifiers of the type of data itself and astronomical device.

On this as the basis, the archive system can perform automatic identification (recognition) of the arrival information and communicate it to specialized procedures for processing.

The first classification of this type was introduced for the system ODA Version 1.0 in 1989. It supported 5 types of data: 1 for radio astronomy, 3 optical, and 1 arbitrary (Kononov, Evangeli, 1991). In 1996 the number of data types for the next version (1.1) of the system was formally brought to 15: 2 for radio astronomy, and 13 optical (Kononov, 1996b). By the present time 34 types of data have been introduced: 6 for radio astronomy and 28 for optical (Table 1). Part of them refer to the devices with the *superset* already terminated, some of them to future designs (in the column *Superset* of the table these are marked by the total time intervals in years and by the commentary *Project*). It should be stressed that, in contrast to the former, the present classification is more flexible and complete to the utmost, because:

1. It embraces all 4 telescopes.
2. It takes account of all astronomical facilities of the observatory and their associated observing techniques.
3. It does not depend on the status of the devices.
4. The two-symbol type identifiers, on the one hand, show the data to be obtained with the particular telescope (first symbol: R — RATAN–600, B — BTA, Z — Zeiss–1000, E — Zeiss–600), on the other hand, they enhance identification of the devices of a separate telescope (second symbol).
5. Appearance of new telescopes or astronomical devices at the observatory may be associated with formation of new groups of data types or introduction of new types with the already defined prefixes.

Thus, when the heterogeneity of experimental data is spoken about in aspect of establishing an integrated archive, account of their different particularities is implied first of all, which is based on the isolated and fixed global formal features.

3. Information flows

Another important factor that has an effect on the construction of the archiving system is the *dynamics of information streams* from different astronomical devices. These may have one of 3 conditional states:

- presently available and operated (*standard*);
- being designed and placed in service (*new*);
- used formerly (*old*).

New astronomical devices and their associated systems of support are continuously designed and put into operation at the observatory. This results in essential growth of the amount of experimental information. Examples are found in the comparatively new optical CCD systems PFES (Panchuk et al., 1998) and NES (Panchuk et al., 1999; Klochkova et al., 1999) for BTA, and also MOUSER for BTA (Afanasiev et al., 1996) and MARS for RATAN–600 (Berlin et al., 1999), which are being designed.

The updating of the *standard* equipment and perfection of the observing techniques are accompanied by growth of the streams of observational information of the amount of which is increasing 5–10 times and above. In the optical region this can be illustrated by the change over from recording of one-dimensional arrays to two-dimensional based on CCD systems (Borisenko et al., 1990) — or by substitution of CCD chips of 1000×1000 pixels and later on of 2000×2000 pixels, for CCDs with 500×500 pixels, in the radio range, by the introduction of the modes of acquisition in the continuum with an increased time resolution, commissioning additional radiometers, as well as increasing the number of channels through division of the main frequency band into a number of narrower ones (Bulaenko et al., 1995; Berlin et al., 1997; Stolyarov, Tsybulev, 1997; Kononov et al., 1999).

At the same time, individual units and systems are "dying", which is basically due to obsolescence. They acquire the status of *old* devices and their associated sections of the observational data Bank cease to extend, all the data obtained and archived earlier being preserved. Among such systems one may rank IPCS 2*1024 and NPh–1 for BTA (Drabek et al., 1986; Vikul'ev et al., 1991).

Apart from the change of the status of the devices, which normally manifests itself on time scales of a few years, there are another two important factors that influence the dynamics of the information streams. First, as a result of scheduling observational time by the Programme committees, priorities are defined of using particular devices on a half-year basis. Since the different techniques of observation give rise to output data flows differing from one another, then on time scales of a few months the net amounts of experimental information may be essentially dissimilar (Kononov, 1996a,b). In all this, weather conditions may exert significant influence, especially in optical observations (Kononov et al., 1996).

Second, the amount of experimental information obtained depends on the modes a specific device is used, which may be connected with the observing programme content and current meteorological conditions. For subsequent archiving it is of importance to have estimates of maximum streams on scales of 24 hours in the radio range and of a night in the optical.

Thus the dynamics of information streams may be influenced mostly by:

- change of the status of astronomical equipment;

Table 1: *Types of observational data*

Data type identifier	*Instrument identifier*	*Instrument*	*Superset, years*
		RATAN–600	
Rr	Continuum	Continuum radiometers of the feeds No. 1, No. 5 No. 6	1974 →
Ry	Radiolines	Radiospectrometric complex of the feed No. 2	1976 →
Rs	PAS	Panoramic spectrum analyser of the feed No. 3	1991 →
Rh	Radioheliograph	Radioheliograph of the feed No. 6	*Project*
Rw	MARS	Matrix radiometric system of the feed No. 5	*Project*
Rx	MIPAR	Multielement integral focal array	*Project*
		BTA	
Bq	MANIA	Complex MANIA	1976 →
Bm	H–Mag	Hydrogen magnetometer-spectropolarimeter at the prime focus	1981 →
Ba	IPCS 2 *1024	1024-channel TV photon counter system at the Nasmyth–1 focus	1983–95
Bk	Speckle	Digital specle-interferometer at the prime focus	1983 →
Bu	UAGS	Fast spectrograph at the prime focus	1985 →
Bi	IFP	Scanning Fabry-Perot interferometer at the prime focus	1987 →
Bp	NPh–1	Two-channel photometer at the Nasmyth–1 focus	1989–93
Be	ZEBRA	Fast echelle-spectrometer at the Nasmyth–2 focus	1989–93
Bc	PFCCD	CCD photometer at the prime focus	1989 →
Bf	MOFS	Multiobject fiber spectrograph at the prime focus	1989 →
Bg	MPFS	Multipupil spectrograph at the prime focus	1989 →
Bl	LYNX	High resolution echelle spectrometer at the Nasmyth–2 focus	1991 →
Bj	MINIPOL	MINIPOL polarimeter at the prime focus	1991 →
Bo	MSS	Main stellar spectrograph at the Nasmyth–2 focus	1993 →
Bb	PMCCD	Moderate resolution spectrograph SP-124 with CCD at the Nasmyth–1 focus	1995 →
Bs	PFES	Echelle spectrometer at the prime focus	1996 →
Bn	NES	Echelle spectrograph with a large diameter collimated beam at the Nasmyth–2 focus	1998 →
Bt	FSP	Fiber spectropolarimeter at the Nasmyth–2 focus	1998 →
Br	MOUSER	Multiobject universal spectrograph at the prime focus	*Project*
Bd	Crab	Moderate resolution echelle spectrograph at the Nasmyth–1 focus	*Project*
		Zeiss–1000	
Ze	CEGS	Echelle spectrometer at the coude focus	1992 →
Zq	MANIA	MANIA complex at the Cassegrain focus	1992 →
Zc	CFCCD	CCD photometer at the Cassegrain focus	1994 →
Zj	MINIPOL	MINIPOL polarimeter at the Cassegrain focus	1996 →
Zu	UAGS	Long-slit spectrograph at the Cassegrain focus	1997 →
Zl	LRS	Low resolution spectrometer at the Cassegrain focus	1998 →
Za	AFST	Automatic photometer for small telescopes at the Cassegrain focus	1997 →
		Zeiss–600	
Ea	AFST	Automatic photometer for small telescopes	1997 →

- distribution of observing time;
- content of observational programmes.

Each of the factors enumerated above makes a contribution, and the final amount of the information obtained in a certain time interval is the result of their net effect. A knowledge of the total body of experimental data acquired in different periods, as well as the streams of information from the telescopes and from the devices separately is extremely useful for the archiving as a technological process. So, based on the variation of current values, one can reveal certain tendencies and make a grounded forcast. Maximum daily information streams can, for instance, serve as a criterion for safe space allocation on the direct-access units, first of all for the data archiving procedure, or, at least, for their preliminary buffering. The maximum values of data amount per year determine the number of required volumes of archive media of specific type.

Thus, the key points in designing the integrated archiving system are:

1. Provision for taking account of changing status of astronomical devices on the basis of centralized formal descriptions.

2. Use of assessments of experimental data streams for allocation of buffers and archive memory for the archiving systems as a whole and its individual subsystems.

3. Implementation of acquisition procedures of the general archive statistics for automatic reservation of required computational resources with allowance made for the dynamics of circulating information streams.

Experience of many years shows that the over-all amount of experimental information obtained tends to increase at SAO. For instance, in 1996 the body of experimental data for 15 types of devices was 18 Gb/yr: 2 Gb in the radio range and 16 Gb in the optical range (Kononov, 1996b). The forcast made in 1996 for the nearest 2–3 years, that is for the present time, was 57 Gb/yr (the rise is due to optical data, up to 55 Gb). It should be noted that the prediction was based on the highest estimate of the amount of data and the progress in instrument making, however the change of the status of a number of devices was not taken into account: *standard* → *old*. Considering the above-said, the amount of data is actually about 60 Gb/yr: 5.5 Gb in the radio range, 54 Gb of optical data. That is, the forcast of 1996 can in general be regarded satisfactory, and the figures confirm an over 3-fold growth of information flows during the past 3 years.

Table 2 gives current estimates of the amount of information for all *standard* pieces of apparatus being presently operated at the SAO telescopes. The *old* BTA devices: IPCS 2 *1024, NPh–1, ZEBRA, and also H–Mag, which is being improved, have been excluded from consideration. Note that maximum "reasonable" values are indicated in Table 2 as daily amounts of information, which may be lower than the corresponding peak ones. The annual amounts of information are computed for the specific telescope time distribution over the devices (column "*Days (nights) per year*") and favourable weather conditions. Accurate account of the number of calendar dates is impeded, since, for instance, up to three devices may be indicated in the BTA schedule for a night. Besides, when estimating, the distribution of the BTA engineering time for a particular device was disregarded.

Thus the daily amount of data vary over a very wide range, from 1 Mb to 2 Gb (and even to 40 Gb for Speckle), depending on the devices, which emphasizes the specific character of different observational techniques that must be taken into account by the integrated archiving system. The total annual information amounts at present to about 380 Gb: 6 Gb in the radio range for 3 types of devices and 370 Gb in the optical range for 22 types (maximum values).

4. Storage medium

Option for an observational data *storage medium* is extremely vital, especially in connection with continuous growth of data amount. At the present stage of data archiving technology development this problem bears directly on the hardware and, first of all, on the types of media and units being used.

It is natural that a constantly expanding archive should be located on the external memory volumes. One should take into account both the specific features of the original information itself and the characteristics of the media.

The basic criteria in selection of archive media are as follows:

- existing observational data volumes;
- predicted growth of information streams;
- current computing facilities;
- capacity of volumes of media;
- memory – cost ratio;
- distinctions of access to the archive information;
- time of access and data exchange speed;
- reliable performance of drives;
- safe duration of data storage;
- organizational points of formation of the archive base.

Although the criteria just enumerated are overt enough, attention should be paid to the fourth of them. Requests for access to the data for subsequent processing may in the general case provide for scanning of sufficiently wide time ranges of archive data

Table 2: *Estimates of volume of observational data*

Data type identifier	*Instrument identifier*	*Observation capacity*	*Day (night) capacity*	*Days (nights) per year*	*Year capacity*
Rr	Continuum	0.9 Mb	14 Mb	360	5.0 Gb
Ry	Radiolines	0.8 Mb	2 Mb	90	180 Mb
Rs	PAS	0.4 Mb	1.2 Mb	360	440 Mb
RATAN–600					6 Gb
Bq	MANIA	100 Mb	2 Gb	5	10 Gb
Bk	Speckle	1.5 Gb	40 Gb	7	280 Gb
Bu	UAGS	0.7 Mb	60 Mb	50	3.0 Gb
Bi	IFP	15 Mb	180 Mb	20	3.6 Gb
Bc	PFCCD	2.4 Mb	250 Mb	60	15 Gb
Bf	MOFS	2.1 Mb	210 Mb	20	4.2 Gb
Bg	MPFS	2.1 Mb	210 Mb	50	10.5 Gb
Bl	LYNX	2.4 Mb	70 Mb	15	1.1 Gb
Bj	MINIPOL	0.5 Mb	5 Mb	10	50 Mb
Bo	MSS	2.4 Mb	100 Mb	30	3.0 Gb
Bb	PMCCD	350 kb	70 Mb	50	3.5 Gb
Bs	PFES	2.4 Mb	120 Mb	10	1.2 Gb
Bn	NES	2.4 Mb	70 Mb	25	1.8 Gb
Bt	FSP	0.8 Mb	16 Mb	6	100 Mb
BTA					337 Gb
Ze	CEGS	2.8 Mb	140 Mb	70	9.8 Gb
Zq	MANIA	100 Mb	2 Gb	5	10 Gb
Zc	CFCCD	0.6 Mb	50 Mb	130	6.5 Gb
Zj	MINIPOL	0.5 Mb	5 Mb	20	100 Mb
Zu	UAGS	0.2 Mb	10 Mb	75	0.8 Gb
Zl	LRS	2.4 Mb	120 Mb	15	1.8 Gb
Za	AFST	0.6 Mb	30 Mb	15	450 Mb
Zeiss–1000					30 Gb
Ea	AFST	0.6 Mb	30 Mb	100	3.0 Gb
Zeiss–600					3 Gb
Radio					6 Gb
Optical					370 Gb
Total					376 Gb

and demand frequent remounting of individual volumes. In so doing the use of volumes of small capacity can result in substantial loss of time and even complete neutralization of the advantages of the media in such parameters as the time of access and the speed of exchange. Of course, everything depends on the *local archive* itself and on how intensively it is used. This is why the option of the medium must take account of particularities of observational information, for instance, its daily amount, quantum of access, frequency of references, etc.

At the present time, the following media are utilized at SAO for long-duration storage of experimental data on a centralized basis and also inside individual subdivisions:

- DAT cassettes (Digital Audio Tape) (4 mm) 1.3–2.4 Gb in capacity;
- EXATAPE cassettes (8 mm) of 5–7 Gb;
- optical disks CD–R (Compact Disks, Recordable) of 650 Mb;
- magneto-optical disks MO–RW (Re-Writable Magneto Optical Disks) 250 Mb to 2.6 Gb in capacity.

In the mid-1990s the most popular for this purpose were DAT and EXATAPE cassettes, whereas now extensive use is made of optical disks CD–R (Denisenko, Balabanov, 1998) which are cheap enough, have a short access time, and high speed of exchange; however they are second to tapes in terms of volume capacity. In the context of the above-said the capacity of volumes is significant for a number of *local archives*, for example for such devices as MANIA, PFCCD, MPFS, CEGS, CFCCD, Continuum, to say nothing of Speckle. For this reason, when generalizing all the types of observational data used at the observatory, we consider the optical disk DVD (Digital Versatile Disk) of 4.7–17 Gb to be most promising for storage of observational data. The manufacturing technology of this disk is being intensively developed (Egarmin, 1998), and it will gradually displace the CD.

The compelled many-volume character of the archives can partly be compensated by the use of stacks of the type of CD-changer or DVD-changer which provide for dozens of disk positions. These variants, however, will not solve the problem. Most likely they will help provide access to the *operative archives* that are subsets of the *local archives* and contain new (recent) data or the information connected with particular observational programmes. A similar result can also be obtained by using hard disks HD of 10–36 Gb or else RAID (Redundant Array of Inexpensive Drives) devices dozens of Gb in capacity.

It should be noted that the problem of storage medium cannot be settled by the choice of the carrier type. A significant part is played here by the environment in which the observational data archive as an information system is supported, and also by the hardware to implement the software. The system Linux on the Pentium-type computers (Petersen, 1997) is most promising and convenient in this respect.

5. Form of representation

By the *form of representation* is meant the formats in which observational data are stored on the archive media.

The introduction of a format pursues generally two objectives:

- taking account of specific peculiarities of data;
- provision of convenience in subsequent processing.

Additional factors that determine use of particular formats may be, for instance, the requirement for physical saving of the carrier, application of the standard (ready) software, desire to provides for visualization of logical structures, etc. At the same time, as experience has shown, the form of archive data representation is mostly affected by the output data format of the acquisition system. Most commonly these formats are interpreted as archive and find further extension in the technological chain as far as the processing system.

The existence of a large number of diversified acquisition systems at SAO gives rise to formats largely distinguishing from one another. What is more, some acquisition systems use several forms of their output data representation depending on the modes of operation. For this reason, *dozens* of different formats are presently numbered at the observatory, each of them suggesting its interpretation and appropriate software support.

Thus the adoption of formats from the acquisition system results in extension of heterogeneous forms of representation to the observational data archives, which is not optimum from the point of view of joint processing of different type data and also in terms of exchange of information with other research centres. The more so, the acquisition systems themselves function under specific conditions of observations, therefore their information outputs mirror in many cases only the simplest alternatives of data "saving". In connection with this the way of *unification* of observational data formats with simultaneous preserving of specific features of their different types appears to be much more effective. First of all, this refers to the archiving systems.

It is expedient to have also a wider view on the matter of unification of formats: for any links of the technological chain of obtaining astrophysical results, beginning with the acquisition system. Actually the question is one of establishment of a unified interface between the systems of different types: acquisition, archiving and processing systems. The ideology of this FLEX interface (FLexible EXchange) was discussed and grounded by Kononov (1995b). The progressive installation of the FLEX interface, which orders communication between the systems, would simplify and speed up the establishment of the observatory's integrated archive. Note that the concept of FLEX interface even now forms the basis of creation of the RATAN–600 observaional data Bank (Kononov, Mingaliev, 1998).

6. Information completeness

Experimental data reperesent a collection of numerical arrays. When locating this kind of data into the archive and organizing access to them, there arise natural questions:

- to which device does a particular array refer?
- which instrument was used and when?
- what astronomical object was observed?
- who was the observer?

- what is the structure of the data? etc.

All these questions can be answered provided that the numerical arrays are accompanied by appropriate parameters.

The concept of *information completeness* of archive data consists in their utmost possible parameterization in all respects. The accompanying parameters are supposed to be an indispensable part of the archive itself.

The optimum moment of *basic* parameterization of observational data is the stage of operation of the acquisition system, that is, when the data are being formed. Subsequently the archiving systems may introduce their *additional* characteristics, which are of importance from the point of view of organization of the archive as an information system.

The most effective form of implementation of the *basic* parameterization is a self-documented format which represents the structure

$$< Header >< Data\ array > .$$

In this case even simple sampling of an archive object preserves automatically the accomanying description. The FITS format (Flexible Image Transport System) for data exchange between astronomical centres (Wells et al., 1981), which is supported by the International Astronomical Union (IAU, 1983; 1989) may be treated as this self-documented format. Note that the FITS format is commonly applied now at many observatories to represent archive data themselves. Unfortunately, this convenient and flexible format came to be used at SAO as late as the middle of the 1990s.

All possible reference databases, dictionary directories (Leong-Khong, Plagman, 1986) established as superstructures over the basic archive and containing different meta-characteristics, may be the forms of implementation of *additional* parameterization of archive data.

7. Access to archive data

The problem of *access* to archive data has always been associated with two questions: what is being sought and what is the way of doing it? The former point bears a direct relation to the formulation of astronomers' inquiries for retrieval, whereas the latter is generally an internal task of the archiving system itself.

Requests for access to the archive information are based on specifying a number of parameters that reflect the features of observational data of interest and identifying them in an appropriate manner. Such parameters are subsets of all parameters archived and considered as *primary keys*. The possibility of indicating specific parameters as primary keys determines the degree of flexibility of the archive as regards the access procedures. On the other hand, this possibility depends also on the solution of the experimental data parameterization problem.

With a sufficient extent of parameterization of the archive data, the choice of particular parameters as key ones for the organization of retrieval must be determined by the specific character of observational data and, in the final analysis, by the enterprise of the astrophysical experiment the astronomer deals with. In other words, the tasks that the astronomer conceives to perform when planning the experiment determine his possibilities of further use of the data. This can be illustrated by certain universally adopted parameters regarded as key ones: *date of observation, object name, instrument used, astronomical device, observational programme, observer, information channel (filter, radiometer)* etc.

The matter of implementing the procedures of access to archive data depends on both the archiving system architecture and the computer hardware (although the former is connected to a certain extent with the latter). In this respect an important role may be played by the superstructures of the type of dictionary directories, which are intended, in particular, for storage of the knowledge of physical location of specific objects in the archive and capable to provide for fast retrieval of the necessary information by way of indexing of individual groups of data.

The effectiveness of access is influenced also by the capabilities of basic algorithms used by the archiving systems in the internal analysis of varions kinds of logical structures. At last, the performance characteristics of the external devices (tape-recorders, disk drives), for instance, the time of access and the speed of exchange have a direct bearing on the functioning of the archive as an information system as a whole.

We do not deal here with other aspects of access to the archive data such as authorization, capabilities of users' interfaces, application of standard systems of database management, net facilities etc. Of importance is what all this can be based on. For this reason, in designing the integrated archiving system, a developed and flexible global *infological* enterprise model reflecting all aspects of possible astrophysical experiments is of fundamental importance (Kononov, 1991; 1995a).

8. Homogeneity of archive bases

Experience of design of various archiving system at SAO has shown that special attention should be paid to the *homogeneity* of archive bases. Even to a greater extent this refers to the integrated information system.

The very concept of *homogeneity* may be concerned with a number of areas among which the most

important is the construction of an observational data models. It is the characteristics of such models that determine the organization of any archive base as a certain set of objects to which further access is provided by means of particular software. The homogeneity of the archive base in this sense must manifest itself in:

- representation of observational results in the form of sets of files;
- file formats;
- file identification;
- extent of data parameterization;
- identification of parameters;
- determination of values of parameters;
- data structures;
- organization of archive volumes.

Disregard of these principles will ultimately result in inadequate consumption of labour and resources, unjustified complication of software, increase of the time of its development and/or adjustment, bulky interfaces, inconsistency of requests, restriction of functional capabilities, and difficulties in organization of support.

All the enumerated refers not only to the *main* archive database of a certain *type*, which contain numerical experimental arrays proper, but also to its superstructure in the form of *auxiliary* reference databases oriented to different fields of application. That is, the question is the homogeneity of a separate *local archive* as a multibase system corresponding to a particular section of the observational data Bank.

On the other hand, for the integrated archiving system that will support a few local archives of different *types* it is advantageous for the same reasons to adhere to the homogeneity of the archive bases also at the level of the system as a whole. This characterizes the highest degree of informational integration which, depending on the functional orientation of the archiving system, may refer to both the level of one telescope and the level of the whole observatory (Kononov, Lipovetsky, 1994; Kononov, 1996a; Kononov, Mingaliev, 1998).

One of the optimum ways to achieve a maximum homogeneity is the adoption of a system of conventions or standards which specify the description of data streams, individual objects, parameters and their values, characteristics and interrelations of the system etc., which is ultimately related to the global *infological* model. An additional and very efficient means here is the introduction of homogeneity into identification of different structures, for instance, names of archive volumes, files, fields and parameters. This allows a certain element of versatility to be introduced at the stage of formation of requests with the aid a system of *patterns* and their lists (Kononov, 1995c).

9. Standardization in different areas

An effective solution of the problem of homogeneity of the archive bases can be illustrated by the concept of unified FLEX interface providing for standardization in 3 areas:

- data interpretation;
- data identification;
- data representation.

By the data *interpretation* (not astrophysical, of course) is implied the examination of the information output of an individual data acquisition system based on a hierarchical model of descriptors of a particular *generalized* acquisition system corresponding to a generalized astrophysical experiment. The output data sets of a specific system are particular cases of the generalized system output, and any output file is interpreted as a *descriptor file* referring to a certain level of the model. The streams of descriptor files of different levels correspond to concurrent processes of data acquisition, including the processes of logging of astrophysical experiments. The description of the model which formalizes the structure of the information output of any acquisition system and at the same time the input of the integrated archiving system represents a *common* FLEX–I standard of data interpretation (Kononov, 1995a).

Universal *identification* of data is ensured by a common system of conventions for naming output files of any acquisition systems. Compound names with a hierachical structure, which include object identifiers of the enterprise of the generalized astrophysical experiment, for instance, the *type of data, date of observation, group* are used. The description of these common rules of file naming represents a *common* FLEX–N standard of data identification (Kononov, 1995d).

The unified form of data *representation* is provided for by using a common and flexible format of exchange that takes account of specific features of observational data of any type. The FLEX format, which is a FITS-like format and is based on the application of the standard extension FITS Binary Table (Implementation of the FITS, 1991; Cotton et al., 1995), is taken as such a format. The description of such a FLEX format is considered as a *common* FLEX–F standard of data representation (Kononov, 1995e).

Thus the *common* standards FLEX–I, FLEX–N and FLEX–F define common rules which must be applied by the integrated archiving system for coordinated interaction with any acquisition system. To take more detailed account of the distinctions of specific types of data and systems, the *common* FLEX standards are extended by *local* standards in the same 3 regions. For example, FLEX–N–Rr, FLEX–N–Bl and FLEX–F–Rw denote the local standards of iden-

tification of observational data of the types Rr and Bl (Continuum and LYNX) and representation of data of the type Rw (MARS). Note that the first versions of the local standards FLEX–I–*kk* and FLEX–N–*kk*, where *kk* is the type of data, were introduced formally for 15 types of devices as early as 1996 (Kononov, 1996c); however, at present they need certain correction.

Despite the necessity for forming the most homogeneous archive bases under real conditions such processes can be implemented only step by step. The cause of this is that many established and continuously expanding local archives embrace tremendous time intervals and include earlier data which have not been properly reorganized. Should the notions of homogeneity and standardization (as perspective direction of development) be combined the degree of homogeneity of the current *main* archive base of the type Rr (Continuum, RATAN–600), for instance, could then be illustrated by the following figures:

Representation of observations by one multifrequency file	—	91 %
Data format	—	100 %
Identification of files	—	33 %
Parameterization level	—	85 %
Structure of data	—	33 %

Here it should be taken into account that the archive base of the type Rr is now the largest, ordered and being developed database of SAO (18-year' period, 94000 multifrequency observations) (Kononov et al., 1999).

10. Old data recovery

In the general case the change of the content of a certain local archive may be connected with two points:

1. Regular filling of the archive with current new data.

2. Gradual involvement of "old" data that have not been archived.

That is, the archive base can develop in time "in two directions". The former point is quite clear, whereas the latter needs a certain elucidation.

Indeed, many *standard* astronomical devices have been in use at SAO for 7–10 years and above. An anormous amount of material has been accumulated during this time, but only part of it is stored in sufficiently ordered and supported archive bases and to a certain extent accessable to astronomers. Part of the results are at best spead over different buffer media in different (sometimes unknown) formats and can be potentially lost. Such information refers to the category of *"old"* data; the question of its recovery has lately been raised with increasing frequency (Boyarchuk, 1994).

Experience of many attempts to recover "old" data and bringing them to a certain systematized form has shown that this process is extremely labour consuming and is getting more and more complicated. Let us enumerate here only major problems:

- Numerical data on old media are either illreadable or the readers are inoperable or already lacking at the observatory.
- Decoding of old formats is impeded since many of them were not properly described in due time or their descriptions are merely nonexistent.
- Unsatisfactory degree of parameterization of "old" data, which inhibits their identificaion even with the observation logs involved.
- Analog form of a part of "old" data (diagram tapes in the radio range and photographic plates in the optical range) requires additional procedures to convert to the digital form.
- Organizational difficulties due to the preliminary collection, systematization and analysis of data inside the observatory, recovery of data passed earlier to other observatories and research centres, and also to the necessity for enlisting the co-operation of experts to perform a great amount of work.

At the same time, taking into account the importance of recovering "old" data, all these procedures can be accomplished by individual steps with gradual inclusion of the recovered information into the appropriate local archives. It is advantageous that the data recovery should be made coincident with the reduction of the observational information to the current standards to achieve a maximum homogeneity of the archive bases.

The results of this kind of work done for the archive of the type Rr are given as an example. The *main* archive observational database of the continuum radiometers began to be established in a virtually standard form in May 1995. Four years later (May 1999) it included 53000 observations (Kononov, 1999). From November 1989 to May 1995 the process of archiving was supported by the first version of the ODA (Observational Data Archive) system, which resulted in preserving 18500 observations (Kononov, 1995f; 1996a). In 1998 this part of data was reorganized, reduced to the standards and switched successfully to the current archive (Kononov et al., 1998). At the second stage the data for the period 1982 to 1989 (nearly 8 years), which had been buffered on magnetic tapes in different formats, were collected and analysed. This part of data containing about 13000 observations was sequentially decoded, reorganized, reduced to standards in 1998-1999 and ultimately switched to the current archive base (Kononov, Pavlov, 1999). As a result of restoration thus performed, the radio astronomers were provided homogeneous access to the archive information

covering the 18-year' period and including, as of 2000 January, about 94000 observations. It should be noted that at the present time the archive base of the type Rr supports 85–90 % of experimental data which are obtained with RATAN–600.

The same as for the *standard* astronomical devices, the problem of recovery of the data obtained earlier will also refer to the *old* nonexistent devices. Here too the work can possibly be broken up into individual stages, if the information has not been preliminary systematized and is not stored in a reliable enough form as, for example, for the devices IPCS 2∗1024, NPh–1 and ZEBRA (Kononov, 1994). To avoid a similar problem in the future for *new* devices, which are currently placed in service or just being in the state of designing, coordinated decisions should immediately be made concerning standardization of output data of the appropriate acquisition systems for their orientation to the integrated archiving system.

11. Authorization of access

The question of *authorization* of access to the archive information is one of the most important and complicated aspects of functioning of the archive as an information system. The urgency of its solution is dictated in part by the abrupt enhancement of exchange of observational results (and their processing) between different astronomical institutions aided by the Internet over the past few years (Krol, 1994). The results of unique investigations can be published fast enough in the traditional printed form and represented electronically. On the other hand, the present-day technology of observing and further processing of data is expected to involve large teams of experts of different professions whose mutual relations can be controlled somehow. In all this it is the copyright that must form the basis of authorization of access to the observational data archives of any types since the latter concentrate the results of work done by different groups of people.

From a more detailed treatment of the problem it follows that the authorization of access must be based on the following four factors:

1. General legal documents regulating copyright in science and computer technology.
2. Internal documents of SAO which take into account distinguishing features of functioning of the observatory as a research centre.
3. International experience in organizing authorized access to experimental data.
4. Special conventions inside the integrated archiving system, which will mirror the above points as formal rules of interrelation of astronomers with archive databases.

Unfortunately the problem of authorization of access to the archive information has come to be discussed at the observatory most recently. In all likelihood this is associated with the appearance of a number of legal disputes and the concurrent desire to solve the problem. By the present time a general Proposition for the observational data archive at SAO have been formulated, in which there are references to the relevant Russian legislation, and a period of 2 years is indicated for exclusive access to archive data on astrophysical objects on the part of applicants. The calibration material and subsidiary data are open to general use. From the practical point of view the formulated general Proposition gives rise to a number of questions. For instance, the specified time priorities are inconsistent with the concept of long-term programme (such programmes may cover as much as 5 years of observing).

To realize clear-cut and flexible authorized access, a system of priorities and appropriate conventions on their application both for individual groups of users and concerning the experimental information of particular character has to be introduced. In other words, it is necessary:

- To define the *categories* of potential users of the observational data Bank of SAO.
- To introduce *classification* of experimental data from the point of view of application of the copyright.
- To determine *time* parameters of access to the archive information on default.
- To take account of the *status* of the existing techniques of observations on the telescopes.
- To establish standard *priorities* of access based on the interrelations of the user categories, data classification and status of the observational techniques.
- To lay down *rules* for changing the standard priorities of access both at the desire of users having privileges for this and in the case of unsanctioned departures from the schedule approved by the Programme committees of the telescopes.

Undoubtedly, effective authorized access to the archive information is possible only with a sufficiently high degree of parameterization of data and correctness of the key parameters. It is only under these conditions that rigorous formal descriptions allowing for all the requirements enumerated above can be introduced into the integrated archiving system.

12. Organizational support of the archive

By the *organizational support* of the archive is implied management (maintenance, administration) of the archive as the Base of observational data. The maintenance should be provided by people because

1. There will always exist the procedures which cannot be automated, and human attendance is an obvious necessity.

2. Depending on the degree of automation of the archive as a technological link, some procedures can be complemented by manual operation which should be reasonably minimized step-by-step.

The role of management is of vital importance at SAO where various and time variable streams of experimental data circulate vigorously both between individual computers and systems, and between scientific sites of the observatory. For this reason, ordering of all this process is an indispensable requirement to be met in establishing and supporting the integrated archive.

Maintenance of the archive includes first of all:

- Control of incoming observational data with the aid of the buffers on the HD (matching of acquisition and archiving systems).
- Rejection of data as to their content (filtering of data) coordinated with the observers.
- Control of conformity of data to the standards (correctness of data).
- Distinguishing and preparation of archive volumes of a particular type.
- Check up of correct filling of the archive (recording of information on media).
- Continuous doubling of data (making copies of archive volumes).
- Necessary reorganization of archive data.
- Control of whether the users' demands for retrieval and fetch of data are met properly.
- Collection of the general archive statistics.

The above-enumerated refers to any local archive base, and in the case of integrated archive it should be added by a set of procedures of general character to control the system as a whole.

13. Conclusion

Summarizing the most important aspects of observational data storage, let us note the key points:

1. The objective of archiving is storage of the unique observational information and provision of automated authorized access of different categories of users to it.

2. By data storage is implied location of data on computer media and knowledge of location of particular objects and their main characteristic features.

3. The archive as a database is supported by the archiving system, which is a programme controlled system functionally oriented.

4. The most efficient way of setting up an experimental fund of SAO is the introduction of an integrated archiving system which would provide flexible support of observational data of any types.

5. Such a system must realize a conveyer to communicate information from different acquisition systems to the systematized archive bases, and to meet requests for retrieval and fetch of data for subsequent processing.

6. The optimum form of interaction of various acquisition systems with the unified archiving system is a flexible unified interface based on the standards.

7. Observational data must be parameterized to the utmost, and the performance of the archive media must be adequate to the existing information streams.

All this reflects the essence of the concept of *automated archive* of observational data, and the most developed version of it is defined as the *observational data Bank*.

Acknowledgements. The work has been accomplished with partial support from the RFBR (project 99-07-90296).

The authors are pleased to acknowledge the cooperation of their colleagues:

I.V. Gosachinskij, Laboratory of Radiospectroscopy,
N.A. Nizhelskij and P.G. Tsybulev, Laboratory of continuum radiometers,
T.N. Kazanina and L.V. Opejkina, Group of solar investigations,
V.B. Khajkin, Laboratory of antenna measurements,
V.L. Plokhotnichenko, Laboratory of relativistic astrophysics,
V.A. Vasyuk, Laboratory of high resolution astronomy methods,
A.N. Burenkov, A.V. Moiseev and S.N. Dodonov, Laboratory of spectroscopy and photometry of extragalactic objects,
S.V. Ermakov and V.G. Klochkova, Laboratory of stellar spectroscopy,
V.D. Bychkov, Laboratory of stellar physics,
A.Yu. Knyazev, Laboratory of galaxy investigations,
G.A. Chountonov, Laboratory of stellar magnetism,
G.A. Galazutdinov, Coude group,
V.G. Shtol', consultant on hydrogen magnetometer,
V.R. Amirkhanian from SSI
for discussion of questions concerning operation of astronomical devices and data acquisition systems at SAO and for the estimates of obtained amounts of information.

References

Afanasiev V.L., Gazhur Eh.B., Dodonov S.N., Drabek S.V., Markelov S.V., Perepelitsin E.I., 1996, MOUSER, SAO Report, **254**

Berlin A.B., Bulaenko E.V., Gol'nev V.Ya., Lovkova I.M., Nizhelskij N.A., Timofeeva G.M., Tuzenko S.V., Fridman P.A., Tsybulev P.G., 1997, in: XXVII Radioastron. Conf., St.Petersburg, 158

Berlin A.B., Chmil V.M., Pilipenko A.M., Bogdantsov A.V., Meshkov Yu.N., Nizhelskij N.A., Timofeeva

G.M., 1999, MARS (MAtrix Radiometric System) project, Gamov Memorial International Conference (GMIC'99), St.Petersburg, Abstracts
Borisenko A.N., Vitkovskij V.V., Zhelenkova O.P., Kopylov A.I., Markelov S.V., Ryadchenko V.P., Shergin V.S., 1990, Astrofiz. Issled. (Izv. SAO), **32**, 157
Boyarchuk A.A., 1994, Frontier of space and ground-based astronomy, Eds.: W. Wamsteker, M.S. Longair, Y. Kondo, A. & Sp. Sci. Library, **187**, 440
Bulaenko E.V., Tuzenko S.V., Fridman P.A., 1995, in: XXVI Radioastron. Conf., St.Petersburg, 308
Cotton W.D., Tody D., Pence W.D., 1995, Astron. Astrophys. Suppl. Ser., **113**, 159
Denisenko A., Balabanov A., 1998, Radio, **6**, 24; **7**, 26
Drabek S.V., Kopylov I.M., Somov N.N., Somova T.A., 1986, Astrofiz. Issled. (Izv. SAO), **22**, 64
Egarmin N., 1998, Komp'yutern. ezhenedel'nik, **7**, 33
Klochkova V.G., Ermakov S.V., Panchuk V.E., Tavolzhanskaya N.S., Yushkin M.V., 1999, Preprint SAO RAS, **137**
Kononov V.K., 1991, Soobshch. Spets. Astrofiz. Obs., **67**, 48
Kononov V.K., 1994, Preprint SAO RAS, **105**
Kononov V.K., 1995a, Preprint SAO RAS, **110T**, 1
Kononov V.K., 1995b, Preprint SAO RAS, **110T**, 13
Kononov V.K., 1995c, Preprint SAO RAS, **108**, 11
Kononov V.K., 1995d, Preprint SAO RAS, **110**, 24
Kononov V.K., 1995e, Preprint SAO RAS, **111T**
Kononov V.K., 1995f, Preprint SAO RAS, **108**, 1
Kononov V.K., 1996a, Ph.D. Thesis, SAO RAS, 248
Kononov V.K., 1996b, Preprint SAO RAS, **112T**
Kononov V.K., 1996c, Preprint SAO RAS, **114T**, 15
Kononov V.K., Evangeli A.N., 1991, Soobshch. Spets. Astrofiz. Obs., **67**, 87
Kononov V.K., Klochkova V.G., Panchuk V.E., 1996, Preprint SAO RAS, **115T**
Kononov V.K., Lipovetsky V.A., 1994, Preprint SAO RAS, **105**, 16
Kononov V.K., Mingaliev M.G., 1998, Preprint SAO RAS, **129T**
Kononov V.K., Pavlov S.V., 1999, Preprint SAO RAS, **130T**
Kononov V.K., Panchuk V.E., 2000, Bull. Spec. Astrophys. Obs., **49**, 110, (this issue)
Kononov V.K., Pavlov S.V., Mingaliev M.G., Verkhodanov O.V., 1998, Preprint SAO RAS, **128T**
Kononov V.K., Pavlov S.V., Mingaliev M.G., Verkhodanov O.V., 1998, Preprint SAO RAS, **131T**
Krol E., 1994, The Whole Internet. User's Guide & Catalog, O'Reilly & Associates, Inc.
Leong-Khong B., Plagman B., 1986, M.: Finansy i statistika, 311
IAU Information Bulletin, 1983, **49**, 14
IAU Information Bulletin, 1989, **61**, 10
Panchuk V.E., Najdenov I.D., Klochkova V.G. et. al., 1998, Bull. Spec. Astrophys. Obs., **44**, 127
Panchuk V.E., Klochkova V.G., Najdenov I.D., 1999, Preprint SAO RAS, **135**
Petersen P., 1997, LINUX: manual, K.: BHV, 688
Stolyarov V.A., Tsybulev P.G., 1997, in: XXVII Radioastron. Conf., St.Petersburg, 182
Verkhodanov O.V., Kononov V.K., Chernenkov V.N., 1996, Bull. Spec. Astrophys. Obs., **39**, 146
Vikul'ev N.A., Zin'kovskij V.V., Levitan B.I., Nazarenko A.F., Neizvestny S.I., 1991, Astrofiz. Issled. (Izv. SAO), **33**, 158
Wells D.C., Greisen E.W. and Harten R.H., 1981, Astron. Astrophys. Suppl. Ser., **44**, 363
Implementation of the Flexible Image Transport System (FITS), November 6, 1991, Draft Standard, NOST 100-0.3b, NASA/OSSA Office of Standards and Technology