Open Access Online Journal on Astronomy and Astrophysics

Acta Astrophysica Taurica

www.astrophysicatauricum.org

Acta Astrophys. Tau. 4(4), 9–16 (2023)

doi:10.34898/aat.vol4.iss4.pp9-16

Information system of the general observation archive of SAO RAS. Current state

O.P. Zhelenkova, V.V. Vitkovskij, T.A. Plyaskina, V.S. Shergin, G.A. Malkova, V.N. Chernenkov, M.M. Kyabishev, A.M. Velichko

Informatics Department, SAO RAS, Nizhnij Arkhyz 369167, Russia
e-mail: zhe@sao.ru

Received 28 September 2023

ABSTRACT

There is no doubt about the need for long-term storage of observations since they do not lose their scientific significance when studying the variability of celestial objects and other tasks. Maintaining archival systems requires software development, metadata management, fault tolerance, and periodic migration of files to modern storage media, which is important to ensure long-term safety of information. Providers of small collections need to join forces and share experiences so as not to "reinvent the wheel". According to the standards accepted in the astronomical community, the SAO RAS observation archive is a small collection compared to the collections of modern astronomical data centers. Here we present the basis and methods of our archive system development. The archive includes more than two dozen digital collections with observations obtained on various observation instruments. It consists of three interconnected components that provide accumulation, long-term storage and access to data. The volume of collections is approximately 2.5 TB. The information system is implemented on the basis of the PostgreSQL software. Access to archived files is provided through a web interface that supports queries by observation instrument and date, coordinates, object name, program applicant, observer. The database contains more than 4 million records. The archive system uses two servers with directly connected storage systems. The work server implements data access, and the test server is used for development and testing. This configuration provides both permanent access to files and the ability to develop the information system.

Key words: observational data, long-term storage, observation archive, information system, database management system

1 Introduction

SAO RAS is a ground-based astronomical center that has the largest Russian astronomical instruments – the 6-m optical telescope BTA and the RATAN-600 radio telescope with the 600-m diameter antenna. The observatory's telescopes produce unique observational material. For observations, various equipment is used that works with specific computer hardware systems or, in other words, data acquisition system. These systems produce data that have different structure and descriptive parameters.

Since the beginning of the 80s of the last century, the accumulation of data obtained on RATAN-600 and BTA began to be carried out in the digital form. This was the impetus for the development of the concept of a digital archive of the observatory (Vitkovskij et al., 1987; Kononov et al., 1990), as well as work on standardizing observation file formats. The first digital archive on magnetic tapes was organized for observations of the Cold survey (Berlin et al., 1984), carried out on RATAN-600 for a number of years, starting in 1980, in order to study fluctuations of cosmological microwave background.

The self-documenting FITS (Wells et al., 1981) format is the de facto astronomical standard for data presentation. At the observatory, this standard was first applied to files (Vitkovskij et al., 1988) obtained using a CCD camera (Borisenko et al., 1990). The FLEX format similar to FITS for continuum radiometer data was then developed by Verkhodanov et al. (1993, 1995).

Since 1987, archival observations have been stored on a variety of media, ranging from streamer tapes, data cassettes to magneto-optical disks. The storage experience on these media was not very successful. When data were transferred from large reel magnetic tapes to strimmer tapes, then to dat-cassettes and magneto-optical disks, a small part of the observations was lost. By 1994, optical disks began to be used to store information.

The observatory archive consists of local archives, where the local archive is a digital collection obtained through the data acquisition system used for a particular observing in-

strument. The data in the general archive are presented in the following formats: FITS and FLEX formats, internal file format of the MIDAS system (Warmels, 1992), as well as binary files with text description, archived in tar-files.

By 1999, the archive contained approximately 60 GB of compressed data, and there was a need to develop an information system for accessing archived files. Based on the local archive of continuous spectrum radiometers in the feed cabin No. 1, a prototype of an information system with web access based on the Oracle DBMS was developed (Vitkovskij et al., 2000). The database server was hosted at the South Russian Regional Informatization Center of the then Rostov State University. For each file in the database, a block of service information was formed from observing parameters from file headers, which also included information to identify the file.

Upon further work with digital collections, it turned out that the service information block contains an excessive number of parameters (namely, all the parameters from the FLEX format header). In other collections, apart from main parameters, many parameters were not available. This was especially true for parameters of radiation receivers. For this reason, it was necessary to limit ourselves to those parameters that would be present in all digital collections in an information system. One such parameter, first of all, is the observation date. Taking this into account, as well as the experience gained during the development of the prototype database, in 2003 an information system was developed and implemented based on a trial version of the Oracle software (Zhelenkova et al., 2003), which was by that time hosted on the observatory server. It provided access by observation date to 14 digital collections.

Note that in 2003, the IAU General Assembly adopted a resolution[1] on observations obtained by observatories that are financed from the state budget. It noted that after the expiration of the copyright of applicants of the observational programs, the observatory is obliged to place the data in open access.

Later, a new version of the search system was developed based on the free PostgreSQL software, which provided data search using a standard set of queries. Up to 30 attributes were used to describe each file in the database, including its identification and location in the storage area.

Since 2013, using the software developed by Shergin (2014), we have been performing an automatic coordinate calibration of direct images and/or correction of file headers. Files processed in this way are written to special local archives. To date, more than 300 thousand files from 4 local archives have been processed in this way.

If new digital collections appear when new observational instruments are put into operation, they are added to the archival system.

In 2015–2018, radio data were added to the archive system. They needed to be transferred to our system from another archive system that is no longer supported. This doubled the number of files in the information system. At first, access to files was implemented only by observation date. Since 2023, the search for radio observations has been carried out using a standard set of parameters.

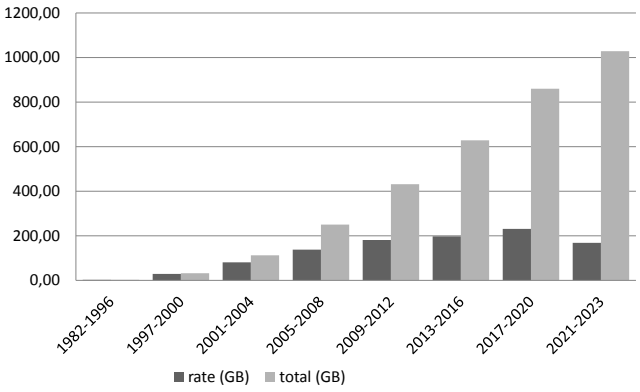
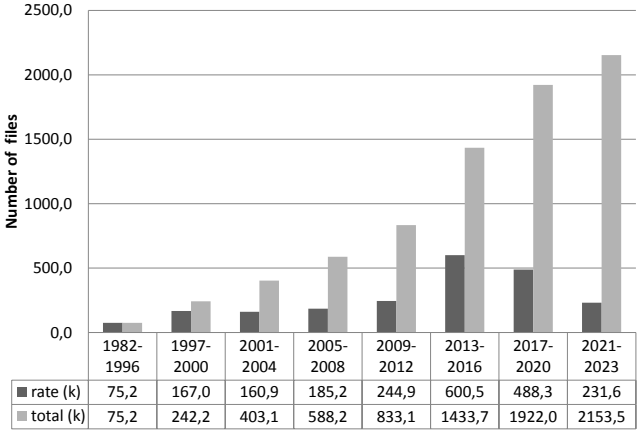


Fig. 1. Total volume and rate of data accumulation in the general archive of SAO RAS (in GB).



	1982-1996	1997-2000	2001-2004	2005-2008	2009-2012	2013-2016	2017-2020	2021-2023
rate (k)	75,2	167,0	160,9	185,2	244,9	600,5	488,3	231,6
total (k)	75,2	242,2	403,1	588,2	833,1	1433,7	1922,0	2153,5

Fig. 2. Number of archived files in the general archive of SAO RAS (in thousands).

Currently, the general observation archive includes 24 local archives, amounting to approximately 2.5 million files, more then 4 million records in the database and more than 2.5 TB of files. The figures show the total volume of data (Fig. 1) and the number of files (Fig. 2) in the general archive of the observatory from 1982 to the first half of 2023.

Note that out of 24 local archives, 7 collections are being replenished with new observations. The remaining collections are not replenished since the observational instruments with which they are associated have been removed from regular observations.

2 Archive system

The general archive stores data obtained with observation methods used at the observatory's telescopes. These data

[1] https://www.atnf.csiro.au/people/Ray.Norris/WGAD/Resolution.htm

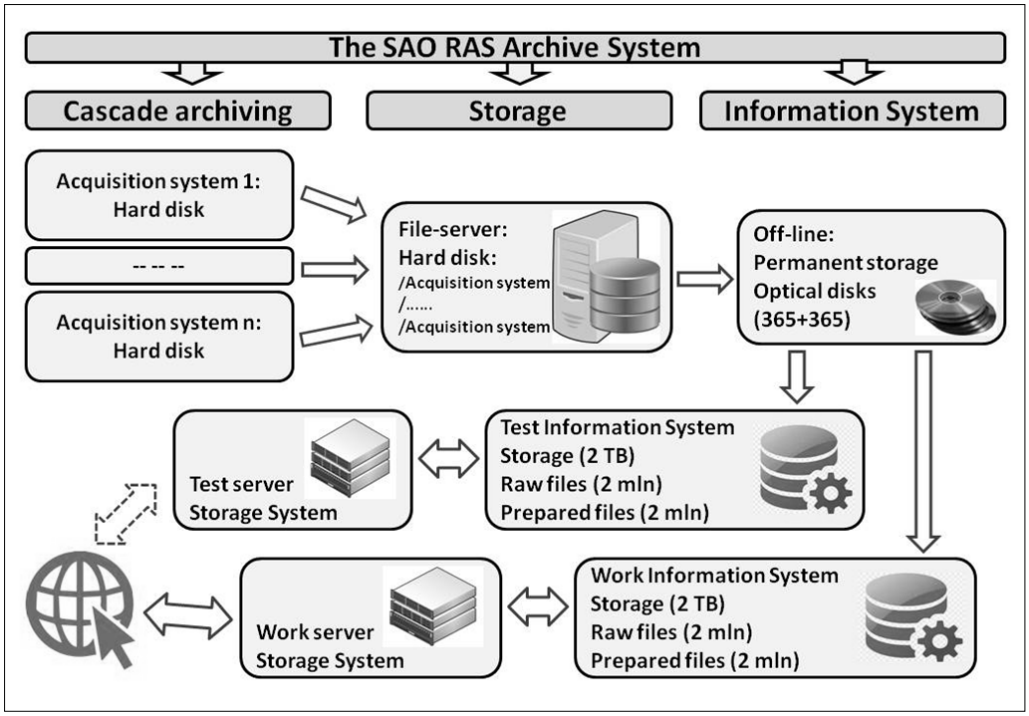


Fig. 3. The diagram represents components of the archive system and interaction between them.

include observations and additional information. Observational files are files of celestial object observations, as well as service files used in processing raw data. Additional information includes observation logs and files prepared by the observer during observations.

When developing the archival system, the following rules were used as a basis:

1) the logical unit in the archive is an observation;
2) files are provided upon request in the same format in which they were received in the archive;
3) the archive stores raw observational data recorded on optical disks;
4) the data have a two-year period of exclusive copyright for observation program applicants; after this period they are open for free access.

The archival system consists of three components:

(I) accumulation – cascade scheme of archiving;
(II) storage – hardware and software for permanent data storage;
(III) search and access – information system with web access.

The archive system has two databases located on two servers (Zhelenkova et al., 2017, 2019). Each server storage area has similar structure and content. One server supports the working version of the archive, and the second one supports the test version with which we carry out and test all new developments. Support for two databases provides additional data safety.

Figure 3 shows the components of the archival system, their interaction with each other, as well as the direction of data flows from the telescope to the permanent storage.

Data management in the archive includes data verification, ingestion, access, curation, long-term preservation and timely migration to new storage media. Taking into account the existing equipment, the prospect of increasing data volumes and modern technologies for supporting persistent archives, we are considering the possibility of organizing software-defined storage for our archive based on the iRODS[2] software (Zhelenkova et al., 2020) or similar systems.

2.1 Cascade scheme

The cascade archiving scheme consists of the process of moving data from the data acquisition system to permanent storage in the archive. Thus, the data accumulated on the computer of the data acquisition system during an observation set are copied to a dedicated file-server in the directory of the corresponding observation method. Then, from the observations obtained by one method, an archive disk is formed, which is then recorded on a CD/DVD media.

We use the following rules that determine the structure of an archive optical disk:

1) a label with the disk number and the name of the local archive and data directories is placed on the optical archive disk;
2) each directory contains data from one night or day (for radio data) of observations;
3) the directory name includes the observation date;
3) one observation is one file.

[2] http://irods.org

These rules do not impose any restrictions on the file format. In this way, we ensure that the archival system is scalable in terms of adding new collections, not only new local observatory archives but also digital collections from other observatories.

2.2 Storage area

To ensure the integrity of data in the case of destruction of physical media or the input/output errors, we have two levels of the archive storage. The low offline level is closed to user access. It includes two copies of optical disks with archival data, and it is used by the archive administrator when there is need to restore data in the case of emergency failures.

The archive storage area on the dedicated server includes the PRIME directory in which copies of optical disks are located and the DATA directory in which copies of disks are located after verification by our software tools.

Each disk is copied into the PRIME and DATA areas into directories with a name corresponding to the disk number in the offline archive. Next are the catalogs named by the date of observation and including observations for that date. The physical structure of directories is superimposed on the logical structure, reflecting the distribution of disks among local archives.

The PRIME and DATA areas are physically located on the RAID array of the storage system. The DATA area is an online storage and provides the functionality of the information system.

3 Information system

The information system of the archive is supported by the open-source DBMS PostgreSQL. Each observational file described in the database tables about 30 attributes. They are used for dynamic formation of the web interface, for mapping FITS parameters and identifying files.

The database schema includes two dozen tables and views. The tables of the information system can be divided into three groups based on a frequency of adding new records. The first group includes the tables which are filled when creating a database. New records can only be added into these tables when a new local archive is added into the information system. Such uploads do not occur often, so such tables can be considered as static.

The second group includes tables that are updated when analyzing a new CD/DVD added to the system. For example, new observation programs, new observers, etc., are added to these tables.

The third group includes the tables with information about each archived file. Entries are added to the tables while ingesting a new disk. In this process, the tables of the first two groups are used when analyzing a new disk. A special place is occupied by a table linking the attributes describing the observation with keywords from the FITS file header and UCD (Unified Content Descriptor) (Derriere et al., 2004).

3.1 Uploading disks into the archive system

Placing a disk in a persistent repository on the RAID array begins by copying it to the PRIME zone. Next, its contents are checked for compliance with the accepted rules; and after necessary corrections, the data are written to the online storage area.

When a disk is written to storage, the following checks are performed:

1) if several local archives are written to a disk, then the disk is processed in several scans – as many methods as are written on the disk. Then a symbolic link is established to the same drive in several corresponding directories of the logical structure of the storage area;
2) if copying the disk into the system it is found that observations of the night are in two directories, then they are copied to one directory;
3) if the name of the catalog does not contain the date of observations, then it is renamed so that the name contains the date;
4) observations are extracted from tar archives, if available;
5) files are recompressed if the compression method is different from bzip2;
6) files are converted from the internal format of the MIDAS system (files with the bdf extension) into the FITS format;
7) FITS headers are parsed; if the headers indicate a different instrument than the one to which the disk belongs, then the file is assigned to a different local archive.

Messages about errors and actions with disks are saved in log files. Additional operations with each disk are recorded in a bash program, which is used to fully or partially restore the storage and information system, if necessary.

After parsing the contents of the disk according to the given rules, we prepare a list of disk files that is used when uploading tables. Lists of the contents of optical disks are stored in text files in a special directory. These lists are part of the archive and can be used to recover information.

3.2 Standard queries

As can be seen from the analysis of file headers, there are common parameters of observations in different data acquisition systems. These parameters include information about the object, observation program, observer, program applicant, meteorological parameters, while the characteristics of the device usually differ. The values of these parameters are formed in the telescope control systems, as well as in the data acquisition system. Some parameters are included automatically in the file header, while others can be entered by the observer.

Based on common descriptive parameters, we selected the following types of queries to search for files, namely, by observation date, instrument, file types, coordinates of the observed field/object, name of the astronomical object, observation program, program applicant and observers who took part in observations.

The difficulties of uploading database tables of the necessary parameters are as follows:

1) when modernizing tools and data acquisition systems, file formats change, and, as a rule, the local archive has several versions of the format, differing in the set of keywords, as well as the form in which their values are recorded;

2) different acquisition systems form file headers with keywords that differ in name but denote the same physical quantity. For example, the date of an observation in different digital collections can be obtained from the following keywords: “DATE”, “DATE-OBS”, “Observation Date”, “OBS-DATE”.

For these reasons, a software filter for parsing and extracting parameter values from file headers is implemented using a special table in the database schema that links the keywords in the FITS file header and the observation attributes.

The observation date appears to be the most resistant to observer input errors. When analyzing files programmatically, we determine the date by the name of the directory with observations. Observation dates obtained from file headers are less reliable because they may contain errors introduced by the observer.

Based on the filename extension, we classify files into observational, log, auxiliary, and unclassified types. A file is assigned to one of these categories using a table containing all filename extensions found in local archives. The extensions .bdf, .mt, .fts, .fits, .tar are assigned to observation files; .tbl, .log, .plog, .base, .pro, .dbf, .lst are assigned to observation logs; and the rest are auxiliary and unclassified files. Since there are log files with .mt extension, the FITS file headers are further parsed to recognize tables and images, and table files are marked as observation logs.

Observation files include both observations of celestial objects and files used in data processing, namely: bias and dark frames, flat fields, standards. Separating files into these types is done by analyzing file names and keywords in file headers. If data type information is contained in both the file name and keywords, then the file name takes precedence.

The instruments used on telescopes generate data of varying structure. These can be direct images, echelle spectra, multi-object spectra, etc. The database lookup table establishes the relationship between the acquisition system and the observation mode.

The coordinate request is performed only for files with observations of celestial objects. The coordinates are extracted from the file header using the corresponding keywords and recalculated to epoch 2000.0.

The names of observers and principal investigators are extracted from the keywords OBSERVER and AUTHOR. As a rule, this information is entered by observers, which often leads to different spellings of the surname. Thus, some surnames have up to a dozen spelling options. To deal with the multiple synonyms, we had to extract a list of all possible observer name entries from the FITS headers of the observation files and place them in a special database table.

The difficulty in organizing a request by the name of an observation program is that the headers of files related to the same program may contain different names, for example, “GRB monitoring”, “GRB”, etc., but none of these names may be the same with what is stated in the observation schedule. A table has been compiled from the archive of observation schedules where, in addition to the name, each program is assigned an identifier, including the year, the number of the half-year, and the number of the program in the half-year. The selection of data related to the observing program is carried out using such an identifier.

4 User web interface

Using the web interface of the information system, open access to observation data is provided[3]. The web interface is created by the Perl script and dynamically displays the date range for local archives included in the database.

The interface implements a set of standard queries by the following attributes: observation date, coordinates, source name, name of the observing program, program applicant, observer, filter, file type, type of observations. If the parameters used for the request are not present in the file header, then it will not be included in the request result.

The start form lists the local archives available for the request, as well as the parameters that form the database request. The query results are displayed in a new browser window (see Fig. 4). This form displays a list of files obtained by the telescope for the selected dates. Files for one date can be downloaded as a tar archive. When one goes to the selected date, a window opens with a form that displays the parameters of the observation night files (see Fig. 5). One can view file headers and also visualize the file contents. In the window displaying the direct image, a coordinate correction can also be performed.

Searching for files by observation date can be implemented for the entire archive. If the query specifies only dates, then the observations, log files, and auxiliary files falling within the date range are selected.

When specifying coordinates, the search is performed in a square with a side equal to the specified double search radius. The center of the area is specified by the entered coordinates (right ascension and declination at epoch 2000.0). As a result of the request, files are returned whose coordinates specified in the FITS header parameters and reduced to 2000.0 fall within the specified area. If only the right ascension coordinate is specified, then the search for files is performed in the declination strip [−90°−+90°]; and if only declination is entered, then the file search is performed in the right ascension strip [0°−360°].

When requesting data by object name, its coordinates are retrieved using the Sesame name resolver[4] in Simbad and NED databases.

It should be noted that if the coordinates and name of the object are specified, then the search in the archive will be performed using the coordinates; the name of the object is ignored in this case.

When searching for observations by the last name of the observer or program applicant, an incomplete last name entry is allowed. The archive is searched by the first surname (in sorting order) corresponding to the entered template. For example, “mon” is entered, and this pattern matches “Monin”, “Montmerle”. The search is carried out using “Monin”.

One can also select the last name from the list that appears in a new window when going to “Program Author”.

The search for observations obtained according to any observation program is carried out using the identifier, which is a string that includes the year, half-year number, and the serial number of the program in the half-year. For reference,

[3] http://www.sao.ru/oasis/cgi-bin/fetch?Z&user&ru

[4] https://vizier.cds.unistra.fr/vizier/doc/sesame.htx

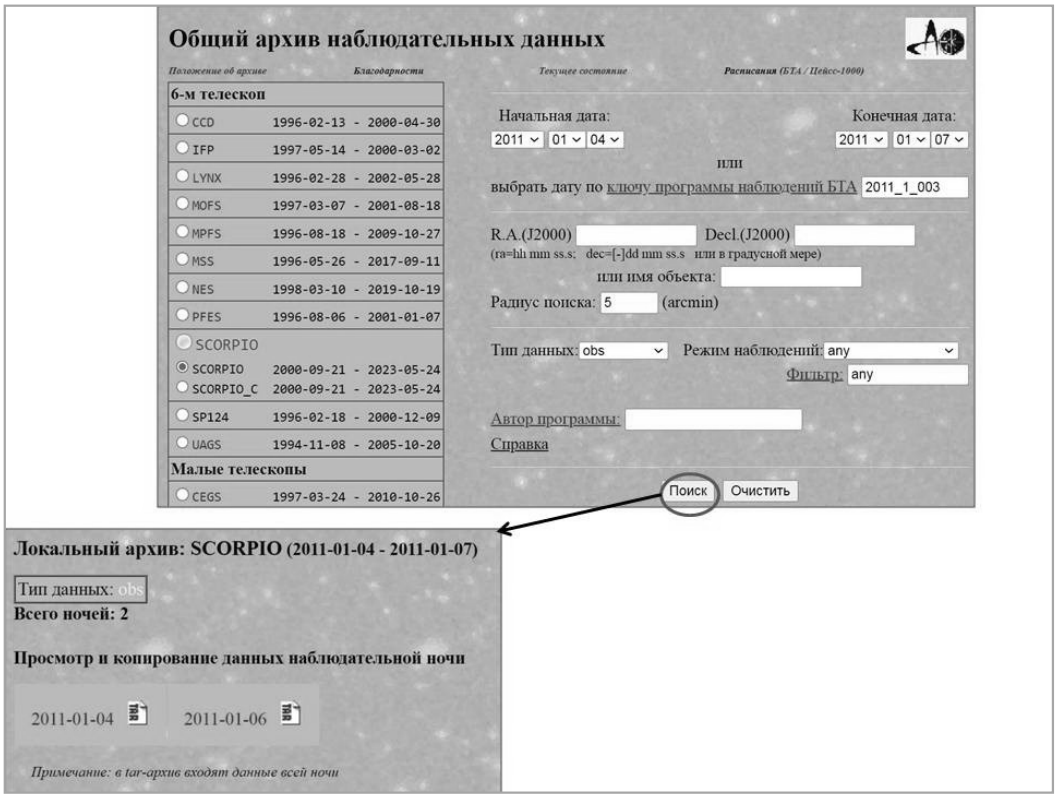


Fig. 4. Example of the starting web form of the archive system and a form with the result of the request.

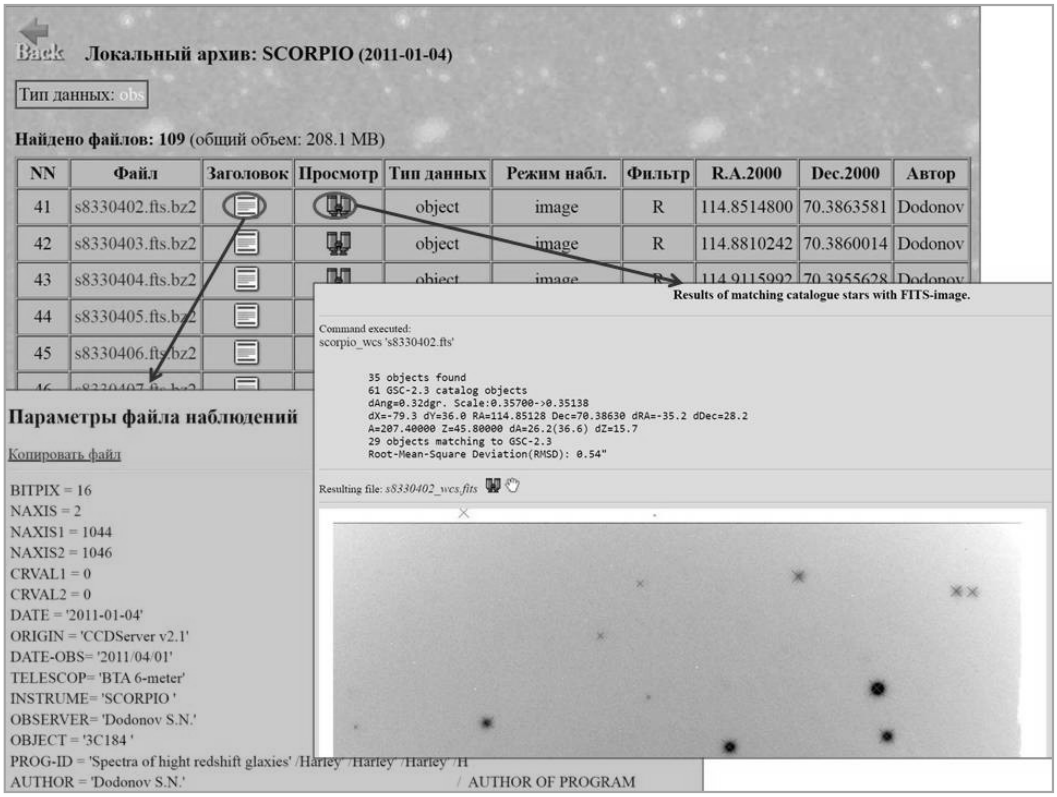


Fig. 5. Example of a query result for the selected date. This displays a table with a set of parameters for each file for the selected date. One can view the entire file header and also its contents using an on-the-fly generated image.

a list of programs with corresponding keys is displayed in a separate browser window.

Free access to archived files is provided only after the copyright of the program applicants will have expired, i.e., in two years from the date of observation. The authors of programs are granted access without limitation for a two-year period. To do this, they must register in the archive system by contacting the archive administrator, who will provide them with a login and a password to download the data.

5 Conclusions

Since astronomical observations do not lose their scientific significance over time, their preservation over a long period of time is one of the main tasks of a digital archive.

Permanent archival storage means storing digital data collections for decades, or better yet, without time limits. This requires mandatory storage, along with the digital collection, of a description of its organization and metadata that allows the information to be interpreted. A permanent archive must find, access, and display the digital objects it stores, even with changing storage technologies. Obsolescence of media coupled with large volumes of data can make persistent storage nearly impossible.

When reading, writing and storing data, errors may occur, leading to a loss of information. Although the reliability of read/write devices improves, and errors occurring during read/write can be controlled by software, the life of such devices is estimated to be 5–10 years, which is less than the service life of the storage medium itself. Recognition of the digital files depends on the software used, the life cycle of which is estimated to be 5–7 years. It implies that during long-term storage it is necessary to monitor the state of the archive and periodically rewrite the data to new media.

Since 2003, we have been taking these circumstances into account when developing and maintaining the archival system of the observatory. This experience of management of various digital collections allowed us to develop the following approach to data preservation: data duplication on various types of digital media. To ensure the survivability of the information system, duplication of the database was implemented on two dedicated servers. If one server fails, the second server can provide access to data. This also allows the archival system to be developed both in terms of software and in terms of adding new collections, because these actions are performed on the test server, and after verification they are transferred to the work server. We also note that this organization of the archival system allows one to rewrite digital collections onto a new type of media.

Below we provide some statistical information about the files stored in the archive. Thus, 98.5% of files are classified as observational data, 0.2% are observation logs, 1.1% are auxiliary files, and 0.2% are files not classified as the above types.

Forty percent of files are observations in the optical range. Of these, 62.2% are celestial object observations, 17.0% are bias frames, 1.5% are dark frames, 8.7% are flat fields, 9.0% are standards, and 1.5% are files not classified as listed types. The division by observation modes is as follows: direct images, 20.4%; spectra (echelle, log-slit, multi-slit), 11.3%; the Fabry–Perot cube, 5.9%; polarization photometry and spectra, 1.7%; and files whose type is not defined programmatically, 0.2%.

The request by date of observation is implemented for all archived files. Other standard queries are implemented for a part of files with observational data due to a lack of parameters in the file headers. Thus, 85% of the files have coordinates; 88%, the name of the observer who conducted the program; 25%, the name of the object; 30%, the observation program; 28%, the applicant of the program.

References

Berlin A.B., Gassanov L.G., Gol'nev V.Y., Korol'kov D.V., Parijskij Y.N., 1984. Soobshcheniya Spetsial'noj Astrofizicheskoj Observatorii, vol. 41.

Borisenko A.N., Vitkovskij V.V., Zhelenkova O.P., et al., 1990. Bulletin of the Special Astrophysics Observatory, vol. 32, pp. 128–134.

Derriere S., Gray N., McDowell J.C., et al., 2004. In F. Ochsenbein, M.G. Allen, D. Egret (Eds.), Astronomical Data Analysis Software and Systems (ADASS) XIII. Astronomical Society of the Pacific Conference Series, vol. 314, p. 315.

Kononov V.K., Monosov M.L., Vitkovskij V.V., Lipovetskij V.A., 1990. Soobshcheniya Spetsial'noj Astrofizicheskoj Observatorii, vol. 65, pp. 32–44.

Shergin V.S., 2014. Tech. rep., Software for automatic calibration of coordinates of direct images. SAO RAS. Available at: https://www.sao.ru/hq/vsher/FITS/fits_wcs.html.

Verkhodanov O.V., Vitkovskij V.V., Eruhimov B.L., et al., 1993. Data presentation format in the recording and processing system on the 1st feed cabine of the RATAN-600 radio telescope, Preprint SAO RAS. St. Petersburg branch.

Verkhodanov O.V., Chernenkov V.N., Kononov V.K., Trushkin S.A., Tsybulev P.G., 1995. Tech. rep., Description of the format for recording multi-frequency RFLEX data coming from broadband receivers of the 1st RATAN-600 feed.

Vitkovskij V.V., Kononov V.K., Lipovetchkij V.A., Monosov M.L., 1987. Tech. rep., Development of the SAO AS USSR Archive. SAO AS USSR.

Vitkovskij V.V., Zhelenkova O.P., Ryadchenko V.P., Shergin V.S., 1988. Soobshcheniya Spetsial'noj Astrofizicheskoj Observatorii, vol. 59, p. 60.

Vitkovskij V., Zhelenkova O., Kalinina N., et al., 2000. Baltic Astronomy, vol. 9, pp. 578–582.

Warmels R.H., 1992. In D.M. Worrall, C. Biemesderfer, J. Barnes (Eds.), Astronomical Data Analysis Software and Systems I. Astronomical Society of the Pacific Conference Series, vol. 25, p. 115.

Wells D.C., Greisen E.W., Harten R.H., 1981. Astron. Astrophys. Supplement, vol. 44, p. 363.

Zhelenkova O.P., Vitkovskij V.V., Kalinina N.A., Shergin V.S., Chernenkov V.N., 2003. In Proceedings of the 5th All-Russian Conference Digital Libraries: Advanced Methods and Technologies, Electronic Collections. pp. 359–363.

Zhelenkova O.P., Vitkovskij V.V., Plyaskina, et al., 2017. In Proceedings of the All-Russian Scientific Conference

"System Synthesis and Applied Synergetics". pp. 143–149.

Zhelenkova O.P., Vitkovskij V.V., Plyaskina T.A., Shergin V.S., Chernenkov V.N., 2019. In M. Molinaro, K. Shortridge, F. Pasian (Eds.), Astronomical Data Analysis Software and Systems XXVI. Astronomical Society of the Pacific Conference Series, vol. 521, p. 128.

Zhelenkova O., Vitkovskij V., Chernenkov V., Shergin V., Plyaskina T., 2020. In R. Pizzo, E.R. Deul, J.D. Mol, J. de Plaa, H. Verkouter (Eds.), Astronomical Data Analysis Software and Systems XXIX. Astronomical Society of the Pacific Conference Series, vol. 527, p. 69.