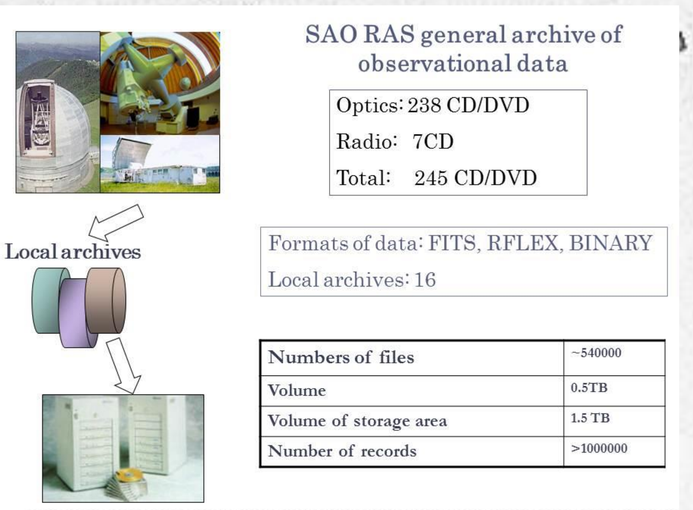
P4.21 Long-term data management in the SAO RAS archive system

Zhelenkova O.P.[1,2], Vitkovskij V.V.[1,2], Plyaskina T.A.[1], Shergin V.S.[1], Chernenkov V.N.[1]

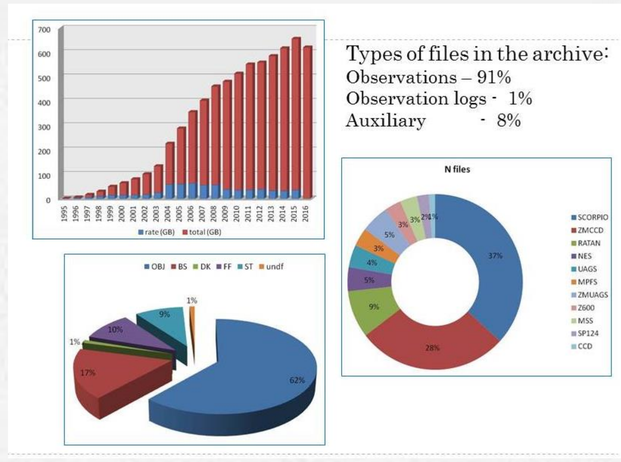
[1] - SAO RAS, Nizhnij Arkhyz, Russia; [2] - ITMO University, Saint-Petersburg, Russia

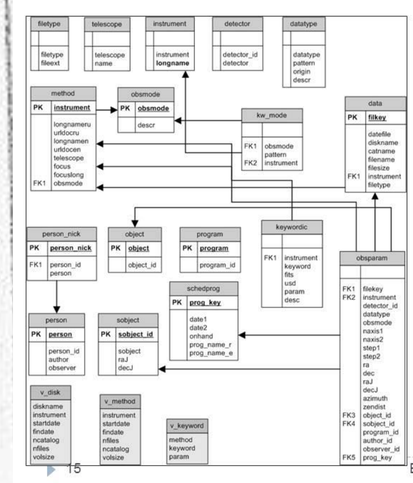
Abstract. The SAO RAS has the archive facility, which contains heterogeneous digital collections with the observations, obtained on the different instruments of SAO RAS since 1994. Within so long term of existence the archive underwent substantial changes in the data formats, methods of data processing and storage. There is no doubt in the need for long-term keeping of astronomical data. A timely migration of digital files on the modern carriers is required to ensure long-term storage of data.

It's been already about 40 years since two the largest Russian instruments - BTA optical telescope, 6/m diam. and RATAN-600 radio telescope, 600/m diam. started to act in the Special Astrophysical Observatory of the Russian Academy of Sciences (SAO RAS). The history of digital collections in the observatory had been started in 1980 after realization of a deep blind radio survey of. Then with the introduction of CCD cameras into observation in the late 80s we started to develop an archive system, which united the different digital collections with observations from diverse acquisition systems.



The observatory has the archive facility, which contains heterogeneous digital collections with the observations, obtained on the different instruments of SAO RAS since 1994. Within so long term of existence the archive underwent substantial changes in the data formats, methods of data processing and storage media. These changes are tightly connected with the progress of computer and observation facilities and with the needs of collaborative scientific research. The distinctive feature of the archive is heterogeneity of data formats, data quality and methods of data processing. In the archive there are data of 40 acquisition systems, which are grouped in 16 digital collections. They include raw observations and calibration data. files with observations. The most of them (96\%) are represented in FITS-format. The system supports public access to 16 local archives with data of different instruments used, or been used on the SAO RAS telescopes. Our archive system is difficult to compare on the data volume with the largest data centers as the ESO Archive Facility, but problems with the support of the same level of an archive system is typical.





OASIS table schema

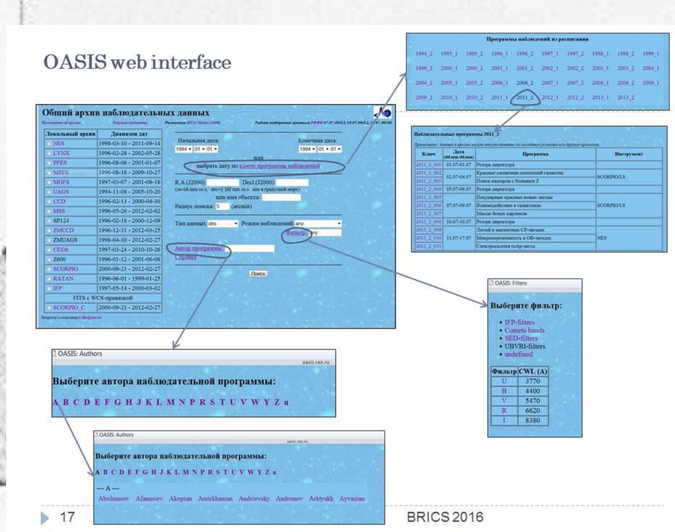
Each file is described with about 60 parameters.

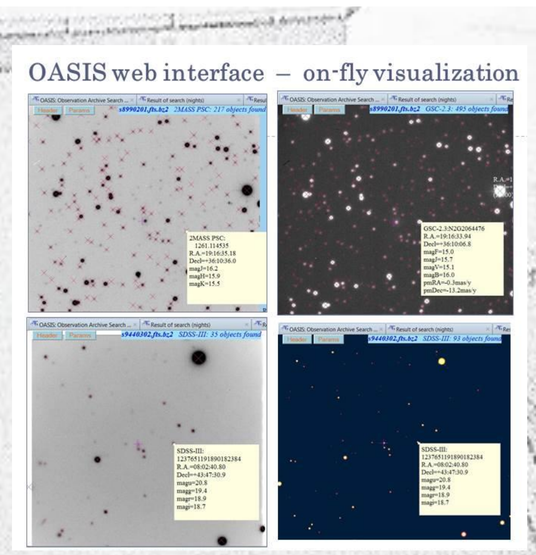
They are used for the dynamic formation of the web interface, for mapping parameters FITS-header and attribute of SIS tables, identify files, specify a file type, etc.

Some tables are vocabularies, others contain parameters

When developing an archive system we conducted an analysis of the completeness of the parameters that are necessary for web access to each local archive, as well as correction, where possible, missing or erroneous values in observation files. For a description of observations on different instruments BTA, we have identified sets of parameters that have a common part, which includes information about the object, program, etc., and characteristics inherent to a particular instrument. The description contains up to 100--300 parameters. These parameter values are formed in telescope control systems and acquisition systems. Some of them supplied automatically in the header file, and other is recorded by an observer. It notes that the query by observation date is implemented to the whole archive, but other types of requests can be made only to the part of the local archives because of the lack of the necessary parameters in the file header. Difficulty of filling of the information system tables with parameters, which necessary for the standard query, is that:

- when upgrading of instruments and acquisition system, data formats are changed. Usually a local archive format has several versions that differ in the set of keywords, their formats and values;
- different acquisition systems form the keywords in a file header with different names but denoting the same physical quantity. So, for an observation date in different digital collections we can obtain a value from a set of keywords -- ``DATE'', ``DATE-OBS'', ``Date of observation'', ``OBS-DATE'' and etc.





For safety and integrity of data with the different situations like physical destruction of media or input/output errors we have two copies of the archive data on CD/DVD disks. They consist an off-line persist level of a storage. An on-line level of storage is a RAID array. The archive facility has two databases located on two servers. Each server data storage area has a similar structure and content. The hard disk copy of archive will accelerate the re-writing of data with migration into modern carriers. One server supports and contains the working version of the system, the second one supports the test version, with which we carry out and test all the new developments. The supporting of two instances provides addition data persistent. The life cycle of modern digital carriers is usually 5-10 years, that also applies to the read-write hardware and software. A timely migration of digital files on the modern carriers is required to ensure long-term storage of data. It is not possible to completely abandon from external carriers for long-term data storage and store the information to the database only.