IRSTI 41.51.02, 41.51.41 https://doi.org/10.26577/RCPh20994326

I.M. Izmailova[1*], N. Akbar[1,2], M.A. Makukov[1], D.V. Yurin[1]
[1]Fesenkov Astrophysical Institute, Almaty, Kazakhstan
[2]Farabi University, Almaty, Kazakhstan
*e-mail: izmailova@fai.kz

A VO-CENTRIC DATA LAKE WITH AUTOMATED EMBARGO FOR FAI DATA ARCHIVE

The study aims to modernize the scientific and technical infrastructure of the Fesenkov Astrophysical Institute and the services of the Kazakhstan National Virtual Observatory to integrate heterogeneous astronomical data into the global research workflow. The methodology is based on designing a scalable file storage split into logical sections (partitioned Data Lakes), complying with International Virtual Observatory Alliance standards, implementing automated streaming FITS reduction pipelines, and applying the 'Scientific Showcase' access control model. The key result of the work is the development and deployment of an end-to-end automated pipeline for astronomical data ingestion, sorting, calibration, and cataloging. A software suite for metadata indexing was established, and a four-class data embargo policy was administratively institutionalized via a public license agreement. Pilot testing of the pipeline on the photometric archive of the 1-meter Zeiss-1000 telescope validated the architecture's stability, successfully cataloging over 800,000 files totaling more than 8 terabytes. The novelty of this work lies in the creation of a hybrid model that separates the physical storage layer from international publishing protocols while reliably protecting the priority rights of principal investigators. The practical significance of the results is that we reorganized and structured our local observational archives: fragmentation was eliminated, unified standardization was introduced, and open access to the data was provided. The created software and hardware infrastructure is easily scalable, allowing new datasets to be integrated without changing the core architecture.

Keywords: astronomical databases, virtual observatory, information storage and retrieval systems, data management, pipeline processing, data lake, data policy.

И.М. Измайлова[1*], Н. Ақбар[1,2], М.А. Макуков[1], Д.В. Юрин[1]
[1]В.Г. Фесенков атындағы Астрофизика институты, Алматы, Қазақстан
[2]әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан
*e-mail: izmailova@fai.kz

АФИ мәліметтер архивіне арналған автоматтандырылған эмбаргосы бар во-бағдарланған мәліметтер көлі

Зерттеудің мақсаты – әртекті астрономиялық деректерді жаһандық ғылыми айналымға интеграциялау үшін В.Г. Фесенков атындағы Астрофизика институтының ғылыми-техникалық инфрақұрылымын және Қазақстан ұлттық виртуалды обсерваториясының қызметтерін жаңғырту болып табылады. Жұмыстың методологиясы логикалық секцияларға бөлінген ауқымды файлдық қойманы жобалауға, Халықаралық виртуалды обсерваториялар альянсының стандарттарына сәйкестікке, FITS-файлдарын ағындық автоматты редукциялау конвейерлерін іске асыруға және «Ғылыми витрина» қолжетімділікті шектеу моделін қолдануға негізделген. Зерттеудің негізгі нәтижесі – астрономиялық деректерді қабылдау, сұрыптау, калибрлеу және каталогтауға арналған тұйық автоматтандырылған конвейерді әзірлеу және енгізу болып табылады. Метадеректерді индекстеуге арналған бағдарламалық кешен құрылды, сондай-ақ жария оферта келісімі арқылы деректер эмбаргосының төрт класты саясаты әкімшілік тұрғыдан бекітілді. Конвейерді 1 метрлік Zeiss-1000 телескопының фотометриялық архивінде пилоттық сынақтан өткізу әзірленген архитектураның тұрақтылығын дәлелдеді; бұл жұмыс аясында жалпы көлемі 8 терабайттан

 © 2026 Al-Farabi Kazakh National University

астам болатын 800 000-нан астам файл сәтті каталогталды. Жұмыстың жаңашылдығы — біз деректерді физикалық сақтау деңгейін халықаралық жариялау хаттамаларынан бөлектейтін гибридті модель құрдық, бұл ретте бас зерттеушілердің басымдық құқықтары да сенімді қорғалған. Нәтижелердің практикалық маңыздылығы — біздің локальды бақылау архивтеріміз ретке келтіріліп, жүйеленді: олардың шашыраңқылығы жойылды, бірыңғай стандарттар енгізілді және деректерге ашық қолжетімділік қамтамасыз етілді. Құрылған бағдарламалық-аппараттық инфрақұрылымды оңай кеңейтуге болады, бұл негізгі архитектураны өзгертпей-ақ жаңа деректер массивтерін қосуға мүмкіндік береді.

Түйін сөздер: астрономиялық мәліметтер базасы, виртуалды обсерватория, ақпараттық-іздестіру жүйелері, мәліметтерді басқару, мәліметтерді конвейерлік өңдеу, деректер көлі, деректер саясаты.

И.М. Измайлова[1]*, Н. Акбар[1,2], М.А. Макуков[1], Д.В. Юрин[1]
[1]Астрофизический институт им. В.Г. Фесенкова, Алматы, Казахстан
[2]Казахский национальный университет им. аль-Фараби, Алматы, Казахстан
*e-mail: izmailova@fai.kz

Во-центричное озеро данных с автоматизированным эмбарго архива данных АФИФ

Целью исследования является модернизация научно-технической инфраструктуры Астрофизического института имени Фесенкова и сопутствующих сервисов Казахстанской национальной виртуальной обсерватории для интеграции разнородных астрономических данных в глобальный исследовательский оборот. Методология работы базируется на проектировании масштабируемого файлового хранилища, разделенного на логические секции (partitioned Data Lakes), соответствии стандартам Международного альянса виртуальных обсерваторий, реализации конвейеров потоковой автоматической редукции FITS-файлов и применении модели разграничения доступа «Научной витрины». Ключевым результатом исследования является разработка и развертывание сквозного автоматизированного конвейера для захвата, сортировки, калибровки и каталогизации наблюдательных материалов. Создан программный комплекс для индексации метаданных, а также административно закреплена четырехклассовая политика эмбарго данных через соглашение публичной оферты. Пилотное тестирование конвейера на фотометрическом архиве 1-метрового телескопа Zeiss-1000 подтвердило стабильность разработанной архитектуры, позволив успешно каталогизировать более 800 000 файлов общим объемом около 8 терабайт. Новизна данной работы заключается в создании гибридной модели, изолирующей уровень физического хранения от протоколов международной публикации при одновременной защите приоритетных прав главных исследователей. Практическая значимость результатов заключается в реорганизации и усовершенствовании структуры локальных наблюдательных архивов: ликвидирована их фрагментарность, обеспечены единая стандартизация и открытый доступ к данным. Созданная программно-аппаратная инфраструктура обладает масштабируемостью, что позволяет интегрировать новые массивы данных без изменения базовой архитектуры.

Ключевые слова: астрономические базы данных, виртуальная обсерватория, информационно-поисковые системы, управление данными, конвейерная обработка данных, озеро данных, политика данных.

Introduction

Modern astronomy is undergoing a major methodological shift: the classical paradigm focused on targeted observations of individual objects is gradually giving way to wide-field sky surveys and continuous petabyte-scale data streams [1]. The exponential growth of data volumes has exposed the decentralized nature of storage infrastructures at many regional observatories. According to estimates [2, 3], the lack of unified data management

approaches and reliance on manual logging led to the loss of up to 50% of numerical scientific results, keeping them from entering global scientific circulation.

To address this systemic issue, the International Virtual Observatory Alliance (IVOA)[1] developed a three-tier architecture comprising the Resource Layer, the VO Core (middleware), and the User Layer. This framework ensured technical data interoperability long before the institutionalization of modern storage standards [4]. The open access and cross-service search principles historically embedded in the Virtual Observatory (VO) largely anticipated the global FAIR data principles [5], the retrospective alignment and practical adaptation of which to astronomical repositories have been thoroughly investigated in [6]. Budgetary analysis of successful projects (e.g., SDSS, 2MASS) confirms that up to 50% of modern observational complex development costs should be allocated to software and data management systems to minimize the risks of diminishing scientific return on instruments [3].

Within this paradigm, the functional role of the scientific archive is shifting from a static repository to an active component of the computational pipeline. Implementing a VO-centric archive architecture based on an invariant metadata model like CAOM (Common Archive Observation Model) [7] demonstrates that observation descriptions can be effectively abstracted from hardware specifications. Furthermore, using the Table Access Protocol (TAP) and the Astronomical Data Query Language (ADQL) ensures that client interfaces remain decoupled from the backend. This enables the server infrastructure to scale and modernize without requiring modifications to end-user application software [7–8].

Beyond structural optimization, open science requires ensuring the verifiability and reproducibility of results, which is achieved through end-to-end data provenance tracking [9]. The resulting conflict between the concept of Open Access and protecting the priority rights of Principal Investigators (PIs) is resolved by introducing automated proprietary periods (embargos). This approach is formally established in the ESO Science Archive [10–11], the ESO Data Access Policy [12], as well as in the data rights regulations of the Hubble [13], Chandra [14], NRAO [15], and Keck [16] missions. Global statistics indicate that the operation of open archives increases publication activity by a factor of 3 to 5 due to the secondary reuse of data by the external scientific community [3].

The technical implementation of this embargo policy is based on the "Scientific Showcase" pattern. To publish data within the KazVO infrastructure, we utilize the DaCHS (Data Center Helper Suite) integrated software solution [17]. DaCHS dynamically filters the exposed datasets: if the observations are under an active proprietary period, only the FITS file metadata (headers) are broadcasted into the open registry. The actual binary files remain isolated within a secure segment of the Data Lake [18] until the calculated expiration timestamp is reached.

This paper presents the interim results of modernizing the scientific and technical infrastructure of the Fesenkov Astrophysical Institute (FAI) and the associated services of the Kazakhstan National Virtual Observatory (KazVO)[2]. The primary objective of this study is to describe the developed architecture of a unified canonical archive, the principles of its integration with the provenance model, and the operational algorithms of the automated embargo pipeline built upon the "Scientific Showcase" concept.

Materials and Methods

The operation of the KazVO archival infrastructure is based on the end-to-end automation of processes: from the primary ingestion of observational materials to their deterministic publication. The overall architectural logic of the developed end-to-end pipeline, which links the processes of inter-server replication, preprocessing, physical partitioning, and access control rights, is illustrated in Figure 1.

A detailed description of the developed methods, algorithms, data storage structure, and functional pipeline components is provided in the respective subsections below.

Data Storage Architecture and Partitioning

The physical infrastructure of the KazVO data repository is designed around the principle of a Partitioned Data Lake. The file space organization logic accommodates the future scaling of the system to the level of a regional data center capable of integrating external heterogeneous datasets.

The storage is divided into two macro-contours: /observations/ (observational results) and /simulations/ (modeling and synthetic spectra). The directory topology within the observational contour has a strict hierarchical structure, as shown in Figure 2.

[1] https://www.ivoa.net/

[2] https://vo.fai.kz/

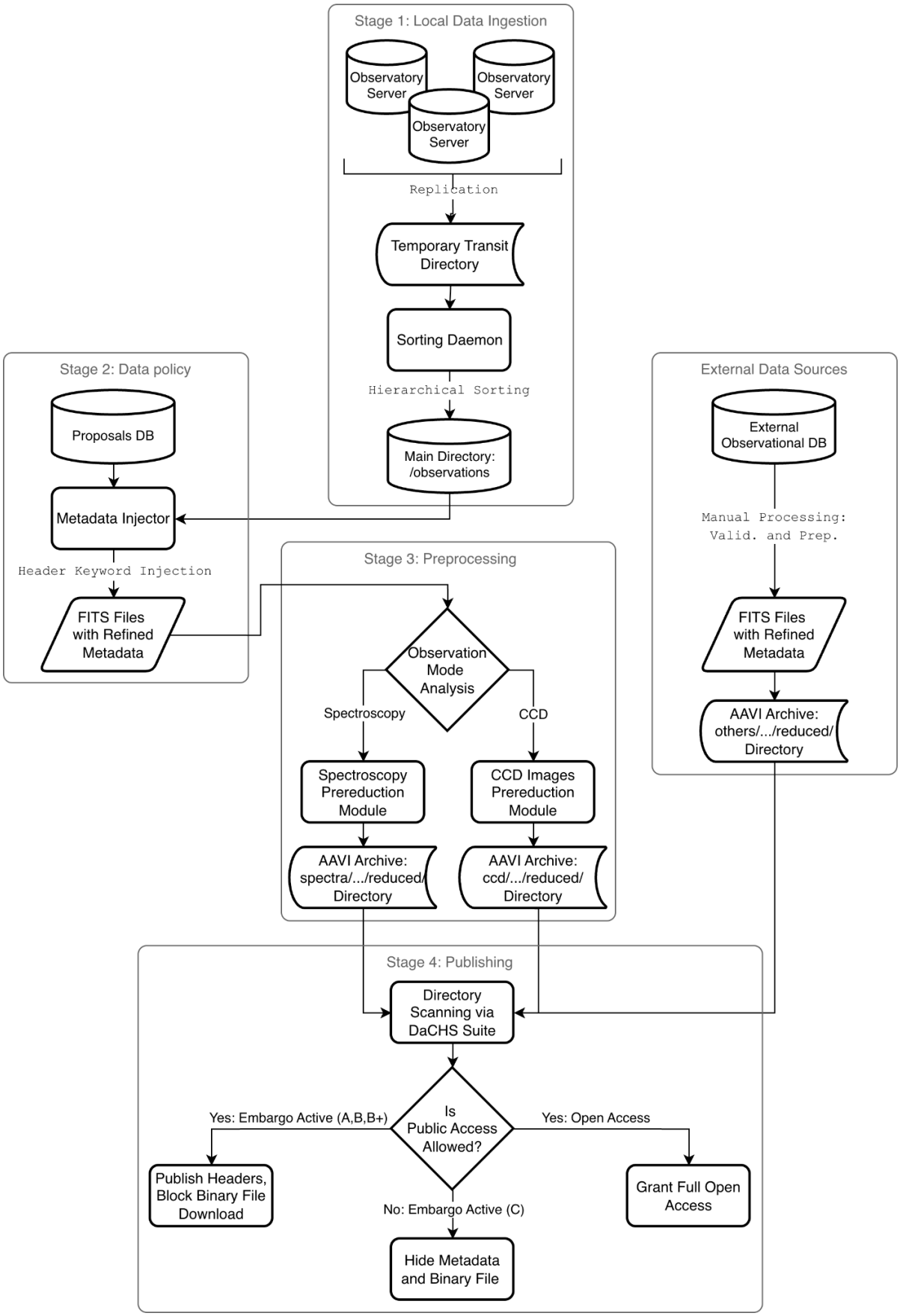


Figure 1 – End-to-end architectural diagram of the KazVO data ingestion, processing, and publication pipeline

```
/fai/
|--observations/
|   ‘-- <observatory_id>/
|       ‘-- <instrument_id>/
|           ‘-- <observ_mode>/
|               ‘--calibration
|                   ‘-- <YYYY>/
|                       ‘-- <YYYY-MM-DD>/
|                           ‘-- bias/
|                           ‘-- dark/
|                           ‘-- flat/
|                           ‘-- lamp/
|                           ‘-- masters/
|                           ‘-- meta/
|               ‘--targets
|                   ‘-- <YYYY>/
|                       ‘-- <YYYY-MM-DD>/
|                           |-- raw/
|                           |-- prereduced/
|                           |-- reduced/
|                           ‘-- meta/
/--simulations/
```

Figure 2 – Partitioning topology of the FAI observational Data Lake.

Deep partitioning isolates data by observatory (<observatory_id>), telescope (<instrument_id>), and observation mode (<observ_mode>: CCD imaging, long-slit spectroscopy, or echelle spectroscopy). Within each mode, the data are split into two pools:

1. Calibration frame pool (calibration): contains date folders (YYYY-MM-DD) segmented by frame types (bias, dark, flat, lamp). It also includes a masters/ directory for integrated master frames.
2. Scientific frame pool (targets): structured by observation years and dates. FITS files are distributed across processing stages:
   - raw/ — raw files directly from the telescopes;
   - prereduced/ — intermediate frames after geometric correction. This stage is critically important for working with digitized archival spectral data, the preprocessing of which is performed using the ARXSP software [19];
   - reduced/ — calibrated science products matching Phase 2 requirements;
   - meta/ — local service files and pipeline execution logs.

The hierarchical directory structure isolates heterogeneous datasets at the physical level, reduces filesystem overhead, and establishes a predictable topology for automated indexing.

To organize a huge amount of data into the required directories according to this hierarchical structure manually would require a lot of time. To solve the time-consuming problem of data sorting, a special file sorting program [20] was created. The program's sorting algorithm begins by checking the specified directory for folders with a date. Within these folders, the "*.fits/*.fit" files are categorized into calibration and scientific frames and moved to their respective destinations. Then the sorted files are placed in the previously mentioned data structure according to the corresponding year and date catalogs. Calibration files are in turn sorted by frame type (bias, dark, flat, lamp). Other service and log files are placed in the "meta" folder. The visual algorithm of the program code is shown in Figure 3.

With this tool, sorting several thousand files into the required categories takes a short period of time. At the same time, possible errors during program execution are also recorded.

Metadata Modeling and Software Layer Architecture

To facilitate cataloging, rapid search, and knowledge base generation for FITS files, the specialized software suite FFDB [21] (FAI FITS database) was developed. Its modular server architecture consists of three main components:

1. Metadata Registry (DB): a relational database that stores physical filesystem attributes (file paths, sizes, access permissions, and SHA-256 hashes) alongside scientific and technical parameters extracted from FITS headers (keywords, values, and comments).

2. Indexing Daemon (Scanner): a background Python process that continuously monitors the Data Lake. Upon detecting new or modified files, the scanner parses their headers using the astropy library, updates the database records, and automatically flags corrupted datasets.

3. Access Gateway (API Server): a web server built with the FastAPI framework that handles JSON data exchange and access control. All operations—including multi-parameter queries, LDAP authentication, tagging, and secure temporary download link generation—are unified via HTTP API.

The FFDB architecture completely automates background cataloging and FITS file analysis, establishing a centralized search environment for extracting accumulated astronomical data.

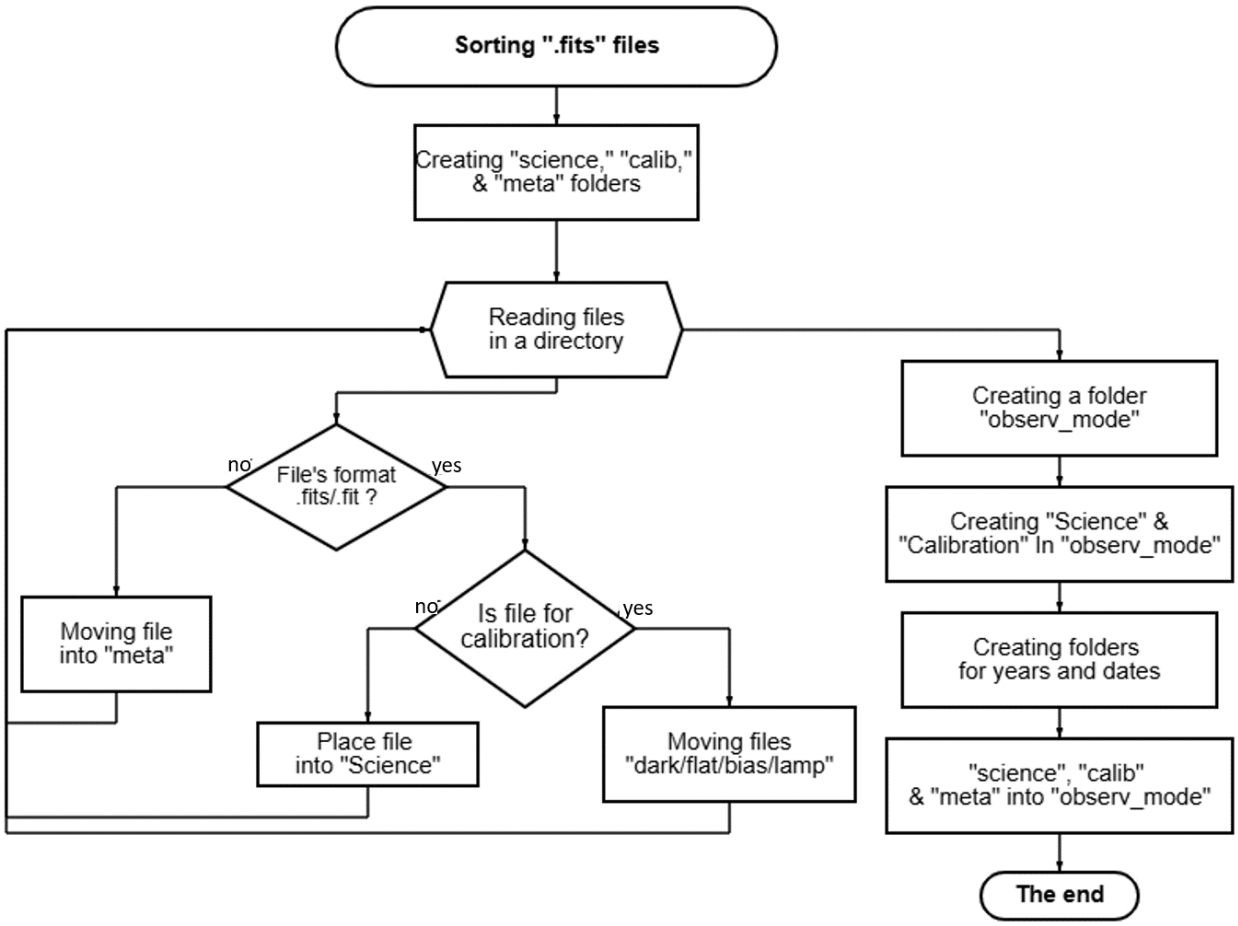


Figure 3 – Sorting algorithm for "fits/fit" and other files

Automated Embargo Algorithm

The technical implementation of the "Scientific Showcase" concept is based on an end-to-end automated pipeline comprising inter-server synchronization, preprocessing, and dynamic publishing. The pipeline runs through the following sequential stages:

1. Ingestion (Capture): A software module replicates primary archives from remote observatory servers to the institute's server, placing them in a temporary transit directory. Subsequently, the sorting daemon routes the files to the main /observations storage directory in accordance with the aforementioned hierarchy (see Figure 1, Stage 1).

2. Preprocessing: A background script parses the original raw file, performs primary reduction, and calculates the data class along with the embargo expiration timestamp. These parameters, combined with the proposal metadata (PI_NAME, COI_NAME, PROPOSID, CONTACT, REFERENC), are written back to the header of the newly generated FITS file as metadata keywords: OBS-CLSS (observation class) and RELEASE (the public release date).

3. Routing: Processed frames are saved to the reduced/ directory of the corresponding observation mode. Currently, the pipeline is fully functional for photometric data, while spectroscopic data processing modules are undergoing testing. The source code and configuration scripts are deposited in a GitHub repository [20].

4. Indexing (FFDB Integration Layer): At this stage, the FFDB suite is integrated into the pipeline. Its background indexing daemon (Scanner) continuously monitors the Data Lake, extracts the newly injected FITS headers, and immediately imports them into the centralized relational Metadata Registry.

5. Publishing: The DaCHS software suite scans the directories of respective observation modes, extracts the RELEASE embargo tags from the FITS headers, and exposes the metadata to the public IVOA registry while blocking binary file downloads until the specified open-access date is reached.

To automate access rights management across the pipeline, we have implemented a strict categorization of observational programs, presented in Table 1. The schematic workflow of this end-to-end ingestion, processing, and publishing pipeline is shown in Figure 1.

Table 1. Observation classes and access levels in the KazVO archival infrastructure

Observation Class	Data Category	Embargo Period	Status in KazVO after Embargo
A	Standard scientific observations	12 months	Public
B	Monitoring observations	Up to 36 months	Public
B+	Long-term monitoring projects	Up to 60 months	Public
C	Strategic / Commercial data	Indefinitely	Restricted / Unlisted

The assignment of these classes is strictly determined by the type of observational program:

- Class A: Observations of point and extended sources obtained through regular proposals, including standard photometric series and routine spectral observations.
- Class B: Short- and medium-term monitoring programs targeting variable objects and dynamic stellar systems.
- Class B+: Large-scale sky surveys, multi-year spectral monitoring campaigns, and international collaborative programs.
- Class C: Positional, photometric, and spectral monitoring of near-Earth space (including active satellites and space debris), as well as proprietary commercial data. These datasets are completely excluded from automatic release cycles, and their distribution is governed strictly by individual non-disclosure agreements.

This established categorization and the "Scientific Showcase" algorithm fully automate the data life cycle, guaranteeing proprietary confidentiality during the initial phase and a structured, deterministic transition of scientific results to the public domain.

Results and discussion

Regulatory Framework and Embargo Policy Implementation Status

A key milestone of this work is the formal legal and administrative institutionalization of the developed data embargo policy. The established access control regulation has been officially approved as an internal regulatory document of the FAI.

Starting from the third quarter of 2026, the proposal submission procedure for observing time on FAI instruments has been fully synchronized with the developed data classification scheme. A public license agreement has been published on the official KazVO web platform [22]. When submitting an observing proposal, the Principal Investigator (PI) is now required to explicitly declare the requested data class (A, B, B+, or C). This selection automatically predetermines the subsequent ingestion parameters and embargo duration timestamps to be written into the FITS headers.

The implementation of this regulatory framework has successfully legitimized the archiving workflow, eliminating potential legal risks associated with the division of researchers' intellectual property rights and the institute's commitment to advancing open science principles.

Pipeline Testing and Verification on Real Datasets

To validate the developed Data Lake architecture and automated processing algorithms, a pilot deployment of the pipeline was conducted using a dataset of archival and current observational data

obtained with the 1-meter Zeiss-1000 telescope at the Tien-Shan High-Altitude Observatory (TShAO).

During testing, the developed sorting module successfully processed the accumulated repository of photometric data. The pipeline execution logic was verified according to the following scenario:

1. Ingestion: Using the developed sorting software, the raw test frames were structurally distributed into the target directories of the Data Lake in strict compliance with the /observations/tshao/zeiss1000_east/ccd/ topology.

2. Reduction: The automated photometric data processor performed primary reduction (including master frame assembly, flat-field correction, and dark-current subtraction), saving the calibrated science-ready outputs directly to the reduced/ directory.

3. Embargo Tagging: The embargo module calculated the precise expiration timestamp and integrated the corresponding metadata attributes into the FITS file headers (see Figure 4).

AS446-001_R.fit

```
SIMPLE  =                    T
BITPIX  =                   16 /8 unsigned int, 16 & 32 int, -32 & -64 real
NAXIS   =                    2 /number of axes
NAXIS1  =                 3056 /fastest changing axis
NAXIS2  =                 3056 /next to fastest changing axis
BSCALE  =   1.0000000000000000 /physical = BZERO + BSCALE*array_value
BZERO   =   32768.000000000000 /physical = BZERO + BSCALE*array_value
DATE-OBS= '2026-07-05T20:53:06.78' /YYYY-MM-DDThh:mm:ss observation start, UT
EXPTIME =   40.000000000000000 /Exposure time in seconds
EXPOSURE=   40.000000000000000 /Exposure time in seconds
SET-TEMP=  -20.000000000000000 /CCD temperature setpoint in C
CCD-TEMP=  -20.000000000000000 /CCD temperature at start of exposure in C
XPIXSZ  =   12.000000000000000 /Pixel Width in microns (after binning)
YPIXSZ  =   12.000000000000000 /Pixel Height in microns (after binning)
XBINNING=                    1 /Binning factor in width
YBINNING=                    1 /Binning factor in height
XORGSUBF=                    0 /Subframe X position in binned pixels
YORGSUBF=                    0 /Subframe Y position in binned pixels
READOUTM= '8 MHz (RBI Flood)' / Readout mode of image
FILTER  = 'R_johnson' /         Filter used when taking image
IMAGETYP= 'Light Frame' /       Type of image
OBJECT  = 'AS446    '
OBJCTRA = '20 50 58' /          Nominal Right Ascension of center of image
OBJCTDEC= '+49 00 24' /         Nominal Declination of center of image
OBJCTALT= ' 83.8733' /          Nominal altitude of center of image
OBJCTAZ = '351.3878' /          Nominal azimuth of center of image
OBJCTHA = '  0.0933' /          Nominal hour angle of center of image
SITELAT = '43 03 26' /          Latitude of the imaging location
SITELONG= '76 58 18' /          Longitude of the imaging location
JD      =   2461227.3702173610 /Julian Date at start of exposure
JD-HELIO=   2461227.3718680711 /Heliocentric Julian Date at exposure midpoint
AIRMASS =   1.0056812301035576 /Relative optical path length through atmosphere
TRAKTIME=   1.0000000000000000 /Exposure time used for autoguiding
FOCALLEN=   6583.0000000000000 /Focal length of telescope in mm
APTDIA  =   1040.0000000000000 /Aperture diameter of telescope in mm
APTAREA =   849486.67716979980 /Aperture area of telescope in mm^2
SWCREATE= 'MaxIm DL Version 5.24 130322 0SH3X' /Name of software that created
        the image
SBSTDVER= 'SBFITSEXT Version 1.0' /Version of SBFITSEXT standard in effect
TELESCOP= '         ' /         telescope used to acquire this image
INSTRUME= 'FLI      ' /         instrument or camera used
OBSERVER= '         '
NOTES   = '         '
FLIPSTAT= 'Rotate 90 CW'
SWOWNER = 'Maxim Krugov' /      Licensed owner of software
OBS-CLSS= 'A        '          / KazVO data-policy class (A/B/B+/C)
RELEASE = '2027-07-05'         / Embargo release date (YYYY-MM-DD)
PROPOSID=                  170 / Proposal identifier
PI_NAME = 'Nadezhda Vaidman'   / Principal investigator
COI_NAME= 'Anatoly Miroshnichenko' / Co-investigators
CONTACT = 'nvaldmann@gmail.com' / Proposal contact email
HISTORY Data policy: proposal 170; match=object_name; source=submission.csv/user
HISTORY s.csv.
END
```

Figure 4 – Structure of a FITS header with integrated KazVO data policy metadata blocks, administrative fields (OBS-CLSS, RELEASE, PROPOSID), ownership credentials (PI_NAME, COI_NAME, CONTACT), and execution provenance record.

The development of a similar automated pipeline for spectral datasets is currently at the design stage. This is due to the significantly more complex mathematical and instrumental nature of spectroscopy—such as the necessity of automating trace extraction, wavelength calibration, and continuum normalization—which demands the construction of customized instrumental profiles for each specific spectrograph.

Analysis of the Current Archival Data Volume

During the system deployment phase, a comprehensive audit of the institute's available archival funds was conducted. Complete cataloging at this stage is constrained by the physical isolation of

certain datasets, which remain stored locally on remote observatory servers lacking permanent high-speed communication channels to the institute's central repository. The quantitative breakdown of these cataloged archival funds by telescope and observation type is detailed in Table 2.

Table 2. Current cataloging status of FAI observational data (excluding service and log files)

Observatory	Type	Telescope	Observation mode	Total volume (GB)	Total quantitative volume (files)
TShAO	FAI	Zeiss 1000 east	CCD	3540	455198
		Zeiss 1000 west	Echelle spectra	460.06	83176
Assy	FAI	AZT-20	CCD	438.67	89886
			Slit spectra	1.92	1347
		RC-500	CCD	77.24	2089
		WFOS-400	CCD	2810	128730
		WFOS-700	CCD	41.98	649
Kamenskoe	FAI	AZT-8	Slit spectra	186.38	27057
		Maksutov meniscus telescope	CCD	169.52	9790
		Schmidt telescope	CCD	283.96	12296
TCO	Partner	Eshel-TCO	Echelle spectra	0.082	1026
At all				8009.81	809244

To illustrate the readiness of these data for processing and publication, Figure 5 presents the distribution of data volumes categorized by observation type and pipeline processing stage. Automated reduction and calibration of these datasets are performed by an open-source Python pipeline [20], built upon the standard astropy [23–25] and ccdproc [26] astronomical libraries.

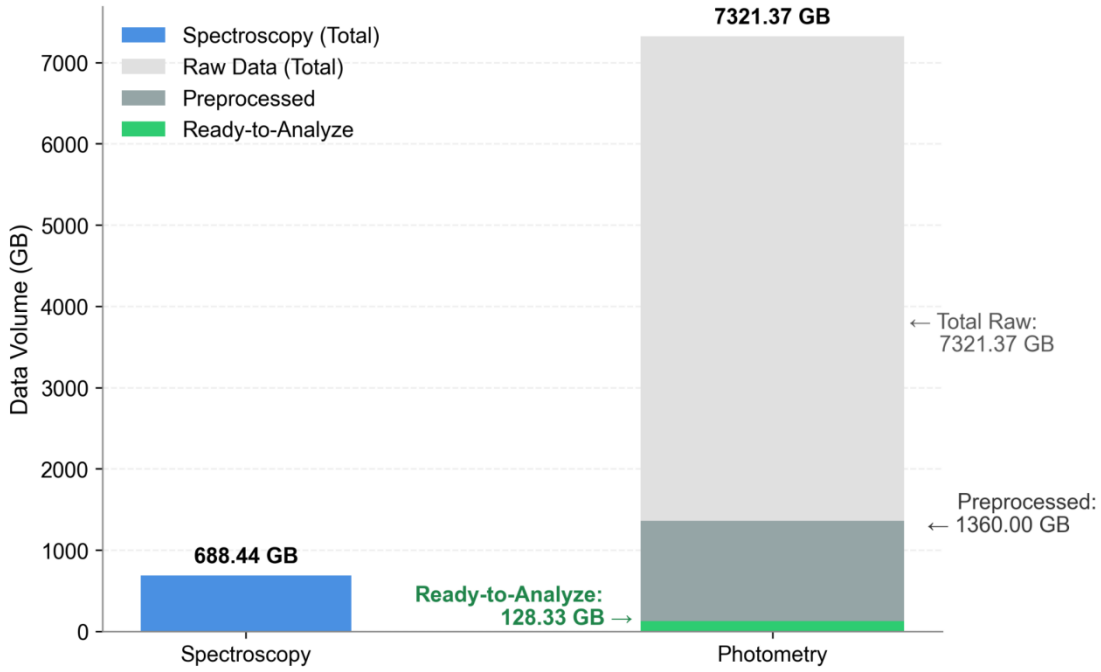


Figure 5 – Distribution of FAI observational data volumes by type (Spectroscopy vs. Photometry) and processing stage (raw, preprocessed, and ready-to-analyze).

The obtained statistical data indicate a substantial volume of accumulated observational material. The primary structuring of these datasets lays a solid foundation for scaling the developed pipeline across other observational instruments within the institute.

The KazVO infrastructure is positioned as a regional data center focused on the integration, long-term storage, and publication of heterogeneous datasets, encompassing both FAI scientific research results and data from external observatories, in full compliance with international IVOA standards. Within this framework, the integration and publication of archival spectroscopic materials obtained with the 0.81-meter telescope at the Three College Observatory (TCO, North Carolina, USA) were successfully carried out. The echelle spectroscopy data of hot supergiants were manually systematized, standardized, and fully published in open access via the Kazakhstan National Virtual Observatory platform [27].

Discussion

The deployment experience and pilot testing results have revealed several infrastructural factors that determine the vector for further optimization of the system:

- Communication Channels and Network Security: Remote FAI observatories are connected to the internet via satellite communication (specifically, Starlink), which provides sufficient bandwidth. However, to minimize the risks of network latency and prevent collisions during interactive telescope control, the inter-server data exchange is intentionally organized in a buffered batch mode according to a fixed schedule (during daytime or technical hours). Configuring and fine-tuning these regular replication cycles without affecting the instrument control loops remains a priority task.
- Architecture of Publication Services: The current configuration of the DaCHS publishing layer is being expanded toward creating globally separated services grouped by observation modes (photometry, spectroscopy, and, in the longer term, polarimetry) as unified end-to-end resources for the entire institutional fund, regardless of the specific instrument or CCD matrix used. This approach aggregates data into unified conceptual pools, simplifying cross-identification for end users and facilitating ADQL queries via TAP servers to file headers (including those under active embargo). The publication of highly specialized or unique datasets will be handled through isolated thematic services.
- Deployment of the FFDB Scanner: The full-scale launch of the background indexing daemon has been postponed until the complete reorganization and cleanup of the central institute server's disk arrays are finalized. This preventive measure is necessary to eliminate parser false positives triggered by improperly structured temporary files and duplicates accumulated during previous periods of unmanaged storage.

The step-by-step removal of these technological barriers will allow the developed system to transition from the pilot testing phase to a fully automated operations mode for the entire KazVO archival fund.

Conclusion

This paper presents the scientific, methodological, and software-hardware architecture of the Kazakhstan National Virtual Observatory (KazVO) data repository, deployed on the server infrastructure of the Fesenkov Astrophysical Institute (FAI). Based on the system design and pilot testing, the following key results have been achieved:

- Architectural: A physical Data Lake topology with deep partitioning was implemented, ensuring secure isolation and predictable addressing of heterogeneous datasets (photometry and spectroscopy). This approach minimizes filesystem overhead and simplifies the integration of new observational instruments.
- Technological: An end-to-end data processing and automated access control pipeline was developed based on the "Scientific Showcase" concept. The DaCHS software suite successfully serves as an end-to-end metadata publishing gateway, ensuring data discoverability within IVOA protocols without premature exposure of proprietary binary content.
- Administrative: A strict classification of observational programs (Classes A, B, B+, and C) was developed and legally implemented (effective from the third quarter of 2026) via a public license agreement on the KazVO web portal, legitimizing the balance between the interests of Principal Investigators (PIs) and open science principles.

Pilot testing of the pipeline on the photometric archive of the 1-meter Zeiss-1000 telescope (TShAO) confirmed the validity of the selected technological

solutions. Future development of the project will focus on optimizing inter-server communication channels for real-time data replication, transitioning the FFDB indexing daemon into full commercial operation, and deploying specialized modules for automated spectroscopic data extraction and reduction.

Acknowledgments

The author expresses sincere gratitude to all colleagues who participated in the development, testing, and fruitful discussions of the data storage architecture and observation data policy. Special thanks are extended to the engineering and observatory staff, as well as IT and technical support teams, for their invaluable support in facilitating data access and maintaining the technical infrastructure.

This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24992807).

Author Contributions

I.M. Izmailova: Conceptualization, Methodology, Software, Validation, Investigation, Data Curation, Writing – Original Draft, Visualization, Project Administration. **N. Akbar**: Software, Writing – Original Draft, Visualization. **M. Makukov**: Software, Methodology. **D.V. Yurin**: Supervision, Funding Acquisition, Conceptualization, Writing – Review & Editing.

References

1. M.A. C. Johnson, M. Paradies, M. Dembska, K. Lackeos, H.-R. Klöckner, D. J. Champion, and S. Schindler, Astronomical Pipeline Provenance: A Use Case Evaluation, *arXiv:2109.10759* (2021). https://doi.org/10.48550/arXiv.2109.10759
2. R.P. Norris, The Management of Astronomical Data, In *Proceedings of the 26th meeting of the IAU, Special Session 6, 22 August, 2006* (Prague, Czech Republic), id.1.
3. R. Norris, How to Make the Dream Come True: the Astronomers - Data Manifesto, *Data Sci. J.* **6**, 116 (2007). https://doi.org/10.48550/arXiv.astro-ph/0701361
4. J. Evans, M. Molinaro, P. Dowler, C. Arviset, S. Gaudet, and IVOA Technical Coordination Group, IVOA Architecture Version 2.1 (IVOA Endorsed Note, 2024).
5. M.D. Wilkinson et al., The FAIR Guiding Principles for scientific data management and stewardship, *Sci. Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
6. S. O'Toole and J. Tocknell, FAIR standards for astronomical data, In Astronomical Data Analysis Software and Systems XXXI, *ASP Conf. Ser.* **535**, 265 (2024).
7. F. Economou, S. Gaudet, T. Jenness, R. O. Redman, S. Goliath, P. Dowler, M. J. Currie, G. S. Bell, S. F. Graves, J. Ouellette, D. Johnstone, D. Schade, and A. Chrysostomou, Observatory/data centre partnerships and the VO-centric archive: The JCMT Science Archive experience, *Astron. Comput.* **11**, 161 (2015). https://doi.org/10.1016/j.ascom.2014.12.005
8. F. Bonnarel, M. Louys, G. Mantelet, M. Nullmeier, M. Servillat, K. Riebe, and M. Sanguillon, ProvTAP: A TAP Service for Providing IVOA Provenance Metadata, In Astronomical Data Analysis Software and Systems XXVII, *ASP Conf. Ser.* **523**, 313 (2019).
9. M. Servillat, C. Boisson, J. Lefaucheur, K. Kosack, M. Sanguillon, M. Louys, and F. Bonnarel, Provenance as a Requirement for Large-scale Complex Astronomical Instruments, In Astronomical Data Analysis Software and Systems XXVII, *ASP Conf. Ser.* **522**, 199 (2020). https://doi.org/10.48550/arXiv.1806.00447
10. A. Zanichelli, C. Knapic, S. Righini, M. Nanni, M. Stagni, F. Bedosti, M. Bartolini, M. Sponza, A. Orlati, and R. Smareglia, A Public Data Archive for the Italian Radio Telescopes, In Astronomical Data Analysis Software and Systems XXVII, *ASP Conf. Ser.* **522**, 463 (2020).
11. M. Romaniello, M. Arnaboldi, M. Barbieri, N. Delmotte, A. Dobrzycki, N. Fourniol, W. Freudling, J. Grave, L. Mascetti, A. Micol, J. Retzlaff, N. Rosse, T. Tax, M. Vuong, O. Hainaut, M. Rejkuba, and M. Sterzik, The ESO Science Archive Facility: Status, Impact, and Prospects, *Messenger* **191**, 29 (2023). https://doi.org/10.18727/0722-6691/5338
12. ESO Data Access Policy, European Southern Observatory (2023), https://archive.eso.org/cms/eso-data-access-policy.html (accessed 20 July 2026).
13. Hubble Space Telescope Science Operations and Data Rights, NASA (2024), https://science.nasa.gov/mission/hubble/observatory/science-operations/ (accessed 20 July 2026).
14. Chandra Data Archive Operations and Public Access Policy, Chandra X-ray Center (2025), https://cxc.harvard.edu/cda/public.html (accessed 20 July 2026).
15. NRAO Data Delivery and Proprietary Periods, National Radio Astronomy Observatory (2023), https://science.nrao.edu/observing/policies/data-rights (accessed 20 July 2026).
16. Keck Observatory Archive Proprietary Policy, W. M. Keck Observatory (2024), https://koa.ipac.caltech.edu/UserGuide/proprietary_policy.html (accessed 20 July 2026).

17. M. Demleitner, M. C. Neves, F. Rothmaier, and J. Wambsganss, Virtual observatory publishing with DaCHS, *Astron. Comput.* **7**, 27 (2014). https://doi.org/10.1016/j.ascom.2014.08.003

18. C. Li, C. Cui, B. He, D. Fan, J. Wang, S. Li, L. Mi, W. Wan, J. Chen, H. Zhang, C. Yu, J. Xiao, C. Wang, Z. Cao, Y. Fan, Z. Hong, J. Wang, S. Yin, L. Liu, and X. Chen, The Design and Application of Astronomy Data Lake in China-VO, In Astronomical Data Analysis Software and Systems XXV, *ASP Conf. Ser.* **512**, 157 (2017).

19. I.M. Izmailova, A.Zh. Umirbayeva, M.K. Khassanov, L. Aktay, S.A. Shomshekova, ArxSP: A Python-Based Modular Application for the Reduction of Digitized Archival Spectra, *Astronomy and Computing* **55**, (2026). https://doi.org/10.1016/j.ascom.2025.101050

20. Data Sorting and Reduction Pipeline, GitHub Repository (2026), https://github.com/fai-kz/obsdata_sort_analyse (accessed 20 July 2026).

21. FFDB API Server for FAI FITS Database Infrastructure. GitHub repository (2026). Available at: https://github.com/fai-kz/ffdb-api-server (Accessed: 20 July 2026)

22. KazVO Observation Data Use Policy and License Agreement, Fesenkov Astrophysical Institute (2026), https://vo.fai.kz/term_of_use_obs.php (accessed 20 July 2026).

23. The Astropy Collaboration, T.P. Robitaille et al., *Astron. Astrophys.* **558**, A33 (2013). https://doi.org/10.1051/0004-6361/201322068

24. Astropy Collaboration, A. M. Price-Whelan, B. M. Sipőcz, H. M. Günther, P. L. Lim, S. M. Crawford, S. Conseil, D. L. Shupe, M. W. Craig, N. Dencheva et al., *Astron. J.* **156**, 123 (2018). https://doi.org/10.3847/1538-3881/aabc4f

25. The Astropy Collaboration, A.M. Price-Whelan et al., *Astrophys. J.* **935**, 167 (2022). https://doi.org/10.3847/1538-4357/ac7c74

26. M. Craig, S. Crawford, M. Seifert, T. Robitaille, B. Sipőcz, J. Walawender, Z. Vinícius, J. P. Ninan, M. Droettboom, J. Youn et al., astropy/ccdproc: v1.3.0.post1, Zenodo (2017), https://doi.org/10.5281/zenodo.1069648

27. KazVO SSA Service: TCO Echelle Spectra of Hot Supergiants, (Fesenkov Astrophysical Institute, 2026), https://vo.fai.kz/obs_data.php?path=/tco_hot_supergiants/q/web (accessed 20 July 2026)

Information about authors:

Ildana Izmailova (corresponding author) – MSc, Researcher, Fesenkov Astrophysical Institute, (Almaty, Kazakhstan, e-mail: izmailova@fai.kz).

Nurken Akbar – Bachelor, Engineer, Fesenkov Astrophysical Institute, (Almaty, Kazakhstan, e-mail: akbar@fai.kz).

Maxim Makukov – Senior Researcher, Fesenkov Astrophysical Institute, (Almaty, Kazakhstan, e-mail: makukov@fai.kz).

Denis Yurin – PhD, Lead Researcher, Fesenkov Astrophysical Institute, (Almaty, Kazakhstan, e-mail: yurin@fai.kz).

Авторлар туралы мәлімет:

Ильдана Измайлова (автор-корреспондент) – магистр, ғылыми қызметкер, В.Г. Фесенков атындағы астрофизика институты, (Алматы, Қазақстан, e-mail: izmailova@fai.kz).

Нұркен Ақбар – бакалавр, инженер, В.Г. Фесенков атындағы астрофизика институты, (Алматы, Қазақстан, e-mail: akbar@fai.kz).

Максим Макуков – аға ғылыми қызметкер, В.Г. Фесенков атындағы астрофизика институты, (Алматы, Қазақстан, e-mail: makukov@fai.kz).

Денис Юрин – PhD, жетекші ғылыми қызметкер, В.Г. Фесенков атындағы астрофизика институты, (Алматы, Қазақстан, e-mail: yurin@fai.kz).

Информация об авторах:

Ильдана Измайлова (автор-корреспондент) – магистр, научный сотрудник, Астрофизический институт им. В.Г. Фесенкова, (Алматы, Казахстан, e-mail: izmailova@fai.kz).

Нуркен Акбар – бакалавр, инженер, Астрофизический Институт им. В. Г. Фесенкова, (Алматы, Казахстан, akbar@fai.kz).

Максим Макуков – старший научный сотрудник, Астрофизический институт им. В.Г. Фесенкова, (Алматы, Казахстан, e-mail: makukov@fai.kz).

Денис Юрин – PhD, ведущий научный сотрудник, Астрофизический институт им. В.Г. Фесенкова, (Алматы, Казахстан, e-mail: yurin@fai.kz).

Article history: *received: 04 May 2026; revised: 01 August; accepted: 03 October 2026.*

Мақала тарихы: *түсті: 14 маусым 2026; түзетілді:01 тамыз; қабылданды: 03 қазан 2026.*

История статьи: *поступила: 04 мая 2026; после доработки: 01 августа; принята: 03 октября 2026.*