Postprint of Core Functional Requirements Survey and Analysis for China Virtual Observatory

Authors: Xu Yunfei, Fan Dongwei, Cui Chenzhou, Boliang He, Li Changhua, Yu Ce, Xiao Jian, Li Shanshan, Mi Linying, Han Jun, Tao Yihan, Wang Chuanjun, Zhang Hailong, Liu Liang, Li Zheng, Han Xu, Yang Sisi, Yang Hanxi, Netherlands, Zhang Lei, Cui Shun, Wang Chuanzhong, Su Liying, Chen Li, Qiao Cuilan, Xu Qian, Li Xiangru, Yang Haifeng, Cao Zihuang, Wang Junfeng, Qiang Zhenping, Bai Chunhai, Weihong Zhou, Yuan Guowu, Bai Zhengyao, Li Ji, Zheng Zipeng, He Shousheng, Qu Caixia, Sun Jixian, Lu Dengrong, Zhao Yongheng

Date: 2019-07-30T00:00:00+00:00

Abstract

After more than ten years of continuous construction and development, the China Virtual Observatory (China-VO) has become an important technical and resource platform supporting astronomical observation, research, and education. With the advent of the era of multi-messenger astronomy and time-domain astronomy, the virtual observatory also needs to upgrade its core capabilities to provide more precise services and technical support for astronomers. To this end, the China-VO team compiled a list of core technical requirements based on the development direction of astronomy and information technology trends, and conducted a survey among experts and users in the field in the form of a questionnaire. Through statistical analysis of the survey results, China-VO has clarified its main direction of effort and objectives for the coming period, planning to adopt a platform-based development model and open third-party development interfaces to attract more interested developers to create practical tools based on virtual observatory resources, thereby better achieving the rapid transformation of resources and technology into services.

Full Text

Investigation and Analysis of Core Functional Requirements of China-VO

Yunfei Xu[1,2], Dongwei Fan[1], Chenzhou Cui[1], Boliang He[1], Changhua Li[1], Ce Yu[3], Jian Xiao , Shanshan Li[1], Linying Mi[1], Jun Han[1], Yi-

han Tao[1], Chuanjun Wang , ,[2], Hailong Zhang , Liang Liu , Zheng Li[1], Xu Han[1], Sisi Yang[1], Hanxi Yang[1], Lan He[1], Lei Zhang[1],[2], Shun Cui ,[1], Chuanzhong Wang ,[1], Liyin Su , Li Chen[1] , Cuilan Qiao[11], Qian Xu , Xiangru Li[12], Haifeng Yang[13], Zihuang Cao[1],[2], Junfeng Wang[1] , Zhenping Qiang[1] , Chunhai Bai , Weihong Zhou[1] , Guowu Yuan[1] , Zhengyao Bai[1] , Ji Li[1] , Zipeng Zheng[1] , Shousheng He , Caixia Qu[13], Jixian Sun[2] , Dengrong Lu[2] , Yongheng Zhao[1]

[1]National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]School of Software, Tianjin University, Tianjin 300350, China
School of Computer Science and Technology, Tianjin University, Tianjin 300350, China
Yunnan Astronomical Observatory, Chinese Academy of Sciences, Kunming, Yunnan 650216, China
Key Laboratory of Astronomical Structure and Evolution, Chinese Academy of Sciences, Kunming, Yunnan 650216, China
Xinjiang Astronomical Observatory, Chinese Academy of Sciences, Urumqi, Xinjiang 830011, China
Purple Mountain Observatory, Chinese Academy of Sciences, Nanjing, Jiangsu 210008, China
School of Mechanical Engineering and Applied Electronics, Beijing University of Technology, Beijing 100124, China
[1] Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 200030, China
[11]Central China Normal University, Wuhan, Hubei 430079, China
[12]South China Normal University, Guangzhou, Guangdong 510631, China
[13]Hebei Normal University, Shijiazhuang, Hebei 050024, China
[1] Shanxi Normal University, Linfen, Shanxi 041004, China
[1] Yunnan University of Finance and Economics, Kunming, Yunnan 650221, China
[1] Shanghai Jiao Tong University, Shanghai 200240, China
[1] Kunming University of Science and Technology, Kunming, Yunnan 650500, China
[1] Hebei University of Technology, Tianjin 300401, China
[1] Qinghai University, Xining, Qinghai 810016, China
[2] Qinghai Observatory, Purple Mountain Observatory, Chinese Academy of Sciences, Delingha, Qinghai 817000, China

Abstract: After more than ten years of construction and development, the China Virtual Observatory (China-VO) has become an important technology and resource platform supporting astronomical observation, research, and education. With the advent of multi-messenger astronomy and time-domain astronomy, the virtual observatory must also upgrade its core capabilities to provide astronomers with more precise services and technical support. To this end, the China-VO team compiled a list of core technology requirements based on the development direction of astronomy and information technology trends, and conducted a survey among experts and users in the field via questionnaire. Through statistical analysis of the survey results, China-VO has clarified its main development directions and objectives for the coming period, planning to adopt a platform-based development model and open third-party development interfaces to attract more interested developers to create practical tools based on VO resources, thereby better achieving the rapid transformation of resources and technology into services.

Keywords: Virtual Observatory; Functional Requirements; Multi-messenger Astronomy; Time-domain Astronomy

0 Current Status of Virtual Observatory

As astronomy continues to develop and scientific collaboration becomes increasingly extensive, traditional research models must evolve. Around 2000, astronomers proposed a cross-disciplinary concept spanning astronomy, computer science, and information science—the Virtual Observatory (VO)—to standardize all processes of astronomical data access. The VO is an environment that seamlessly and transparently connects astronomical research resources worldwide through advanced information technology for data-intensive networked astronomical research and science education [1]. By leveraging the most advanced computer and information technologies, the VO integrates various astronomical research resources—including observational data, astronomical literature, computing resources, storage resources, software services, and even observational facilities—into a physically distributed yet logically unified system through a unified service model. After nearly two decades of development, the VO has unified access standards across various data service systems and has become an important information technology support platform for astronomical research.

The concept of China-VO was proposed against this backdrop. The key research and development areas of China-VO include the development of the China-VO system platform, unified access to domestic and international astronomical research resources, VO-based astronomical research demonstrations, and astronomical science education [2]. The goal of China-VO is to serve as a bridge and link between astronomy and information technology, promoting the application of advanced information technologies to astronomical research.

In 2013, China-VO began developing its public service platform based on the CloudStack cloud computing middleware, deploying cloud nodes in multiple locations across China. Starting in the second half of 2016, leveraging the strategic partnership between the National Astronomical Observatories and Alibaba Cloud, China-VO comprehensively explored a hybrid architecture solution combining "public cloud + private cloud." The China-VO cloud service platform fully utilizes existing network, storage, and computing infrastructure at various observatories and within the Chinese Academy of Sciences. By employing advanced key technologies from the information and VO fields, and focusing on the pipeline of astronomical observation data submission, sharing, and usage, as well as time application and approval for domestic core astronomical observational facilities, the platform integrates resources required for astronomical observation and research activities, such as computing, storage, and software, to form a physically distributed yet logically unified networked scientific research platform.

Currently, the China-VO public service platform has developed seven nodes, including the National Astronomical Observatories (Beijing), Purple Mountain Observatory (Nanjing), Shanghai Astronomical Observatory (Shanghai), Yunnan Astronomical Observatory (Kunming), Xinjiang Astronomical Observatory (Urumqi), Nanjing University (Nanjing), and Alibaba Cloud (Hangzhou). It provides data management and open sharing services for over ten observational facilities, including the Xinglong LAMOST, the 13.7-meter millimeter-wave telescope, the Sheshan 1.56-meter telescope, the Lijiang 2.4-meter telescope, and the Nanshan 25-meter telescope. It also provides time application services for telescopes such as the Lijiang 2.4-meter and Xinglong 2.16-meter telescopes, offers cloud-based teaching environments for the University of Chinese Academy of Sciences and Nanjing University, provides virtual machine services for over a thousand scientific researchers, has discovered 19 supernovae and novae based on domestic amateur observational data, and has more than 21,000 registered users. In early 2015, China-VO was selected as one of the "Top Ten Excellent Cases of Informatization for Scientific Research of the Chinese Academy of Sciences." In the era of "Internet+" and big data, China-VO is playing an increasingly important role in providing information technology support for the development of astronomy.

1 New Requirements and Challenges

With the discovery of the binary neutron star merger gravitational wave event GW170817 and its electromagnetic counterpart [3], as well as the identification and characterization of the BL Lac blazar associated with the high-energy neutrino IceCube-170922A [4], astronomy has entered a new era of multi-messenger scientific discovery. At the same time, time-domain survey projects represented by LSST and SKA have ushered in a new era of exploration and discovery in time-domain astronomy, while also bringing enormous challenges. The development of multi-messenger astronomy and time-domain astronomy strongly relies on information technology, requiring a powerful network system architecture that includes data processing pipelines, archiving, automatic event detection, classification and release, and efficient data fusion. Astronomy is transitioning from a discovery-driven science to a data-driven, science-guided science—namely, data-intensive science. The goal of the VO is precisely to transform astronomy from the old data-poor research paradigm to a new data-rich research paradigm, fully leveraging information technology to address the challenges of the data-intensive multi-messenger astronomical discovery era.

The challenges brought by this new era impose higher demands on the technical support capabilities of the VO. Although China-VO has accumulated considerable resources and technologies, there remains a significant gap between its current capabilities and the research capabilities and technical requirements needed for time-domain astronomy and multi-messenger astronomy. The PB-level data streams generated by new-generation observational equipment present new challenges for data processing and mining, with a key aspect being real-time processing and analysis of big data streams [5][6]. Since many multi-messenger and time-domain observational events are transient, event detection and timely follow-up observations are crucial. This requires data processing to begin simultaneously with data acquisition from telescopes, automatically and reliably detecting and analyzing transient events, and arranging timely follow-up observations based on priority. Additionally, there is a need for programmatic telescopes to replace manual scheduling and execution of time-domain observations, and to automatically complete the discovery and classification of time-domain events. These requirements can be summarized as "efficient fusion of multi-band, multi-messenger massive data." Data fusion and interoperability are core scientific objectives and technical capabilities of the VO. Previous research on data fusion has primarily focused on multi-band aspects, while research on multi-messenger data fusion is just beginning. Moreover, traditional data fusion and interoperability research has mostly been conducted to meet the needs of conventional research projects rather than the requirements of time-domain astronomy scientific programs or even scientific engineering projects, without serious consideration of performance issues such as completion deadlines. Therefore, China-VO currently does not fully possess the conditions for "efficient fusion of multi-band, multi-messenger massive data" and urgently needs to enhance its core capabilities.

2 Core Capabilities Survey

Based on this understanding, the China-VO team compiled a list of key technical requirements related to time-domain astronomy and multi-messenger astronomy, and solicited feedback from experts on the China-VO Academic Committee and a broad user base through a survey questionnaire. The questionnaire aimed to determine the main development directions and objectives for the coming period. The survey covered various aspects of services provided by the China-VO team, including file storage services, data retrieval services, data visualization, cross-matching services, online computing services, remote observatory automation control, automated data classification methods, and platform expansion, totaling eight sections. Each section contained several specific key technical points, amounting to 23 specific requirements. The questionnaire described the application scope and technical characteristics of each key technology, as detailed in Table 1 .

We conducted a survey of cutting-edge technologies for each of these eight sections. Achieving efficient file storage services and data retrieval services has always been a research hotspot in the VO field. In addition to high-speed local data reading and writing, with the continuous construction of multi-node data centers, remote data retrieval and tiered storage have also become current research priorities. Astronomical data storage formats are relatively diverse, with mainstream formats including FITS, CASA, and HDF5. Large amounts of fragmented data result in extremely high I/O consumption, making it urgent to implement parallel reading and writing strategies and distributed storage strategies for massive data, and to provide seamless data access services. Currently, popular distributed parallel storage systems in the industry include Lustre, Haystack, HDFS, and TFS. Their design concepts are basically similar, primarily comprising an underlying file storage system and a directory service that enables efficient file storage and retrieval through directory indexing. In addition to distributed file storage systems, efficient data storage and retrieval also rely on efficient indexing techniques. The right ascension and declination of celestial objects are two-dimensional vectors, and indexing celestial coordinates can be achieved by mapping two-dimensional vectors to one-dimensional data. A feasible approach is to divide the celestial sphere into multiple sub-regions and assign unique identifiers to these sub-regions to achieve spatial indexing. Commonly used methods include the Hierarchical Triangular Mesh (HTM), HEALPix, and Zones algorithm, which have been employed in projects such as SDSS, MUSER, and LSST.

Data visualization includes image layered visualization and catalog layered visualization, primarily used for displaying all-sky maps and all-sky catalog distributions. The emphasis here is on layered visualization, which refers to progressive visualization of massive data based on conditions such as resolution and magnitude. Commonly used methods include Level of Detail (LOD) techniques, with main implementations including quadtree and octree algorithms. The Skyview system from the U.S. MAST and the Aladin system from the French CDS have both achieved all-sky image mosaicking under big data conditions.

The core method of cross-matching is calculating the distance between two targets; if the distance is less than a threshold, the two targets are considered the same celestial object. Research on cross-matching has achieved a series of results both domestically and internationally, such as the nearest neighbor algorithm based on HTM index partitioning and kd-tree [7], the distributed cross-matching method based on MapReduce [8], and the GPU cluster algorithm based on HEALPix and the Zones Algorithm [9]. However, many key technologies still need to be breakthrough in multi-band, multi-messenger data cross-matching and confidence estimation.

Automated data classification methods primarily target massive data analysis in the time-domain astronomy field. With the continuous deployment of new-generation observational equipment, massive amounts of transient source observational data in different bands will be obtained, and existing data classification methods can no longer meet the processing requirements of the big data era. Leveraging artificial intelligence methods, especially deep learning methods, will be a feasible path. For image data classification, deep learning methods have developed a series of networks that have been tested in practical applications, such as ResNet, AlexNet, and VGG in convolutional neural networks, all of which can be applied to the classification of image data and even spectral data in different bands.

The main objective of online computing services is to facilitate convenient and efficient data processing for researchers, eliminating the need to download large amounts of data from online storage to local machines or deploy complex data processing environments on local computers. Instead, data preprocessing, computation, and analysis can be performed directly through an online platform. CasJobs is such an online work platform for large scientific databases, providing powerful database retrieval and operation functions in the form of a web application. Initially developed for online processing of SDSS data, it has now been integrated by the China-VO team into the processing system for LAMOST survey catalog data. Jupyter Hub has also been widely used for online processing of astronomical data. Its usage is similar to Jupyter notebook, providing an online platform for writing and running Python programs. It can support multi-user login, with a powerful computing cluster in the backend and strong scalability, making it an excellent platform for implementing online computing services.

Remote observatories have developed rapidly in recent years due to their unique advantages in continuous observation and rapid response. A series of time-series survey programs based on remote observatories have been carried out both domestically and internationally, such as ROBONET, ROTSE-III, and BOOTES, which have been well applied in optical observation and research of gamma-ray burst afterglows and other transient sources. Remote observatories consist of telescopes, equatorial mounts, domes, computers, weather stations, all-sky cameras, and other equipment, with multiple automated software and hardware subsystems that achieve programmatic observation functions through coordinated control. After years of investigation and research, China-VO has proposed and designed an embedded hardware integration system that serves as a bridge between users and observatory equipment to facilitate integrated control and extensible transplantation of various subsystems [10]. Future work will further optimize this system, focusing on intelligent observation applications.

Platform expansion and construction primarily cover the improvement of China-VO' s own service capabilities, including more flexible user permissions, integration of practical research tools, convenient third-party application development interfaces, and platform implementation based on private or public clouds. These infrastructure-based functional expansions can provide users with better service experiences and further enhance user engagement.

3 Survey Results Analysis

The survey was conducted through both email and online questionnaires, receiving 12 responses from China-VO Academic Committee experts and 65 responses from China-VO mailing list and web users. The user group primarily included astronomy researchers and graduate students, all of whom had experience using the China-VO public service platform.

3.1 Statistical Method

When compiling feedback statistics, expert responses were weighted three times higher, equivalent to 101 questionnaires in total. Based on the questionnaire score multiplied by the proportion of the score count to the total, scores were calculated for the eight major categories and 23 specific requirements in the questionnaire.

3.2 Survey Results

Figure 1 [Figure 1: see original paper] shows the specific feedback results for the eight major categories, sorted by score (scored 0-5, with higher scores indicating greater importance). The highest-scoring item was file storage service (3.89 points), followed by data retrieval service (3.87 points), data visualization (3.68 points), cross-matching service (3.52 points), online computing service (3.39 points), remote observatory automation control (3.36 points), automated data classification methods (3.21 points), and platform expansion (3.11 points).

Figure 2 [Figure 2: see original paper] shows the scores for the 23 specific requirements. Additionally, responses from China-VO Academic Committee experts were analyzed separately. Expert scores were not significantly different from the overall scores: file storage service (4.11 points), data retrieval service (3.96 points), data visualization (3.92 points), automated data classification methods (3.59 points), cross-matching service (3.47 points), online computing service (3.44 points), platform expansion (3.03 points), and remote observatory automation control (2.86 points). The difference between expert scoring and overall scoring is shown in Figure 3 [Figure 3: see original paper]. Experts gave slightly higher scores across all categories, indicating strong recognition of the VO team' s requirement compilation. However, expert scores for automated data classification methods and remote observatory automation control differed significantly from the overall scores, suggesting that experts and users hold different opinions on these two types of requirements.

3.3 Analysis and Reflection

The feedback results indicate that data accessibility and ease of data operation remain core requirements for the VO. Astronomers' primary concern has always been how to analyze and process data more conveniently. Currently, the Chinese Astronomical Data Center maintained by China-VO provides archiving and query services for 21 datasets, allowing users to perform online data retrieval and download directly. The public service portal provides virtual computing and storage resource services, enabling users to customize required computing and storage resources through the cloud platform' s online console. Additionally, based on the strategic partnership between the National Astronomical Observatories and Alibaba Cloud, users can also utilize Alibaba Cloud' s computing resources. However, the VO has not yet achieved integrated online services for data retrieval and processing. Improving service usability and convenience will be a key focus of future work.

Data visualization, cross-matching services, and online computing services are extensions of data processing requirements and breakthrough points for providing convenient services. The China-VO team has already conducted some preliminary work, including the "WorldWide Telescope" astronomical data visualization platform developed for astronomical science education, and the release of commonly used data processing environments in the form of cloud virtual machine templates, such as the virtual machine processing template developed for the multi-band data processing course at the University of Chinese Academy of Sciences, which includes scientific data processing program libraries such as CFITSIO, Graphviz, and GSL. Users can conveniently use these by generating virtual machines based on the template. In the future, while improving service usability and convenience, China-VO will also strengthen the promotion of relevant resources and services, enabling users to better understand existing China-VO functions and thereby providing higher-quality services.

Achieving integrated online services for data storage, retrieval, and processing has always been the goal of China-VO. However, significant difficulties remain in meeting the VO' s core requirements, primarily the contradiction between the high difficulty and large scale of requirements and the insufficient number of VO R&D personnel. Converting China-VO' s accumulated data, storage, and computing resources into efficient and convenient services requires a large number of R&D personnel, but under the current system, further increasing researcher positions is unrealistic. Therefore, China-VO plans to adopt a platform-based development model by opening VO third-party development interfaces to attract more interested developers to create practical tools based on VO resources, achieving rapid conversion of resources and technology into services.

Figure 4 [Figure 4: see original paper] shows the architecture design for the next phase of the China-VO platform. The bottom layer is the resource layer, which includes not only hardware resources such as storage servers and computing servers but also various datasets archived by China-VO. Above the resource layer is the key technology layer, which connects downward to invoke hardware resources in the resource layer through virtualization technology and data resources through data retrieval algorithms and database tools. It also encapsulates various key technologies involved in the VO into libraries to enable on-demand computing based on hardware and data resources. These encapsulated key technologies include data retrieval methods, cross-matching methods between different catalogs, various astronomical data visualization methods, and machine learning-based data classification methods, which can be gradually expanded as R&D progresses.

Above the key technology layer is the interface layer, which will be the focus of

China-VO' s next-phase construction. By encapsulating VO hardware and software resources and key technologies into Application Programming Interfaces (APIs) and releasing them to the broad astronomical community, along with comprehensive documentation and training support, astronomers will be able to build new applications according to their research needs. Based on the survey results, APIs related to file storage, data retrieval, data visualization, and cross-matching will be prioritized for release, with corresponding documentation and sample application code published on the China-VO website.

Applications built by researchers through the China-VO API can be registered on the China-VO platform for use by more users. These applications will constitute the application and service layer of the China-VO platform. Based on unified application development standards and frameworks, these applications can run cross-platform in mobile and desktop browsers, and under unified user account management, users can seamlessly access different platform applications and data.

4 Summary and Outlook

In summary, through this core functional requirements survey, China-VO has clarified its construction priorities and objectives for the next phase. The China-VO team will focus on improving data accessibility and ease of data operation, and strive to enhance online services such as data visualization, cross-matching, and online computing. However, the contradiction between high demand for the VO and insufficient R&D personnel is becoming increasingly prominent. To address this, China-VO will adopt a platform-based development model by opening VO interfaces to attract developers to create more practical services and tools based on VO resources and key technologies, and share them through the China-VO platform. Through this approach, China-VO will continuously enhance its core capabilities, forming an application ecosystem oriented toward research needs, thereby better addressing the challenges brought by the era of time-domain astronomy and multi-messenger astronomy.

This paper acknowledges the data resources and technical support provided by the National Astronomical Data Center. We thank all experts and users who participated in this survey and analysis discussions, and we thank the National Astronomical Observatories-Alibaba Cloud Joint Research Center for Astronomical Big Data for its support of this work.

References

[1] Zhao, Y., & Cui, C. (2011). From Virtual Observatory to Astronomical Informatics. *e-Science Technology & Application*, 2(3), 3-12.

[2] Cui, C., Xue, Y., Li, J., et al. (2013). Virtual Observatory—An e-Science Environment for Astronomical Research. *Bulletin of Chinese Academy of Sciences*, 28(4), 511-518.

[3] Abbott, B. P., et al. (2017). Multi-messenger Observations of a Binary Neutron Star Merger. *Astrophysical Journal Letters*, 848(2), L12.

[4] IceCube, Fermi-LAT, et al. (2018). Multimessenger Observations of a Flaring Blazar Coincident with High-energy Neutrino IceCube-170922A. *Science*, 361.

[5] Graham, M. J., et al. (2012). Data Challenges of Time Domain Astronomy. *Distributed and Parallel Databases*, 30(5-6), 371-384.

[6] Bloom, J. S. (2011). Technical and Observational Challenges for Future Time-Domain Surveys. *Proceedings of the International Astronomical Union*, 7(S285), 165-170.

[7] Gao, D., Zhang, Y., & Zhao, Y. (2005). Implementation of Cross-Identification for Massive Multi-band Catalog Data. *Astronomical Research & Technology*, 3.

[8] Jia, X., Luo, Q., & Fan, D. (2015). Cross-Matching Large Astronomical Catalogs on Heterogeneous Clusters. In *2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS)* (pp. 617-624). Melbourne, VIC.

[9] Fan, D., Budavari, T., Szalay, A. S., Cui, C., & Zhao, Y. (2013). Efficient Catalog Matching with Dropout Detection. *Publications of the Astronomical Society of the Pacific*, 125, 218.

[10] Han, J., Wang, C., Fan, D., Cui, C., Li, S., Mi, L., Li, Z., Xu, Y., He, B., Li, C., & Yang, S. (2018). Amateur Public Observatory I: The Observatory and Hardware Integration System. *Astronomy and Computing*, 25, 89-93.

[11] Cui, C., Yu, C., & Xiao, J. (2015). Astronomical Research in the Era of Big Data. *Chinese Science Bulletin*, 5, 445-449.

[12] Li, C., Cui, C., & Zhao, Y. (2012). Concept, Design and Application of Virtual Observatory. *China Education Network*, 6.

Note: Figure translations are in progress. See original paper for figures.