

## Design and Implementation of an Expert Music Score Recognition Platform: Postprint

**Authors:** Lei Guohong, Xu Yang, Chenhui Niu, Tian Haijun, Zhang Yanxia, Cui Chenzhou, Zhao Yongheng

**Date:** 2018-05-29T00:00:00+00:00

### Abstract

When utilizing LAMOST survey data for searching special celestial objects or constructing samples, and when processing the continuously accumulating low signal-to-noise ratio spectra from LAMOST, scientists typically need to spend a significant amount of time completing manual spectral identification. To address such issues, we have designed and implemented an expert spectral identification platform. This platform is an integrated scientific and educational research platform that combines functions such as spectral visualization, spectral analysis, multi-band image fusion, and various data mining algorithms. Utilizing this platform will make it relatively easier for astronomers to conduct scientific research such as celestial object searches and sample construction; through this platform, university teachers can carry out various astronomy-themed teaching and research activities; with the power of the community, the platform will be able to gradually digest the continuously accumulating unknown spectral data labeled as “UNKNOWN” from LAMOST.

### Full Text

## Design and Implementation of an Expert Platform for Spectral Inspection

**Guohong Lei<sup>1</sup>, Yang Xu<sup>2</sup>, Chenhui Niu<sup>2,3</sup>, Haijun Tian<sup>1</sup>, Yanxia Zhang<sup>2</sup>, Chenzhou Cui<sup>2</sup>, Yongheng Zhao<sup>2</sup>**

<sup>1</sup> China Three Gorges University, Yichang 443002, China

<sup>2</sup> National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China

<sup>3</sup> Central China Normal University, Wuhan 430079, China

**Abstract:** When searching for special celestial objects or constructing samples using LAMOST survey data, and when processing the continuously accumulat-

ing low signal-to-noise ratio (SNR) spectra from LAMOST, scientists typically need to invest substantial time in manual spectral inspection. To address this challenge, we have designed and implemented an expert spectral inspection platform. This integrated scientific and educational platform combines spectral visualization, spectral analysis, multi-band image fusion, and various data mining algorithms. Using this platform, astronomers can more easily conduct scientific research such as object searches and sample construction; university teachers can carry out various astronomy-focused teaching and research activities; and with the power of crowdsourcing, the platform can gradually process the growing collection of LAMOST spectra labeled as “UNKNOWN.”

**Keywords:** LAMOST; manual spectral inspection; data mining; astronomical teaching

## 1. Introduction

Since its formal commissioning in 2012, China’s Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST) has obtained over 7 million spectra, providing a rich data foundation for scientific research. Scientists have published nearly a hundred high-quality academic papers using LAMOST data, contributing significantly to advancing Chinese astronomy to the forefront of international research.

From a scientific output perspective, LAMOST’s large sample advantage has attracted numerous scientists, with many achievements focusing on searches for special objects or construction of special samples. Examples include: searches for metal-poor star samples by Li Haining et al. [1]; quasar sample construction by Wu Xuebing et al. [2]; searches for quasars near M31/33 by Huo Zhiying et al. [3]; searches for white dwarf-main sequence binaries by Ren Juanjuan et al. [4]; identification of DA white dwarf samples and related research by Zhao Jingkun et al. [5], Zhang Yueyang et al., and Rebassa-Mansergas et al.; Am star sample construction by Hou Wen et al. [6]; carbon star candidate searches by Si Jianmin et al. and Ji Wei et al. [7]; RR Lyrae variable star and Balmer line distribution statistics by Yang Fan et al.; searches for double-peaked narrow emission line galaxies and AGN samples by Shi Zhixin et al.; E+A galaxy searches by Yang Haifeng et al.; Be star searches by Lin Jianzheng et al.; and galaxy pair sample identification by Shen Shiyin et al. [8]. Due to LAMOST’s enormous sample size, these research projects typically require first establishing complex selection criteria and using various data mining algorithms or statistical methods to identify target candidates, followed by candidate verification through manual spectral inspection combined with multi-band photometric information. Machine-selected candidates often number in the thousands, tens of thousands, or even hundreds of thousands, creating enormous workloads for manual verification.

Regarding data quality distribution, among LAMOST’s released 7+ million spectra (using DR4 as an example), as shown in [Figure 1: see original paper],

while 6 million brighter stars can be automatically processed by the pipeline, 118,000 galaxies and over 40,000 quasars suffer from low SNR, making automatic pipeline processing ineffective, with a substantial portion requiring purely manual, one-by-one spectral identification. Of particular concern is that five years of survey observations have accumulated over 620,000 spectra labeled “UNKNOWN” awaiting classification. Most of these objects are too faint, yielding spectra with SNR too low for traditional algorithms to process; a few may be extremely unusual objects with spectra too peculiar for traditional methods to recognize. For instance, Mao Xiaoyan et al. proposed a weighted filtering algorithm for denoising low-SNR LAMOST fiber spectra [9]. As survey observations continue, these unknown spectra will only accumulate, presenting a challenge for modern statistics and data mining techniques.

In response to LAMOST spectral processing, data mining expert teams have joined spectral processing research under the leadership of the LAMOST project team, achieving significant results. However, even when data mining experts can provide classifications or parameter estimates for these low-SNR spectra, astronomers remain skeptical and often abandon using such data. Only after manual inspection can the credibility of these spectra be enhanced. Yet processing such massive volumes of data through manual inspection by a few experts with limited time and energy is impossible.

To improve this situation, we propose building a data mining-based expert spectral inspection platform. Using advanced IT and data mining technologies to integrate common algorithms, we have developed a scientific and educational expert spectral inspection system that combines spectral data management, visualization, and analysis.

## 2. Goal Analysis and Structural Design

Based on the vision of astronomers and astronomy educators, the expert spectral inspection platform should satisfy four main requirements: (1) Integrate various visualization and analysis tools and data mining algorithms to make it easier for astronomers and even ordinary students to search for special objects and make new discoveries; (2) Conduct data mining algorithm research on low-SNR spectra to extract maximum valuable information through machine assistance, reducing manual inspection difficulty; (3) Support collaborative work modes to assist astronomy educators in conducting special object searches and manual spectral inspection teaching activities, enabling astronomy-focused research-oriented quality education; and (4) Like the international Galaxy Zoo project [Raddick J, Lintott C J, Schawinski K, et al. Galaxy Zoo: an experiment in public science participation. *Advances in Atmospheric Sciences*, 2007, 39:892], mobilize public participation to complete astronomical spectral data processing requiring substantial manual intervention, gradually digesting the continuously produced unknown spectra from LAMOST.

From these requirements, we designed the expert spectral inspection platform.

As shown in [Figure 2: see original paper], the platform comprises three main components: the Data-mining Layer, Data Node Layer, and Expert Layer. (1) The Data-mining Layer conducts research on various data mining algorithms for LAMOST low-SNR spectra to maximize their scientific value, focusing on classification of low-SNR spectra, feature extraction of stellar spectra (providing reference values for some stellar atmospheric physical parameters), and redshift measurement for low-SNR galaxies and quasars. Data mining results are uploaded to the Data Node Layer for service. (2) The Data Node Layer manages user-uploaded data, including database files storing metadata and FITS files storing spectral data. Regarding data access permissions, it includes three levels: PublicDB visible to all registered users, GroupDB visible to group members, and MyDB for private user data. This layer interacts with the Expert Layer through ODBC or JDBC. (3) The Expert Layer establishes multi-platform clients (Web, desktop application, mobile APP) for online spectral data visualization and analysis, including catalog data visualization, real-time interactive visualization of FITS spectral images, multi-band data fusion visualization, real-time redshift/radial velocity calculation, convenient template matching, spectral line identification, data querying, and user feedback functions.

The Data-mining Layer has relatively loose interfaces, allowing users to customize algorithms around specific scientific goals, requiring only that outputs be in platform-compatible formats. The platform will gradually integrate various common data mining algorithms such as neural networks, support vector machines, and principal component analysis for specific needs like photometric redshift, spectral classification, and feature extraction. This advanced functionality is currently mostly reserved as interfaces for gradual implementation based on user demand and is not the focus of this paper.

The Data Node Layer and Expert Layer have strong coupling and are implemented under the lightweight J2EE Spring framework, using the astronomically common MySQL database with ORM plugins for dynamic catalog data management. Spectral data analysis and visualization interfaces are implemented using Javascript, chart processing components, and Mallet wavelet denoising plugins. These two layers constitute the basic modules of the inspection platform and are the focus of this paper.

### 3. Functional Analysis and Implementation

This section describes the functional decomposition and implementation of the Data Node Layer and Expert Layer. Data management in the Data Node Layer is primarily achieved through database and network technologies, while online visualization and analysis of one-dimensional spectral data in the Expert Layer is mainly implemented through Java chart processing technologies. The expert spectral inspection platform currently features eight main functions: user management, data access permission management, data upload, data retrieval and download, multi-band information fusion, one-dimensional spectral data visualization and analysis, expert feedback and knowledge base management, and

Virtual Observatory services.

### 3.1 User Management

The platform fully simulates real scientific research team collaboration modes, dividing users into four levels: System Administrator, Group Administrator, Expert User, and Anonymous User. (1) Anonymous users can browse open information and learn relevant knowledge, such as Learning modules, FAQs, and public data, but have no data import permissions. (2) Expert Users can upload their private data (including spectral and catalog data), use visualization tools to identify and mark spectra, and write results to the database. (3) Group Administrators have permissions to create new users, add members, and remove members, automatically becoming members of new groups. (4) System Administrators have all Group Administrator and Expert User permissions, can create groups, determine whether users are Group Administrators or ordinary users, and allocate maximum available storage space.

### 3.2 Data Access Permission Management

Based on actual scientific research needs, the platform divides data into three levels: MyDB (Personal Data), GroupDB (Group Data), and PublicDB (Public Data), as shown in [Figure 3: see original paper]. (1) MyDB: Data tables uploaded by users themselves, private data visible only to the uploader. (2) GroupDB: Shared data tables within groups, submitted by the uploader for group sharing and approved by the Group Administrator to be visible to all group members. (3) PublicDB: Data tables accessible to all registered users, submitted by Group Administrators and approved by System Administrators.

### 3.3 Data Upload

The platform's data types are mainly divided into two categories: catalog data and one-dimensional spectral data, with the upload interface shown in [Figure 4: see original paper]. (1) Catalog Data: The system primarily supports CSV format. Users upload CSV files to the database, and the system generates corresponding data tables in MyDB. The first row of the CSV file is automatically recognized as field names, and the CSV filename becomes the default table name (users can also specify a custom name). (2) Spectral Data: The system primarily supports FITS files, obtained through two methods: for large surveys like LAMOST and SDSS with online data releases, users only need to provide unique spectral identifiers (mjd, plateID, fiberID) in the uploaded catalog data, and the system automatically locates the corresponding FITS files; for small-sample spectral data from telescopes like Xinglong 2.16m and Lijiang 2.4m, users must package and upload FITS files themselves.

### 3.4 Data Retrieval and Download

Data retrieval has two meanings: catalog file retrieval, and retrieval of selected catalog contents and user feedback records. As shown in [Figure 5: see original paper], when managing many catalogs, users may need to match table names, keywords, data sources, and other conditions to find desired catalogs. Since users may not be familiar with database query languages, the system provides dynamic generation of common query condition expressions that can be combined using “AND” / “OR” relationships to form complex tree-like logical structures, as shown in [Figure 6: see original paper]. For query results, users can select all or some checkboxes on the left and click the “Download” button on the right to package and download data, as shown in [Figure 7: see original paper].

### 3.5 Multi-band Information Fusion

During scientific research, scientists typically need to integrate multi-band imaging or spectral information in addition to examining a celestial object’s spectrum to accurately determine object types or measure physical parameters. Therefore, the platform provides interfaces to fuse photometric imaging data from surveys like SDSS and 2MASS. The query interface is shown in [Figure 8: see original paper], with query results shown in [Figure 9: see original paper].

### 3.6 Spectral Data Visualization and Analysis

This component includes basic operations such as local zooming, spectral line identification, equivalent width measurement, wavelet filtering denoising, template matching for over 170 spectral types, automatic redshift measurement, image saving, and printing. Human-computer interaction for spectral visualization is completely browser-based without requiring any plugin installation, with the display interface shown in [Figure 10: see original paper].

Pipeline batch processing programs typically perform automatic template matching for survey spectra, determining types and measuring parameters based on best matches. However, pipeline processing performs poorly on low-SNR spectra, resulting in many LAMOST unknown spectra that should not be forgotten. The platform provides comprehensive templates for stars, galaxies, and quasars to facilitate interactive template matching. Users select templates, mask unreliable spectral portions, complete rough matching, and the system automatically seeks the best match position and calculates astrophysical parameters (redshift or radial velocity, etc.).

### 3.7 Expert Feedback and Knowledge Base

For each spectrum, experts can complete manual processing through the spectral visualization and analysis module and feed results back to the database. Each spectrum in GroupDB and PublicDB may be analyzed by multiple users, and

the platform records all analysis results in a list for scientist reference. After long-term operation, the system will accumulate a series of expert knowledge bases.

### 3.8 Virtual Observatory (VO) Services

The International Virtual Observatory Alliance has long been committed to seamlessly and transparently connecting global research resources to form a data-intensive networked astronomical research platform. After more than a decade of development, it has accumulated rich data and service resources, including full-band information from high-energy gamma rays to radio, and has developed many excellent software tools such as Topcat, Aladin, NED, SkyServer, and SciServer. Our platform will integrate with these services and data through backend scripts.

## 4. Application Cases

This section demonstrates the platform's effectiveness through teaching activity designs for several university astronomy elective courses. (1) Before class, we mixed the 318 white dwarf-main sequence binary samples selected by Ren Juanjuan et al. [7-8] with ordinary star samples, imported them into the platform, and batch-registered student accounts by student ID as members of the same group. (2) After students gained preliminary understanding of the "Life of Stars" chapter, teachers introduced special stellar systems and their scientific significance, demonstrated various white dwarf-main sequence binary photometric images and spectral features on the platform, and finally distributed cross-group "white dwarf-main sequence binary search" tasks. (3) In the following class session, to motivate students, we arranged an astronomical observation activity of "Albireo" using ordinary telescopes—it appears as a yellow-blue binary star similar to but fundamentally different from the white dwarf-main sequence binaries being searched; under high magnification, yellow Albireo A is itself a binary system, while blue Albireo B is a rapidly rotating Be star. (4) Each student was required to manually inspect at least 50 different spectra and submit a simple research report after one month.

[Figure 12: see original paper] shows photometric images (left) and LAMOST spectra (right) of white dwarf-main sequence binary candidates. Both images are captured from the expert spectral inspection test platform, where students need only three to five mouse clicks to obtain similar images. Ordinary students can effectively judge binary candidates through the left-side object images, while further confirmation requires analyzing the right-side spectral images, which are dynamically displayed in real-time by directly reading FITS files with interactive operation support. The blue curve is the LAMOST observed spectrum, smoothed through wavelet filtering (one mouse click) to eliminate most noise signals, while the black curve is a white dwarf spectrum template retrieved from the system (two mouse clicks). Comparing the blue and black curves, students can easily see that the red end of the blue curve beyond approximately 6500

Å does not match the template well, primarily because this red portion is contributed by the companion star (the redder dwarf in the left image), which can be well-matched using a dwarf star template.

The platform records each student's inspection results, enabling teachers (Group Administrators) to easily retrieve objects confirmed as white dwarf-main sequence binaries by multiple students—the more confirmations from different students, the higher the probability the object belongs to this category. Through such research-oriented teaching activities, we found students highly interested in invisible but objectively existing cosmic objects. Through interest stimulation and grade incentives, most students actively participated to collectively identify the vast majority of white dwarf-main sequence binaries. Teachers can then extract the most suspicious candidates for physics students or genuinely interested astronomy students to further investigate which binaries belong to common envelope post-binary candidates through university science project funding or undergraduate thesis opportunities.

Beyond this application case, the platform can also be effectively applied to manual quasar searches (as shown in [Figure 10: see original paper]), galaxy pair searches, carbon star searches, and other scientific objectives, supporting both individual and collaborative team modes. Due to space limitations, these are not detailed here.

## 5. Summary and Outlook

This paper outlines the development of an expert spectral inspection platform based on advanced IT technology to address LAMOST spectral data processing needs. Released as a website service, it helps scientists improve manual inspection efficiency, assists astronomy educators in enriching teaching methods, and helps LAMOST process unknown spectra.

Currently, the platform continues to integrate various data mining algorithms. We are collaborating with the authors of the internationally renowned one-dimensional spectral visualization software SPLAT-VO to make SPLAT-VO one of the platform's main clients. Through continuous improvement, we hope this platform will become an indispensable spectral data processing platform for China's LAMOST major scientific project and the preferred platform for astronomical spectral science education.

### References

- [1] Li H N, Zhao G, Christlieb N, et al. Test observations that search for metal-poor stars with the Guoshoujing Telescope (LAMOST)[J]. *Research in Astronomy and Astrophysics*, 2010, 10(8): 753-760.
- [2] Wu X B, Chen Z Y, Jia Z D, et al. A very bright ( $i = 16.44$ ) quasar in the 'redshift desert' discovered by the Guoshoujing Telescope (LAMOST)[J]. *Research in Astronomy and Astrophysics*, 2010, 10(8): 737-744.

- [3] Huo Z Y, Liu X W, Xiang M S, et al. The LAMOST survey of background quasars in the vicinity of the Andromeda and Triangulum galaxies -II. results from the commissioning observations and the pilot surveys[J]. The Astronomical Journal, 2013, 145(6): 159-168.
- [4] Ren J J, Luo A L, Li Y B, et al. White dwarf - main sequence binaries identified from the LAMOST pilot survey[J]. The Astronomical Journal, 2013, 146(4): 82-93.
- [5] Zhao J K, Luo A L, Oswalt T D, et al. 72 DA white dwarfs identified in the LAMOST pilot survey[J]. The Astronomical Journal, 2013, 145(6): 169-187.
- [6] Hou W, Luo A L, Yang H F, et al. A large sample of Am candidates from LAMOST Data Release 1[J]. Monthly Notices of the Royal Astronomical Society, 2015, 449(2): 1401-1409.
- [7] Ji W. Carbon star candidates identified from LAMOST DR2[C]// Abstracts of the 2015 Annual Meeting of the Chinese Astronomical Society. 2015: 67.
- [8] Shen S Y, Maria A F, Chen L, et al. A sample galaxy pairs identified from the LAMOST spectral survey and the Sloan Digital Sky Survey[J]. Research in Astronomy and Astrophysics, 2016(3): 63-72.
- [9] Mao Xiaoyan, Zhang Bo, Ye Zhongfu. Using weighted filtering to denoise low-SNR spectra observed through the LAMOST fiber optics[J]. Astronomy Research and Technology, 2015, 12(4): 447-454.
- [Raddick J, Lintott C J, Schawinski K, et al. Galaxy Zoo: an experiment in public science participation[J]. Advances in Atmospheric Sciences, 2007, 39: 892.]
- [Szalay A S, Gray J, Thakar A R, et al. The SDSS skyserver: public access to the sloan digital sky server data[C]// Proceedings of the 2002 ACM SIGMOD international conference on Management of data. 2002: 570-581.]
- [Bonnarel F, Fernique P, Bienaymé O, et al. The ALADIN interactive sky atlas: a reference tool for identification of astronomical sources[J]. Astronomy & Astrophysics Supplement, 2000, 143: 33-40.]
- [Helou G, Madore B F, Bica M D, et al. The NASA/IPAC extragalactic database[C]// Proceedings on the 6th Workshop of the Advanced School of Astronomy of the Ettore Majorana Centre. 1991: 89-106.]
- [Škoda P, Draper P W, Neves M C, et al. Spectroscopic analysis in the virtual observatory environment with SPLAT-VO[J]. Astronomy and Computing, 2014, 7-8: 108-120.]

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

*Source: ChinaXiv – Machine translation. Verify with original.*