

LAMOST天体光谱数据开放共享的回顾与展望

李珊珊 樊东卫 崔辰州 何勃亮 陶一寒 霍志英 米琳莹 罗阿理 陈建军 侯文
孔 啸 李荫碧 郭炎鑫 李 双 李长华 许允飞 韩 军 杨丝丝 杨涵溪 赵永恒
(中国科学院国家天文台, 北京 100101)

摘要: 大天区面积多目标光纤光谱天文望远镜(LAMOST, 也称郭守敬望远镜)是国家重大科技基础设施。其每天观测获取的原始数据及经过处理产生的次生数据, 可以通过LAMOST数据发布平台实现开放共享。本文回顾近10年来LAMOST数据开放共享历程。对LAMOST数据发布批次、数据概况、发布系统架构及功能等进行概述, 并总结相关科学成果, 展望未来发展。LAMOST数据开放共享促进了中国虚拟天文台技术的进步, 符合未来科学数据管理的发展方向, 推动了科研模式的变革。

关键词: 天文数据; 开放共享; 虚拟天文台; 郭守敬望远镜; 科学平台

DOI: 10.3772/j.issn.1674-1544.2022.01.006

CSTR: 15994.14.issn.1674-1544.2022.01.006

中图分类号: P11

文献标识码: A

Review of LAMOST Open Data Access and Future Prospect

LI Shanshan, FAN Dongwei, CUI Chenzhou, HE Boliang, TAO Yihan, HUO Zhiying, MI Linying, LUO Ali, CHEN Jianjun, HOU Wen, KONG Xiao, LI Yinbi, Guo Yanxin, LI Shuang, LI Changhua, XU Yunfei, HAN Jun, YANG Sisi, YANG Hanxi, ZHAO Yongheng

(National Astronomical Observatories, Chinese Academy of Science, Beijing 100101)

Abstract: Guoshoujing Telescope (Large Sky Area Multi-Object Fiber Spectroscopic Telescope, LAMOST)

作者简介: 李珊珊(1984—), 女, 中国科学院国家天文台助理研究员, 研究方向为天文信息学; 樊东卫(1985—), 男, 中国科学院国家天文台副研究员, 研究方向为天文信息学(通信作者); 崔辰州(1976—), 男, 中国科学院国家天文台研究员, 研究方向为天文信息学(通信作者); 何勃亮(1979—), 男, 中国科学院国家天文台副研究员, 研究方向为天文信息学; 陶一寒(1989—), 女, 中国科学院国家天文台助理研究员, 研究方向为基于机器学习的光谱分类; 霍志英(1977—), 女, 中国科学院国家天文台副研究员, 研究方向为星系宇宙学; 米琳莹(1987—), 女, 中国科学院国家天文台助理研究员, 研究方向为天文信息学; 罗阿理(1969—), 男, 中国科学院国家天文台研究员, 研究方向为天文数据处理; 陈建军(1974—), 男, 中国科学院国家天文台高级工程师, 研究方向为光谱数据处理方法; 侯文(1987—), 女, 中国科学院国家天文台助理研究员, 研究方向为海量数据的分析与处理; 孔啸(1986—), 男, 中国科学院国家天文台工程师, 研究方向为海量光谱数据处理和挖掘; 李荫碧(1982—), 女, 中国科学院国家天文台副研究员, 研究方向为海量数据的分析与处理; 郭炎鑫(1983—), 女, 中国科学院国家天文台高级工程师, 研究方向为海量光谱数据处理和数据挖掘; 李双(1986—), 女, 中国科学院国家天文台工程师, 研究方向为信息和信号处理; 李长华(1977—), 男, 中国科学院国家天文台副研究员, 研究方向为天文信息学; 许允飞(1988—), 男, 中国科学院国家天文台副研究员, 研究方向为天文信息学; 韩军(1985—), 男, 中国科学院国家天文台助理研究员, 研究方向为天文信息学; 杨丝丝(1989—), 女, 中国科学院国家天文台助理研究员, 研究方向为天文信息学; 杨涵溪(1985—), 女, 中国科学院国家天文台科学传播专员, 研究方向为天文科普教育; 赵永恒(1964—), 男, 中国科学院国家天文台研究员, 研究方向为天文数据处理、天文信息学。

基金项目: 国家自然科学基金项目“基于深度学习等机器学习算法的星系光谱自动分类方法研究”(11803055); 国家自然科学基金项目“望远镜网络协同观测引力波事件的策略优化研究”(12103070); 国家自然科学基金委员会—中国科学院天文联合基金项目“面向时域天文学的虚拟天文台核心能力建设与科学应用”(U1731243); 国家自然科学基金委员会—中国科学院天文联合基金项目“面向大视场时域巡天观测的大数据检索与融合方法研究”(U1931132); 中国科学院网信专项2022年度应用示范项目“LAMOST天体光谱权威科学数据库”(CAS-WX2021SF-0204)。

收稿时间: 2021年9月30日。

is a National Major Scientific Project built by the Chinese Academy of Sciences. Data obtained through daily observation and through standardized reduction pipelines, will be released through LAMOST data release platform under the guidance of the Scientific Committee and according to the data policy of LAMOST. This paper reviews the development and construction process of LAMOST open data access in the past ten years. Briefly introduce the versions of LAMOST data, the basic functions of data release system, scientific achievements and future outlook. The open data access of LAMOST promoted the technological progress of Chinese Virtual Observatory, in line with the development direction of scientific data management and will continue to reform the way of scientific research.

Keywords: astronomical data, open data access, virtual observatory, LAMOST, science platform

0 引言

大天区面积多目标光纤光谱天文望远镜 (Large Sky Area Multi-Object Fiber Spectroscopic Telescope, LAMOST), 也称郭守敬望远镜, 是一架由我国自主研制兼具大口径、大视场 (5°)、高光谱获取率等特点的反射式施密特望远镜。其利用焦面上 4 000 根精确定位的光纤, 可以同时观测 4 000 个左右的天体目标, 通过分布式可控光纤定位系统, 将数据记录在 16 台光谱仪上的 32 个 CCD 中 (每台光谱仪分红、蓝端)。LAMOST 在每观测日可获取约 10 GB 原始数据以及处理分析后产生的更多的次生数据^[1-2]。

LAMOST 于 2008 年建设完成并进行了第一次测试观测, 于 2011 年 10 月开始先导巡天观测^[3]。从那时起, 这架国家重大科学望远镜开启了它 10 年巡天观测的历程, 像一只坐落在我国北方壮丽群山之巅的观天巨眼, 以惊人的速度和效率捕获来自宇宙中天体的光谱信息, 不断刷新人类在巡天观测领域的光谱获取记录, 孕育大量科研成果, 助力国内外天文学家在恒星、星族、银河系结构及运动学研究等领域的不断开拓和进取。

在天文学界, 望远镜观测获得的科学数据经过数据保护期后, 通常会面向全世界范围公开发布, 使任何人 (不仅仅是天文学家) 都能通过一定途径方便地获取数据、使用数据, 进行科学研究。保护期内的天文观测数据通常只面向观测及项目团队内部使用, 保护期时长根据不同望远镜不同项目的规定各有不同, 通常为 1 ~ 2 年。因

此, LAMOST 的巡天观测数据最终向全世界开放是一种必然。LAMOST 的观测能力和数据获取能力是毋庸置疑的, 必须为它构建一个与之相匹配的数据发布系统和天文数据库, 使其具备世界级的数据管理、数据检索与开放共享能力。但是, 我国天文学界尚未建立如此海量天文数据的开放共享体系。于 2001 年, 在中国首届虚拟天文台技术研讨会上提出了建设中国虚拟天文台 (China-VO) 及面向虚拟天文台的 LAMOST 计划^[3]。LAMOST 经过 10 余年的建设与发展, 推动了我国天文数据相关技术的发展。国家天文科学数据中心正在应用国际先进的虚拟天文台技术为 LAMOST 提供全生命周期的数据服务, 并与 LAMOST 运行和发展中心共同形成了相对完整稳定的数据发布流程及系统, 成为支撑望远镜从观测到数据发布再到天文学家下载使用整个流程不可或缺的组成部分。

本文将回顾近 10 年来 LAMOST 数据发布系统的开发建设情况, 并对系统架构及功能进行简要介绍, 探讨 LAMOST 数据的影响力, 展示国际发布工作及科普教育延伸应用开发取得的成果, 展望未来建设与发展。

1 LAMOST 数据发布

LAMOST 先导巡天观测于 2011 年 10 月 24 日正式启动, 并于 2012 年 6 月结束, 覆盖了 9 个月亮周期, 获取光谱约百万条。最终经过信噪比等筛选流程, 保留了 319 000 条光谱。FITS 格式的光谱数据可以通过当时的 LAMOST 官方网站 (<http://www.lamost.org/>) 进行索引和下

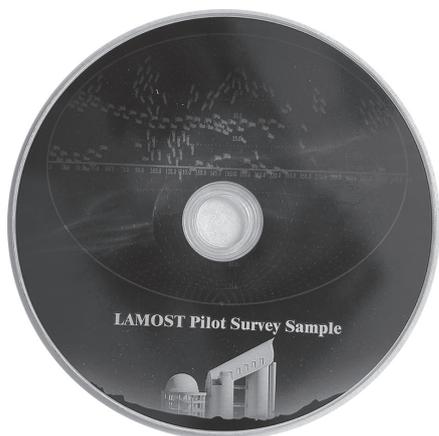


图1 LAMOST 先导巡天数据光盘

载^[4]。这是LAMOST最早的数据发布。为了进行数据的宣传和推广,研究团队在第28届国际天文学联合会(IAU)大会上设置展台,制作包含LAMOST先导数据的光盘(图1)供天文学家获取和研究数据。IAU会刊《天问》报道了LAMOST建设、观测与数据发布的相关情况^[5]。

此后,基于先导巡天得到的经验,LAMOST科学委员会分析了观测数据并制定了5年巡天观测计划。该计划主要观测目标为河内恒星,也包括相当一部分的河外源^[5]。2012年9月28日,LAMOST第一年巡天正式开始,至2013年6月结束。在此期间,LAMOST汲取了斯隆数字化巡天(Slona Digital Sky Survey, SDSS)的经验,高度重视数据管理与发布工作,正式开启了LAMOST数据发布系统的建设及数据开放共享的序幕。

1.1 LAMOST 数据发布系统的源起

2015年3月19日,正式对全世界发布LAMOST取得的首批巡天光谱数据,即完全开放共享。此次公开发布的数据包含220万条光谱信息,完成了于2011年9月至2013年6月近两年的巡天观测任务(先导巡天及第一年正式巡天)。其中,信噪比大于10的恒星光谱有172万条,超过了当时世界上所有已知恒星巡天项目的光谱总和。同时,发布数据中还包括一个108万颗恒星的光谱参数星表,是当时世界上最大的恒星光谱参数星表^[6]。这批数据最早于2013年8月对国内天文学家和国际合作者优先发布,并根据国际天文学界的惯例,在数据保护期之后对全世界开放共享。

为更好地服务此次国际发布,中国虚拟天文台团队基于自身在虚拟天文台领域和天文数据领域的多年研究经验和技術积累,投入全部研发力量与LAMOST数据团队合作,开发了为项目量身定制的全新的数据发布在线服务系统(网站平台)^[7-8]。研发过程主要参考了国际上备受肯定的SDSS检索界面和发布模式^[9],借鉴了国际先进的数据管理发布理念,尊重天文科研人员的使用习惯,不仅支持数据下载,还提供LAMOST先导巡天和第一年巡天数据(DR1)的在线检索和浏览,是国内首个支持国内天文大科学装置的专业数据发布平台。这也形成了现在LAMOST数据发布系统的雏形。

平台实现了多项技术的融合应用,主要包括:光谱数据可视化,支持在线查看光谱信息、修改红移、自定义光谱发射线吸收线等;打通系统和云平台的通道,检索结果可直接导入云空间存储;每个版本释放前预先做好LAMOST数据与40余个常用星表的融合工作并提供检索服务;实现国际虚拟天文台联盟制定的锥形检索、光谱检索等数据访问协议,拓宽数据传播范围和使用渠道。这种设计开发思路不但保证了科研学者使用时的流畅性和连贯性,也最大限度地向全世界开放,并为对接国际化需求和标准铺平了道路。此次LAMOST数据发布服务工作是在中国科学院“十二五”科研信息化专项“天文学科技领域云”和国家地球系统科学数据共享平台天文数据中心的支持下完成的。

1.2 LAMOST 巡天观测数据发布

郭守敬望远镜是国家投资的大型科学设备,按照天文界惯例,其观测数据需通过“数据发布”最终向所有公众开放。截至2020年年底,LAMOST光谱巡天观测数据已经进行了26次不同批次、不同版本的发布(其中部分发布批次中包含多个星表和光谱数据),设计开发全新数据发布网站平台10余个,总发布数据量超过29.74TB。为方便相关数据的使用和溯源,LAMOST数据发布在线服务平台除提供最新巡天观测数据的浏览、检索查询和下载服务外,还保留了所有

已发布批次数据的公开检索服务。

LAMOST巡天观测获得的原始数据只有经过复杂的数据处理流程 (PIPELINE) 后才能对外发布。处理流程中的关键环节会形成中间数据、次生数据等。为了保证数据的可追溯性,方便质量控制,这些数据也会入库存储供后续使用。所有观测和处理流程均有日志记录。目前的LAMOST原始观测数据包括低分辨率、中分辨率两部分,需经过分类处理进行参数测量、绘制光谱图、人工检查等环节,再经数据质量控制,最终制作成数据对外发布。其简化后的流程如图2所示。根据《LAMOST光谱巡天数据政策》,发布数据主要包括以下3种类型:一是原始数据及原始来源信息(如观测日志文件、校准文件等),由科学委员会酌情公布;二是—维光谱数据,在每年观测

季数据获取后18个月内进行增量发布;三是星表数据,于一维光谱数据发布后的6个月内发布。在实际发布中,星表数据通常与光谱数据同期发布。为保护参与项目的机构和个人成员利益,只有获得项目科学委员会批准的成员才能提前使用未发布的观测数据。原则上除—维光谱数据、星表数据外,原始数据不会公开发布,将进行长期保存^①。

LAMOST一期巡天观测(2011—2017年)仅提供低分辨率光谱及星表数据,中国虚拟天文台团队为LAMOST巡天观测数据提供在线检索、查询及下载服务的时间、发布网址等信息,如表1所示。

LAMOST二期巡天观测(2018—2022年),其数据发布以《LAMOST光谱巡天第六次数据

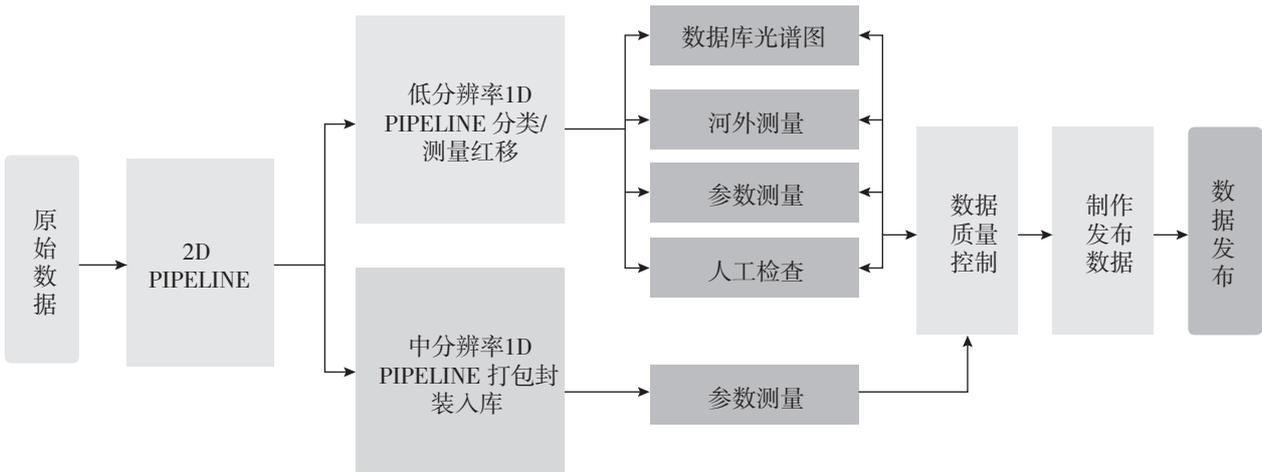


图2 LAMOST巡天观测数据处理流程示意

表1 LAMOST一期巡天(2011—2017年)数据发布情况

| 数据集名称 | 观测天区/个 | 总光谱数/条 | 观测时间 | 发布时间 | 发布网址 |
|----------------------|--------|-----------|-----------------------|---------|---|
| LAMOST光谱巡天第一次数据发布 | 1 202 | 2 204 696 | 2011-10-24—2013-06-03 | 2013-08 | http://dr1.lamost.org |
| LAMOST光谱巡天第二次数据发布 | 1 934 | 4 132 782 | 2011-10-24—2014-06-05 | 2014-12 | http://dr2.lamost.org |
| LAMOST光谱巡天第三次数据发布 | 2 667 | 5 755 126 | 2011-10-24—2015-05-30 | 2015-12 | http://dr3.lamost.org |
| LAMOST光谱巡天第四次数据发布第一版 | 3 459 | 7 661 651 | 2011-10-24—2016-06-02 | 2016-12 | http://dr4.lamost.org/v1/ |
| LAMOST光谱巡天第五次数据发布第一版 | 4 154 | 9 017 844 | 2011-10-24—2017-06-16 | 2017-12 | http://dr5.lamost.org/v1/ |
| LAMOST光谱巡天第四次数据发布第二版 | 3 461 | 7 620 612 | 2011-10-24—2016-06-02 | 2018-06 | http://dr4.lamost.org/v2/ |
| LAMOST光谱巡天第五次数据发布第二版 | 4 154 | 9 027 634 | 2011-10-24—2017-06-16 | 2018-12 | http://dr5.lamost.org/v2/ |
| LAMOST光谱巡天第五次数据发布第三版 | 4 154 | 9 026 365 | 2011-10-24—2017-06-16 | 2019-06 | http://dr5.lamost.org/v3/ |

① LAMOST光谱巡天数据政策 <http://www.lamost.org/lmusers/cms/article/view?id=1>。

发布第〇版第一季度》为起点。与一期巡天仅进行低分辨率观测不同,二期巡天开启了中、低分辨率光谱巡天交替进行的观测模式。二期巡天观测数据的释放更加注重发布版本的规范性。以第〇版(V0)对应季度数据,该版本数据为测试版数据,于观测结束后一个半月内发布,仅限于工作组成员和获得项目科学委员会批准的成员使用;以第一版(V1)对应国内正式版数据,于观测结束后一年发布(每年夏季),可供国内科研工作者及国际合作者使用;第二版(V2)对应国际正式版数据,于国内正式版数据发布一年

半后发布(每年冬季),该版本完全向全世界公开开放,任何人都可下载使用。其他数据更新和修正以子版本号标记,如V1.1、V1.2等。由于天文数据的特殊性,每个发布版本都有科研人员使用并产出相应的成果,不能随意替换删除。因此,LAMOST释放的所有数据版本在国家天文科学数据中心长期保留,并允许科研人员根据需求进行检索和下载。截至目前,LAMOST二期巡天已发布数据的详细情况见表2。从先导巡天到一期巡天,再到二期巡天,LAMOST持续进行光谱巡天观测工作,不断扩大观测天区范

表2 LAMOST二期巡天(2018—2022年)数据发布情况

| 数据集名称 | 分辨率 | 观测时间 | 发布时间 | 发布网址 |
|---------------------------|-----|-----------------------|---------|---|
| LAMOST光谱巡天第六次数据发布第〇版 第一季度 | 低 | 2017-09-01—2017-11-30 | 2018-01 | http://dr6.lamost.org/v0/ |
| LAMOST光谱巡天第六次数据发布第〇版 第二季度 | 低 | 2017-12-01—2018-02-28 | 2018-04 | http://dr6.lamost.org/v0/ |
| LAMOST光谱巡天第六次数据发布第〇版 第三季度 | 低 | 2018-3-12—2018-05-18 | 2018-07 | http://dr6.lamost.org/v0/ |
| LAMOST光谱巡天第六次数据发布第〇版 | 低 | 2017-09-01—2018-05-18 | 2018-07 | http://dr6.lamost.org/v0/ |
| LAMOST光谱巡天第七次数据发布第〇版 第一季度 | 低 | 2018-10-05—2018-12-31 | 2019-02 | http://dr7.lamost.org/v0/ |
| | 中 | | | |
| LAMOST光谱巡天第六次数据发布第一版 | 低 | 2011-10-24—2018-05-18 | 2019-03 | http://dr6.lamost.org/v1/ |
| | 中 | 2017-09-01—2018-06 | | |
| LAMOST光谱巡天第七次数据发布第〇版 第二季度 | 低 | 2019-01-01—2019-03-31 | 2019-05 | http://dr7.lamost.org/v0/ |
| | 中 | | | |
| LAMOST光谱巡天第七次数据发布第〇版 第三季度 | 低 | 2019-04-01—2019-06-30 | 2019-07 | http://dr7.lamost.org/v0/ |
| | 中 | | | |
| LAMOST光谱巡天第七次数据发布第〇版 | 低 | 2018-10-05—2019-6-30 | 2019-07 | http://dr7.lamost.org/v0/ |
| | 中 | | | |
| LAMOST光谱巡天第八次数据发布第〇版 第一季度 | 低 | 2019-10-23—2019-12-31 | 2020-02 | http://www.lamost.org/dr8/v0/ |
| | 中 | | | |
| LAMOST光谱巡天第七次数据发布第一版 | 低 | 2011-10-24—2019-06-08 | 2020-04 | http://dr7.lamost.org/v1/ |
| | 中 | 2017-09—2019-06 | | |
| LAMOST光谱巡天第八次数据发布第〇版 第二季度 | 低 | 2020-01-01—2020-03-31 | 2020-05 | http://www.lamost.org/dr8/v0/ |
| | 中 | | | |
| LAMOST光谱巡天第六次数据发布第1.1版 | 低 | 2011-10-24—2018-05-18 | 2020-06 | http://dr6.lamost.org/v1.1/ |
| | 中 | 2017-09—2018-06 | | |
| LAMOST光谱巡天第七次数据发布第1.1版 | 低 | 2011-10-24—2019-06-08 | 2020-06 | http://dr7.lamost.org/v1.1/ |
| | 中 | 2017-09—2019-06 | | |
| LAMOST光谱巡天第八次数据发布第〇版 第三季度 | 低 | 2020-04-01—2020-06-30 | 2020-08 | http://www.lamost.org/dr8/v0/ |
| | 中 | | | |
| LAMOST光谱巡天第八次数据发布第〇版 | 低 | 2019-10-23—2020-06-30 | 2020-08 | http://www.lamost.org/dr8/v0/ |
| | 中 | | | |
| LAMOST光谱巡天第六次数据发布第二版 | 低 | 2011-10-24—2018-05-18 | 2020-10 | http://dr6.lamost.org/V2/ |
| | 中 | 2017-09—2018-06 | | |
| LAMOST光谱巡天第七次数据发布第1.2版 | 低 | 2011-10-24—2019-06-08 | 2020-12 | http://dr7.lamost.org/v1.2/ |
| | 中 | 2017-09—2019-06 | | |

续表

| 数据集名称 | 分辨率 | 观测时间 | 发布时间 | 发布网址 |
|-----------------------------|--------|-----------------------|---------|---|
| LAMOST 光谱巡天第九次数据发布第〇版 第一季度 | 低 中 | 2020-09-18—2020-12-31 | 2021-02 | http://www.lamost.org/dr9/v0/ |
| LAMOST 光谱巡天第八次数据发布第 1.0 版 | 低 中 | 2011-10-24—2020-05-27 | 2021-03 | http://www.lamost.org/dr8/ |
| LAMOST 光谱巡天第七次数据发布第 1.3 版 | 低 中 | 2011-10-24—2019-06-08 | 2021-04 | http://dr7.lamost.org/v1.3/ |
| LAMOST 光谱巡天第九年数据发布第〇版 第二批数据 | 低 中 | 2021-01-01—2021-03-31 | 2021-06 | http://www.lamost.org/dr9/v0/ |
| LAMOST 光谱巡天第九年数据发布第〇版 第三批数据 | 低 中 | 2021-04-01—2021-06-30 | 2021-08 | http://www.lamost.org/dr9/v0/ |
| LAMOST 光谱巡天第九次数据发布第〇版 | 低 中 | 2020-09-18—2021-06-30 | 2021-08 | http://www.lamost.org/dr9/v0/ |

围，累积获得观测光谱数量逐年增加，如图 3 所示。截至 2021 年 3 月，在 LAMOST 光谱巡天第八次数据发布第一版时，低分辨率总光谱数已达 10 388 423 条，中分辨率总光谱数达到 6 038 218 条。

1.3 LAMOST 数据发布系统的架构及功能

随着 LAMOST 巡天观测的持续进行，数据发布较为频繁，版本复杂多样，数据内容丰富。为更好地方便天文学家检索使用数据，在每次发布前还需要将发布数据与其他国内外光学、射电、红外、紫外等常用星表融合，实现多种数据在线检索和分析等功能，最终形成具有可用性、易用性和实用性的在线服务平台。LAMOST 数据

发布系统架构既能满足数据持续稳定对外发布需求，也能为科研人员最大限度地提供便捷。该系统由硬件层、系统层、在线服务、宣传推广 4 个层面构成。其具体架构如图 4 所示。

其中，在线服务模块是用户直接体验和使用的平台，由下面 5 个部分组成（以 DR8 V1 为例）。

(1) 数据发布系统首页：LAMOST 巡天数据以年度巡天为单位进行网站首页风格的升级和设计开发，至今已正式发布的系统平台累计达 10 个版本。在每个系统的首页上展示版本名称、版本号、常用功能连接、数据统计（低分辨率、中分辨率）、致谢模板等。

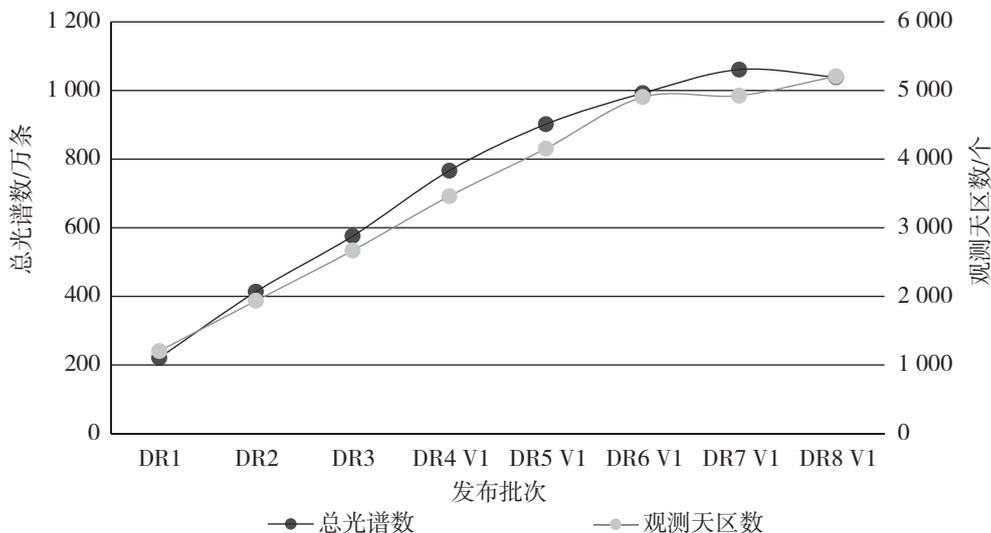


图 3 LAMOST 低分辨率巡天观测总光谱数及观测天区数

(2) 数据在线检索：用户可通过该功能对低分辨率和中分辨率光谱进行在线检索。该模块是参考 SDSS 检索功能并结合 LAMOST 数据自身特点设计的。用户可通过位置区域、指定目标、观测 ID、红移、观测信息、恒星参数、类别、信噪比等多种限制方式进行检索。检索后即返回如图 5 所示的数据列表。

用户可将检索结果返回列表以 CSV 或 VOT-table 格式下载。光谱文件 (FITS、PNG 等格式) 可下载到本地计算机上, 也可存储至中心提供的云空间内进行在线科研。基于 HTML5 开发的光谱查看功能, 可实现放大缩小检查光谱细节、查看光谱分类、修改红移值、自定义光谱射线吸

收线, 可支持用户进行在线光谱可视化和数据分析。

(3) 数据获取：LAMOST 巡天观测发布数据包括星表数据、光谱数据等。系统允许用户通过 SQL 查询数据、使用 FTP 下载、离线等方式获取数据。前两者符合条件的用户可在线自行操作, 离线获取数据需要提交申请表, 获得批准后即可至数据中心线下拷贝数据。

(4) 文档：系统中的文档栏目包括该版本数据释放文档、低分辨率数据描述 (中英)、中分辨率数据描述 (中英) 和 LAMOST 数据政策、发表政策等。

(5) 帮助与联络：系统提供数据获取介绍文

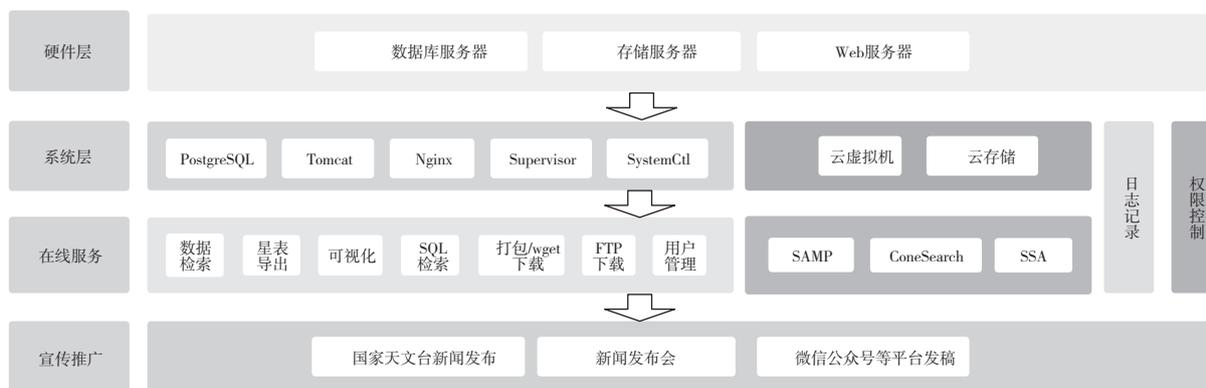


图 4 LAMOST 数据发布系统架构

| LAMOST | | | | | | | | | | | | | 李珊珊 | Signout |
|--|------|-----|------|------------|--------|------------|-------|-------|--------|------|---------|---------|----------|-----------------|
| Low Resolution Search Result | | | | | | | | | | | | | | |
| Save As Tools Download Download All Help | | | | | | | | | | | | | | |
| | FITS | PNG | View | CDS Portal | obsid | obsdate | lmid | mjd | planid | spid | fiberid | class | subclass | z |
| 1 | | | | | 101001 | 2011-10-24 | 55859 | 55858 | F5902 | 1 | 1 | STAR | G7 | -0.0001386960 |
| 2 | | | | | 101002 | 2011-10-24 | 55859 | 55858 | F5902 | 1 | 2 | STAR | K7 | 0.0001191700 |
| 3 | | | | | 101005 | 2011-10-24 | 55859 | 55858 | F5902 | 1 | 5 | STAR | G0 | -0.0002506401 |
| 4 | | | | | 101007 | 2011-10-24 | 55859 | 55858 | F5902 | 1 | 7 | STAR | dM4 | -0.0001687600 |
| 5 | | | | | 101008 | 2011-10-24 | 55859 | 55858 | F5902 | 1 | 8 | STAR | F9 | 0.00000873938 |
| 6 | | | | | 101009 | 2011-10-24 | 55859 | 55858 | F5902 | 1 | 9 | STAR | G0 | -0.0001152130 |
| 7 | | | | | 101010 | 2011-10-24 | 55859 | 55858 | F5902 | 1 | 10 | Unknown | Non | -9999.000000000 |
| 8 | | | | | 101012 | 2011-10-24 | 55859 | 55858 | F5902 | 1 | 12 | STAR | G7 | -0.0000182600 |
| 9 | | | | | 101013 | 2011-10-24 | 55859 | 55858 | F5902 | 1 | 13 | STAR | A2IV | -0.0010664700 |
| 10 | | | | | 101014 | 2011-10-24 | 55859 | 55858 | F5902 | 1 | 14 | STAR | K5 | 0.0000746200 |
| 11 | | | | | 101015 | 2011-10-24 | 55859 | 55858 | F5902 | 1 | 15 | Unknown | Non | -9999.000000000 |
| 12 | | | | | 101016 | 2011-10-24 | 55859 | 55858 | F5902 | 1 | 16 | STAR | K3 | -0.0000437303 |
| 13 | | | | | 101017 | 2011-10-24 | 55859 | 55858 | F5902 | 1 | 17 | STAR | G0 | 0.0000057707 |
| 14 | | | | | 101018 | 2011-10-24 | 55859 | 55858 | F5902 | 1 | 18 | STAR | A9V | -0.0024628100 |
| 15 | | | | | 101019 | 2011-10-24 | 55859 | 55858 | F5902 | 1 | 19 | STAR | A2V | -0.0003525200 |
| 16 | | | | | 101020 | 2011-10-24 | 55859 | 55858 | F5902 | 1 | 20 | STAR | F6 | 0.0000519026 |
| 17 | | | | | 101021 | 2011-10-24 | 55859 | 55858 | F5902 | 1 | 21 | STAR | F5 | -0.0000822335 |
| 18 | | | | | 101023 | 2011-10-24 | 55859 | 55858 | F5902 | 1 | 23 | STAR | G3 | 0.0001614117 |
| 19 | | | | | 101024 | 2011-10-24 | 55859 | 55858 | F5902 | 1 | 24 | STAR | K5 | 0.0000123752 |
| 20 | | | | | 101026 | 2011-10-24 | 55859 | 55858 | F5902 | 1 | 26 | STAR | F9 | -0.0000756857 |

图 5 LAMOST DR8 数据检索结果页面

档、常见问题 (FAQ) 等帮助信息, 允许用户在线提交联系 LAMOST 运行和发展中心的申请。

2 LAMOST 数据的影响力

2015年3月至2020年9月, LAMOST DR1至LAMOST DR6的数据逐年向全世界公开发布。以LAMOST DR6数据为例, 2020年9月发布版本为《LAMOST光谱巡天第六次数据第二版》(LAMOST DR6 V2), 包含先导巡天及正式巡天前6年的光谱数据, 如图6所示。这一数据发布标志着在国际上LAMOST光谱数据正式迈入千万量级时代, 成为世界上第一个获取光谱数超千万的光谱巡天项目。

2.1 国际推广

除面向全球的数据释放外, 国家天文科学数据中心利用自身在国际虚拟天文台领域的影响力, 积极推广LAMOST数据的国际化应用, 提高中国天文数据知名度。2016年3月, LAMOST第一批巡天数据(LAMOST DR1)正式被法国斯特拉斯堡天文数据中心(CDS)的VizieR数据库系统收录, 获得永久星表编号“V/146”。从这一刻开始, LAMOST光谱数据完整地融入了国际天文数据库系统。其星表和光谱数据均可按照国际虚拟天文台标准接口和检索方式访问, 可方

便地与其他数据交叉融合使用。此后, LAMOST DR2、LAMOST DR4数据分别在2017年3月和2018年9月被VizieR数据库系统收录, 并分别获得了“V/149”和“V/153”的永久星表编号。LAMOST数据被VizieR收录, 提升了我国天文数据在国际上的显示度, 可以更大程度地发挥数据的科学价值。

2020年6月, 欧洲空间局与中国科学院国家天文台、国家天文科学数据中心合作为其交互式天体图谱平台ESASky打造简体中文版, 并收录了LAMOST DR5 V3数据。通过该平台, 来自全世界的天文学家和天文爱好者能够随时随地浏览LAMOST巡天观测数据。

2.2 社会影响力

大数据时代, 天文数据的开放性已引起越来越多来自其他领域的关注和兴趣。海量真实的天文观测数据不仅能够帮助天文学家探索未知, 也在云计算、人工智能技术、机器学习等方面展现了跨领域应用的潜力。2018年2月, 国家天文台—阿里云天文大数据联合研究中心启动了“天文数据挖掘”天池大赛, 以郭守敬望远镜巡天光谱分类为课题, 通过阿里云天池数据众智平台征集高效、高准确率自动化算法。2020年1月, 由北京智源人工智能研究院主办, 国家天文台、

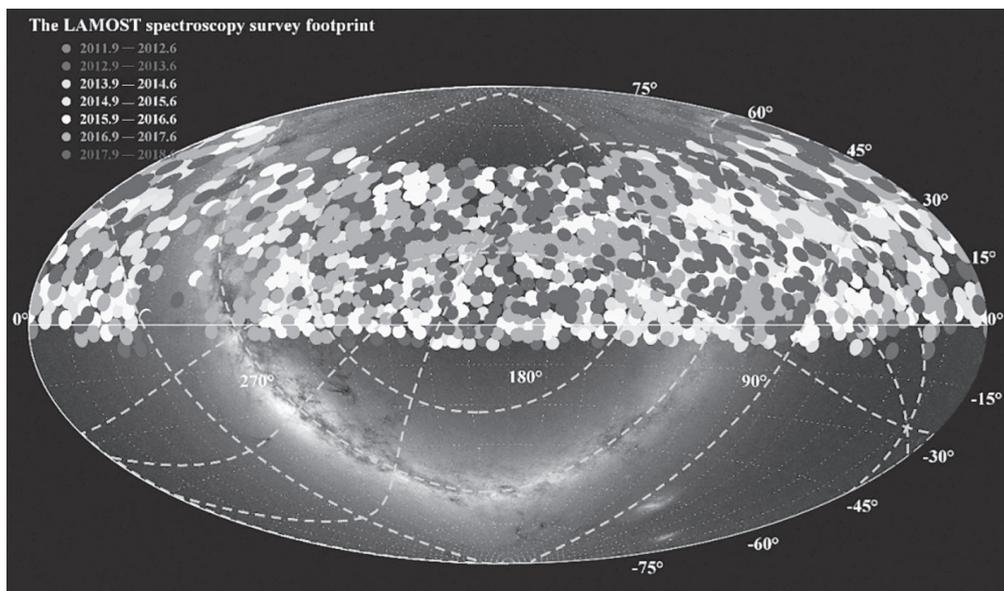


图6 LAMOST 先导巡天和正式巡天前6年的天区覆盖图

数据评测平台 biendata 联合举办的天体分类数据竞赛“智源杯天文数据算法挑战赛”正式启动。参赛选手使用 LAMOST DR4 数据集中近 100 万条天体光谱数据, 利用人工智能技术进行分类算法优化。此类竞赛不仅能为解决天文研究中的实际光谱分类问题提供新的思路, 也为人工智能算法提供了天然的练兵场。大量参赛选手为高校学生, 他们通过实际科学问题的求解, 有机会更加深入地学习相关天文知识, 增加了对天文数据的了解。国家天文科学数据中心将总结相关比赛组织经验, 提炼标准数据集, 继续拓宽 LAMOST 巡天数据更广泛的应用场景。

3 成果与展望

3.1 成果与荣誉

LAMOST 天体光谱数据开放共享的 10 年收获颇丰。目前, 该系统正在服务国内几乎所有天文研究教学单位的专业用户和大量国际合作伙伴, 累积超过千人。系统获得了国内外天文学家的广泛认可, 帮助他们在银河系结构与演化、恒星物理、特殊天体及致密天体、类星体等重要前沿领域取得了一系列有国际影响力的研究成果。其中,《发现迄今银河系中自转最快恒星》《发现宇宙中锂丰度最高恒星》《银河系发现新的移动星群》等基于 LAMOST 数据取得的研究成果在中央电视台相关栏目中播出, 影响广泛。LAMOST 用户委员会对数据服务的评价连年为优。

截至 2020 年年底, LAMOST 数据发布系统支持科学用户在线检索达 3 179 139 次, 支持科研学者使用 LAMOST 巡天数据产出 SCI 论文 700 余篇, 利用 LAMOST 数据发表的 SCI 论文引用数超过 7 000 次并呈现快速增长趋势。发表于 RAA 的 LAMOST DR1 数据发布论文由于高被引而荣获“第三届中国科协优秀论文奖”和英国皇家物理学会出版社 2018 年“中国高被引文章奖”。2019 年 5 月,《郭守敬望远镜数据发布系统》入

选“2018 年度中国科学院信息化优秀案例”^①。由国家天文科学数据中心和 LAMOST 运行和发展中心共同研制的《海量天体光谱数据分析与产品发布系统的研制与应用》获得北京市科技进步二等奖。

3.2 未来展望

从打包下载到在线检索, 从对国内释放到面向全世界, 从天文学家专用到服务于跨领域研究应用, LAMOST 观测数据的开放共享不断完善改进, 数据发布方式与内容从稚嫩探索走向成熟丰富。其中, 国家天文科学数据中心全程参与, 使天文学家使用 LAMOST 数据更加便捷, 不仅能够“用好数据”而且能够“好用数据”。数据全生命周期管理模式为国内其他天文望远镜的数据管理提供了示范。

追溯过往, 国家天文科学数据中心的前身中国天文数据中心已经成立 30 多年。但是在 LAMOST 开始正式观测并持续生产天文数据之前, 我国缺少自己生产的大规模天文观测数据, 中国的天文学家和天文数据工作者也没有开展大型天文数据库的建设以及学习并实践国际虚拟天文台联盟相关标准的机会。LAMOST 的出现和其稳定的数据产出, 对我国天文数据中心建设起到了至关重要的推动作用, 为中国虚拟天文台技术的发展提供了机会。

LAMOST 数据的开放共享, 缩短了中国天文数据开放共享工作与世界先进水平的距离, 打开了全新交流窗口和渠道。10 年中, 大量相关研究人员和技术人才参与发布系统建设开发, 不仅为这个世界上最大的天文光谱库添砖加瓦, 也锻炼了队伍, 积累了经验。科学数据管理指导原则 FAIR (可发现, 可访问, 可互操作, 可重用) 自初见雏形到 2016 年正式提出, 已成为科学数据管理的国际准则^[10]。LAMOST 数据发布系统完全符合这一数据管理原则, 其成功更是验证了该原则的有效性, 说明其符合未来科学数据管理的发展方向。

(下转第 82 页)

① 《郭守敬望远镜数据发布系统被评为中科院信息化优秀案例》https://nadc.china-vo.org/article/20190509090152?locale=zh_CN。

- [16] 张云中, 祝蕊. 面向知识问答系统的图情学术领域知识图谱构建: 多源数据整合视角[J]. 情报科学, 2021, 39(5): 115-123.
- [17] 严佩敏, 唐婉琪. 基于改进 BERT 的中文文本分类[J]. 工业控制计算机, 2020, 33(7): 108-110.
- [18] REIMERS N, GUREVYCH I. Sentence-bert: sentence embeddings using siamese bert-networks[J]. arXiv preprint arXiv: 1908.10084, 2019.
- [19] DONG L, YANG N, WANG W, et al. Unified language model pre-training for natural language understanding and generation[J]. arXiv preprint arXiv: 1905.03197, 2019.

(上接第55页)

LAMOST二期巡天即将完成, LAMOST的观测数据管理和开放共享将进入崭新阶段。未来LAMOST数据团队将继续优化数据质量, 提升数据的可追溯性, 规范数据入库过程, 完善数据发布系统, 进行更深更广的国际化推广, 打造国际权威的科学数据库系统, 并继续推动LAMOST巡天数据在可视化与可视分析、人工智能、机器学习、科普教育等领域的应用。同时, 国家天文科学数据中心将以国际化先进理念为指导, 打造科学平台, 实现数据与科研要素的深度融合, 对天文数据的开放共享进行全新探索, 推动科研模式的变革^[11]。

致谢

郭守敬望远镜 (Large Sky Area Multi-Object Fiber Spectroscopic Telescope, LAMOST) 是中国科学院建设的国家重大科学项目。该项目资金由国家发展和改革委员会提供。LAMOST由中国科学院国家天文台运营和管理。本文得到了中国虚拟天文台、国家天文科学数据中心、中国科学院科学数据中心体系提供的数据资源和技术支持。感谢国家天文台—阿里云天文大数据联合研究中心对本项工作的支持。

参考文献

- [1] CUI X Q, ZHAO Y H, CHU Y Q, et al. The Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST)[J]. 云南天文台台刊, 1999, 12(S1): 257-260.
- [2] 张海龙, 冶鑫晨, 李慧娟, 等. 天文数据检索与发布综述[J]. 天文研究与技术, 2017, 14(2): 212-228.
- [3] 崔辰州, 赵永恒, 赵刚, 等. 虚拟天文台的技术进展[J]. 天文学进展, 2002, 20(4): 302-311.
- [4] LUO Ali, ZHANG Haotong, ZHAO Yongheng, et al. Data release of the LAMOST pilot survey[J]. Research in astronomy and astrophysics, 2012(9): 1243-1246.
- [5] The Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST)[N]. Inquiries of Heaven, 2012-08-24(2).
- [6] LUO A, ZHAO Y, ZHAO G, et al. The first data release (DR1) of the LAMOST regular survey[J]. Research in astronomy and astrophysics, 2015, 15(8): 1095-1124.
- [7] 崔辰州, 赵永恒. 中国虚拟天文台研发策略与重点[J]. 天文研究与技术—国家天文台台刊, 2004(3): 203-209.
- [8] HE B, FAN D, CUI C, et al. The LAMOST data archive and data release[C]. San Francisco: Astronomical Society of the Pacific, 2016(512):153.
- [9] SZALAY A, GRAY J, THAKAR A, et al. The SDSS sky server, public access to the sloan digital sky server data[C]. the 2002 ACM SIGMOD international conference. ACM, 2002.
- [10] WILKINSON M D, DUMONTIER M, AALBERS-BERG I J, et al. The FAIR guiding principles for scientific data management and stewardship[J]. Scientific data, 2016, 3(160018 (2016)): 167-172.
- [11] CUI C, TAO Y, LI C, et al. Towards an astronomical science platform: experiences and lessons learned from Chinese virtual observatory[J]. Astronomy and computing, 2020(32):1-8.