doi: 10.3978/j.issn.1000-4432.2019.09.06
View this article at: http://dx.doi.org/10.3978/j.issn.1000-4432.2019.09.06

亚太眼科学数据中心建设初探

陈有艺[1]，苏超[2]，李扬杵[1]

(1. 中山大学中山眼科中心信息科，广州 510060；2. 南京慧目信息技术有限公司，南京 210000)

[摘　要] 随着亚太眼科学会(Asia-Pacific Academy of Ophthalmology，APAO)永久会址和秘书处落户在中山大学中山眼科中心珠江新城院区眼科学国家重点实验室大楼(中国广州)，在大数据时代背景下，建设先进的亚太级眼科学科学数据中心已成为亚太眼科创新战略发展的需要。通过建设科学数据中心，一方面缓解眼科学国家重点实验室项目科研数据增长压力，一方面为眼科大数据与人工智能等新兴科研项目开展提供保障平台，为亚太眼科学科建设打下坚实的数据基础。科学数据中心设计建设遵照国家《科学数据管理办法》实施，以加强和规范科学数据管理，保障科学数据安全，提高开放共享水平，更好支撑国家科技创新、经济社会发展和符合国家信息安全建设方针。

[关键词] 亚太眼科；科学数据；数据管理

Construction of Asia-pacific ophthalmology data center

CHEN Youyi[1], SU Chao[2], LI Yangchu[1]

(1. Department of Information, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510060; 2. Nanjing Brighteyes Information Technology Co,. Ltd. Nanjing 210000, China)

Abstract As the Asia-Pacific Academy of Ophthalmology (APAO) secretariat permanently located in the building of State Key Laboratory of Ophthalmology (Guangzhou, China). In big data time, it's a problem for the development of the Asia-pacific ophthalmology innovation strategy to construct the Asia-pacific science data center of ophthalmology. Through the construction of science data center, alleviate the pressure of the state key laboratory of ophthalmology data growth, for big data and new research fields such as artificial intelligence to provide guarantee, provided a basis for the Asia-pacific eye subject construction. Science data center construction in accordance with national "scientific data management method", to strengthen and standardize the scientific data management, to ensure the safety of scientific data, improve the level of open sharing, better support scientific and technological innovation, economic and social development and national security as a policy.

Keywords Asia-pacific ophthalmology; scientific data; data management

收稿日期 (Date of reception)：2019–04–11

通信作者 (Corresponding author)：陈有艺，Email: 38245767@qq.com

2018年1月23日中央全面深化改革领导小组第二次会议审议了《科学数据管理办法》，其中第十条规定，科学数据中心是促进科学数据开放共享的重要载体，由主管部门委托有条件的法人单位建立，主要职责是：1)承担相关领域科学数据的整合汇交工作；2)负责科学数据的分级分类、加工整理和分析挖掘；3)保障科学数据安全，依法依规推动科学数据开放共享；4)加强国内外科学数据方面交流与合作[1]。由此显见，建设亚太眼科科学数据中心符合国家相关规定，各眼科学科带头人应当遵照国家政策和规划，积极努力推动国家乃至亚太地区眼科学的科研数据基础性建设。

1 建设目标

把建设一个便于管理、安全规范、合理共享、持续创新和促进科研成果转化的新型科学数据中心作为亚太眼科科学数据中心的建设目标。实现眼科学数据集中存放、集中管理、分类管理和分级管理，避免意外丢失、秘密泄露和危害国家安全等事件发生。通过科学数据管理信息平台，分权限、分类别、分阶段和分级别管理眼科学数据，构建有序共享、无偿共享和有偿共享相结合的共享模式。

2 中心科学数据概况

2.1 科学数据管理概况

中山大学中山眼科中心(以下简称“中心”)目前尚未建设科学数据中心统一管理科研数据。随机抽取中心课题组相关人员32名(其中15男，17女；35岁以上8人，35岁以下24人；技术员18人，研究生学历或中级职称以上14人；表1)，进行一次中心科学数据情况问卷调查。其中问卷选择“中心院内个人存在自行保存或处理科学数据行为的”占比53.13%，选择“有在院内部门存放或集中管理科学数据行为的”占比65.63%，选择“将科学数据存放在院外机构的”占比8.38%。78%选择永久保存科学数据，56%的研究人员发生过科学数据丢失事件(表2)，59%的研究人员曾经利用过他人的原始科学数据。在获取他人原始科学数据困难选项调查中，选择“没有统一的科学数据管理平台”占比最高，达到53.13%，但仅有41%的研究人员选择“同意将个人或者项目组的原始数据交由中心专门的数据管理机构管理”，94%的研究人员同意在适当情况下将科学数据交由中心专门的数据管理机构管理(表3)，38%的研究人员愿意使用“中心数据管理机构提供的他人的原始数据”，56%和53%的研究人员支持和信任中心建立科学数据中心统一管理科研数据，仅有19%(6人)的研究人员了解过国务院办公厅2018年3月17日印发的《科学数据管理办法》。

表1 问卷调查-人员构成统计分析表

Table 1 Personnel constitute statistical analysis

选项	例 (%)
性别	
男	15 (46.88)
女	17 (53.13)
年龄	
25 岁及以下	4 (12.50)
26~35 岁	20 (62.50)
36~45 岁	5 (15.63)
46~55 岁	1 (3.13)
56~65 岁	1 (3.13)
65 岁以上	1 (3.13)
职称	
教授 (正高级研究员)	2
副教授 (副高)	1
讲师 (中级)	3
博士后	3
博士研究生	5
其他	18

表2 问卷调查-数据丢失情况统计分析表

Table 2 Statistical analysis of data loss situation

选项	小计	比例
非常频繁	0	0%
频繁	1	3.13%
一般	4	12.5%
偶尔发生	13	40.63%
从未发生	14	43.75%

表3 问卷调查-愿意将原始数据交由中心专门的数据管理机构管理情况统计分析表

Table 3 Statistical analysis of willing to put the raw data to a specific data management department

选项	小计	比例
非常同意	3	9.38%
比较同意	10	31.25%
无所谓	7	21.88%
不太同意	9	28.13%
非常不同意	3	9.38%

2.2 科学数据存储概况

定向访谈40个中心的课题组，访谈内容主要包括：课题组现有科学数据总量，未来3年科学数据增长量预测，有关数据来源、数据生成类型和数据管理类型等问题(图1，2)。目前中心科学数据存放和管理主要由各自课题项目负责人和相关课题技术人员负责，大多存放在移动硬盘和各自项目经费购置的服务器和存储设备上，尚无科学数据统一管理机构和配置集中存储设备。

2.3 科学数据计算能力和环境概况

中心目前建设有16台曙光CB60-G16双路刀片和2台曙光I840r-GP四路SMP服务器组成的高性能计算集群，CPU整体峰值性能达到8.192万亿次/s(Tflops级运算能力)，中心超级计算能力处于国内中小型平台排名下游靠后水平(图3)。集群平台主要提供给中心3个课题组和中大精准医学科学中心等部门使用(表4)。

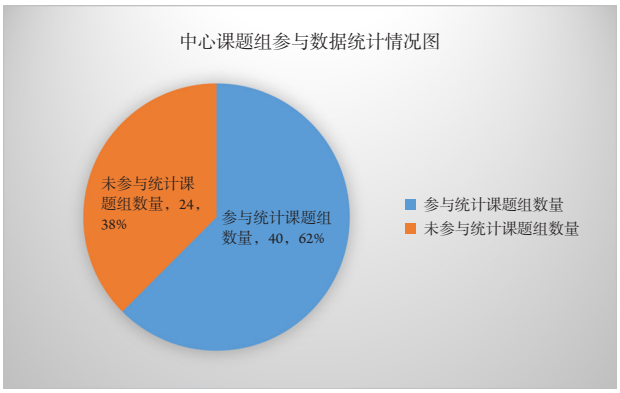

图1 中心课题组参与数据统计情况图

Figure 1 Research group participation analysis

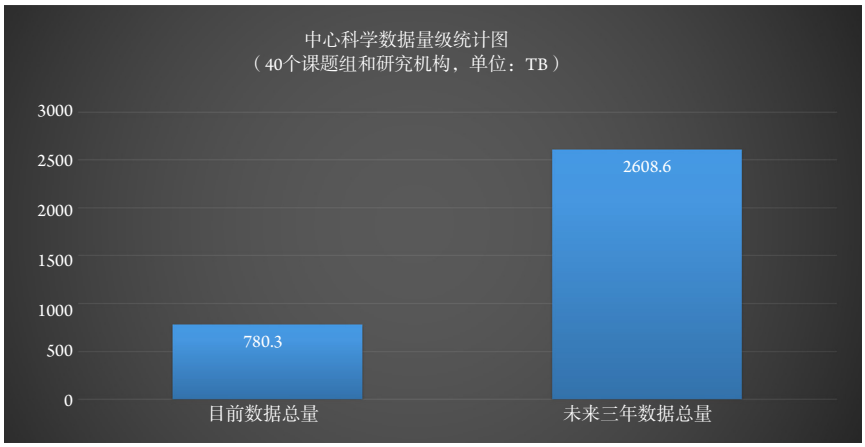

图2 中心科学数据量级统计图

Figure 2 Scientific data level statistics

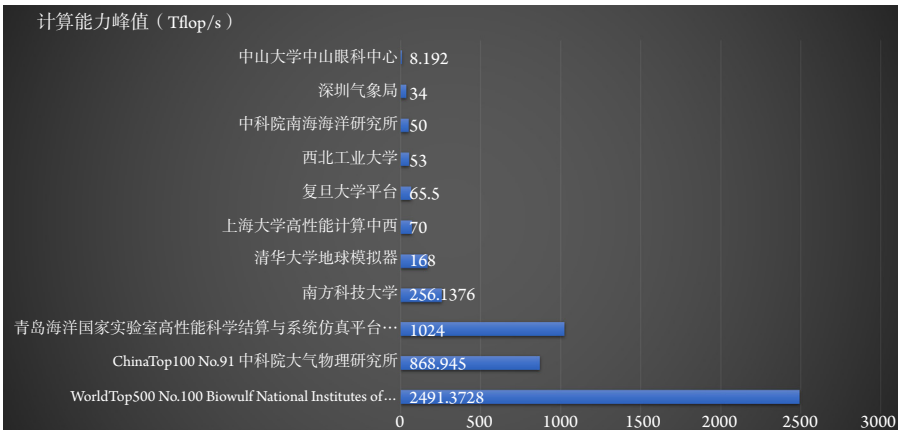

图3 国内外知名中小型超级计算平台计算能力对比图(数据来源：互联网2018年11月)

Figure 3 The comparison of domestic and foreign small and medium-sized super computing platform computing power (source: Internet in November 2018)

表4 常用计算软件列表

Table 4 Commonly used computing software list

软件类别	软件名称
比对软件	Blast，blat，lastz，BWA，SMALT，Bowtie，Mummer，Clustalw，Muscle，SOAP，MAQ，Genewise
基因组组装软件	Newbler，SOAPdenovo，ALLPATH-LG，Velvet，CAP3，ABySS，SSPACE，SOPRA，MIP scaffolder，Opera
转录组组装软件	Trinity，Cufflink，SOAPdenovo-trans
基因预测	AUGUSTUS，EVM，PASA，Glimmer，GenScan，GeneMark
基因注释	Blast2GO，Interproscan
重复序列识别软件	RepeatMasker，PILER，TEClass，TRF
进化分析软件	STRUCTURE，MEGA，PHYLIP，PAML
多态性检测	SOAPsnp，Samtools，MAQ，GATK

2.4 临床数据中心概况

在医学领域上的科学研究是依托于良好的试验设计方案及研究方法上开展的，这要求临床研究者花费大量人力物力精力来进行检查试验、随访记录、收集数据、整理数据、数据分析，得出研究结果。在中心信息化不断完善及临床科研工作不断发展的形势下，如何有效利用医院信息系统中宝贵的临床数据为临床科研服务，是目前中心临床科研工作发展面临的挑战和未来发展的主要方向。

目前，中心大部分临床研究中的科研工作处于原始的手工整理阶段，没有一套统一的科研管理工具去实现和规范临床科学数据的信息化管理。大多数研究者依然在使用EXCEL文件等简单文本处理的方式获取和整理数据，再利用相关统计软件进行分析，并且手工录入数据产生的的低效率和高错误率，会数据时效性和准确度，极大的影响临床科研的工作进程。通过对现有的数据业务进行统计发现，眼科辅助检查仪器产生的大量原始数据(图4)，因存储空间不足和缺乏有效的科学数据管理手段等原因，这些数据并不能直接为临床科学研究提供帮助，通过在原有仪器设备上进行人工分析从而获得结构化数据的落后模式普遍存在，采用移动硬盘等方式保存原始数据而造成数据意外丢失的情况并不罕见。眼科检查仪器数量多，种类杂，中心现有的PACS(Picture

Archiving and Communication Systems)系统只能通过DICOM(Digital Imaging and Communications in Medicine)接口获得报告，却无法将报告中的临床数据转换成符合临床科学数据规范的格式进行保存。各仪器厂商软件数据标准的不统一，接口标准的不统一和各自为政的服务模式，注定需要建立一个能够实现临床科学数据符合标准的互联互通管理平台，通过建立临床科学数据管理平台，提高和促进眼科临床科研工作效率和信息化发展。

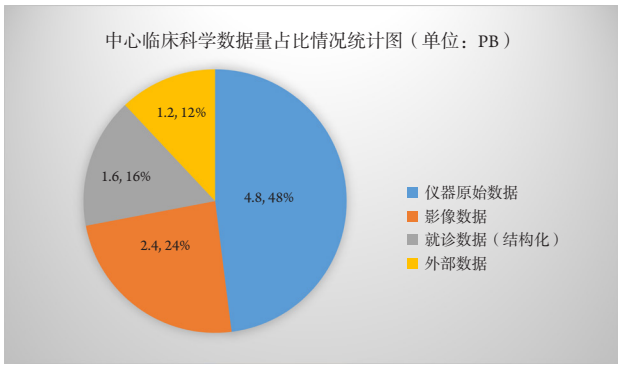

图4 中心临床科学数据量占比情况统计图

Figure 4 The percentage of clinical science data statistical analysis

3 眼科学数据中心管理体系建设

随着大数据时代的到来和开放存取运动的蓬勃发展，科学数据管理和共享逐渐成为学术界的重要命题。各高校通过成立跨部门、多学科的专门机构，实现了对科研数据进行集中建设，从而更好地实现了科研数据的共享[2]。在国际上，美国联邦政府已通过统一的数据开放平台data.gov向社会开放数据库，此外还通过开源程序平台GitHub启动在线开放数据工程(Open Data Project)共享各种面向政府开放数据库的创新解决方案[4]。欧盟也将基础设施建设作为推进欧洲科学数据共享的重点。欧盟第七框架计划、地平线2020计划资助了多个科学数据共享的基础设施项目，主要包括EUDAT(European data infrastructure)、DASISH项目(Date Service Infrastructure for the Social Sciences and Humanities)和DwB项目(Data Without Boundaries，数据无国界)。亚太眼科学数据中心的建设符合世界主流与国家当前信息化发展形势和要求，其意义将影响深远，积极探讨为推动亚太眼科学发展有着重要的现实意义。

3.1 眼科学数据管理理论体系

科学数据管理的目的是为实现更大程度和范围的共享，从而推动科研发展。为促进科学数据共享，国家科学数据管理公共政策、国家或科学团体层面共享机制的实践、科学数据出版、科学数据管理人力资源等因素从不同角度共同构建科学数据管理体系[3]。眼科学数据管理的目标是实现国家级，乃至亚太范围的眼科学数据共享服务，从宏观层面的眼科学数据管理体系还包括眼科学数据的商业出版与机构产出的发布和眼科学数据人力资源的培养等(图5)。

3.1.1 人力资源和组织规划

科学数据的管理对应信息资源管理过程涉及到采集、整合、存储、元数据、保存和分析等[3]。科学数据管理者是科研者本身与科学数据管理专业人员两者的有机结合，两者密不可分，承上启下，而科学数据管理专业人员是科学数据管理人力资源建设的主体。科学数据管理专业人员首先应具备数据管理专业能力，日常主要负责科学数据的保存和管理、与科研人员的数据沟通以及科学数据中心基础设施维护等工作，专业素质方面需要具备信息技术，尤其是数据库、信息系统项目管理和网络安全等专业技能。而作为眼科学方面的科学数据管理专业人员，还应了解生物信息学(bioinformation)所涵盖的生物学、医学、行为与健康学等专业领域。从目前国外主流的数据管理和专业培养发展方向来看，“data curation”已经成为信息科学及图书馆学专业教育的一个重要方面[3]。科学数据管理员(数字图书馆员等)核心重点业务技能侧重在数据内容管理，包括元数据、数据保存、信息数据格式标准和规范、检索策略等工作技能。

结合眼科中心的组织建设和科室设置等实际情况，应当将现有的图书管理、科学出版、大数据及信息统计等相关职能部门和专业从业人员，系统的整合为科学数据中心管理部门，全面开展眼科学科学数据管理工作，与科技和科教等相关部门形成合力，探索建设适应亚太眼科学数据中心发展的管理和运行模式。

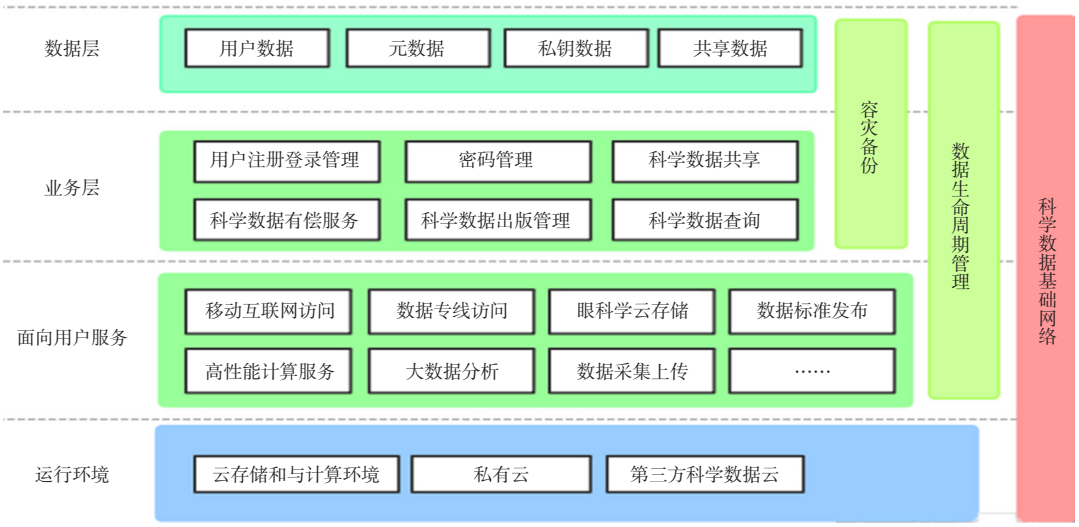

图5 亚太眼科学数据管理理论体系架构图

Figure 5 Asia-pacific ophthalmology data management theory system architecture

3.1.2 眼科学大数据生命周期管理

高校科研活动具有大科学背景下的“当代小科学”以及教学科研相结合的特点，高校科学数据具有以课题组为“根部”的树状结构和具有非规划性与庞杂两大特征[3]。目前眼科学数据亦如各高校科学数据一样，各个眼科学科的科学数据和临床数据记录组成相关的科学数据集合，各个科学数据集合归属于某个课题组或学科带头人，学生和导师的科学数据形成对应课题的数据集合，绝大多数科研活动和课题研究开展前并没有制定相应的科学数据管理计划，科研活动过程中科学数据散落在个人的计算机和移动存储中，“个性化”的管理数据模式不仅仅是无法对眼科学数据实现共享与再利用，还存在经常出现数据丢失的情况，互相私下拉关系求数据已成常态。

目前，眼科学数据类型主要有观测型数据(Observational data)、计算型数据(Computational data)、实验型数据(Experimental data)、记录型数据(Records)和特殊的临床数据(Clinical data)几种。截止2012年4月，世界范围内已有55个不同的科学数据管理生命周期模型建立[4]。

3.1.3 眼科学数据质量评价体系

科学数据一般是由科学试验和理论计算获得，在其产生过程中，试验的原理或方法、使用的试验设备、原始信息采集与记录仪器的精密度、数据分析的理论方法成熟度、试验标准、内外部的环境条件以及数据产生过程中的人为因素，均会对数据产生过程客观性、真实性和准确性产生影响[5]。科学数据质量存在于数据的整个生命周期，与其全寿命周期的任何一个环节有关，从数据生产的全过程考虑，眼科学数据从生产、收集、加工整理、存储和发布等环节、步骤都会产生质量影响，所以建立眼科学数据质量评价体系(图6)，以监控和把握数据从产生到使用的每一个阶段的数据质量水平。

3.1.4 科学数据出版

随着科学数据与科学文献的关联性加强以及科学数据逐渐成为科研对象而不仅仅是科研最终结果，科学数据出版将成为推进学术交流与共享的重要方式[3]。眼科学数据出版的最终目的就是“一切为了数据”，眼科学各学科科研项目众多，各学科的项目数据量有大有小，眼科学术出版部门既要着眼大数据存储管理，又要兼顾小数据精细化管理，大挖掘的宏观分析和小挖掘的微观研究都要做好，要从数据出版的源头元数据管理工作开始，到最后数据出版和提供数据服务，最终实现“内容出版”向“知识服务”的转型。在大力推进眼科学电子刊物出版的同时，要尽早建立眼科学数据论文唯一标识符(digital object identifier，DOI)并与眼科学实体数据及元数据对应绑定，实体数据的引用格式符合国家标准《信息技术科学数据引用》(GB/T35284-2017)相关规定[6]。

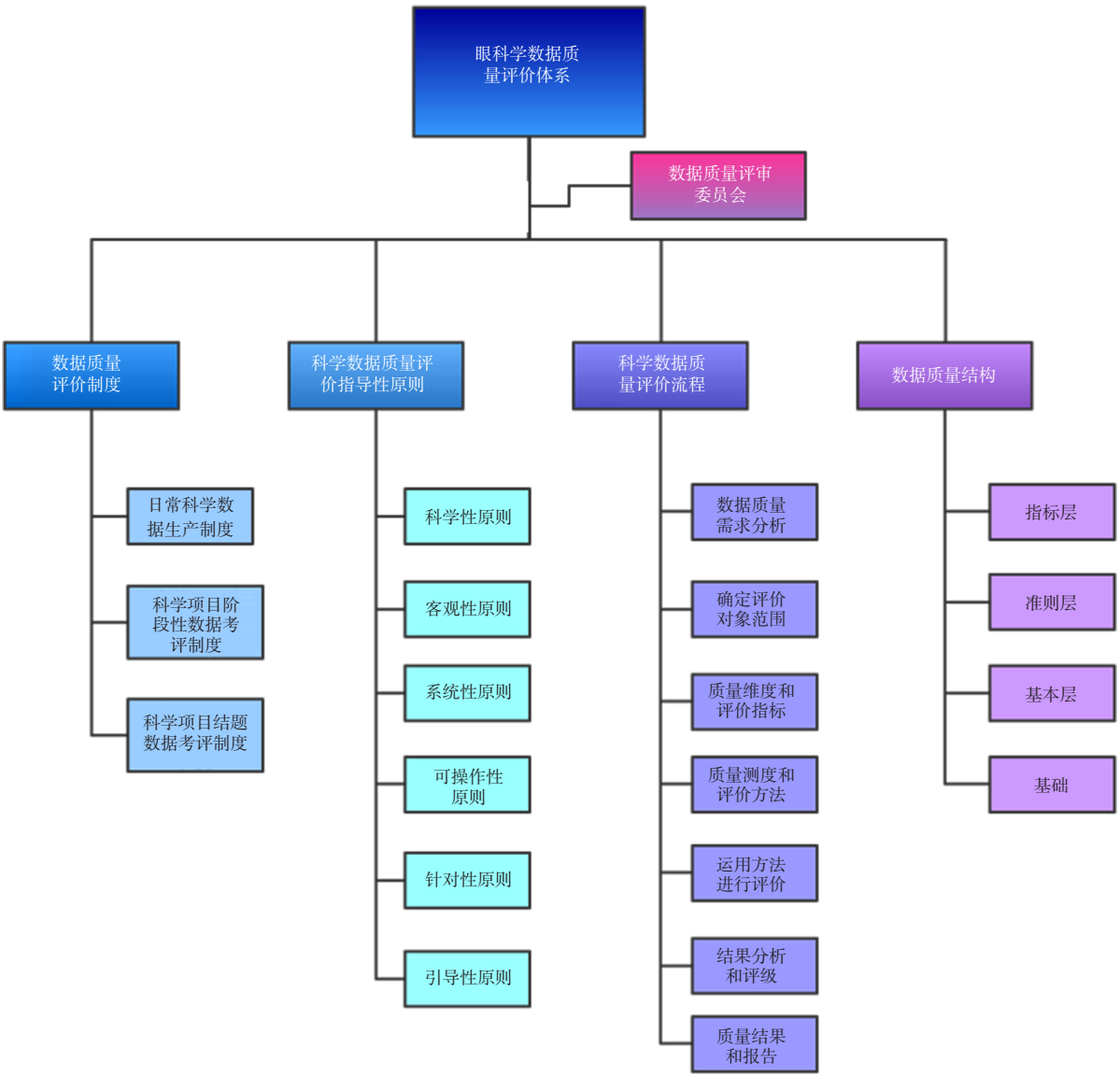

图6 眼科学数据质量评价体系示意图

Figure 6 Ophthalmology data quality evaluation system

目前中心眼科学数据出版人才紧缺，要实现亚太眼科学数据中心建设和完成出版任务，需要引进和培养数据采集人才、算法师、分析师、优秀数据论文评审人员，建立眼科学数据论文审稿人专家库。在无法及时取得人才的情况下，还要灵活运用服务外包等手段以取得智力支持。

3.2 眼科学数据中心管理平台建设

眼科学数据中心管理平台建设的目的，是为了更好的实现眼科学数据共享，有效开发和利用科学数据，消除“数字鸿沟”，让各眼科学科发展形成合力，从小学科精细化研究跨越至大数据宏观发展。平台建设要充分考虑眼科学科研人员自身的数据素养、各学科科研机构的共享氛围和国家制定的相关法规政策及管理制度等各方面因素，通过平台规范数据生产、存储和利用，建立眼科学数据分级安全管理和合理有序的审批共享机制，使眼科学数据管理符合国家标准。最终通过平台管理形成眼科学数据资产，通过对外数据共享、数据查询和数据分析等手段，探索眼科学

数据有偿服务模式，实现眼科学数据价值和产值等利益最大化，从而激发和提高科研人员对科学数据的共享意愿。

3.3 临床数据管理平台建设

中心临床数据管理信息平台需要将原有系统的数据进行分析和梳理，在物理汇聚的基础上，遵循行业或卫生部技术标准所定义的数据模型重新进行建模，按照以患者为中心的方式对临床业务数据进行集中管理。目前临床科研数据管理中心迫切需要通过建立统一的数据接口，将各个系统和仪器设备中的数据进行整合汇总，即将所有患者的医疗结果和其他临床数据集中存储并作为一个中心存储仓库，持续性地形成一套完整的可供临床研究使用的患者临床信息数据，并关联到中心其他各临床业务系统数据库中去(图7)。中心临床数据管理信息平台建设核心就是以患者为中心，实现重新构建新的一层数据存储结构这一必要条件。为下一步更高效地利用临床科学数据，夯实临床数据基础，必需加快基于眼部单病种大数据库的临床科研数据平台等相关数据平台建设工作(包括单病种大数据库建设、眼部检查图像分析、临床科研项目数据管理、临床科研项目管理)。

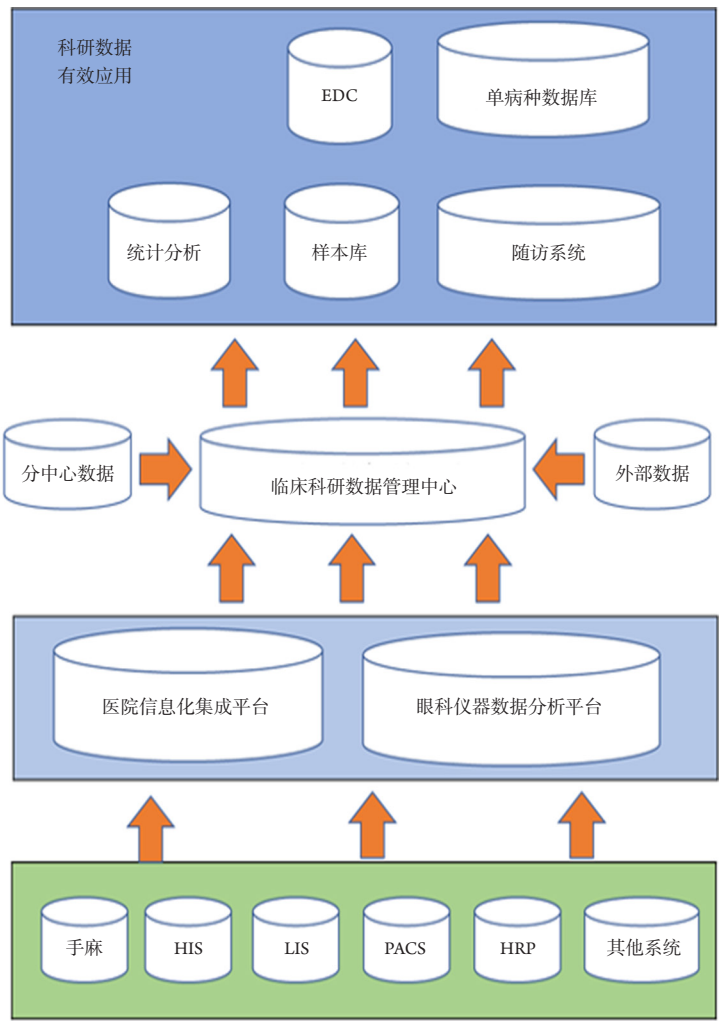

图7 亚太眼科临床科学数据管理平台架构示意图

Figure 7 Ophthalmology clinical science data management platform architecture

3.4 高性能计算平台建设

2015年3月11日，科技部召开国家首次精准医学战略专家会议，并决定在2030年前政府将在精准医疗领域投入600亿元，其中中央财政支付200亿元，企业和地方财政配套400亿元。2015年2月，习近平总书记批示科技部和国家卫生计生委，要求国家成立中国精准医疗战略专家组，共19位专家组成了国家精准医疗战略专家委员会。2015年1月20日，美国总统奥巴马在国情咨文演讲中启动精准医疗计划。一期投入2.15亿美元。奥巴马“精准医疗”计划的第一步是希望招募100万名志愿者进行基因组测序。精准医疗计划将推动精准医疗向临床实践提供科学依据[7]。

随着国家精准医疗战略的全面启动，以及生物基因测序等新技术的飞速发展，眼科学精准医疗也将以大数据为基础，以基因测序为工具，为特殊眼科疾病和特定病人研究出更具针对性的治疗方法。眼科大数据是精准医疗实施的基础，必须通过建立适应眼科学发展研究的高性能计算平台，为眼科学基因组测序、眼科学生物医学分析和大数据分析工具提供高性能运算和驱动能力，才能够不断的推动眼科学精准医疗快速发展。

3.5 数据安全管理体系

眼科学数据安全管理体系的第一要务是保证目前亚太眼科学领域的科学数据安全，要尽早建立眼科科学数据应急管理和容灾备份机制，并使之符合《中华人民共和国网络安全法》第三十四条第三款规定要求[8]，即“对重要系统和数据库进行容灾备份”。同时，还要根据《科学数据管理办法》第二十八条和第二十九条[1]，建立完整的亚太眼科学数据网络安全保障体系，灵活运用各种安全手段和采用安全可靠的产品及服务，完善数据管控、属性管理、身份识别、行为追溯、黑名单等管理措施，健全防篡改、防泄露、防攻击、防病毒等安全防护体系，积极研究和探索使用区块链等安全新技术、新手段，在科学数据管理上实现去中心化、可溯源、防篡改、隐私保护等功能。

4 科学数据中心建设发展规划前景展望

亚太眼科学数据中心的起步可以算是从零开始，人、财、物的基础性建设投入力度需要足够大，平台的建设目标才能定的足够远。私有云和公有云的建设规模，各科学云间互联的安全性标准，以及科学数据平台的管理能力和制度建设都急需进行相关的可行性研讨。眼科科学数据中心建设要走可持续发展的道路，就必须实现科学数据共享、大数据挖掘和数据出版等有偿服务的盈利模式，其中科学数据出版提供了一个很有前景发展方向。科学数据中心建设是国家信息化发展战略的一个重要组成部分，应抓住有利政策时机，尽早启动眼科科学数据信息化工作。

参考文献

1. 国务院办公厅. 国办发〔2018〕17号《科学数据管理办法》[Z]. 2018-03-17.
   General Office of the State Council. "Scientific data management solution"[2018] no. 17[Z]. 2016.
2. 刘兹恒, 曾丽莹. 我国高校科研数据管理与共享平台调研与比较分析[J]. 情报资料工作, 2017(6): 90-95.
   LIU Ziheng, ZENG Liying. Investigation and Comparative Analysis of Scientific Research Data Management and Sharing Platform of Universities in China[J]. Information and Documentation Services, 2017(6): 90-95.
3. 钱鹏. 高校科学数据管理研究[D]. 南京: 南京大学, 2012.
   QIAN Peng. A study on scientific data management in university[D]. Nanjing: Nanjing University, 2012.
4. 杨林, 钱庆, 吴思竹. 科学数据管理生命周期模型比较[J]. 中华医学图书情报杂志, 2016, 25(11): 1-6.
   YANG Lin, QIAN Qing, WU Sizhu. Comparison of scientific data management life cycle models[J]. Chinese Journal of Medical Library and Information Science, 2016, 25(11): 1-6.
5. 李正.科学数据质量水平的分析与评估[R]. 中国科学数据大会, 2014.
   LI Zheng. Scientific data analysis and evaluation of quality level[R]. China science data conference, 2014.
6. GB/T35294—2017. 中国科学院计算机网络信息中心. 信息技术

科学数据引用[S].
GB/T35294-2017. The computer network information center, Chinese Academy of Sciences. Information technology science data references[S]. 2017.
7. The Precision Medicine Initiative Cohort Program – Building a Research Foundation for 21st Century Medicine[EB/OL]. https://acd.od.nih.gov/documents/reports/DRAFT-PMI-WG-Report-9-11-2015-508.pdf.
8. 民政部信息中心. 中华人民共和国网络安全法[Z]. 2017-05-15.
The Ministry of Civil Affairs information center. Network security law of the People's Republic of China[Z]. 2017-05-15.

本文引用： 陈有艺, 苏超, 李扬杵. 亚太眼科学数据中心建设初探[J]. 眼科学报, 2020, 35(1): 33-42. doi: 10.3978/j.issn.1000-4432.2019.09.06
Cite this article as: CHEN Youyi, SU Chao, LI Yangchu. Construction of Asia-pacific ophthalmology data center[J]. Yan Ke Xue Bao, 2020, 35(1): 33-42. doi: 10.3978/j.issn.1000-4432.2019.09.06