



开放科学
(资源服务)
标识码
(OSID)

上海科学数据管理和共享需求分析及对策

普丽娜 殷晓 谢文娴

上海市研发公共服务平台管理中心(上海市科技人才发展中心) 上海 200031

摘要: [目的/意义] 为全面了解上海科学数据资源现状, 以及不同利益主体对上海科学数据管理与共享政策的需求。[方法/过程] 从数据治理视角, 遵循信息资源全生命周期规律, 设计问卷内容, 面向各利益相关者开展调研。[结果/结论] 总结科学数据全生命周期各阶段不同利益相关者的需求特点, 提出上海科学数据管理与共享政策建议。

关键词: 科学数据; 数据共享; 数据全生命周期; 科学数据政策

中图分类号: G35; TP391

Demand Analysis and Countermeasures of Scientific Data Management and Sharing in Shanghai

PU Lina YIN Xiao XIE Wenxian

The Administration Center of Shanghai R&D Public Service Platforms, Shanghai 200031, China

Abstract: [Objective/ Significance] In order to understand the current situation of scientific data resources in Shanghai and the needs of stakeholders for scientific data management and sharing policies in Shanghai. [Methods/Process] The author designs the survey questionnaire for various stakeholder, from the perspective of data governance, following the law of the whole life cycle of information resources. [Results /Conclusions] This study summarizes the demand characteristics of different stakeholders in each stage of the whole life cycle of scientific data, and put forward policy suggestions on scientific data management and sharing in Shanghai.

Keywords: Scientific data; data sharing; data lifecycle; scientific data policy

基金项目 上海市软科学研究计划项目“加快科学数据立法推进上海综合科学数据中心建设研究”(18692117900); 上海市2021年度“科技创新行动计划”软科学计划项目“面向共享的上海科学数据资源集成发现体系及保障机制研究”(21692109100)。

作者简介 普丽娜(1981-), 博士后, 高级工程师, 研究方向为数据治理与知识组织, 科技创新资源规划与管理, E-mail: lnpu@sgst.cn; 殷晓(1987-), 硕士, 工程师, 研究方向为科技创新资源管理与研究; 谢文娴(1987-), 硕士, 工程师, 研究方向为科技创新资源管理与研究。

引用格式 普丽娜, 殷晓, 谢文娴. 上海科学数据管理和共享需求分析及对策 [J]. 情报工程, 2021, 7(6): 88-100.

引言

科学数据作为重要的科技战略资源，其开放共享是更大限度挖掘和发挥科学数据价值的重要方式之一。科学数据共享的本质是让数据资源能够在利益相关者之间合理有效配置^[1]，伴随着信息化技术的进步，科学数据的发展越来越表现出多学科领域的交叉融合、多主体参与和大规模协作的特点，科学数据的治理和开放共享制度的制订需要更加重视和协调科学数据各利益主体的权责关系，以更好细分确认数据的价值^[2-4]。与此同时，科学数据从产生到消亡的全生命周期^[5]，具有不同阶段和利益主体特点^[6,7]，因此科学数据共享和管理政策的制定，还应结合科学数据全生命周期特征，充分考虑多元利益相关者的权益，作出合适的制度安排。

我国在2018年3月发布了首个国家层面的《科学数据管理办法》(国办发〔2018〕17号)，随后，十余个各省、市、自治区以及中科院积极响应，并出台实施细则。上海是全国较早启动科学数据开放共享工作的省市，早在2002年就建设“一网两库”，推进科学数据开放共享，对推动上海市研发活动起到了积极作用。随着经济社会的发展、科研组织方式发生了重要变革，对科学数据资源管理、共享、利用的目标和需求有了新的变化，产生了新的特征。本研究立足上海市科学数据工作现状，深入调研科学数据管理与利用主体，全面了解掌握新形势下研发活动对于科学数据的现实需求，总结梳理科学数据管理与共享中急需解决的主要矛盾

和问题，为上海落实国家《科学数据管理办法》，推进科学数据更高效的管理与利用提供参考。

1 科学数据管理与开放共享需求分析框架

科学数据管理是科学数据高效利用的重要基础，科学数据共享是科学数据高效利用的重要方式。国内外学者根据科学数据资源特征和管理实践提出了数据全生命周期理论，并提出了一些经典的科学数据管理全生命周期模型^[8-10]，可以看到，在科学数据管理的过程中，数据采集、数据保存、数据利用等是核心环节^[7]，同时在生命周期不同阶段需要采用不同的管理策略^[11-13]。

科学数据共享本质上就是不同角色围绕科学数据这一资源，从满足数据用户需求出发，利用各种方式实现科学数据从数据提供者到数据用户对于其拥有和使用权利的让渡，达到科学数据资源满足科研活动需求的状态^[14]。由于，科学数据生产和利用的跨区域、跨部门、跨学科的大协作，参与主体的规模和类型越来越多元，科学数据共享利益相关者不仅包括数据提供者和数据用户，还包括研究人员、研究机构、数据中心、用户、资助者、出版者等^[15-17]，共享和管理政策的制定更聚焦到这些多利益相关者之间权益的确认和平衡。

为深入了解上海市科学数据管理利用需求，本研究基于科学数据全生命周期理论和利益相关者理论^[18-23]，构建了科学数据管理与开放共享需求分析框架(详见图1)。

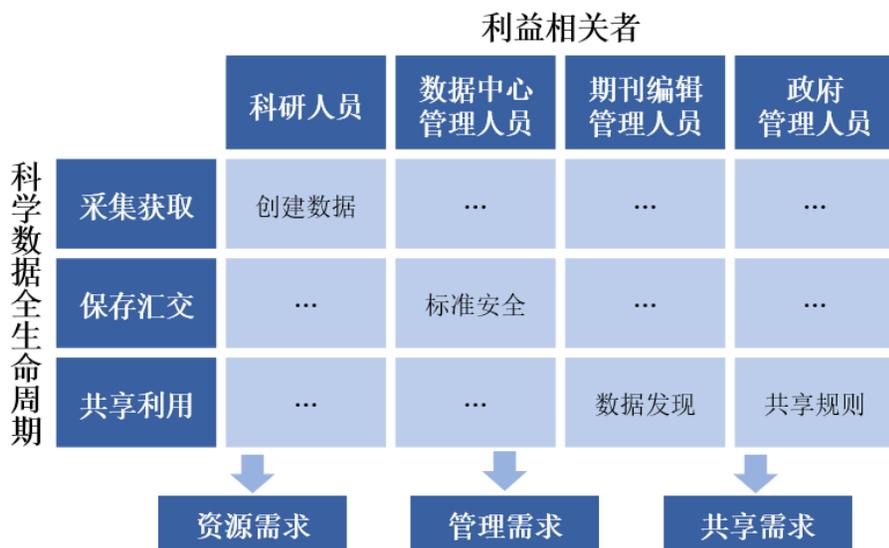


图1 科学数据管理与开放共享需求分析框架

在获取相关需求的过程中，重点抓住科学数据全生命周期中采集与获取、汇交与保存、共享与利用等重要阶段，以这些阶段的主要利益相关者对科学数据共享和管理政策制定的需求和诉求为目标，针对科学数据开放共享中涉及的主要利益相关者，设计了面向科研人员（数据使用者和利用者）、数据中心管理人员（传播者）、期刊编辑管理人员（出版者和传播者）、政府管理人员（管理者和组织者）的调研内容，主要包括：

- （1）科学数据采集和获取的方式、途径、可能遇到的困难等。
- （2）当前从事工作的科学数据保存与管理现状、可能遇到的困难，以及原因。
- （3）对科学数据共享及现状的态度，包括：是否从其他途径获得过科学数据、是否向其他人提供过科学数据，以及对科学数据共享的动机、态度和意愿。
- （4）对科学数据管理与共享政策的认知、

态度与需求，包括：是否赞成制定科学数据共享政策，对政策的担心、顾虑，以及建议政策应包含的内容等。

- （5）对科学数据管理平台、科学数据中心（库）布局建设的认识、需求和建议。

2 科学数据管理与开放共享需求调研

本研究面向上海市科研人员、数据中心管理人员、期刊编辑管理人员、政府管理人员开展问卷调查，为保障问卷调查的有效性，对重点关注的调研单位开展了预调研，并根据预调研情况完善问卷，通过上海科技、上海市研发公共服务平台的官方网站、官方微信公众号发放了问卷，开展了较为广泛的调查。本次问卷调查时间为2019年3月28日—2019年4月27日，收回有效问卷571份。

从问卷回收情况来看，上海科学数据生产、

利用的参与主体呈现出多样化特征，高校、科研院所和医院最多，接近一半的调研对象来自高校，其次是科研院所（23.1%）和医院（9.3%），同时各类企业也积极参与到科学数据的应用中，成为不容小觑的重要力量。

从具体调研对象来看，以26~40岁（含40岁）青年人员为主，其中科研院所这一年龄段的比例最高（76.5%），其次是医院（70%），企业（65.2%），高校（59.1%）。这一年龄阶段的从业者，有活力、有拼搏的动力，也积累了相当的科研/管理经验，是承担和推动科学数据

工作的重要力量。

从调研对象的学科分布来看，工学、理学、医学最多，而管理学、哲学、经济学、法学等人文学科参与较少，这与各学科产生、利用科学数据的关联程度基本吻合，自然科学和工程学科是大规模产生、利用科学数据的主要学科。

从调研对象的工作性质看，主要从事科研（85.9%），教学（22.2%）和管理（22.4%）工作，受访者身兼多职的情况比较普遍，其中医院和高校最为显著，科研院所和企业受访者的岗位设置则较为固定（详见图2）。

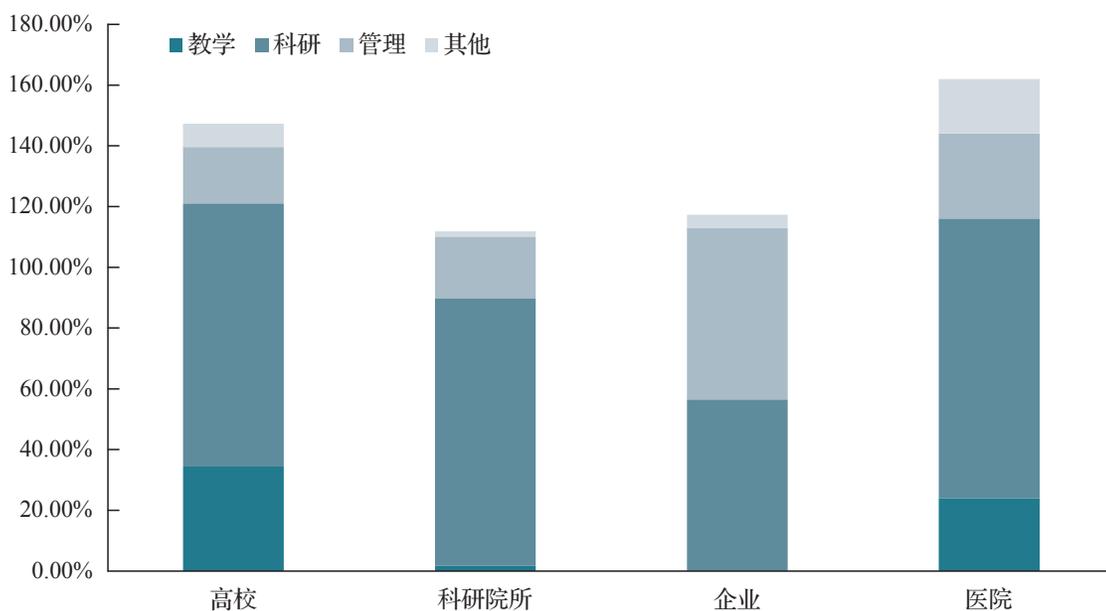


图2 单位性质与调研对象工作性质

从调研对象身份来看，本次调查主要包括七类人员：教师、学生、研究人员、图书管理员、实验人员、科研管理人员、医生等。其中，占比较多的人员是教师、学生和研究人员。

总体而言，从科学数据管理与共享利益相关者年龄、学科、工作性质分析，上海开展科

学数据生产和应用的机构中，科研院所基础要更好，集聚了更专业的从业人员、从业人员普遍比较年轻，形成了更有活力的有生力量，而且科研院所的科研活动主要集中在科学数据普及推动较好的自然科学领域，也更容易积累和推进科学数据的使用。同时，科研院所更专业

化和细分的组织管理方式使明确的数据管理岗位设置成为可能,更有利于科学数据管理工作的专业化、规范化。

3 上海市科学数据管理与开放共享需求分析

3.1 科学数据采集获取

目前科研人员生产数据的方式多种多样,主要有收集后加工处理(65.9%),人工采集(61.2%),仪器设备自动采集(54.9%)几种方式,

而且科研人员通常采用多种方式生产数据。从采集方式来看,主要以机构/课题组自主生产为主(占77.7%),科学数据由公开渠道免费获得的也占有较大比例(占47.8%),进而导致目前科学研究中使用的数据主要通过小范围的共享来实现,科学数据共享的意识需进一步提升。

不同学科领域的科研人员收集获取科学数据的主要方式有所不同,工学、理学获取方式更为多样化,以机构/课题组自主产生、公开渠道免费获得数据为主;医学、农学以机构/课题组自主产生为主;管理学以公开渠道免费获得数据为主(详见图3)。

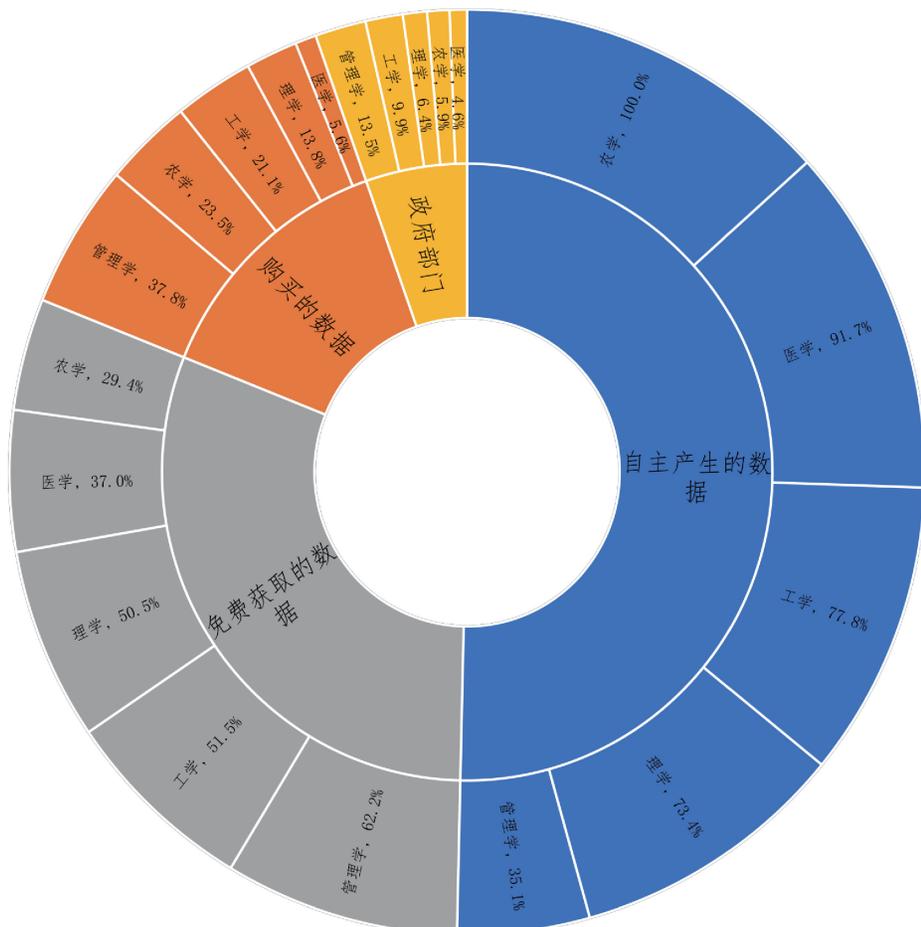


图3 不同学科领域获取数据的方式

科研人员获取数据过程中的主要困难包括：数据质量参差不齐（63.9%），没有统一的科学数据平台（50.9%），数据收集成本或获取成本太高（48.9%），公开发表文献中找不到原始数据（47.8%）等（详见图4），这与数据中心管理人员认为数据获取描述中数据质量及控制方式（完整性、准确性）是科研人员获取数据时最重要的参考标准（详见图5）的观点不谋而合。

进一步调查研究发现，高校、科研院所、医院认为数据获取的最大困难是数据质量参差不齐

不齐，而企业认为最大的困难是没有统一的科学数据平台。究其原因，高校、科研院所、医院以科研为主，科研合作对象较为固定，获取的科学数据的领域相对较集中，在长期积累过程中形成了一批领域科学数据中心，而企业需要通过数据共享获得技术研发新动态，促进科技成果转化和项目落地^[1]，需要的数据所涉及的领域更加交叉综合，没有明确的合作对象，但目前能满足跨学科、多领域需求的数据发现平台目前还较为缺乏。

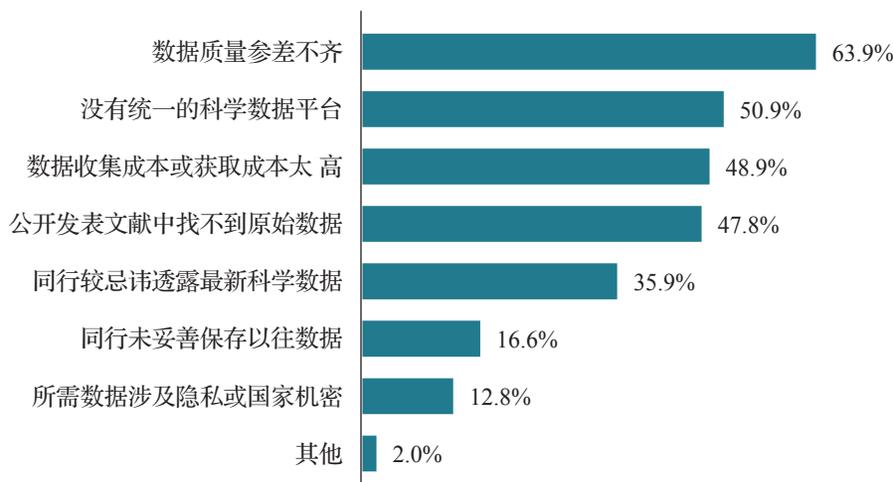


图4 科学数据获取的主要困难



图5 数据获取的重要指标

3.2 科学数据汇交保存

从数据存储方式来看,首先上海科研人员科学数据存储的方式分散,汇交到统一公共平台的还较少,主要是存储在自己/团队的计算机中(89.5%);其次是U盘或移动硬盘(57.4%),图6可以看出目前科研人员主要采取的存储方式较为分散,存在数据丢失的隐患,进一步调研发现近一半的科研人员发生过科学数据的丢

失情况。从丢失数据的科研人员年龄分布来看,26~40岁(含40岁)数据丢失的比例最高,同时这一年龄段也是科研高产出的年龄段,数据的丢失会对科研人员持续的科研活动产生不利影响,增加重复科研的可能性。从数据丢失的方式来看,发现造成数据丢失的主要原因是存储设备故障(占比75.6%),其次是误操作或误删除(占比48.8%),第三是计算机病毒(25.9%)。

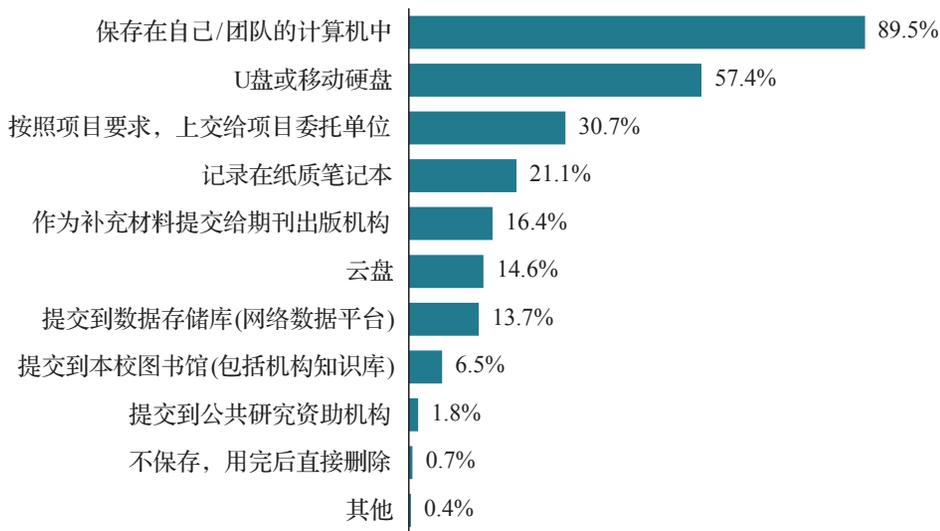


图6 科学数据存储方式

在数据保存的期限方面,55.8%的科研人员会将数据保存5年以上,27.6%的科研人员会将数据保存3-5年。从机构视角,92%的医院科研人员会将科学数据保存3年以上,70%的医院科研人员保存期限达到5年以上;高校、科研院所、企业的数据保存3年以上的均超过80%。由此可以看出,各主体对于科学数据长期保存的需求较为强烈。

从数据保存的规模来看,上海科学数据存储较为分散,数据规模较小,目前上海课题组/个人每年产生的原始数据总量主要以Gb级(41%)

的为主,其次是500Mb-1Gb(22.9%),再次是Tb级(占15.5%)。工学、理学、医学是产生科学数据较多的领域,且其TB级的数据总量占本领域的比例均达到了20%左右,超过了平均水平。在医学领域,超过70%的科研人员会将数据保存5年以上。从机构来看,各机构还是以GB级的数据存储总量为主,TB级数据总量最多的主体依次是企业(21.7%)、科研院所(18.5%)。

从数据备份情况来看,科研人员 and 期刊编辑人员对于数据备份有较强的意识和需求(详

见图 7 和图 8)，87.7% 的科研人员会对科研活动中产生和收集的数据进行备份，尤其以高校的科研人员意识最强，达到 92.3%；进一步对备份方式调研，发现首选备份到课题组 / 个人备份设施(占 62.4%)，其次是定期磁带的备份方式(占

30.9%)。期刊方提供的科学数据库是目前科研人员存取科学数据的主要方式之一^[24]。期刊编辑人员首推将论文数据递交至学科领域内的专业数据中心(占比 71.1%)，其次是政府管理部门或资助机构建立的数据平台(占比 55.6%)。

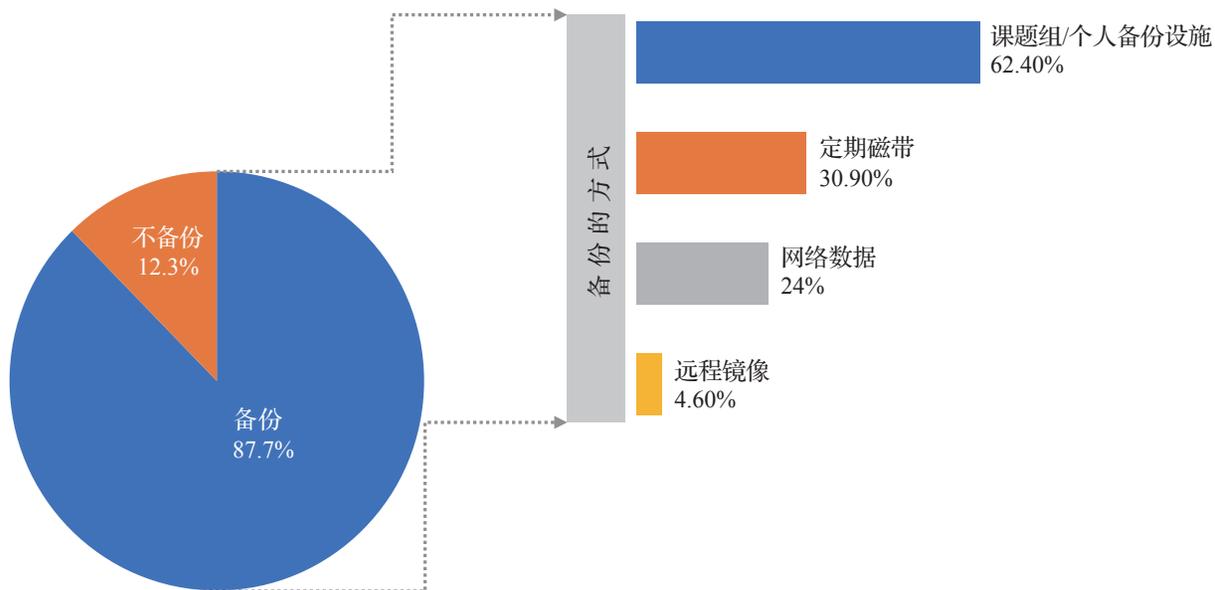


图 7 科研人员数据备份情况及备份方式

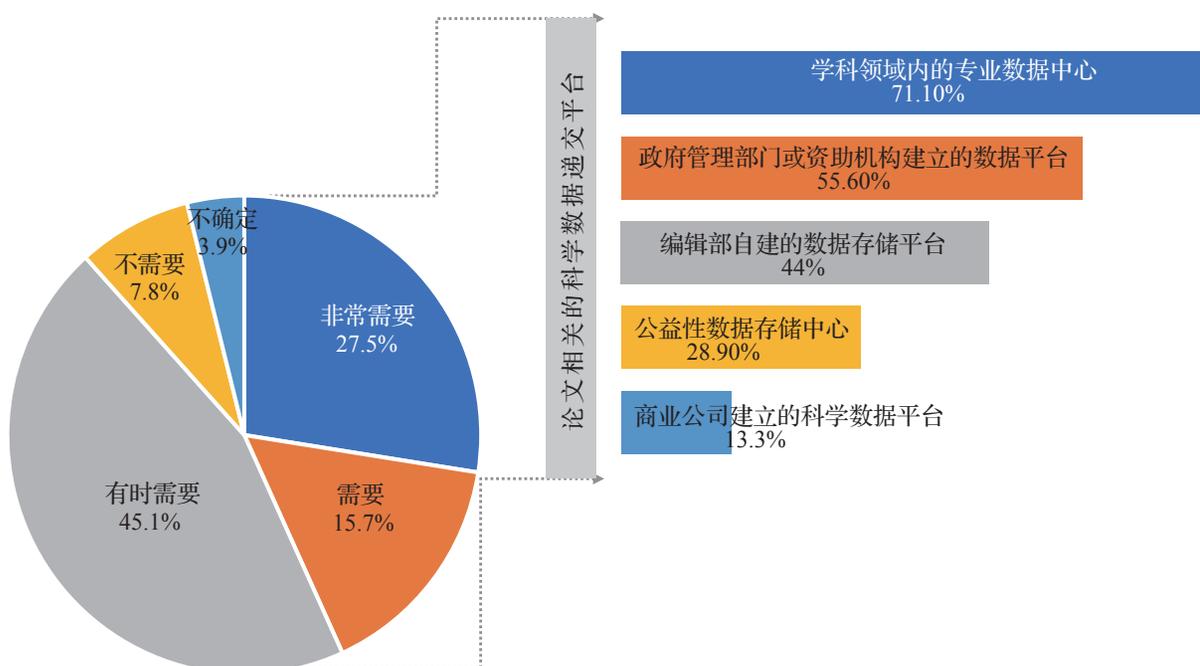


图 8 期刊编辑人员认为集成发现的必要性和对集成发现平台的建议

从科学数据建设主体来看, 科研院所和企业的科学数据库建设已超 30%, 高校、医院的数据库建设率虽只有 15% 左右, 但这些机构都充分认识到科学数据库的重要作用, 绝大多数已在积极筹建科学数据库。

3.3 科学数据共享利用

从科学数据共享意愿来看, 上海科研人员具有很强的科学数据共享与利用的意愿(详见图 9), 94.8% 的科研人员愿意或者征得同意后愿意共享数据。从机构视角, 医院的科研人员共享意愿最为强烈, 达到了 98%, 其次是科研院所(95.2%), 不愿意共享的主要原因是出于数据安全、隐私、知识产权等方面的考虑, 以及希望后续进行数据价值挖掘等, 可以看出, 数据权益保护是科研人员共享数据的重要前提。

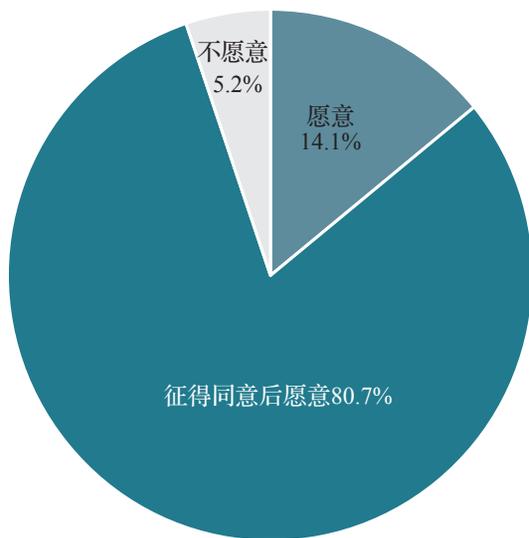


图 9 科学数据共享意愿

从科学数据政策需求来看, 各机构的科研人员、数据中心管理人员最关心的政策内容一

致, 体现了对各自数据权益的充分确权 and 保障的重视(详见图 10), 包括: 明确各个利益相关方职责(科研人员 82.3%, 数据中心管理人员 87.5%), 科研数据的归属权(科研人员 78.3%, 数据中心管理人员 83.9%), 知识产权以及科研数据的共享方式和范围的确认(科研人员 71.5%, 数据中心管理人员 85.78%)。其中, 高校更关注利益平衡和数据权属问题, 这与高校科研人员是科学数据的主要创建者, 希望通过明确共享权利义务获得合理的回报相关; 科研院所、企业、医院更希望通过政策制订明确科学数据的共享方式和范围。

从科学数据基础设施建设需求来看, 建设覆盖数据全生命周期的科学数据中心是各利益相关者共同的需求(详见图 11 和图 12), 科研人员、数据中心管理人员、政府工作人员都认为数据管理和数据存储功能是科学数据中心最基础、最重要的功能, 且对数据安全保护的需求最为突出。各主体都认为查找与获取是数据库最重要、基础的服务, 高校和科研院所的科研人员对数据价值交换和共享具有较强需求, 企业和医院的科研人员则更需要数据分析等增值服务。

在科学数据共享载体选择方面, 不同利益主体的选择具有明显的差异性, 科研人员较多选择自己信任的人进行共享, 各主体选择课题组内部的数据库最多, 高校和科研院所更多选择学科领域公共数据库, 企业和医院优先选择本机构的数据库(详见图 13)。由此可以看出, 可信任的、权威的、公益的共享载体是各类主体的共性需求, 通过此共享载体认证认可的数据, 科研人员才愿意放心使用。

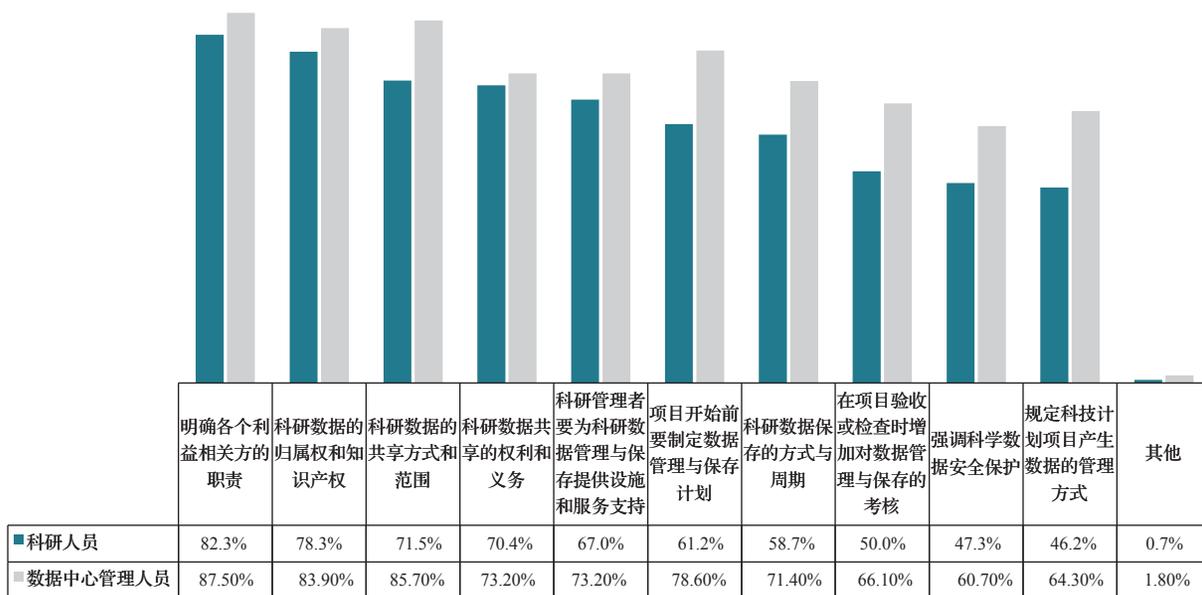


图 10 科学数据政策应包含的内容

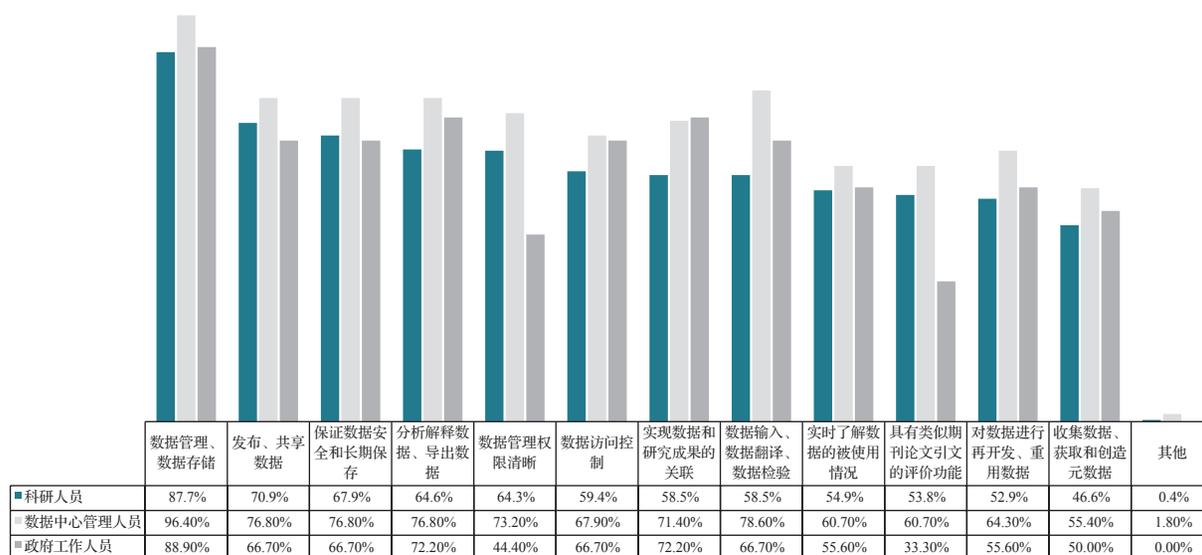


图 11 不同利益相关者对科学数据共享利用平台功能的选择

从图 14 可以看出，各利益主体对于数据中心建设功能需求各有侧重，基础性的功能包括数据浏览、查询、下载和复制功能（94.6%），其次是数据调用和分析服务（85.7%）以及数据保存、检索、统计服务（82.1%）。

通过调研发现，当前上海科学数据管理与共享已经取得重要进展，但是与各类主体和个人强烈的共享意愿和强劲的数据需求相比，

与对数据多元价值更深挖掘的要求相比，也存在一些需要不断提升的空间。从科学数据本身来看，其产生量逐年增大，参与主体更加多元复杂，但存储备份方式较为分散，数据安全得不到根本保障；公益、专业、安全的汇交平台缺乏，难以提供数据可信度和安全的有效支撑，尚未形成科学数据的开放共享合力。

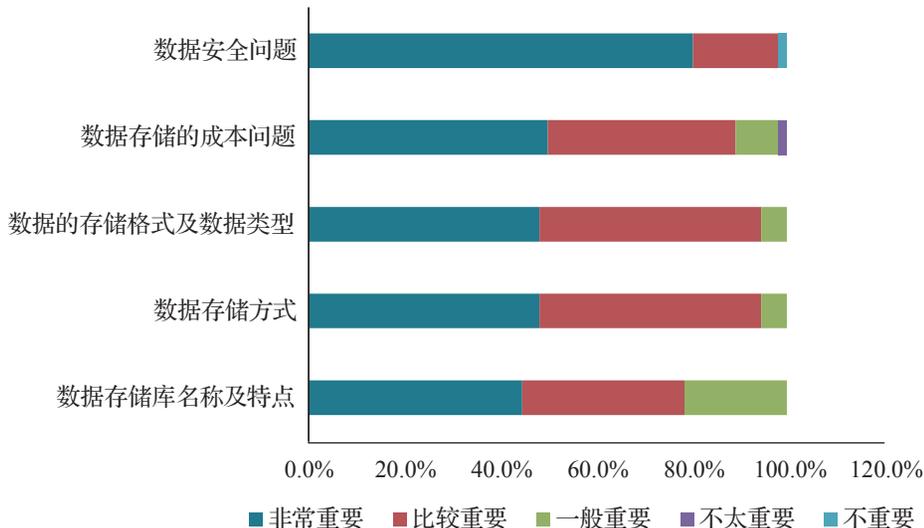


图 12 数据中心对于数据保存相关的重要性指标评价

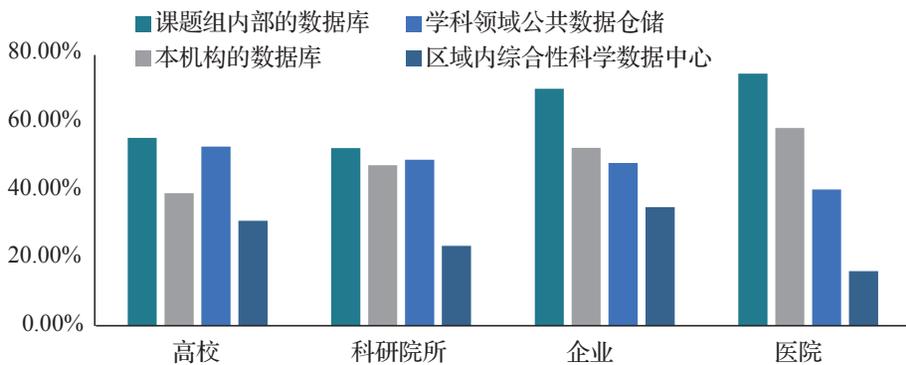


图 13 科研人员共享载体的选择



图 14 科学数据中心建设的重点

4 上海科学数据管理与共享对策建议

通过对上海科学数据管理与共享现状的调

研，我们发现上海各类主体对科学数据支撑科学研究已有较深刻的认识，需求非常强烈，已形成在各自机构和小范围的数据存储、共享使用的局面，为上海制定全市层面的科学数据共

享和管理政策奠定了良好的基础,但需要看到数据的生产和存储还很分散,数据安全和权益保障缺乏,规范的数据价值确认规则缺失,极大制约了上海科学数据价值的更好发挥。鉴于此,建议在政策制订中强化以下几个方面。

(1) 注重科学数据共治的体系化安排。重视构建多元主体对话共谋机制,以更好激发利益相关主体推动科学数据开放共享的积极性和创造性;强化政策保护和规范数据利益相关者权责;系统布局基础设施(制度、标准、技术、平台载体),倡导数据管理计划,鼓励数据出版,组织数据竞赛,鼓励各类利益相关主体先行先试推进开放共享。

(2) 充分重视科学数据的经济社会价值。科学数据作为价值最高的新型生产要素,其经济和社会的多维价值将不断放大,建议以目标为导向,打破科技、经济、社会领域边界,促进数据流动和重构,使其成为创新生产方式、重构生产关系、提升生产能级的重要要素。

(3) 加强科学数据全生命周期管理。在数据采集获取阶段,应重视数据质量保障问题,建立统一标准,强化主体责任,从源头上指导数据生产单位建立科学数据质量控制体系;在数据保存汇交阶段,要指导主体筑牢数据安全意识,确保数据得到完整和安全的保存;在数据共享利用阶段,应规范科学数据开放共享目录建设,鼓励采用在线下载、离线共享、定制服务等多种形式的共享方式,推动数据使用便利化,保障科学数据使用的可持续发展。

(4) 加强科学数据统一集成发现平台的建设。科学数据发现作为科学数据管理和价值提升的重要手段,是实现数据共享的第一步也是

最重要的一步,也是各类主体呼吁强烈的需求,应加强布局建设具有公益性、可信度、专业化的集成发现平台,作为共享枢纽,链接各类科学数据主体和数据库,推动科学数据的使用和开放共享。

(5) 大力推进科学数据中心建设。应充分发挥科学数据中心在领域内的专业作用,调动其开放共享的积极性,建议依托具有优势资源、已有较好基础的机构建设上海市级科学数据中心,授权先行形式,探索一套本领域可推广使用的科学数据管理和开放共享实践经验,形成可以借鉴参考的模式,以实践示范推动政策的落实。

参 考 文 献

- [1] 章琰,杨一图,吴健,等.我国科学数据共享运行机制模式创新探讨——以产业技术联盟为例[J/OL].科学学研究:1-21.[2021-03-29 08:30].<https://doi.org/10.16192/j.cnki.1003-2053.20210325.005>.
- [2] 梅宏.数据治理之论[M].北京:中国人民大学出版社,2020:62-66.
- [3] Mahanti R. Data Governance and Data Management[M]. Strathfield: Springer, 2020.
- [4] Janssen M, Brous P, Estevez E, et al. Data governance: Organizing data for trustworthy Artificial Intelligence[J]. Government Information Quarterly, 2020:101493.
- [5] 钱鹏.信息生命周期管理两重性辨析:以科学数据管理为例[J].情报理论与实践,2013,36(3):11-14.
- [6] 柴会明,张立彬,赵雅洁.国内图书馆科学数据研究述评[J].图书情报工作,2019,63(7):116-126.
- [7] 李伟绵,崔宇红.研究数据管理生命周期模型及在服务评估中的应用[J].情报理论与实践,2015,38(9):38-41.
- [8] 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.

- [9] 丁宁, 马浩琴. 国外高校科学数据生命周期管理模型比较研究及借鉴 [J]. 图书情报工作, 2013, 57(6):18-22.
- [10] Ball A. (2012). Review of Data Management Lifecycle Models (version 1.0)[EB/OL]. [2021-03-25]. <https://arxiv.org/ftp/arxiv/papers/2110/2110.00888.pdf>
- [11] 邢文明, 洪芳林, 李晓妍. 科学数据管理体系的二维视角——《科学数据管理办法》解读 [J]. 图书情报工作, 2019, 63(23):30-37.
- [12] 张洋, 肖燕珠. 生命周期视角下《科学数据管理办法》解读及其启示 [J]. 图书馆学研究, 2019(15):37-43.
- [13] 高瑜蔚, 石蕾, 朱艳华, 等. 《科学数据管理办法》实施细则比较研究——以正式发布的 11 份细则为例 [J]. 中国科技资源导刊, 2019, 51(3):1-10+17.
- [14] 屈宝强, 彭洁, 刘蔚, 等. 科学数据共享及其发展趋势 [J]. 情报学进展, 2020(13):381-420.
- [15] Lyon L. Dealing with Data: Roles, Rights, Responsibilities and Relationships [R/OL]. [2021-10-11]. http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf
- [16] 崔宇红. 机构研究数据管理实践探析: 模型、核心服务和优先战略 [J]. 情报理论与实践, 2017, 40(8):19-22+29.
- [17] 孟祥保, 高凡. 利益相关者视角下科研数据战略规划研究 [J]. 图书情报工作, 2016, 60(9):38-44.
- [18] 司莉, 辛娟娟. 科学数据共享中的利益平衡机制研究 [J]. 图书馆学研究, 2015(1):13-16+12.
- [19] 储节旺, 夏莉. 嵌入生命周期理论的科学数据管理体系构建研究——牛津大学为例 [J]. 现代情报, 2020, 40(10):34-42.
- [20] 顾立平. 科学数据权益分析的基本框架 [J]. 图书情报知识, 2014(1):34-51.
- [21] 盛小平, 吴红. 科学数据开放共享活动中不同利益相关者动力分析 [J]. 图书情报工作, 2019, 63(17):40-50.
- [22] 盛小平, 王毅. 利益相关者在科学数据开放共享中的责任与作用——基于国际组织科学数据开放共享政策. 图书情报工作, 2019, 63(17):9
- [23] 陈琳. 精简、精准与智慧政府数据治理的三个重要内涵 [J]. 国家治理, 2016 (27):28-39.
- [24] 邱春艳. 期刊文献与科学数据的关联服务研究 [J]. 情报资料工作, 2014(2):63-66.