科技文献数据资源建设模式数智化转型研究 *

——中国科学院文献情报中心的实践探索

钱力[1,2,3] 刘志博[1,2] 刘细文[1,2] 张智雄[1,2,3] 常志军[1,2,3] 李欣[1,3] 刘峥[1,3]

[1] 中国科学院文献情报中心 北京 100190
[2] 中国科学院大学经济与管理学院信息资源管理系 北京 100190
[3] 国家新闻出版署学术期刊新型出版与知识服务重点实验室 北京 100190

摘 要：[目的/意义] 面对开放互联网数字经济与以 ChatGPT 为代表的人工智能新技术的快速发展趋势，建设面向大模型需要的新型科技文献数据资源，研究设计面向智慧科研与智能情报的科技文献数据资源建设数智化转型体系框架，加速 AI4S 科研范式与智能情报模式的快速发展。[方法/过程] 通过文献调研与概念分析方法，从发展定位、服务取向、价值内涵与建设内容 4 个视角分析科技文献数据资源内涵演进特征，归纳科技文献数据资源建设的三大典型场景，分析科技文献数据资源建设数智化转型的迫切需求与发展趋势。[结果/结论] 提出从专业基础理论研究、政策机制及信息技术支撑、智慧知识底座建设 3 条主线构建科技文献数据资源建设数智化转型的体系框架，设计包含专业基础理论研究、数据权益体系、数据安全策略、先进信息技术应用、科技文献智慧数据中心、面向大模型应用的新型科技文献数据资源、自主可控的智慧科研数据平台、智能化服务引擎、情报场景化的智慧数据产品 9 个科技文献数据资源建设核心要素，形成具备未来科技水准和旺盛创新活力的科技文献数据资源建设及服务生态。

关键词：科技文献数据资源 数智化 AI-Ready AI4S 智慧知识底座 科技文献语料
分类号：G25 TP18
CSTR：32305.14.CN11-1541.2025.10.001
DOI：10.13266/j.issn.0252-3116.2025.10.001
引用本文：钱力，刘志博，刘细文，等．科技文献数据资源建设模式数智化转型研究——中国科学院文献情报中心的实践探索 [J]. 图书情报工作，2025, 69(10): 4-13. (Citation: Qian Li, Liu Zhibo, Liu Xiwen, et al. Research on Digitalization & Intelligentization Transformation of Scientific and Technological Literature Data Resources Construction: Practical Exploration of National Science Library, CAS[J]. Library and Information Service, 2025, 69(10): 4-13.)

1 引言 /Introduction

数智时代开放互联网数字经济发展背景下，科技文献数据资源的范围及形态正在发生本质变化：传统纸质科技文献数据资源已经不再是主流，更多地作为战略保存资源存在，数字化、数据化与知识化成为科技文献数据资源的主要形态。而当前以 ChatGPT 为代表的人工智能新技术的深入应用，给科技文献数据资源的获取、组织以及利用等方面带来重大机会和全新挑战，例如：大语言模型的本质是数据驱动的 AI 系统，科技文献数据资源正是大语言模型预训练所需要的高质量数据。然而面对智能化情报服务、AI4S（AI for Science，AI4S）新科研范式及知识增强大模型研发等服务场景，现行的科技文献数据资源内容迫切需要解决如下问题：①支撑智能情报服务的高质量专属精编数据集规模还非常有限；②面向科技文献全文内容的细粒度知识抽取与组织大多是试验性工作，还没有形成规模；③面向科研大模型预训练场景的科技文献语料数据资源需求迫切及如何精细化组织；④建设支撑智慧科研的数据生态体系需要顶层设计并建立；⑤激发科研大模型知识涌现并加速知识发现的提示工程数据集

* 本文系中国科学院文献情报能力建设专项“科技态势感知与分析能力建设”（项目编号：E3290909）研究成果之一。
作者简介：钱力，正高级工程师，博士，博士生导师；刘志博，硕士研究生；刘细文，研究员，博士，博士生导师；张智雄，正高级工程师，博士，博士生导师，通信作者，E-mail：zhangzhx@mail.las.ac.cn；常志军，正高级工程师，硕士，硕士生导师；李欣，副研究馆员，硕士；刘峥，研究馆员，博士，硕士生导师。
收稿日期：2024-02-20 修回日期：2024-11-09 本文起止页码：4-13

还非常缺失等。因此迫切需要专业图书情报机构面向科学研究、情报服务及文化传播等场景，重新思考和研究设计科技文献数据资源建设的新模式与新体系框架，充分重视其科技战略意义，以期回答“如何建设更符合数智时代特点的科技文献数据资源”这一关键发展问题，凸显在人工智能时代专业图书情报机构贡献科技文献知识语料的重要作用。

鉴于上述发展趋势及需求问题，本文采用文献调研和概念分析为主的研究方法，从发展定位、服务取向、价值内涵与建设内容 4 个方面分析了科技文献数据资源内涵演进特征，归纳总结了当前专业图书情报机构面向不同场景的科技文献数据资源典型建设模式，立足数智时代科技文献数据资源建设需求与发展趋向，结合中国科学院文献情报中心十四五规划在数据资源建设方面的实践，提出科技文献数据资源建设数智化转型的体系框架以及 9 个核心建设要素，以期为数智时代专业学术图书馆科技文献数据资源建设提供参考借鉴。

2 科技文献数据资源建设内涵演进 / Connotation evolution of scientific and technological literature data resources

2.1 发展定位演进：由以文献为中心到“信息—数据—知识”三维拓展

科技文献数据资源的概念内涵是一个生长着的有机体，随着信息技术和数字图书馆的发展而发展演化，形成从以篇章为粒度的科技文献数据资源的建设与利用模式，向从科技文献数据资源中进行“信息内容的深度组织、数据采集与挖掘、基于本体的知识提取”等多维度演化方向快速发展，而且当前人工智能技术驱动的智慧图书馆正在呈现以知识为中心的数智化时代特征。同时知识资源建设日益上升为国家数据战略、智慧图书馆建设的重点，场景化、专业化、关联化数据知识资源建设需求迫切，需要解决“面向国家重大科学问题的认知与决策数据知识资源体系泛在化、面向 AI 场域的多模态、多维度和多粒度的大规模计算下数据知识资源体系单一化、面向垂直深度场景的规范化知识智能感知、加工与组织的数据知识体系普惠化、面向国家区域发展及产业创新的全链动能型数据知识资源体系静态化”等问题，建设融合知识图谱（技术图谱、产业图谱、问题图谱、领域图谱、人才图谱等）的大规模、细粒度、高质量、强逻辑的知识资源。

2.2 服务取向演进：由科技文献数据资源服务向智慧知识服务发展

随着图书馆形态由传统图书馆向数字图书馆、智能图书馆及智慧图书馆的多阶演化，图书馆服务呈现由到馆纸本文献服务、数字化信息服务、智能智慧化知识服务的迭代升级 [1]，作为支撑图书馆服务的科技文献数据资源建设的基本取向也由纸本资源建设、数字化信息资源建设向智能智慧化数据知识资源建设转型。就具体建设模式而言，智慧知识服务需求下的智慧科技文献数据资源建设具有鲜明的数智时代特征：①在建设内核上，以大数据为原理，以分析和挖掘为方法，以用户为中心，以实时洞察客户为辅助，以转变被动服务向主动服务为取向，实现数据和决策支撑正循环，驱动图书馆服务智能化和用户服务体验的提升 [2]；②在建设目标上，以精准和最大化释放数据价值为目标，从大规模的科技文献数据资源中挖掘出具“规模化、可信任、场景化、可认知、可识别、可预测、可使用”特征的数据子集，支撑快速精准获得重要见解和洞察力需求；③在建设方式上，依托智能化技术方法实现非结构化海量数据汇聚、智能化处理、知识化萃取与可及化应用；④在建设过程上，在数据涌现产生、加工处理、流通利用的全生命周期中融合人的智慧，实现科技文献数据资源建设的智慧化。

2.3 价值内涵演进：由辅助机器智能向赋能智慧应用转变

与数字图书馆、智能图书馆范式下的科技文献数据资源辅助数字技术、智能技术的作用价值不同，数智时代智慧科技文献数据资源价值具有更强的目的性、更高的质量、更高级别的定制化服务、对机器学习有更大用处、促进更自信的决策，进而更大限度地赋能用户的智慧化应用。此价值内涵下的智慧科技文献数据资源包括如下特征 [2]：①面向应用业务场景，实现数据从智能计算跃升到智慧决策，数据本身不具有智慧，但通过嵌入到业务场景流程之中融合智能技术方法，便可实现计算、推理与决策；②拥有高质量的数据要素，即高质量的基础数据、知识实体及知识关系；③具有丰富的标签体系，拥有来源业务场景的多维度标签目录体系，实现对数据与知识的多维度、多尺度描述，面向个性化需求可以快速、精准、高质量遴选出数据资源体系；④具有多粒度的语义知识单元，从句子级、术语级挖掘与组织海量的语义知识单元，特别是面向专业化、垂直化领域方向，构建以语义知识单元为实体的知识图谱，实现支撑专业领域的

知识发现、知识推理及认知计算；⑤具有支撑计算、关联与推理特征的关联信息网络，实现支撑问题的全景化复原、关联化推理、循证化溯源。

2.4 建设内容演进：由数字化应用向支撑 AI 就绪转变

如今，以大语言模型为代表的 AI 工具已经开始在各个领域的科学研究与业界实践中扮演重要角色，其发展受到各界高度重视。为了在生产实践中进一步为快速迭代的先进 AI 技术落地应用打好基础，“AI 就绪”（AI-Ready）概念应运而生，开展 AI 就绪建设成为数智时代科技文献数据资源建设的必需。

立足人工智能驱动的科研新范式以及智能情报服务模式的需求，科技文献数据资源建设不能局限于传统文献数据资源既有的应用场景，还需要面向未来智能应用场景建设发展全面升级，为先进 AI 技术的应用提供充分的准备，重点从两个方面建设：一方面，建设包括高质量的基础数据、多维度标签的增值数据、专业人员加工与治理的规范数据、多粒度语义化知识元的语义数据、异质信息关联网络的图谱数据、面向 AI 大模型智能问答及生成启示的提示指令集数据等在内的智慧科技文献数据资源内容；另一方面，建设保障智慧科技文献数据资源应用的智慧架构体系，打造基于业务的主数据生成能力、多类型数据增值关联能力、基于中台的数据服务能力，加强智慧数据分级分类、安全使用、描述标准、计算引擎等内容的建设。

3 场景驱动下科技文献数据资源建设模式演进 /Evolution of scenario-driven construction of scientific and technological literature data resources

面对科技文献数据资源建设内涵的演化历程，在大数据与人工智能技术的加持下，科技文献数据资源建设模式也呈现出从馆藏纸质化，到数字化、数智化的发展趋势，逐步形成服务传统馆藏图书馆、数字图书馆、智慧图书馆三大场景的科技文献数据资源典型建设模式，着力建设印本馆藏的科技文献数据资源、数字化数据库的科技文献数据资源与多来源多类型多粒度的科技文献数据资源。

3.1 面向传统馆藏的科技文献数据资源建设模式

传统馆藏通常指传统的以不同实物载体形式而存在的非数字科技文献数据资源，如印刷型图书、期刊等纸质文献 [3]，也包括报纸、报告、微缩复制品、地图、手稿、唱片、磁带及其他视听资料和机读资料，具有较强的馆藏刚性 [4-5]，在部分研究中也被称为原始馆藏。传统馆藏是图书馆资源的根基，具有信息载体稳定、阅读环境的人文性和对经济发展不平衡的适应性等优势，更是当前数字图书馆的重要发展基础。然而，传统馆藏的数字化加速及大数据技术快速发展，催生了数字人文等领域研究成为文献情报领域的新热点，传统馆藏资源的价值也进一步凸显出来。

传统馆藏科技文献数据资源建设主要包括 4 个方向：①建设经良好组织的及有相对明确的应用场景传统图书馆馆藏资源，按照馆藏稀缺程度和收藏价值一般分类为重点馆藏（需要受到保护或具有高价值的珍贵馆藏）、基础馆藏、一般馆藏；按照实际的文献载体可以分类为图书馆藏、器物馆藏、报刊馆藏、手稿馆藏、胶片馆藏等；其他根据具体的建设场景，可以进行多样化的资源组织建设方法。②维护和保障传统馆藏科技文献数据资源持续建设及提供文献情报服务的配套硬件设施，分为馆藏保护与修复模块、馆藏评价模块、馆藏购置与更新模块、硬件设施保障模块与应急管理模块。③开展传统馆藏资源底座与用户相连接的组织方法研发，使传统馆藏资源底座能够更充分地发挥文献情报服务功能，其中面向传统馆藏的组织方法主要包括分类法、文献编目法、标引法、检索法等以实物文献为基本组织对象的方法。④系统化开展馆藏建设预算管理、图书馆人员管理、知识产权管块、图书馆战略发展管理及对外交流合作管理等传统馆藏管理。

3.2 面向数字图书馆的科技文献数据资源建设模式

作为图书馆的高级发展阶段，数字图书馆科技资源建设既要连接传统馆藏科技文献数据资源建设成果，也要逐步顺应数字技术发展，内容类型几乎覆盖了电子期刊、电子图书、电子报纸、学位论文、科技报告、工具书等几乎所有电子文献数据库类型 [6]。因此，数字图书馆场域下科技文献数据资源建设典型模式是在传统馆藏数字化工作的基础上，汇聚与集成建设原生数字化馆藏资源、海量第三方数字化资源及拓展其他来源多类型的数字化资源。其中，在文献数据资源的数字化建设方面，对于内部馆藏资源，一般采取特色馆藏与高价值馆藏、高频使用馆藏及紧急保护馆藏，优先进行数字化，同时充分集成外部数字化信息资源，实现快速丰富馆藏的同时也充分发挥数字化服务的最大价值；在数字化工程技术方面，采用扫描技术及光学字符识别技术以及全信息采集策略 [7]，之后进行数字化组织；在数字化存储方面，通常采用光盘（MO、DVD–R、CD-R、CD-ROM、

VCD）、磁带及磁盘等多种存储介质，保障数据安全；在数字化加工组织方法方面，采用信息组织技术、信息检索技术、信息安全与软件开发技术等一系列转化加工技术体系作为支撑[8]，以及语义技术、数据聚类技术、基于大数据的信息分析与检索技术等技术也在数字化馆藏利用和资源服务建设中发挥关键作用；在数字化图书馆管理方面，重点从人才队伍、资金支持、硬件环境、网络环境等方面重点部署，保障相关工作创新、协同、有序及高效。

3.3 面向智慧图书馆的科技文献数据资源建设模式

智慧图书馆是继数字图书馆、智能图书馆等概念之后的数智时代先进图书馆形态，通过高度融合智慧化设备与云计算、人工智能等最新智能化技术为用户提供智慧服务。面向智慧图书馆的科技文献数据资源建设思路是将数字化科技文献数据资源进行知识化、结构化、关联化与语义化组织的知识底座，以支撑知识深度挖掘、知识语义分析、知识多维揭示以及知识内容关联集成。

具体建设内容包括以下三方面：①大数据、物联网与人工智能的技术能力基础设施建设，面向场景的数据智慧、知识智慧以及硬件设施的智慧化逐步凸显其重要性，例如智能信息采集技术库、大数据与物联网及人工智能计算硬件资源、高效信息传输网络、协调互联智慧硬件等；②多来源、多粒度的科技文献数据资源建设，从基础资源（传统馆藏和原生数字馆藏构成）、转化资源（图书馆在数智时代信息技术、智能技术的基础上，借助计算机产生的二次原生资源与外来资源（非自身生产的，以共享、采购或引进等方式获得的其他资源等多种方式加强智慧图书馆科技文献数据资源内容的丰富性；③计算型的科技文献数据资源建设，以知识计算类技术和智能数据分析类技术为主，目前已经见诸智慧图书馆建设中的知识抽取、知识挖掘、知识推理、知识可视化、语义计算、多模态 AIGC 等技术已经开始发挥关键作用，为相应的智慧服务提供支撑并产生转化资源。

4 科技文献数据资源建设模式数智化转型的需求分析与框架设计 / Demand analysis and framework design on digitalization & intelligentization transformation of scientific and technological literature data resources construction

二十届三中全会要求“统筹各类科创平台建设”“加强创新资源统筹和力量组织”“构建科技安全风险监测预警和应对体系，加强科技基础条件自主保障”；国家数据局“数据要素 ×”三年行动计划中明确提出“以科学数据支持大模型开发，深入挖掘各类科学数据和科技文献，通过细粒度知识抽取和多来源知识融合，构建科学知识资源底座”[10]；中国科学院侯建国院长提出“加快推进数据和人工智能驱动的科研范式变革，打造安全可靠的科技文献开放存取基础设施”。面向国家及中国科学院对科技相关数据资源的建设要求，充分考虑复杂的新技术环境以及复杂交叉的科学问题的机遇和挑战，在第三章节分析的“面向智慧图书馆的科技文献数据资源建设模式”的基础上，如何以数智化转型实现高质量建设科技文献数据资源，“以数补智、以智强数、数智优链”，实现科技文献工作全要素生产率提升，发展科技文献情报工作的新质生产力，是当前科技文献情报机构面向未来一段时期迫切关注并解决的问题。

4.1 面向数智化转型的建设需求与发展趋向

（1）科学数据成为科技创新的关键要素，构建自主可控的开放学术交流环境、有序开放共享高质量科技文献数据资源需求迫切。PubScholar 公益学术平台的发布引起中国上千万用户的聚焦关注，通过该平台科研用户可以对高价值科研论文（期刊论文、预发布论文、学位论文、会议论文）、图书、专利、标准等科技文献数据资源进行免费检索与下载，极高的关注度凸显了科研用户对开放获取的高价值科技文献数据资源的迫切需求。新兴科研力量和新技术的融合发展使得科技文献原始数据资源建设向深层次、规模化、多样性、贴近创新需要方向发展，建设包括经由加工形成的专业高价值文献及数据集、特异和高效的研究材料与工具集、跨领域的复合科技文献数据资源，提供支撑科技创新过程的丰富敏捷的创新联结空间。

（2）科技文献与领域知识成为赋能 AI4S 新科研范式的关键核心。随着以虚实交互、平行驱动的 AI 技术为核心的新技术与科学研究实践的深度融合，科学研究工作正进入全新的“AI4S 的第五范式”阶段，即以 AI 驱动科学研究。在数学、物理、生物学等领域，AI4S 已经展现了其具有的强大应用潜力和驱动创新能力，取得了等令人瞩目的成果[9]。相对于经验范式、理论范式、计算范式及数据驱动范式等科学研究范式，AI4S 科研范式的核心是数据驱动型技术的集合，提供能与 AI/ML 新技术相互匹配、共同进化、高效利用的 AI 就绪型数据资源成为当前必需。各类

科技文献数据资源蕴藏着丰富的人类科学研究知识，不仅包括科研实体、科研主题等题录信息，也包括更多有价值的深层知识内容，如科学问题、科学理论、科学方法、技术手段、科学工具、理论模型、研究背景、研究基础、研究假设、研究思路、研究方法、科学实验、实验结果、研究结论等，充分利用好海量科技文献数据资源及其隐含知识，是图书馆赋能 AI4S 新科研范式的关键，也成为数据主管机构战略规划的重要内容。如在国家数据局发布的《“数据要素 ×”三年行动计划（2024—2026 年）（征求意见稿）》中，提出“数据要素 × 科技创新”场景下科学数据资源建设需求，要求深入挖掘包含科技文献在内的各类科学数据，通过细粒度的知识抽取，构建科学知识资源底座，建设高质量、有序开放共享的语料库和基础科学数据集，支持开展通用人工智能大模型和垂直领域人工智能大模型训练 [10]。

（3）智慧型科技文献数据资源成为支撑智能情报服务和场景的关键点。情报服务的泛在化与应用场景的具象化，推动传统科技情报服务向智慧数据驱动型智能科技情服务转变，拥有直接性、准确性、扩展性、知识性和预测性等特点 [11]，而建设融合“面广”“类多”“深度”“精准”“权威”“可信”等基因的智慧型科技文献数据资源成为支撑智能科技情报服务的关键。如驱动情报服务迭代发展的“科技论文、发明专利、学者库、机构、奖项、规划”等基础情报资源和通过多样化的聚合、挖掘等深度分析过程形成的“人才清单、技术清单”等高质量间接的科技文献数据资源；支撑深度情报挖掘和智慧情报服务的知识线索、知识脉络类科技文献内容型资源；面向及时、精准、权威情报场景的高质量精编类科技文献数据资源。

（4）支撑科技文献与情报大语言模型的科技文献语料成为专业图书情报机构数据资源建设的新生增长点。GPT 模型迭代升级的过程进一步证实了人工智能以数据为中心而非模型为中心的特性，有标注的高质量数据是释放人工智能价值的关键驱动。因此，对于科技文献与情报服务领域来说，一方面需要充分认识自身价值和作用，积极发挥科技文献情报资源优势，有效利用知识组织管理专长 [12]，面向大模型在科技创新领域应用需求建设其所需的“种类广、价值齐、领域专”的科技文献数据资源；另一方面，对于科技文献与情报服务领域来说，大语言模型不仅仅是一种技术基础能力，更是一种新型的知识库，将海量的科技文献与知识输入到大语言模型进行预训练，并经过科技文献与情报服务任务场景的有监督微调，直接形成知识问答、知识生成的智能知识服务基础设施，有效支撑智能问答、智能生成情报报告等智能情报服务领域的用户需求，因此，利用海量科技文献基础数据、图情专家标注的人类价值对齐数据、专业领域知识数据进行预训练形成的有监督微调、特定领域应用的“通用科技文献大模型、专业领域科技文献大模型、专业领域情报认知大模型”新型数据知识库资源，可以直接支撑科技文献与情报服务场景的知识问答、知识获取、情报咨询、情报报告生成等场景。

（5）智能化、敏捷化、协同化、可信任成为智慧科技文献数据资源建设的着力点。以智慧科技文献数据资源为关键要素的科技创新加速推进，对与之相关的基础设施能力提出了新的要求，需要融合数据科学、计算科学、学科领域信息学（如 Bio-、Geo-informatics 等）、计算语言学、人工智能、信息安全技术等先进理论和技术，建构适应智慧科技文献数据资源要素特征、促进智慧科技文献数据资源流通利用、发挥智慧科技文献数据资源价值效用的智能化、敏捷化、协同化、可信任数据基础设施，使之具备在海量数据中快速精准发现业务价值、通过数智技术进行验证并变成数据服务、打造人机协同共建共享的智慧型数据生态能力，建设支撑智慧科技文献数据资源服务的“数据清洗厂”“信息加工厂”“知识生成厂”与“决策制定厂”。

4.2 科技文献数据资源建设数智化转型框架设计与实施应用

本文结合中国科学院文献情报中心已经构建的“科技文献智慧数据中心、科技文献数据治理工具、科技文献大模型星火科研助手”等实践应用，面向数智化转型赋能科技文献数据资源建设需求与发展趋势，结合科学研究新范式与智能情报服务的具体应用场景，融合“基础理论—政策制度—生态平台—知识内容—信息技术”等维度，从“专业基础理论研究、政策机制及信息技术支撑、智慧知识底座建设”三大主线，“专业基础理论研究、数据权益体系、数据安全策略、先进信息技术应用、科技文献智慧数据中心、面向大模型应用的新型科技文献数据资源、自主可控的智慧科研数据平台、智能化服务引擎、情报场景化的智慧数据产品”九大核心要素，提出设计科技文献数据资源建设数智化转型的体系框架，加速将科技文献数据资源建成支撑国家科技创新的重要科技基础能力设施。如图 1 所示：

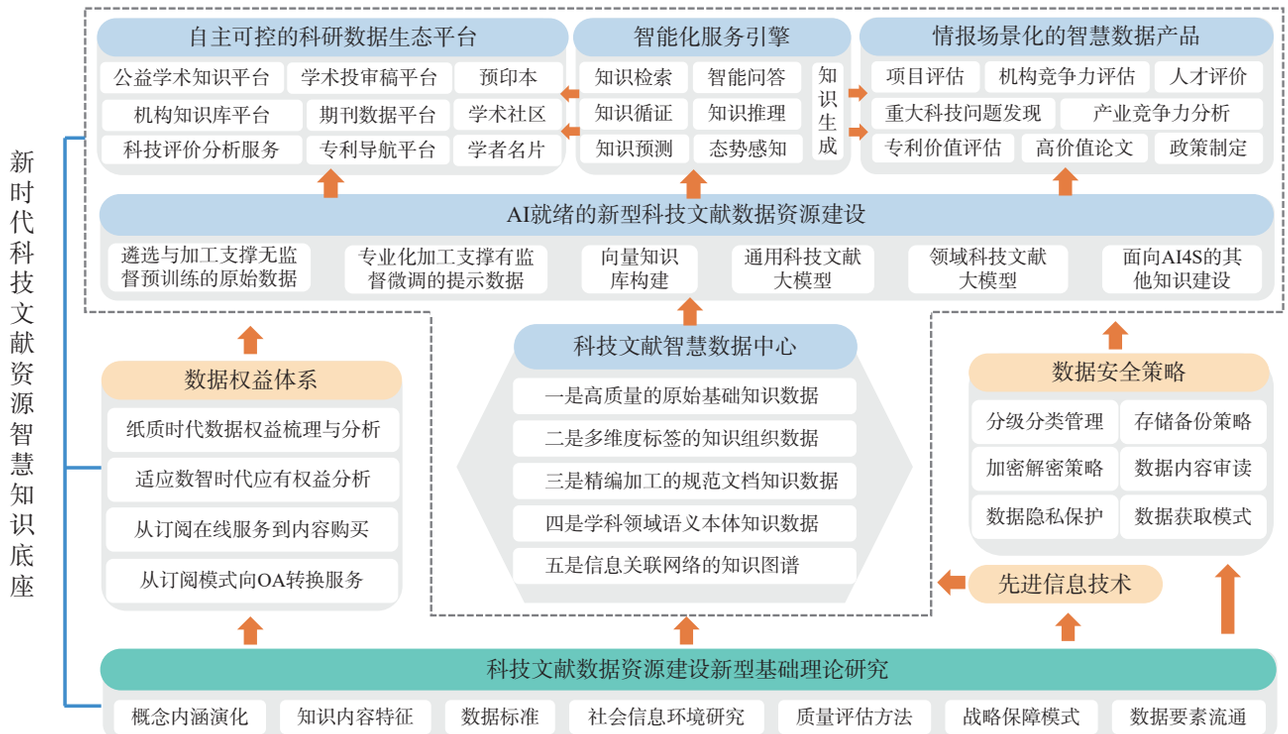

图 1　科技文献数据资源建设数智化转型体系框架

Figure 1　Digitalization & intelligentization transformation framework of scientific and technological literature data resources construction

面向上述数智化转型的体系框架的总体设计及发展愿景目标，中国科学文献情报中心在科技文献数据资源建设方面围绕“专业基础理论研究、政策机制及信息技术支撑、智慧知识底座建设”三大发展主线，面向具体应用场景，进行总体部署与分布协同实施，具体部署内容及部分实施效果如下：

（1）围绕专业基础理论研究建设主线，即以理论创新引领科技文献数据资源建设数智化转型发展方向，部署加强科技文献数据资源建设新型基础理论研究。

新技术和新语境下科技文献数据资源建设理论是科技文献数据资源建设的指引和基础，面向数据赋能型发展态势亟需从理论上廓清情报服务和科学研究场景下科技文献数据资源概念内涵、内容特征、制度标准、质量评估、安全保障、流通机制、技术方法等内容。如新技术环境下科技文献数据资源建设的观念指引，由传统数据资源建设观念转向数值资源观、数据资源观和特色资源观[13]。在“大数据”向“智慧数据”发展迈进过程中，研究设计科技情报智慧数据的建设的方法与框架，包括科技情报智慧数据建设的定位与价值、遵循科技情报智慧数据架构体系、保障科技情报智慧数据质量控制等内容[14]。

中国科学院文献情报中心在该方向重点部署了“科技文献语义内容挖掘、面向 AI4S 科技文献知识底座建设、智能情报关键技术”等方面的前沿理论及技术研究，并且获得两项社科基金重大项目（大数据驱动的科技文献语义评价体系研究、数智转型背景下智能情报关键技术研究）、一项社科基金一般项目（AI4S 科技文献知识底座的理论体系及建设方法研究）的支持，同时也积极申请面向科技文献全文内容多模态数据组织、多粒度语义知识组织的相关标准，为科技文献数据资源建设及科技情报分析场景自动化提供了互操作标准。

（2）围绕政策机制及信息技术支撑主线，即以制度与机制创新保障新时期的科技文献数据资源内容建设及服务模式，部署加强数据权益、商业资源采购模式、安全策略研究以及先进信息技术的深度应用。

——要建设面向新技术及数智时代需求场景的科技文献数据资源建设数据权益体系。数据作为在新时期基于信息技术发展产生重要价值的无形财产，目前仍是处于权利与法益之间的一种权益客体[15]，而相比传统资源，其权益主体和权益构成更加复杂，利用方式更加多元[16]，管理和保护范围更加模糊。

中国科学院文献情报中心面向数智化转型需要，要求所有采购的数据库要具有“元数据清单、全文内容资产化本地存储、全文内容中国本土长期保存”等服务于 AI4S 应用的核心权益，因此调研梳理纸质时代的科技文献馆藏、传播利用及战略保存的相关权益清单，研究分析数据资源建设、存储共享及传播利用等阶段数智化转型的特征，梳理科技文献数据资源与纸质时代要求的权益相违背的清单，进而提出数智时代科技文献数据资源采购应该具有的数据权益，推动知识产权面向社会信息环境的快速发展等一系列措施理应做出改变。

——要加快建设数智时代商业科技文献数据资源的采购策略。AI4S 科研范式的快速兴起以及商业订购费用的持续上涨使传统的科技文献订阅服务已经不能够充分适应科技创新要求与可持续发展。从传统订阅在线服务向订购内容模式转变，这本应是订购单位拥有的权益；从订阅在线服务向 OA 出版转换模式转变，应利用“多维度的大数据驱动的采购方案多维评价方法”，实现商业科技文献数据资源的价格回归“科学、合理、公平与公正”，保障科技文献数据资源购买服务的可持续性，保证科技创新对科技文献数据资源的战略需求。

—— 加强科技文献数据资源的安全策略制定及举措实施。数据安全事关国家安全，在国家整体发展战略中占有重要地位，开展数据安全策略及方法研究是数智时代科技文献数据资源建设的必备保障。应当基于《中华人民共和国数据安全法》，参照机构对科技文献数据资源的安全要求，从“数据分级分类管理、数据隐私保护、加密解密策略方法、存储架构及设施要求、数据内容审读、备份策略要求、数据获取模式”等方面开展研究设计，对战略性科技文献数据资源开展策略性采集与长期保存，确保科技文献数据资源的长期安全使用。2023 年 5 月 17 日国家科技图书文献中心（National Science and Technology Library, NSTL）依托中国科学院文献情报中心，正式成立国家数字科技文献资源长期保存中心，保障采购的数字科技文献资源的长期可用。

——在科技文献数据资源建设的全生命周期中，加强先进信息技术的深度利用。从机器人数据稳定感知与采集，人工智能技术赋能的自动编目、数据清洗、数据挖掘、知识组织以及多语种翻译等数据增值场景出发，应当进一步加强大数据与人工智能技术的深度应用，解放生产力、释放知识化专业组织队伍的想象力。中国科学院文献情报中心研发了“科技文献大模型、星火科研助手、科技情报数据治理加工工具、SCIAIEngine”等实现科技文献知识挖掘、数据加工工作的工程化与智能化、智能化知识服务平台的智能化云基础设施。

（3）围绕智慧知识底座建设主线，即面向数智时代迫切需求，部署加强科技文献智慧数据中心、AI 就绪的新型科技文献数据资源、自动可控的科研数据生态平台、智能化服务引擎以及情报场景化的智慧数据产品五大新型科技文献数据资源。

——建设科技文献智慧数据中心。中国科学院文献情报中心以数据驱动数智时代科技情报工作新范式的实施思路，研究设计 5 层级的科技文献智慧数据体系，即高质量的原始基础知识内容、多维度标签的知识组织内容、精编加工的知识规范内容、领域多粒度语义化的知识本体内容与异质信息关联网络的知识图谱内容，使科技文献智慧数据中心以智慧化方式支撑“数据清洗、信息加工、知识生成、决策制定”等上层知识服务，使科技情报工作能够快速洞悉变化、凝练问题、聚焦目标、形成解决方案，增强人们应对复杂问题与任务的能力。目前中国科学院文献情报中心已经建成覆盖“情报监测类、创新活动类、科研成果类、领域知识类”的特色数据库近 30 多个，数据体量近 10 亿规模。

——抓住 AI 就绪的新型科技文献数据资源的新生增长点，充分发挥文献情报机构对内容的专业化加工与理解能力。以大模型为代表的先进信息技术正在形成类似操作系统的新质生产力，而对未来先进 AI 技术应用准备就绪的新型科技文献数据资源建设将开辟数智时代的资源建设新赛道。“遴选与加工支撑无监督预训练的原始数据、专业化加工支撑有监督微调的提示数据、专业领域科技文献内容的知识图谱、向量知识库构建、通用科技文献大模型、领域科技文献大模型、面向 AI4S 的其他知识建设”等类型的新型科技文献数据资源，是人工智能场景下的科技创新新范式的迫切需求。其中，面向科技文献大模型研发的需要，从“动机解读、贡献解读、方法解读、结果解读、创新点、优缺点分析”等 19 种类型对科技论文全文进行了 Prompt 微调数据标注，实现了科技文献大模型在自动综述、论文研读等方面近 90% 以上

的准确率。

——加快建设自动可控的科研数据生态平台，保障学术科研数据的开放、共享与公益服务。高效高质的科技文献数据资源建设离不开其环绕的数据服务生态平台，加强以数据服务生态平台实现科技文献数据资源的自主建设是数智时代数据资源建设必须重点研究布局的方向，对于进一步实现学术话语权的掌握至关重要。中国科学院文献情报中心已经构建了“以 PubScholar 公益平台为基础的个人科技文献成果发布与管理、以机构知识库平台为基础的机构科技文献发布与管理、以期刊编审系统与发布系统为基础的期刊论文数据、以 ChinaXiv 等预印本平台为基础的预印本论文数据以及其他第三方合作机构的科技文献”等数据服务生态，以 PubScholar 公益平台（https://pubscholar.cn/）为开放数智化云服务基础设施，从底层数据流通方面实现统筹设计、顶层汇聚、各自运营与协作共创的思路，实现中国科技文献数据资源的自动汇聚与主动掌握。

——打造面向科技创新场景的智能化服务引擎，赋能知识服务平台的智能性与精准性、情报分析服务的敏捷性与权威性。充分发挥拥有的高质量、海量的科技文献数据资源的知识底座优势，深度利用大数据与人工智能先进技术，从“知识检索、智能问答、知识循证、知识推理、知识预测、态势感知以及知识生成”等场景研发智能化服务引擎，为科技文献内容深度挖掘与价值利用提供智能化服务。中国科学院文献情报中心已经公开发布了科技文献大模型与 SCIAIEngine，全面支撑科技文献知识挖掘、知识服务及情报分析等场景。

——研发面向情报场景化的智慧数据产品，形成能够快速与精准提供科技情报服务的精编数据体系。基于科技文献智慧数据中心，从“人才评价、机构竞争力评价、研究前沿研判、经费投入、科研成果评价”等情报场景研发新型智慧数据产品，充分发挥数据要素作用，提升我国科技情报服务能力水平。中国科学院文献情报中心已经建成了拥有一定规模数据的“科技人才库、科技奖项库、学术观点评论库、技术清单库、新科技产品库、学术会议库、基金项目库、期刊编委库、科技舆情库、科技政策库、查新报告库”等特色智慧数据产品，提升科技情报分析工作的数智化服务能力。

5 结论 /Conclusion

科技文献数据资源建设数智化转型发展应当着重面向国家重大需求、重大问题、卡脖子问题，以关键场景的科技文献数据资源保障为首要建设目标，以发展科技文献新质生产力为关键基础，以解决科技文献数据资源建设与当前新技术背景下国家重大科研项目与情报需求之间的不平衡、不协调为出发点，紧抓“需求导向”和“体系建设”，面向新一代人工智能技术环境的科学研究与情报服务开展实施科技文献数据资源建设数智化转型的体系框架设计，努力实现“以数补智、以智强数、数智优链”与科技文献工作全要素生产率提升。

未来，科技文献数据资源建设数智化转型工作既要注重科技文献数据资源建设体量与质量“数”的特征，更要充分发挥应用人工智能新技术“智”的特点，实现数智融合、应用导向，并且要创建科技文献数据资源的自主可控的开放学术生态体系，以人才队伍要素所发挥的创新构想涌现为根本、以数据知识等要素所发挥的智慧赋能和结构置换为关键、以生成式人工智能等要素所表现为技术升级为核心，支撑“数据与知识、人才、生成式人工智能、工具与平台、科研发现、情报分析、产业分析、科技政策、科技金融”等创新要素的优化组合。

当然，科技文献数据资源建设是长期成长性事业，并非一蹴而就，还需久久为功。以 ChatGPT 等 AI 产品为代表的大语言模型为科技文献数据资源建设带来了全新的发展机遇，但技术变革必然永不停歇，如何面对日新月异的技术环境，建立长效的科技文献数据资源建设方略，建成能够长期与先进技术发展共存共演，同时保有高度创新活力的战略性科技文献数据资源，是把握科技情报事业未来发展的关键。

参考文献 /References：

[1] 初景利，任娇菡，王译晗．从数字图书馆到智慧图书馆 [J]. 大学图书馆学报，2022, 40(2): 52-58. (CHU J L, REN J H, WANG Y H. From digital libraries to smart libraries[J]. Journal of academic libraries, 2022, 40(2): 52-58.)

[2] 钱力，刘细文，张智雄，等．科技情报智慧数据：方法、体系与应用 [J]. 情报理论与实践，2024, 47(1): 12-21; (QIAN L, LIU X W, ZHANG Z X, et al. Smart data for scientific and technological information: method, framework and

application[J]. Information studies: theory & application, 2024, 47(1): 12-21.)

[3] 章红 . 数字馆藏建设对传统馆藏建设政策的挑战 [J]. 图书馆建设 , 2003(6): 40-42. (ZHANG H. The strategy of digital collection building challenges to the strategy of traditional collection building[J]. Library development, 2003(6): 40-42.)

[4] 王秀芬 . 论网络信息资源与传统馆藏资源的互补性 [J]. 图书馆界 , 2005(1): 11-13. (WANG X F.On the complementarity of network information resources and traditional library resources[J]. Library world, 2005(1): 11-13.)

[5] 董敏斐 . 传统馆藏、数字馆藏和虚拟馆藏一体化建设模式初探 [J]. 浙江万里学院学报 , 2006(1): 115-118. (DONG M F. On integration of the traditional, digital and virtual collections[J]. Journal of Zhejiang Wanli University, 2006(1): 115-118.)

[6] 孙秀丽 . 高校图书馆数字资源建设与利用的调查分析 [J]. 大学图书馆学报 , 2008, 26(6): 45-50. (SUN X L.Investigation and analysis on the development and usage of digital resources in the university library[J]. Journal of academic libraries, 2008, 26(6): 45-50.)

[7] 刘家真 . 馆藏文献数字化的原则与方法（下）[J]. 中国图书馆学报 , 2001(6): 45-48. (LIU J Z. Principles and methods of the digitization of library collections, part II[J]. Journal of library science in China, 2001(6): 45-48.)

[8] 李维纯 , 杨明 . 数字图书馆应用技术研究 [J]. 情报科学 , 2005(5): 750-754. (LI W C, YANG M. Research on the applying technology of digital library[J]. Information science, 2005(5): 750-754.)

[9] 王飞跃 , 缪青海 . 人工智能驱动的科学研究新范式：从 AI4S 到智能科学 [J]. 中国科学院院刊 , 2023, 38(4): 536-540. (WANG F Y, MIU Q H. Novel paradigm for AI-driven scientific research: from AI4S to intelligent science[J]. Bulletin of Chinese Academy of Sciences, 2023, 38(4): 536-540.)

[10] 中华人民共和国国家互联网信息办公室 . 十七部门关于印发《“数据要素 ×”三年行动计划（2024—2026 年）》的通知 [EB/OL]. [2025-01-19]. http://www.cac.gov.cn/2024-01/05/c_1706119078060945.htm. (Cyberspace Administration of China. 17 departments on the issuance of the “Data Elements×” three-year action Plan (2024-2026) notice[EB/OL]. [2025-01-19]. http://www.cac.gov.cn/2024-01/05/c_1706119078060945.htm.)

[11] 丁洁兰 , 钱力 , 常志军 , 等 . 科技情报智慧数据服务体系建设研究 [J]. 情报理论与实践 , 2024, 47(1): 30-37. (DING J L, QIAN L, CHNAG Z J, et al. Research on the construction of service system of smart data for scientific and technological information[J]. Information studies: theory & application, 2024, 47(1): 30-37.)

[12] 张智雄 . 在人工智能时代贡献文献情报领域的智慧和方案 [J]. 农业图书情报学报 , 2023, 35(1): 5-8. (ZHANG Z X. Contribute wisdom and solutions in the field of scientific and technological information in the era of artificial intelligence[J]. Journal of library and information science in agriculture, 2023, 35(1): 5-8.)

[13] 孟祥保 , 高冕 . 图书馆数据资源建设：内涵、价值与路径 [J]. 新世纪图书馆 , 2023(5): 48-54. (MENG X B, GAO M. Library data resources construction: connotation, value and approach[J]. New century library, 2023(5): 48-54.)

[14] 赵昆华 , 李欣 , 章岑 , 等 . 文献情报机构面向智慧数据建设与服务的权益体系设计 [J]. 情报理论与实践 , 2024, 47(1): 38-45. (ZHAO K H, LI X, ZHANG C, et al. Design of the rights and interests system for the construction and service of smart data in documentation and intelligence institutions[J]. Information studies: theory & application, 2024, 47(1): 38-45.)

[15] 王利明 . 论数据权益：以“权利束”为视角 [J]. 政治与法律 , 2022(7): 99-113. (WANG l M. On data rights and interests: from the perspective of bundle of rights[J]. Political science and law, 2022(7): 99-113.)

作者贡献说明 /Author contributions：

钱力：论文选题提出、撰写与修改；
刘志博：相关资料调研、论文撰写与修改；
刘细文：论文选题与修改；
张智雄：论文选题与修改；
常志军：论文资料调研；
李欣：论文资料调研；
刘峥：论文资料调研。

Research on Digitalization & Intelligentization Transformation of Scientific and Technological Literature Data Resources Construction: Practical Exploration of National Science Library, CAS*

Qian Li[1, 2, 3] Liu Zhibo[1, 2] Liu Xiwen[1, 2] Zhang Zhixiong[1, 2, 3] Chang Zhijun[1, 2, 3] Li Xin[1,3] Liu Zheng[1,3]

[1] National Science Library, Chinese Academy of Sciences, Beijing 100190

[2] Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

[3] Key Laboratory of New Publishing and Knowledge Services for Scholarly Journals, Beijing 100190

Abstract: [Purpose/Significance] Facing the rapid development of the open internet digital economy and emerging artificial intelligence technologies represented by ChatGPT, this paper studies how to construct a new type of scientific and technological literature data resources for large language model, and designs a digitalization & intelligentization transformation framework for the scientific literature data resources construction for smart research and intelligence information, and accelerates the rapid development of AI4S research paradigms and intelligence models. **[Method/Process]** Through literature research and concept analysis, this paper analyzed the evolution characteristics of the connotation of scientific and technological literature data resources from four perspectives: development positioning, service orientation, value connotation, and construction content. It summarized the three typical scenarios of the construction of scientific and technological literature data resources, and analyzed the urgent demands and development trends of digitalization & intelligentization transformation. **[Result/Conclusion]** This paper proposes a digitalization & intelligentization transformation framework for constructing scientific and technological literature data resources, based on three main lines: professional foundational theory research, policy mechanisms and information technology support, and smart knowledge bases. The framework includes nine core elements: professional foundational theory research, the data rights system, the data security strategy, advanced information technology applications, smart data of science and technology, new type of science and technology literature data resources for large language model, autonomous and controllable of data platform of smart scientific research, intelligent service engine, and intelligence scenario-based smart data products. This forms an ecosystem for the construction and service of scientific and technological literature data resources that possesses future scientific standards and vigorous innovation vitality.

Keywords: scientific and technological literature data resources digitalization & intelligentization AI-ready AI4S smart knowledge base scientific and technological literature corpus

*This work is supported by the Chinese Academy of Sciences Special Project for Literature and Information Capability Building, titled "Science and Technology Situation Awareness and Analysis Capability Building" (Grant No. E3290909).

Author(s): Qian Li, senior engineer, PhD, doctoral supervisor; Liu Zhibo, master candidate; Liu Xiwen, researcher, PhD, doctoral supervisor; Zhang Zhixiong, senior engineer, PhD, doctoral supervisor, corresponding author, E-mail: zhangzhx@mail.las.ac.cn; Chang Zhijun, senior engineer, master's degree, master supervisor; Li Xin, associate research librarian, master's degree; Liu Zheng, research librarian, PhD, master supervisor.

Received: 2024-02-20 Revised: 2024-11-09 Pages: 4-13