

大数据背景下科技资源信息挖掘与利用的思考

袁伟¹ 石蕾²

(1. 中国科学技术信息研究所, 北京 100038; 2. 国家科技基础条件平台中心, 北京 100862)

摘要: 科技资源信息是科技创新的基础条件, 也是科技创新实力的重要体现。概述国内外科技资源信息的发展态势, 阐述科技资源信息在科研创新性中的关键作用, 总结我国科技资源信息的发展现状, 分析当前科技资源信息发展存在的问题, 探讨大数据背景下加强科技资源信息挖掘与利用的努力方向和思路, 以期为进一步促进科技进步与发展提供参考。

关键词: 大数据; 科技资源; 科技资源信息; 信息挖掘; 信息利用

中图分类号: G203

文献标识码: A

DOI: 10.3772/j.issn.1674-1544.2017.06.001

Reflections on Mining and Utilization of Sci-Tech Resources Information Under Big Data

YUAN Wei¹, SHI Lei²

(1. Institute of Scientific and Technical Information of China, Beijing 100038; 2. National Science and Technology Infrastructure Management center, Beijing 100862)

Abstract: Sci-tech resources information is not only the foundation of sci-tech innovation, but also a significant manifestation of sci-tech creative strength. This paper summarizes the development trend of domestic and international sci-tech resources information, expounds the key role of sci-tech resources information in scientific research and innovation, summarizes the development status of information resources in China, analyzes the existing problems in the development of sci-tech resources information, discusses the mining and utilization of the direction and thinking of strengthening the sci-tech resources information under big data background, in order to provide reference for further promotion of the sci-tech resources information.

Keywords: big data, scientific and technical, resources, scientific and technical resources information, information mining, information utilization

1 引言

大数据是指在合理时间内使用传统方法或工具无法实现分析或处理的规模巨大的数据集^[1]。大数据的应用揭示了过去传统方式无法展现的相关关系, 促进了数据的融合与资源的整合。随着

大数据时代的到来, 对数据的生产加工能力、占有能力以及大数据处理和分析能力, 将成为一个国家能否把握全球竞争优势的关键指标。

科技资源是科技活动中人力、物力、财力以及组织、管理、信息等软硬件要素的总称^[2], 是科技创新和经济社会发展重要的支撑和保障。科

作者简介: 袁伟 (1962—), 男, 中国科学技术信息研究所教授, 研究方向: 科技资源管理战略研究; 石蕾 (1982—), 女, 国家科技基础条件平台中心副研究员, 研究方向: 科技资源管理 (通讯作者)。

基金项目: 国家自然科学基金项目“科技基础条件对创新能力影响的定量研究”(m1551001)。

收稿时间: 2017年8月14日。

技资源的总量间接制约着国家的创新能力提升,世界发达国家无一例外地都拥有强大的科技资源生产、占有和使用能力。作为科技资源的描述性信息,科技资源信息充分反映了科技资源客观状况、特点、发展状态、使用方式等,是促进科技资源开放利用的重要基础,也是重要的科技资源类型。

改革开放以来,我国科技资源状况发生了大幅改善,积累了相当数量的仪器设备、文献、标本、数据及人才资源,有力地支撑了国家科技创新。特别是在大数据时代的影响和推动下,科技创新活动日新月异,科技资源信息的产生途径也异常丰富,通过互联网、移动设备、智能终端产生的科技资源信息更呈爆炸式增长。当前科技资源信息获取也已经从时间、距离上的障碍转变为内容选择上的障碍,如何从大量科技资源中获取有价值的、能够直接用于解决科研问题、攻克技术难点的知识成为主要关注的因素^[3]。此外,随着社会的发展,科学研究的问题也越来越复杂,仅靠单个平台提供科技资源已不能满足当前科技发展的需要,大数据发展为科技创新发展提供了新的思路和技术支撑,通过多渠道、多来源整合科技资源信息,形成海量大数据基础上的多点协同创新,也成为解决复杂科学问题的重要基础。综合来看,对科技资源信息的科学分析、深入挖掘和综合利用,已成为科技创新、经济社会发展和管理决策的重要依据^[4]。

然而,当前我国对科技资源信息的管理和利用能力,与新时期我国建设创新型国家的需求以及实现创新驱动发展的战略目标存在一定的差距,进一步优化科技资源优化配置,提高科技资源利用效率的旺盛需求和挑战仍然并存。因此,通过深入分析大数据背景下科技资源信息的基本特征,以及大数据对科技资源信息挖掘和利用的影响,集成利用好数量大、种类多、增速快的科技资源信息成为我国科学研究及经济社会发展的关键因素^[5],探究大数据背景下我国科技资源信息的挖掘与利用具有重要的意义和价值。

2 我国科技资源信息发展现状

2.1 科技资源信息化程度明显提高

当前,我国进入科技创新发展战略实施的新时期,实现全面提升科技创新能力目标对科技资源信息化水平以及信息挖掘分析能力提出了更高的要求,而提升政府管理决策能力也越来越依赖海量科技资源大数据信息的综合分析。在目标需求的引导和带动下,政府有关部门和科技资源机构大力推进科技资源信息化工作,取得了积极进展和显著成效。聚焦物力科技资源,我国已经在多个领域建成了一批国家科技资源共享服务平台,在科研设施、科学仪器、科学数据、生物种质、科学实验材料所涉及的多个学科领域形成了一大批专门开展科技资源信息化、从事资源利用与信息挖掘并面向全社会开展科技资源公共服务的科技资源中心(库、馆),建成了国家科技资源共享服务平台门户系统,专门用于科技资源信息的整合集成与分析挖掘。这些资源中心和门户系统对国家管理决策、科学研究发展发挥了重要的支撑和保障作用。按照《国务院关于国家重大科研基础设施和大型科研仪器向社会开放的意见》(国发〔2014〕70号)要求建成的重大科研基础设施和大型科研仪器网络管理平台,已对各部门、各高校院所的科研设施和仪器进行信息化整理并向社会公布。

与此同时,各部门、行业、单位也积极推动各类科技资源信息化并用于科学研究或公共服务。如依托于北京大学成立的国家医疗数据中心,收集、整理了近400家医院超过10余年共4800多万条的病案首页数据,并将数据全部进行了数字化,为临床医师和科研人员的医学科学研究提供了基础数据支撑^[6]。再如黑龙江省已完成了700余家单位近4000台(套)20万元以上大型仪器设备的资源信息数字化,此外还收录高端制造服务设备进600台(套),实现了海量资源信息的高度整合^[7]。科研人员利用这些科技资源信息在使用仪器设施过程中改进了操作方法,完善了实验数据,形成了更准确的试验结果。大

量的科技资源信息为科研仪器利用与管理、科学数据分析和科研效率提升提供了强大支撑。

同时，随着国家信息化工作的快速推进，大量统计数据、财政投入数据以及科技人才信息等也通过多种信息化手段成为公共可利用的大数据资源，为科技资源信息大数据挖掘奠定重要的基础。

2.2 科技资源信息来源广泛

大数据是计算机科学、统计学和社会学等多学科交叉渗透所形成的。随着现代信息化、网络化等技术手段的应用，科技资源信息摆脱了地域上的屏障和束缚，人、财、物等各类科技资源信息源源不断地产生，其信息量迅速攀升，覆盖的科技领域范围也不断扩大，通过运用科学方法处理，最终将这些信息整合形成了由结构化数据、半结构化数据和非结构化数据的科技资源信息大数据。同时，科技资源信息来源在种类上趋于多样，不仅包括科技文献资源信息、科学数据资源信息、仪器科技资源信息等传统的科技资源信息，还有实时传播信息（如科技快报、天气预报等）、参考工具信息（如年鉴、手册等）、内部科技资料（如专业技术图片、视听音像等）、网络灰色科技信息（如科技机构发布的科技政策、科技新闻、产品信息等）以及个人在网站、微博、个人网页发布的科技观点、手稿等新的科技资源信息^[8]，这些形式多样的科技资源信息源不断增加，相互影响、结合，又会形成新的科技资源信息来源。此外，物联网智慧城市、智慧医疗、智能电网以及群体感知等新兴技术的应用，也使各行各业在其业务发展过程中形成新的科技资源信息大数据的来源。

一些领域科技资源信息在快速发展的同时，也带动了其他领域的共同发展。例如，气象科技资源信息的发展为农业生产带来了巨大的益处，气象条件与农业生产有着非常密切的关联，农业生产过程中的温湿度、日照强弱、风力风向等气象因子都会影响农作物的产量和品质。现今全国已建成了2000多个气象地面站、120多个高空探测站、300多个雷达站和近百个酸雨监测站^[9]，通过对这些多源站点中的科技资源进行信息化整

合，形成了数量多、覆盖全的气象科技资源信息，通过分析和应用这些气象科技资源信息，为国家管理部门进行农业区域规划提供了参考，同时也为相关科学研究提供了基础数据支撑。

2.3 多渠道推动科技资源信息公开

大数据技术是一种先进的技术，有着信息容量大、传播范围广等优点，大数据技术的运用有利于高价值科技资源信息在多领域、深层次间实现更有效的信息公开与传播。科技资源的高效利用水平取决于科技资源信息的公开程度。经过数年发展，我国已初步实现科技资源信息多渠道公开，如国家科技资源共享服务平台建设、科技文献或数据出版、科技资源报表发布、科技人财物信息公开等。以科技期刊出版为例，目前我国已形成较完善的科技期刊数字化出版模式，利用大数据相关的计算机技术和网络技术实现期刊文献数字信息加工、存储、整合、编辑和出版，并开始探索科学数据出版等多元化出版模式^[10]，以进一步拓展科技资源信息的公开渠道。此外，各部门、各地方围绕科技资源信息公开制定了一系列法规政策，如中国气象局制定了《气象资料共享管理办法》，国土资源部制定了《公益性地质资料提供利用展现办法》和《深部地球物理探测数据共享管理办法》，农业部制定了《农作物种质资源管理办法》等。各省市也通过制定和实施相关科技资源共享管理办法，从政府和科技资源管理单位等多主体的权利、义务和责任等方面对科技资源信息公开做出了明确规定^[11]，有效推动了大数据背景下的科技资源信息公开。

3 加强科技资源信息挖掘与利用的思考与建议

尽快我国在科技资源信息管理、公开与利用方面开展的大量工作取得了巨大进展，但与欧美等发达国家相比仍然存在很多不足与问题。突出表现在以下几个方面：一是科技资源信息生产受到普遍关注，然而大量科技资源由于未得到科学及时的描述并建立索引而石沉大海；二是科技资源信息管理体制机制尚不健全，尚未建立互联互

通的科技资源信息基础保障体系,导致信息分散,利用率不高;三是科技资源信息尚未形成统一的标准化规范体系,科技资源信息量大、异构的特点制约了科技资源信息的快速传播与广泛利用;四是对科技资源信息分析挖掘和利用的技术体系尚需完善,对信息挖掘的需求和目标还不甚明确,从而导致大量科技资源没有得到充分利用。

3.1 夯实科技资源信息基础,促进信息开放

海量、充沛的科技资源信息是开展科技资源信息挖掘与利用的前提和基础,建立完整的适用于大数据背景下的科技资源信息管理体系,有效积累科技资源信息,是提高科技资源利用效率的首要任务。当前,我国通过科技计划、行业业务发展、产学研结合、企业自主研发等方式产生了大量的科技资源,应依托法人单位建立稳定的科技资源数字化机制,逐步丰富和完善科技资源信息基础。同时,借助大数据管理方法与模式,在基础层面对科技资源数据进行多、精、准的分类处理,运用大数据相关技术将资源数据信息化与数字化,并转换成利于后续应用的大数据结构,通过引入先进的大数据处理手段,夯实科技资源信息基础。

同时,在大数据背景下对科技资源信息的充分挖掘与广泛应用有赖于高效的公开机制。首先需要建立科技资源信息开放宏观政策环境,充分发挥政府部门职能作用,摸清各主体对科技资源信息公开的需求,并利用大数据的网络便利性和覆盖广泛性等特性在各个层面上组织和协调各类科技资源信息的公开活动,形成统一管理、分层负责的科技资源信息公开机制。专门制定基于大数据环境下的科技资源信息公开披露政策,包括信息公开权利、责任、义务和总体规划等,树立共建共享理念,解决阻碍科技资源信息公开的体制性问题。同时,建立完善科技资源信息报告制度,定期及时向社会公布现有科技资源的现状、更新和升级情况,编制科技资源信息发布计划,健全科技资源信息开放目录制度。推动建立完备的以大数据应用为基础的科技资源信息库,利用大数据的优势,开辟科技资源信息公开渠道。

3.2 加强科技资源信息集成与汇聚

加强科技资源信息集成与汇聚,需要利用多种渠道、多种方式对这些科技资源信息进行加工、处理、分类和优化。应当充分利用大数据采集技术、存储技术、云存储、云计算等基础架构技术,依托如“中国科技云”等大数据设施环境进行信息集成与汇聚的模式创新、应用创新和外延服务创新,形成基于大数据的科技资源信息管理、信息调度和信息处理的基础环境。此外,公共财政支持的科技计划项目、科研业务工作等也产生大量的科技资源信息,通过国家科技管理信息系统、国家科技报告制度、国家科技资源共享服务平台汇聚了大量科技人、财、物信息。对于这些信息,可探索建立统一的科技资源信息集成和汇聚平台,形成资源信息相对集中、资源实体物理分布的信息集成与汇聚体系。提高科技财政投入效率优化科技资源配置、支撑政府有效决策。

3.3 建设科技资源信息标准化体系

针对当前科技资源信息数量多、来源广、格式多样、增长速度快,如果缺少统一的标准和规范,难以实现科技资源信息的充分传播。为使科技资源信息适应大数据环境下的信息交换模式,需要制定统一的科技资源信息标准,规范各机构的科技资源信息建设行为。在国家、行业标准的基础上,应建立统一的科技资源信息数据标准,结合大数据环境下的信息处理要求与规范,建立一套从科技资源信息采集、信息传输、信息处理、信息存储、信息发布与服务等的完整标准化体系,规范提供相关科技资源信息的背景描述性信息,建设大数据背景下常态化的科技资源信息标准化日常管理规范,保证科技资源信息内容的质量和科技含金量。

3.4 加强科技资源信息挖掘与利用技术研发

科技资源信息涉及国家经济、社会发展等各个方面,将科技资源信息转化为知识甚至智慧,是未来科技创新驱动发展的必然趋势。要充分挖掘和利用科技资源信息,建立高效的信息数据分析和利用的应用技术体系,如将信息进行关联性价值分析的应用技术系统,实现多信息充分融合,

通过系统对科技资源信息进行垂直整合或对相关科技领域进行水平整合，这就需要大数据系统自动采集大量数据，学习大量相关规则。通过利用人工智能、机器学习等领域的理论，以科技资源信息数据分析为核心，融合云计算、物联网、移动互联网等大数据技术，研究开发非结构化数据处理、非关系型数据管理系统和数据可视化软件等先进大数据处理工具系统，为大数据背景下的科技资源信息挖掘与利用提供先进的技术支撑。

3.5 开展面向需求的科技资源信息分析挖掘与利用

在大数据背景下开展科技资源信息分析挖掘与利用的关键问题就是要找准需求。针对科技创新、民生发展、国家安全等关键问题，运用先进的大数据技术进行相关资源信息的收集整理，充分利用大数据的先进管理模式和信息处理手段，对科技资源信息进行有效的分析挖掘，形成可供使用的知识结果，为科技创新和管理决策提供支撑。

同时，利用大数据自身的优势，探寻面向科技创新市场的巨大需求，顺应当前以市场为主体的科技创新决策部署，研究能够满足更广泛市场需求的科技资源信息产品，以支撑科技创新研究与决策市场化发展，将成为大数据背景下科技资源信息挖掘与利用的另一重要领域。

4 结语

随着我国科技资源规模快速增长，科技资源信息海量、复杂、异构等特点愈发明显，大数据的特点和发展趋势对科技资源信息的管理和利用提出了更多更大的挑战，使得以往相对独立、封闭的科技资源应用模式难以适应新形势下的科技创新活动。虽然我国科技资源信息化程度明显提高，资源信息的公开渠道更加多样和便捷，科技资源信息基础设施也得到较快发展，但仍然面临着科技资源信息数量庞杂、信息源头众多、结构内容各异等诸多问题，从而导致科技资源信息内容的利用率较低。

我国科技资源信息的积累和整合工作起步较晚，对大数据的理论研究和技术应用也才开始。然而，当前的大数据环境为科技资源信息的发展

提供了良好的发展契机和广阔发展空间，其他领域的成功经验为科技资源信息的挖掘与利用提供了借鉴。利用先进的大数据理论和方法，深入挖掘科技资源信息价值，提高科技资源信息服务能力，有助于发挥科技资源对科技创新和经济社会发展的支撑作用。

在大数据背景下加强对科技资源信息挖掘和利用，积极面向科技创新发展和管理决策需求，加强科技资源的汇聚、公开与整合，加强科技资源信息标准化体系建设，研发相关软件与工具，加强科技资源信息基础设施建设，深度挖掘与利用相关科技资源信息，实现科技资源信息利用的最优化，保障科技资源受益群体，促进资源效益最大化，是未来相当长时间内的工作目标和主要任务。

参考文献

- [1] 王学琴, 杨剑, 康磊, 等. 大数据驱动科技信息资源市场化开发利用[J]. 中国科技信息, 2015(5): 29-30.
- [2] 丁厚德. 科技资源及其配置的研究[J]. 中国科技资源导刊, 2009, 41(2): 1-7. DOI: 10.3772/j.issn.1674-1544.2009.02.001
- [3] 李扬. 科技图书馆一站式知识产品与服务研究[C]// 全国图书馆学情报学集成创新与发展学术研讨会, 2007.
- [4] 邹佳利, 山红梅. 基于云计算的科技资源共享问题研究[J]. 科技管理研究, 2013, 33(8): 186-189.
- [5] 杨慧, 宋华. 大数据背景下的信息资源管理与利用[J]. 管理观察, 2016(3): 47-49.
- [6] 陈敏, 刘宁. 医疗健康大数据发展现状研究[J]. 中国医院管理, 2017, 37(2): 46-48.
- [7] 王宏起, 程淑娥, 李玥. 大数据环境下区域科技资源共享平台云服务模式研究[J]. 情报理论与实践, 2017, 40(3): 42-47.
- [8] 宋立荣, 齐娜. 我国网络科技信息资源开发中的问题及对策思考[J]. 中国基础科学, 2012, 14(2): 39-44.
- [9] 崔巍. 大数据在气象服务中的应用与分析[J]. 低碳世界, 2016(26): 121-122.
- [10] 张硕, 韩纪富, 史海英. 大数据时代科技期刊数字化发展探析[J]. 科技视界, 2015(33): 109-110.
- [11] 陈志辉, 孙亮, 马欣, 等. 我国科技资源共享立法策略研究[J]. 中国科技论坛, 2013(8): 5-8.