基于内容、平台、范式的科技信息高端交流平台思考

陈煦 徐宏宇 杨荣斌
（上海图书馆（上海科学技术情报研究所），上海 200031）

摘要：高端交流平台的提出是对现有科技情报工作的深化和拓展，应该成为中国科技情报体系的顶层设计，是加强我国战略科技力量的重要举措。本文从高端交流平台的资源内容、平台建设和交流范式3个方面切入，结合国际科学研究趋势，分析国外最新实践案例，提出高端交流平台需要多样化内容、多主体建设、拥抱全新科研范式的思考，以期为我国高端交流平台的建设提供借鉴。

关键词：高端交流平台；科研范式；科学数据；人工智能
中图分类号：G359　DOI：10.3772/j.issn.1673-2286.2021.10.001
引文格式：陈煦，徐宏宇，杨荣斌. 基于内容、平台、范式的科技信息高端交流平台思考[J]. 数字图书馆论坛，2021（10）：2-7.

《中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要》在“强化国家战略科技力量”中，提出构建国家科研论文和科技信息高端交流平台（以下简称“高端交流平台”）的具体要求。高端交流平台的提出与60多年前我国科技情报体系的构建有着相似之处。1956年，科技情报被《1956—1967年科学技术发展远景规划》列为第57个重大科技任务，这是科技情报首次被写入国家规划。2020年，高端交流平台建设被提出，可以看作对现有科技情报工作的深化和拓展。高端交流平台的内涵尚待深化、细化，本文从“高端”“交流”“平台”的字面切入，提出对建设高端交流平台的一些思考。

1　高端交流平台的“高端”需要多样化内容支撑

高端交流平台的内容主要来自高水平科技期刊、高质量的科学数据等科技信息[1]。过去，Elsevier、Springer、Wiley、IEEE、SAGE五大出版商控制了全球约50%的科学出版。21世纪，大数据、深度学习技术取得突破性进展，以期刊论文、会议论文、科技报告、专利为主的传统科技信息资源已经无法完全满足科学研究的需求，科技信息资源逐渐扩大到数据、数据集、代码、图片和视频等。这给我国建设高端交流平台带来了前所未有的机遇。一方面是数据资源的兴起，人工智能技术的快速发展为科技情报服务提供了新方法和新工具，科学研究历经演变逐渐走向数据密集型科学发现的第四范式，数据类资源的重要性相较于传统的文本信息不断增大；另一方面是数据资源跨越语言的流通性，传统文本信息的收集、加工和传播对英语国家有着得天独厚的优势，而数据科技信息则恰巧弱化了此类优势。

美国的科技信息平台建设一直走在国际前沿，很早就开始在科学数据领域布局。2009年，美国总统奥巴马一就任就签署了《透明与开放政府》（*Transparency and Open Government*）备忘录。同年5月，当时刚上任的美国联邦首席信息官（CIO）维维克·昆德拉（Vivek Kundra）宣布建立Data.gov网站。该网站由美国总务管理局下属技术转化服务部（U.S. General Services Administration, Technology Transformation Service）

管理和托管，是联邦、州、地方和部落政府信息的存储库，向公众开放，旨在改善公众对联邦政府行政部门生成的高价值机器可读数据集的访问。2016年，美国政府又进一步发布政府开放代码平台Code.gov，开放定制开发的联邦源代码以供公众重用。据美国能源部科技信息办公室（OSTI）主任布莱恩·希森（Brian A. HITSON）在2019竞争情报上海论坛上的演讲可知，美国联邦政府每年大约投入1 500亿美元用于支持研发，产出的约20万篇期刊论文、7 000份专利、艾字节级别的数据、6 500个开源软件项目都可以在Science.gov（收录所有美国政府机构的科技信息）、Data.gov、Code.gov上检索到，至此，美国联邦政府形成文献、数据、代码三足鼎立的开放科学格局。

除了美国联邦政府层面，在部际层面的科技资源建设近年来也将建设重点放在了数据和代码上。以成立于1947年的美国能源部科技信息办公室为例，虽然与美国国会图书馆、美国国家医学图书馆等老牌图情机构相比成立时间不长，但它在收集、保存、传播能源部开展的研发、示范和商业应用活动所产生的科技信息方面的业绩是被业界认可的[2]，美国四大科技报告之一的能源系统DOE报告就由其发布。据希森介绍，能源部每年投入约120亿美元，在全美17家国家实验室产出超过5万项的科技信息成果，这些成果统一由OSTI对其进行建设维护。OSTI为了更好地利用这些数据集、图片、视频、可视化成果、软件代码等各类资源，给每一种特定资源都定制了特殊的检索工具库，如期刊文章/收录手稿（www.osti.gov/pages）、数据集（www.osti.gov/dataexplorer）、软件（代码）（www.osti.gov/doecode）、专利（www.osti.gov/doepatents）、视频（www.osti.gov/sciencecinema）。除了各类资源，OSTI在这些成果的索引关联上也动足脑筋，每项成果既有独立唯一的DOI，也能在页面上一键外链到与其相关的文献、代码或数据，实现资源的跨库关联。同时，所有的特定资源也可以在总数据库（www.osti.gov）中找到。

伴随着数据时代带来的科研范式转变，数据、软件、视频、编程代码成为科技信息的重要组成部分，将这些信息收集、整合、存储并关联对资源的使用影响深远、意义重大。从近年来美国政府建设科技信息平台的经验做法来看，不难发现科技资源，尤其是包括数据、代码等新兴科技信息的多样化科技资源，是新时代下强化国家战略科技力量必不可少的基础设施。我国的高端交流平台建设在补足高水平科技期刊短板的同时，也要抓住科技数据发展的机遇，将数据、代码无缝纳入科技信息的大图景中，重视科技数据、代码的收集、保存和共享。

2　高端交流平台的“平台”需要多主体参与建设

从上述美国科技信息平台的建设经验可以看出，数据、代码等科技数据是科学研究未来的发展趋势，也是关键的科研基础设施。笔者聚焦科学数据平台进行了初步检索，汇总整理了中国科技信息平台（见表1），主要包括科学数据银行（Science Data Bank）、数据出版学术期刊《中国科学数据》、中国科技云、中国科学院数据云、20家国家科学数据中心和31家国家资源库。其中科学数据银行、《中国科技数据》、中国科技云、中国科学院数据云都由中国科学院计算机网络信息中心建设，前三者都是中英双语界面，面向国内外科研人员。尤其是科学数据银行，是国内唯一一家被施普林格·自然列为推荐的7家通用型数据存储库之一。中国科技资源共享网由国家科技基础条件平台中心建设，汇集了20个国家科学数据中心和31个国家资源库的科技资源。与国外相比，我国科技数据的平台初具规模，但几乎都是国家建设的科技信息平台，很少有企业或民间团体建设的成型的科技数据平台。以国内对标谷歌的百度为例，虽然拥有“百度数据开放平台”，但是页面并无检索功能，提供的也是一些天气预报、列车航班、休闲娱乐等信息，很难说可以为科学研究提供数据支撑。软件代码方面，国内较为知名的大型科技公司，阿里巴巴、腾讯和华为虽有各自的平台（阿里云-天池、腾讯云、华为云），但腾讯云和华为云与真正意义上的科学数据平台相差甚远。阿里云-天池呈现出部分数据平台的雏形，如提供数据集检索下载，与其他企业、科研机构、大学合作举办算法代码竞赛等，但还远未达到能够服务于国内外科学研究的水准。

在美国科技数据平台的生态中，除了联邦政府和机构官方搭建的科技信息平台，企业也是平台建设的重要主体之一，而且在数据集和代码两大领域均有布局。数据集方面，谷歌公司开发的谷歌数据集搜索引擎（Google Dataset Search），可以检索各个来源的数据集并下载，平台与包括OSTI在内的多个政府机构数据库通过DOI相关联，数据集和索引来源都会标注清楚，每一个数据源都会有简介、更新日期、作者、版权、内

表1 中国科技信息平台

平 台	主管机构	详 情	网 址
科学数据银行	中国科学院计算机网络信息中心	2021年1月底正式发布的自主研发、具有国际化服务能力的论文关联数据存储库平台。其前身于2015年开始上线提供服务，能够为论文关联数据的汇聚、管理、开放、共享提供高效的解决方案。2020年被施普林格·自然列为推荐的7家通用型数据存储库之一，也是国内唯一一家	https://www.scidb.cn
《中国科学数据》	中国科学院主管，中国科学院计算机网络信息中心和ISC CODATA中国全国委员会合办	《中国科学数据》是目前中国唯一的专门面向多学科领域科学数据出版的学术期刊，2015年创刊，面向内外公开发行，为中英文季刊，中国科学引文数据库（CSCD）来源期刊	http://www.csd2.1ata.org
中国科技云	中国科学院计算机网络信息中心	基于云计算的国家科研信息化基础设施，人工智能计算及数据服务应用平台，是国内第一个基于云计算的服务科研人员的“云”。平台把科学家要在各个网站上寻找的数据和计算工具整合在一个网站上，分门别类地提供给科学家，供科学家按需使用	http://www.cstcloud.cn/#/
中国科学院数据云	中国科学院计算机网络信息中心	面向科研创新的科学大数据服务平台，可以直接进行科技信息资源检索，下载资源外链到相关数据库	http://www.csdb.cn
中国科技资源共享网-国家科学数据中心（20家）	中国科学院	国家高能物理科学数据中心、国家基因组科学数据中心、国家微生物科学数据中心、国家空间科学数据中心、国家天文科学数据中心、国家对地观测科学数据中心、国家青藏高原科学数据中心、国家生态科学数据中心、国家冰川冻土沙漠科学数据中心、国家地球系统科学数据中心、国家基础学科公共科学数据中心	https://www.escience.org.cn/data-center
	中华人民共和国农业农村部	国家农业科学数据中心	
	中华人民共和国国家林业和草原局	国家林业和草原科学数据中心	
	国家气象局	国家气象科学数据中心	
	中国地震局	国家地震科学数据中心	
	中华人民共和国自然资源部	国家极地科学数据中心、国家海洋科学数据中心	
	国家市场监督管理总局	国家计量科学数据中心	
	中华人民共和国国家卫生健康委员会	国家人口健康科学数据中心	
	中华人民共和国教育部	国家材料腐蚀与防护科学数据中心	
中国科技资源共享网-国家资源库（31家）	中华人民共和国农业农村部	国家热带植物种质资源库、国家作物种质资源库、国家园艺种质资源库、国家家养动物种质资源库、国家海洋水产种质资源库、国家淡水水产种质资源库、国家菌种资源库、国家禽类实验动物资源库	https://www.escience.org.cn/resource-library
	中国科学院	国家重要野生植物种质资源库、国家水生生物种质资源库、国家病毒资源库、国家干细胞资源库、国家植物标本资源库、国家动物标本资源库、国家模式与特色实验细胞资源库、国家非人灵长类实验动物资源库、国家鼠和兔类实验动物资源库	
	中华人民共和国国家林业和草原局	国家林业和草原种质资源库	
	中华人民共和国国家卫生健康委员会	国家寄生虫资源库、国家病原微生物资源库、国家人类生殖和健康资源库、国家发育和功能人脑组织资源库、国家生物医学实验细胞资源库、国家人类疾病动物模型资源库	
	中华人民共和国教育部	国家健康和疾病人脑组织资源库、国家干细胞转化资源库、国家岩矿化石标本资源库、国家遗传工程小鼠资源库	
	国家市场监督管理总局	国家标准物质资源库	
	广东省科技厅	国家犬类实验动物资源库	
	中华人民共和国国家药品监督管理总局	国家啮齿类实验动物资源库	

容说明、下载链接等。深度学习视觉领域常用的开源数据集MNIST、Imagenet、COCO、CIFAR等也能在谷歌平台上[3]找到。在处理数据集的代码方面，软件源代码托管服务平台GitHub在2018年被美国科技公司微软收购。GitHub是一个典型的协作平台，通常用于科学项目的协作管理和代码数据的共享，其中GitHub的Jira用于总体项目管理、研究问题的提出，而OmniPlan则用于创建研究的时间表和跟踪时间[4]。另一家创立于2010年的数据建模和数据分析竞赛平台Kaggle在2017年被谷歌公司收购，是一家为开发商和数据科学家提供举办机器学习竞赛、托管数据库、编写和分享代码的平台，吸引了众多科学家和开发者。

除了各自建设、并购数据平台，企业和政府机构也会合作建设数据平台。2020年3月16日，艾伦人工智能研究所（AI2）与白宫科技政策办公室（OSTP）、美国国家医学图书馆（NLM）、陈扎克伯格倡议（CZI）、微软研究院、数据科学家代码分享平台Kaggle，在乔治城大学安全与新兴技术中心（CSET）的协调下，发布了新冠病毒公开数据集（COVID-19 Open Research Dataset，CORD-19）[5]的第一个版本。该数据集的资源来自WHO、PMC、bioRxiv等数据库，由艾伦人工智能研究所的语义学者团队筛选，组成COVID-19和此前发现的冠状病毒（如SARS和MERS）的出版物和预印本集合，数据经过清洗和规范，统一整理为易于自然语言处理的JSON数据格式共享，数据集每几天更新一次。CORD-19旨在将机器学习社区与生物医学领域专家和政策制定者联系起来，以期为COVID-19确定有效的治疗方法和管理政策。数据集在跨界的用户团体和社区获得了积极反馈，CORD-19数据集在发布的第一个月内被查看超过150万次，下载超过7.5万次。多家团体使用数据集构建了搜索和提取工具。围绕数据集涌现了一个蓬勃发展的用户社区，共同讨论和共享信息、注释、项目以及反馈。Kaggle和白宫科技政策办公室、艾伦人工智能研究所还共同主办了CORD-19 Research Challenge开放式文本挖掘竞赛，参与者的任务是从CORD-19的论文中提取有关COVID-19关键科学问题的答案。另外，还有由艾伦人工智能研究所、美国国家标准与技术研究所（NIST）、美国国家医学图书馆、俄勒冈健康与科学大学（OHSU）和德克萨斯大学健康科学中心（UTHealth）共同组织的TREC-COVID信息检索共享任务。这项任务可以评估检索系统根据查询主题，对CORD-19数据集中的论文进行相关性排名的能力。两项任务都由生物医学领域的专家来评估数据科学家提交的代码，评选出的最优代码会被置顶供大家使用和分享。

我国的科技信息平台建设起步较早，与国外的基础设施平台相比，差距不像科技期刊那么悬殊，随着我国在科技信息、人工智能、大数据中心建设等方面不断发力，也为日后建设科技数据的高端交流平台打下了坚实的基础。同时，也要看到，我国的“国家队”虽然在科技数据的基础设施建设上初见成效，但国内企业在这一领域有所缺席。从美国科技数据、代码平台的情况可以看到，企业的科技信息交流平台和政府创建的科技信息基础设施各自面向不同的用户，在服务科学家的科学发现、科学研究的产业链中各司其职，填补不同的空缺。此外，如何在短时间内围绕某一主题集结各利益相关方（跨领域的企业、政府机构、公益组织等）构建科技信息平台，如何打通政府、企业、非营利组织的跨界合作，以及如何跨越不同学科建立联系，将科技信息利用起来转化为对社会有益的成果，CORD-19数据集的案例为我们提供了一张路线图。我国应当同时加强企业建设的高端交流平台，只有让更多主体参与到高端交流平台的建设中来，才能更好打通企业和各层级政府的检索渠道，进一步扩大科技信息的触及范围，造福科学家，助力科学发现，强化我国战略科技情报力量。

3　高端交流平台的“交流”需要拥抱全新科研范式

科技信息平台最终是为了实现科学家之间快速、有效的交流——获得信息，同时传播自己的研究成果。对于科技信息资源的利用和信息获取，过去科学家主要通过阅读文献来了解最新科研进展并进行创新，科学研究的主体依然是科学家。人工智能的快速发展，加上新冠疫情给科学研究带来的紧迫性，使得当前科学研究呈现出了某种范式上的转变——科学家需要更快速地开展工作，而全球新冠疫情期间产生的科学研究和论文数量惊人，大大超出了任何人吸收消化信息的能力[6]。于是，快速从文献和数据中获取信息的研究变得非常重要，换言之，科学研究不再只是科学家的事，定制化的数据分析处理工具和科技情报人员，尤其是擅长人工智能、机器学习、自然语义处理等技术的情报人员，将在科学研究中扮演更重要的角色，甚至成为科学研究不可或缺的环节。

这一趋势早有端倪，美国化学文摘社的SciFinder，斯普林格-自然出版集团的SpringerMaterials、AdisInsight，爱思唯尔的Reaxys、Knovel、ClinicalKey等，这些学术出版机构的产品经过专业人员的整编后不再是简单的数据库，而是实现了相关知识内容的集成汇聚，成为化学物质、化学反应、材料研究、药物研发等研究领域有用的专业知识工具。有些工具已经经过实践验证，CAS利用自身已有的大规模、高质量的化学反应数据支持Bayer公司对相关化合物合成方案可行性的预测，结果显示预测的准确率提高了32个百分点[7]。

新冠疫情加速了这一趋势。首先是科学家的需求，新冠疫情时期海量文献和争分夺秒的研发进度让科学家不得不借助情报手段。此外，新冠病毒的大流行吸引了全球各行各业的广泛关注，来自生物、医药、临床、人工智能、大数据等不同领域的科学家都从自己的专业角度贡献力量，这种凝聚力从某种程度上构建了一个跨界合作创新的理想环境。两者的共同作用催化了科研范式的大转变。前文提到的CORD-19数据集就是以公开数据集作为平台，吸引人工智能、数据挖掘领域科学家根据需求定制工具方案，“外包”平台检索、索引、挖掘、分析的功能，再通过用户（学科专业人士）反馈选出最优工具的路线，实现科学家、数据学家、人工智能专家之间有效的信息交流。哈佛大学医学院INDRA实验室研发EMMMA（Ecosystem of Machine-maintained models with Automated Analysis），该系统对海量的生物医学文献进行机器自动化阅读并提取250万种知识点之间的关系，与生物数据库（Pathway Commons，SIGNOR和BEL Large Corpus）构成的先验知识网络进行融合，自动化构建了COVID-19病毒生物学的因果机制，用于新药物的研发等[4]。微软在2021年3月推出生物医药搜索引擎，能够让研究人员用自然语言而不是关键词/术语来检索获取专业文献。平台还把预测未来可能会变得重要的文献提到前面展示，平衡旧文献因为被引率高而占优势的弊端[8]。这样的科研范式被证明是成功的，在各界的努力下，从新冠病毒的基因序列在2020年1月发布，到疫苗在多个国家/地区获批紧急使用，前后只有短短不到1年的时间，而通常情况下疫苗的研发需要5~10年。

新冠疫情影响下，另一个科学研究交流范式的转变是预印本平台的加速发展，本文也将其笼统地归纳为科研范式的一种。疫情初期，为了加速科研成果的分享，科学家纷纷选择在预印本平台上公开自己的研究成果，科学家和公众也通过预印本得以更快地获得新冠疫苗、新冠药物全球研发的进展。各大预印本平台中，康奈尔大学的arXiv和爱思唯尔的SSRN都建立于20世纪90年代，威利的Authorea建立于2012年。2016—2019年曾出现预印本平台的建设高峰，非营利性研究和教育机构冷泉港实验室的bioRXiv、medRXiv，施普林格·自然的Research Square，中国科学院的ChinaXiv，美国化学学会的ChemRxiv，瑞士多学科数字出版机构（MDPI）的Preprints在这一时期相继建立。还有许多迹象显示预印本平台对于科研越来越重要，全球65家预印本服务器（preprint server）大约有一半在过去5年内成立，世界排名前十的学术出版商大多建立或收购了一个预印本服务器，科研基金会如比尔及梅琳达·盖茨基金会、英国惠康基金会积极资助预印本平台，一些权威学术文献检索平台如Scopus、Europe PMC、SCIE开始将预印本的内容纳入其索引范围[9]。

数据与人工智能驱动的科学研究也好，预印本也好，都是被讨论和关注较多的科研范式趋势，这些趋势是否真的对科学研究具有变革性的影响仍然存在争议[10]，但最终都是为了更好更快的科学交流与分享。我国高端交流平台的建设需要在对全球科研范式、交流模式进行深入研究的基础上，把握并拥抱全球科研范式的转变。

4 结语

本文从高端交流平台的“高端”内容、“平台”主体、“交流”范式3个方面，结合国际科技情报和科学研究的大趋势，列举了一些国外最新的实践案例，对我国高端交流平台发展与建设提出了思考。笔者认为，高端交流平台的“高端”需要多样化内容支撑，需要扩大科技数据范围，尤其重视数据、代码等新型科学信息；高端交流平台的“平台”需要多主体参与建设，要鼓励个人、中小企业和其他大型企业参与数据科学的平台建设与共享；高端交流平台的“交流”模式需要拥抱全新科研范式，要抓住科研范式的变化趋势，调整平台以满足科学家对于知识获取、知识交流传播的需求，提高科研人员研发的效率和预测的准确性。要从内容建设、平台建设、交流范式等各方突破，才能打造具有科技强国战略视角的“高端交流平台”，这也是我国“强化国家战略科技力量”的必要保障。

囿于字数限制，本文对于国外案例只是介绍性分

析，就现象论现象，没有探究现象背后的深层逻辑和原因。对于高端交流平台的思考也是一种设想，没有落实到具体可执行的建设路径，希望在后续的研究中能有所加强。

参考文献

[1] 陈超. 再议“高端交流平台”[J]. 竞争情报，2021，17（5）：1.

[2] 武夷山. 美国能源部科技信息办公室成立70年了[EB/OL]. [2021-09-22]. http://blog.sciencenet.cn/blog-1557-1052572.html.

[3] 都保杰. 谷歌搜索神器：Dataset Search数据集搜索了解下[EB/OL]. [2021-09-22]. https://www.sohu.com/a/252574133_99970711.

[4] 李广建. 高端交流平台及其情报计算能力建设[J]. 数字图书馆论坛，2021（3）：3-8.

[5] WANG L L，LO K，CHANDRESEKHAR Y，et al. CORD-19：The COVID-19 Open Research Dataset[EB/OL]. [2021-09-22]. https://aclanthology.org/2020.nlpcovid19-acl.1.pdf.

[6] CULLER L. Q&A：Peter Lee on the COVID-19 pandemic，societal resilience and crisis-response science[EB/OL]. [2021-09-22]. https://news.microsoft.com/innovation-stories/peter-lee-resilience/.

[7] 张智雄. 高端交流平台建设需要把握知识服务的发展大势[J]. 智库理论与实践，2021，6（1）：5-6，9.

[8] HORVITZ E. Aiming advances in AI at biomedical search[EB/OL]. [2021-09-22]. https://blogs.microsoft.com/ai-for-business/biomedical-search/.

[9] University of Hong Kong Libraries. The Landscape and Transformative Role of Preprints[EB/OL]. [2021-10-02]. https://hku.zoom.us/j/94746575645?pwd=MVVpc3ZlZjRrK1RzUWRRVXNoMG9lZz09.

[10] BRAINARD J. No revolution：COVID-19 boosted open access，but preprints are only a fraction of pandemic papers[EB/OL]. [2021-10-02]. https://www.science.org/content/article/no-revolution-covid-19-boosted-open-access-preprints-are-only-fraction-pandemic-papers?utm_campaign=news_weekly_2021-09-10&et_rid=511589141&et_cid=3915480.

作者简介

陈煦，女，1990年，助理研究员，研究方向：竞争情报、情报方法。
徐宏宇，女，1981年，硕士，研究员，研究方向：竞争情报、情报方法。
杨荣斌，男，1972年，硕士，研究员，通信作者，研究方向：科技情报、科技趋势，E-mail：rbyang@libnet.sh.cn。

Thoughts on High-end Communication Platform of Science and Technology Information Based on Content, Platform and Paradigm

CHEN Xu　XU HongYu　YANG RongBin
(Shanghai Library (Institute of Scientific and Technical Information of Shanghai), Shanghai 200031, China)

Abstract: The proposal of a high-end communication platform is a deepening and expansion of existing scientific and technological information work. It should be the top-level design of China’s S&T information system. It is an important measure to enhance strategic S&T strength of China. From the three aspect of content, platform and paradigm, this article studies international scientific research trends, analyzes latest foreign best practice, proposes that the high-end communication platform needs diversified content, diversified player, and embrace new scientific research paradigm. The article wishes to provide a reference for the construction of China's high-end communication platform.

Keywords: High-end Communication Platform; Scientific Research Paradigm; Scientific Data; Artificial Intelligence

（收稿日期：2021-10-03）