

教育部工程研究中心年度报告

(2022年1月——2022年12月)

工程中心名称：数字图书馆

所属技术领域：信息与电子工程

工程中心主任：庄越挺

工程中心联系人/联系电话：张寅/13136157369

依托单位名称：浙江大学

2025 年 3 月 19 日填报

一、技术攻关与创新情况

2022年，中心沿着新一代人工智能、数字图书馆、知识中心的发展脉络，围绕知识和数字内容为主的信息服务产业需求以及国家重大战略需求，按照“从数字图书馆走向智慧图书馆，进而升华到知识中心”的技术研发路线，研发数字内容与知识服务核心技术。中心发起成立数字知识服务联盟；发布了AI发展水平、全球智库影响力评估报告，建设工程科教图书、中草药、辅助科研、产业链等知识服务系统，设计并研发OpenKS知识计算引擎系统，支撑构建基于知识图谱的问答、推荐等应用，形成优秀的研究与工程技术团队，培养博士生10余名、硕士生30余名。

2022年，数字图书馆教育部工程研究中心承担国家级、省部级研发任务5余项，企业研发任务10余项，继续支撑CADAL项目运营，中心学术主任潘云鹤院士、中心主任庄越挺教授通过相关学术报告，探索了人工智能的未来走向和前沿应用，为该领域的创新发展提供有力支持；中心技术团队支撑研发了元宇宙数字图书馆原型系统，研发了“与古人对话”原型系统支持自然语言交互，为专业领域知识消费者提供了沉浸式知识获取与交流体验；提出了融入词典知识进预训练模型的DictBERT方法、基于实例的并行查询网络方法以及多粒度多模态视觉文档理解模型。

中心技术团队支撑的CADAL项目本年度完成西文古籍数字化3000册(1,074,505页)，高校人文社科获奖成果专题库36册(17,608页)，俄文专题文献2770册(1,060,239页)，德文文献资源532册(183,100页)，四川大学特色馆藏“张之洞捐俸置书”700册(117,694页)，民国期刊589册(42,864页)，当代图书3册(1050页)，徽州文书17册(4,752页)，红色文献1075册(108,476页)，西泠印社印谱2082册(205,841页)，云南省碑拓特色资源207张，敦煌写本献25,510件，契约文书3,067件。截至2022年12月，CADAL入库总量2,888,088册(件)，在线量2,720,420册(件)。

支持构建中国最大的篆刻数据集，目前共计收藏印谱资源2000余种(全国共8000种)，印章资源约50多万枚。中心研发的2.5维数字化设备从48个不同的角度照射甲骨，拍摄48张照片，全方位覆盖甲骨的每一个刻痕和细节，完整呈现出甲骨表面的深度信息。建成“甲骨数字化”特藏库，让甲骨实物走出象牙之塔，共享于世。

已建成民国文献大全、甲骨文、老照片、现代生活资料等10个特藏库的基础上，新增中国历代墓志数据库，探索通用型特藏平台建设，重点推出红色文献、印章印鉴、写本文献三个主题的数字特藏库建设。

针对海量数字内容造成信息过载，多粒度图文对齐学习，实现不同媒体的跨越式关联与检索，帮助读者快速精准地获取所需内容。中心自主研发了中草药自动问答系统，利用问答系统双向注意力关系匹配模型等算法，实现药材、配方等问诊生成。

针对推荐物品涉及某一特定领域、领域知识的重要性、用户兴趣挖掘的困难等问题，中心

技术团队自主研发了中医医案推荐系统，通过基于图神经网络和多任务学习技术的研发，提出聚合不同层次特征的基于自注意力机制的图神经网络NGAT，解决过度平滑问题。

二、成果转化与行业贡献

1. 总体情况

2022年，中心技术团队研发的OpenKS 系统，以帮助行业快速构建行业知识图谱，提供行业相关的智能规划与决策支持为目标，形成了一整套可服务于知识密集型行业共性需求的知识计算算法库。

中心技术团队研发了AI Paper Collector，生成高效检索和筛选顶会论文的工具，AI Paper Collector和谷歌学术、arXiv和ReadPaper平台等学术平台相比较，性能更加优越，在第一届机器学习算法和自然语言处理大会（MLNLP2022）上分享，获得36000人次在线观看。同时，中心技术团队采集近十年NLP，CV，ML等领域的论文，基于OpenKS进行结构化信息抽取，通过学术实体和关系抽取，形成学术知识图谱和算法脉络图。可以帮助科研人员快速了解海量文献中算法的演化关系，以及算法的极简摘要信息，帮助科研综述的生成。

面向产业链认知决策场景，中心技术团队开展大数据治理与数据平台架构的研究，攻克产业链与创新链知识图谱构建关键技术，最终形成算法模型或平台工具。将产业链图谱的完整构建上线周期从1个月压缩至1周。在德清试点中，已累计推送企业与项目有效信息超过1万条，帮助拜访企业和项目200多个，招商项目综合落地转化率提升30%。团队通过联合量知科技承担浙里攻关在线和产业一链通重大应用建设，入选2021年浙江省数字化改革最佳应用。

2022年度，中心技术团队支撑的CADAL资源总访问量96,211,677次，首开“CADAL”官方微信公众平台以及视频号，总浏览量9691次，关注者244人，通过视频号直播的“基于大数据的个性化推荐：思路与实践”线上技术研讨会共计334人次收看和回看研讨会的直播视频。

由“大学数字图书馆国际合作计划（CADAL项目）”倡议，与武汉大学、北京大学、清华大学、上海交通大学、北京师范大学等高校联合创建数字知识服务联盟，英文名称为Digital Knowledge Services Alliance，缩写为DKSA，是一个非法人性质的组织。本年度，共有85家出版社加盟数字知识服务联盟，可售中文电子图书数量达60,833册，目前除17所发起高校图书馆外，还有43所高校图书馆申请加入联盟。

2022年4月19日，潘云鹤院士基于中美公开数据，做中国人工智能发展评估报告。

2022年7月1日，在中国智库网发表《2021人工智能发展水平评价分析 报告》。

2. 工程化案例

一、CADAL数字图书馆开放赋能

2022年度，中心技术团队支撑的CADAL资源总访问量96,211,677次，其中网站门户访问总量38,317,345次，Open API接口对馆藏资源有效调用请求57,894,332次。新增共建共享签约单位51所，新增用户注册总量48,817个。CADAL项目采用云服务方式，依靠分布全国的服务网络，通过CADAL读者服务协同工作平台、CADAL服务工作群等为全球读者服务，做到“有问必答，有难必解”。

首开“CADAL”官方微信公众平台以及视频号，其中公众号共计推送17条图文消息，总浏览量11,388次，粉丝量2216人，视频号共计发布23条短视频，总浏览量9691次，关注者244人，通过视频号直播的“基于大数据的个性化推荐：思路与实践”线上技术研讨会共计334人次收看和回看研讨会的直播视频。

2022年度，共有85家出版社加盟数字知识服务联盟，可售中文电子图书数量达60,833册；CADAL项目中心不断完善联盟数字平台建设，根据实际需求不断完善平台功能，为联盟成员馆提升学术信息资源建设水平和支撑高校教学科研服务能力提供了可靠保障；积极推广，稳步开展联盟纳新工作，目前除17所发起高校图书馆外，还有43所高校图书馆申请加入联盟。

二、数字内容技术应用推广

针对数字图书馆的资源组成发生显著变化，需要分析多种类型的数字资源，包括文本、图片、视频等需求，中心自主研发了文本信息抽取-嵌套实体识别技术。解决实体存在嵌套和非连续的结构，依赖序列标注的解码方案难以识别，对于长实体，序列标注方案会导致实体断裂现象等问题。

针对海量数字内容造成信息过载，研发多粒度图文对齐学习技术，实现不同媒体的跨越式关联与检索，技术中心自主研发了问一问自动问答系统、专业知识问答系统，利用问答系统双向注意力关系匹配模型等算法，实现药材、配方等问诊生成。

针对推荐物品涉及某一特定领域、领域知识的重要性、用户兴趣挖掘的困难等问题，中心技术团队研发了基于图神经网络和多任务学习技术，提出聚合不同层次特征的基于自注意力机制的图神经网络NGAT，实现了中医医案推荐系统的研发。

三、知识计算引擎辅助产业链、科技大脑

面向产业链认知决策场景，中心技术团队开展大数据治理与数据平台架构的研究，攻克产业链与创新链知识图谱构建关键技术，最终形成算法模型或平台工具。基于OpenKS以及产业链知识图谱构建的大量实践经验，总结形成人机协同知识图谱构建流程标准规范，定义了从产业链数据源到点状知识元素、树状知识体系、网状知识图谱的一整套产业链知识图谱协同构建范式。基于OpenKS设计研发知识在线学习与图谱运维管理工具，辅助研发人员高效地完成知识的获取与图谱的校验，将产业链图谱的完整构建上线周期从1个月压缩至1周。基于OpenKS构建招商推荐指标模型，从企业优质程度、企业动迁意愿、产业协同分

析等方面设计模型，并利用实际招商工作反馈与机器学习算法给出推荐目标清单。德清试点中，已累计推送企业与项目有效信息超过1万条，帮助拜访企业和项目200多个，招商项目综合落地转化率提升30%。团队通过联合量知科技承担浙里攻关在线和产业一链通重大应用建设，入选2021年浙江省数字化改革最佳应用。

按照“数据+知识+决策能力”的一体化设计思路，搭建科创知识中台，支撑重大应用。中心技术团队通过内聚外联、多跨融合国家/省/地市科技业务数据以及跨部门数据，并接入全球专利等第三方数据，构建科技企业、高校院所、平台载体、创新人才等科技创新主题库，形成十联动科技数据仓。

3. 行业服务情况

2022年，中心技术团队支撑的CADAL项目，坚持优化服务模式，提高服务质效，大力提升服务和组织能力。采用云服务方式，依靠分布全国的服务网络通过CADAL读者服务协同工作平台、CADAL服务工作群等为全球读者服务，做到“有问必答，有难必解”。推进与成员馆API对接，提升资源效能，拓展服务范围，让更多的读者了解CADAL并享有CADAL项目的个性化服务，充分揭示CADAL数字资源和服务体系的最大价值。

扩大服务范围，推进API对接，增进高校共享。2022年，CADAL 网站访问 总量达到了33,795,922次，Open API接口有效调用52,680,947次，共建共享单位，931家，资源总数2,852,494册，用户总数575,975个。

三、学科发展与人才培养

1. 支撑学科发展情况

工程研究中心支撑学科发展的举措和成效如下：

（1）推动数字图书馆走向智慧图书馆

联合国内外顶尖智力，将大数据、人工智能等先进技术引入数字图书馆。推动建立高校科研数据管理工作组，支持上海图书馆、复旦大学图书馆、北京大学图书馆举办开放数据大赛。作为中国工程院工程科技知识中心的技术中心，研制KS-Studio普遍应用于各分中心知识服务系统的构建，同时承担“中国工程科技机构与专家库”、“工程科技图书知识服务”、“中草药知识服务”等知识中心项目，成为数字图书馆知识服务的有效示范。

（2）推动人工智能与科教交叉融合—智海科教平台

为了促进优秀教材资源共享，本中心成员主导发布了“智海—新一代人工智能科教平台”（www.wiscean.cn），汇聚了前沿技术和产业资源，联动政校企力量，搭建开源、开放、互通的新一代人工智能生态体系，深度聚焦人工智能人才培养、学科交叉和人工智能生态建设，推动人工智能交叉学科范式变革、赋能场景应用。“智海”的算法实训平台为“智海在线”（<http://www.wiscean.cn/online/>）和“智海-Mo平台”（momodel.cn）。

2. 人才培养情况

2022年，中心与华为、海康威视、阿里巴巴、百度联合培养研究生，在CVPR、ACL、AAAI、ACM MM、SIGIR、WWW、TKDE、TVCG、TNNLS、TMM等顶级学术会议和期刊上发表了20余篇学术论文，申请中国发明专利10余项。

本年度，中心技术团队支撑cadal项目完成合作交流会议 4次；技术交流会议17次；数字化加工培训会议1次；地区服务推广会议4次；数字知识服务联盟会议4次。

3. 研究队伍建设情况

2022年，引进青年教师朱霖潮，担任浙江大学百人计划研究员。2019年博士毕业于悉尼科技大学，2019年7月至2022年7月担任悉尼科技大学讲师，获谷歌学术研究奖（2021）。在IEEE T-PAMI、IJCV、CVPR、ICCV等高水平学术期刊及会议发表论文50余篇。曾获得美国国家标准总局TRECVID LOC竞赛冠军（2016）、THUMOS动作识别竞赛冠军（2015）、EPIC-KITCHENS第一视角动作识别竞赛冠军（2019，2020）、CVPR MABe多智能体行为建模竞赛冠军（2022）等8项国际竞赛冠军。主要研究方向为表征学习，时序建模，模型迁移，多模态感知、预测、生成及交互，人工智能交叉领域研究。

四、开放与运行管理

1. 主管部门、依托单位支持情况

2022年依托单位浙江大学为数字图书馆教育部工程研究中心提供了5200方研发运行场地，拨款专项经费50万元，用于中心购置设备、举行学术活动以及中心日常运营的支出。

2. 仪器设备开放共享情况

2022年度，中心技术团队支撑的CADAL以地区中心为抓手，各地区中心对接区域内高校，积极开展宣传推介活动；以共建共享为理念，与西藏大学图书馆已成功实现和CADAL平台开放检索接口（OPEN API）对接推动CADAL资源进藏区；以服务升级为导向，开放数据助力技术研究，先后支持第七届上海图书馆开放数据竞赛和第四届“慧源共享”全国高校开放数据创新研究大赛，充分释放开放数据价值；以学术研讨为窗口，探索资源建设新路径，主办“基于大数据的个性化推荐：思路与实践”线上主题研讨会，参与第十四次中文文献资源共建共享合作会议、第八届全国外语院校图书馆联盟年会等学术研讨会，广泛开展交流活动；以文化传承为目标，以艺术乡建为形式，分别在国家级传统村落松阳南岱村的问山美术馆、杭州市弥陀寺文化公园、杭州海塘遗址博物馆举办民国门神画像巡展，让中国传统文化与艺术在乡村与城市中得以创造性转化、创新性发展。

2011年购置的按需印刷机和大幅面扫描仪，面向浙江大学内部共享使用，提供数字内容打印和制作服务。

中心技术团队支撑的CADAL项目数据保障落实多副本异地备份及第三方云存储独立备份，数据总量约325TB。2022年运行状况稳定，数据可靠。机房监控大屏完成安装部署，提高了日常运维效率，增强管理多维度、精细度。进行故障处理和配置优化101人次，完成日常巡检、故障预警处理及设备优化等工作。

CADAL门户网站完成国家信息系统安全等级保护三级备案年度复测，验收合格。

3. 学风建设情况

2022年，中心在疫情期间，充分利用钉钉、微信会议等远程协同工具，有效执行了考勤制度，促进了线上线下协同，邀请腾讯、华为、阿里巴巴、大华、讯飞等大企业的管理人员来交流研发管理和激励制度设计，开展了讲座交流10余次，规范了研发计划和进度安排环节，提高了中心人员的研发产出，形成了高质量成果导向的永攀科研高峰的优良学风。

4. 技术委员会工作情况

2022年12月14日，采取线上线下相结合的形式，在浙江杭州召开了“数字图书馆教育部工程研究中心”技术委员会会议，会议由中心主任庄越挺教授主持，会议邀请潘云鹤院士、郑南宁院士、高文院士、朱文武教授等中心专家委员出席，具体参会人员名单如下：

一、专家委员会成员：

潘云鹤	院士	中国工程院
郑南宁	院士	中国工程院
高文	院士	中国工程院
来茂德	教授	中国药科大学
朱文武	教授	清华大学
周傲英	教授	华东师范大学
李建中	教授	哈尔滨工业大学
庄越挺	教授	浙江大学
朱强	教授	北京大学图书馆馆长
景祥祜		香港城市大学图书馆特别顾问
胡国平		科大讯飞研究院 院长
浦世亮		海康威视研究院 院长
黄晨	研究馆员	浙江大学图书馆副馆长

二、浙江大学：

曹阳	部长	浙江大学科研院高新技术部
张寅	副教授	浙江大学计算机科学与技术学院
鲁伟明	副教授	浙江大学计算机科学与技术学院
张引	副教授	浙江大学计算机科学与技术学院

邓晁煌 讲 师 浙江大学计算机科学与技术学院

陆国强 副研究馆员 CADAL项目管理中心秘书处

会议议程：

- 1、 领导致辞
- 2、 “数字图书馆教育部工程研究中心”及“中国工程科技数据和知识技术研究中心”进展汇报
- 3、 从大学数字图书馆国际合作到数字知识服务联盟
- 4、 数字图书馆元宇宙初探
- 5、 听取院士、专家建议
- 6、 总结

五、下一年度工作计划

技术研发方面，持续研发基于大语言模型的多模态知识图谱构建技术，工程化开放概念体系构建算法，跨域泛化实体/事件识别与链接、属性填充、关系抽取算法，提升算法在零样本设定下的性能表现。基于大语言模型，研发可解释、可泛化的思维链自动生成方法，为大模型推理提供高质量的思维链文本，支撑多轮智能问答系统，提升从数据到知识、从知识到服务的自主学习归纳能力。

成果转化方面，智库评价工作经过几年的积累，受到了国内智库及智库研究机构专家的广泛关注。后续在宣传推广方面将加大力度。首先，继续通过各种新媒体渠道对知领名片、创作者与机构名称标识符国家标准和智库评价进行宣传推广；其次，继续加强同商业机构合作，推广知领名片和KID注册平台，引流更多的注册用户；再次，通过参与不同级别的会议进行推广，组织访问国外著名智库机构，深入了解智库运行规律，并组织召开智库评价研讨会议，宣传推广项目创设的评价指标体系和智库数据平台，建立智库评价的合作联盟。

人才培养方面，中心在前期与华为、海康威视、阿里巴巴联合研发项目的基础上，进一步按照中心发展方向规划人才培养计划，借助人才流动和产学研合作，从工业界获取真实的应用场景和需求，推动中心研发的技术成果向产业界转化。

团队建设方面，中心计划根据智慧图书馆、知识中心、新一代人工智能技术研发的需要，引进1名以上固定或流动研究人员，加强研究梯队建设，开拓更多前沿应用，加速创新成果转化。

六、问题与建议

无。

七、审核意见

（工程中心负责人、依托单位、主管单位审核并签章）

<p>工程中心负责人审核意见：</p> <p>情况属实，同意上报。</p> <p>工程研究中心主任：</p> <p>年 月 日</p>
<p>依托单位审核意见：</p> <p>情况属实，同意上报。</p> <p>依托单位： (单位公章)</p> <p>年 月 日</p>

八、年度运行情况统计表

研究方向	研究方向1	新一代人工智能		学术带头人	庄越挺
	研究方向2	数字图书馆		学术带头人	黄晨
	研究方向3	知识中心		学术带头人	张寅
	研究方向4	非结构化数据管理		学术带头人	鲁伟明
工程中心面积	5200.0 m ²			当年新增面积	0.0 m ²
固定人员	52 人			流动人员	10 人
获奖情况	国家级科技奖励	一等奖	0项	二等奖	0项
	省、部级科技奖励	一等奖	1项	二等奖	0项
当年项目到账总经费	2744.31万元	纵向经费	2194.31万元	横向经费	550.0万元
当年知识产权与成果转化	专利等知识产权持有情况	有效专利	5项	其他知识产权	2项
	参与标准与规范制定情况	国际/国家标准	0项	行业/地方标准	0项
	以转让方式转化科技成果	合同项数	0项	其中专利转让	0项
		合同金额	0.0万元	其中专利转让	0万元
		当年到账金额	0.0万元	其中专利转让	0.0万元
	以许可方式转化科技成果	合同项数	1项	其中专利许可	1项
		合同金额	40.0万元	其中专利许可	40.0万元
		当年到账金额	40.0万元	其中专利许可	40.0万元

		以作价投资方式 转化科技成果		合同项数	0项	其中专利作价	0项
				作价金额	0.0万元	其中专利作价	0.0万元
		产学研合作情况		技术开发、咨询 、服务项目合同 数	10项	技术开发、咨询 、服务项目合同 金额	250.0万 元
当年服务情况		技术咨询		15次		培训服务	6人次
学科发 展与人才 培养	依托学科 (据实增删)	学科1	自然语言处理	学科2	软件开发环境 与开发技术	学科3	
	研究生 培养	在读博士	26人		在读硕士		68人
		当年毕业博士	4人		当年毕业硕士		26人
	学科建设 (当年情况)	承担本 科课程	1280学时	承担研究生 课程	420学时	大专院校 教材	1部
研究队 伍建设	科技人才	教授	15人	副教授	20人	讲师	5人
	访问学者	国内		0人	国外	0人	
	博士后	本年度进站博士后		4人	本年度出站博士后		2人