【信息管理】

AI 新时代面向文化遗产活化利用的智慧数据生成路径探析

范 炜 曾 蕾

【摘 要】生成式人工智能引爆 AI 新时代,新技术不断涌现并快速迭代更新,AI 技术应用呈现出百花齐放、百家争鸣的繁荣发展局面。借助 AI 发展东风,智慧数据的生成进入了高效、深化、多模态集成的新阶段,提升了数据驱动的文化遗产活化利用创新手段和创新形式的丰富度与可行性。本文旨在探索面向文化遗产活化利用的智慧数据生成路径。首先,从 AI 技术视角,对文化遗产智慧数据的内涵与价值进行回顾并知新;其次,系统分析从多元异构数据资源中生成智慧数据的典型做法;再次,以羌年为例,探讨非遗活态文化的智慧数据生成思路。最后,总结归纳 AI 赋能文化遗产智慧数据生成路径的四点参考策略:①抓住 AI 赋能机遇,补齐数据基础设施短板,加强数据资源体系建设;②尽快开展馆藏数据资源的"大语言模型 + 知识库"结合工作,实现智能分析与计算增强;③鼓励更广泛的文化遗产数据开放与共享,支持活化利用的创新应用;④确保可信的智慧数据。

【关 键 词】人工智能;文化遗产;活化利用;智慧数据;生成路径

【作者简介】范炜,四川大学公共管理学院信息资源管理系副教授,硕士生导师(四川 成都 610065);曾蕾(通讯作者),美国肯特州立大学信息学院终身教授,博士生导师,E - mail:mzeng@ kent. edu(美国 肯特 44240)。

【原文出处】《中国图书馆学报》(京),2024. 2. 4 ~ 29

【基金项目】本文系国家社会科学基金一般项目"面向文化遗产开放数据的关联索引构建与服务研究"(项目编号:22BTQ088)的研究成果。

0 引言

全球正快速进入人工智能(Artificial Intelligence,AI)新时代,以 ChatGPT 为代表的生成式 AI 与大语言模型应用正在引领新一轮的信息技术变革,诸多新技术不断涌现并快速迭代更新。信息资源管理专业人员应该迅速理解 AI 重大变化并将之纳入工作场景,掌握信息技术应用的技能升级及信息系统的适用性进化,以便及时升级与之相适应的增强型信息服务,更好地满足用户不断变化的需求。

对 AI 的认知,既不能停留在字面解读,也不能仅将其作为问答型对话系统和生成式预训练转换器。AI 不是单指某一项技术,而是以智能化为目标,集成各类新兴技术的组合生态;AI 是技术创新集大成者,其发展符合技术演进一般规律和技术组合效应规律;大数据(数据资源)和云计算(算力资源)为 AI(算法/模型资源)提供了大数据量和大算力,三者相互依存、共促发展,构成了智能化实现的三角形结构。

除了技术层面外,对 AI 的理解还需关注伦理和治理层面的内容。2021 年 11 月 23 日,联合国教科文组织(UNESCO)发布了《人工智能伦理问题建议书》,这是由 193 个会员国参与并一致通过的全球性 AI 治理重要参考文本。该建议书将 AI 定义为:有能力以类似于智能行为的方式处理数据和信息的系统,整合模型和算法的信息处理技术,包括但不限于机器学习和机器推理等[1]。该建议书还强调包容、平等、公正的价值观,并就 AI 系统与

UNESCO 所辖的教育、科学、文化、传播与信息等领域普遍存在的伦理问题提出对策和建议,倡导合理、和平利用 AI,警惕和防止技术带来的直接和衍生危害。2022 年起,UNESCO 每年召集并举办全球 AI 伦理论坛。2023 年 10 月 18 日,第三届“一带一路”国际合作高峰论坛发布了《全球人工智能治理倡议》,该倡议坚持以人为本、智能向善、相互尊重、平等互利,引导人工智能朝着有利于人类文明进步的方向发展,致力于形成具有广泛共识的全球人工智能治理框架和标准规范,贡献中国智慧和解决方案。

文化遗产活化利用是 AI 技术应用的新领域。近些年,我国非常重视文化遗产保护和活化利用,提出“要扎实做好非物质文化遗产的系统性保护”[2],“让更多文物和文化遗产活起来,营造传承中华文明的浓厚社会氛围”[3]。同时,陆续制定和颁布了一系列重要政策、意见与办法,为文化遗产数据资源体系构建与智慧数据生成提供了实施依据和行动参考。2022 年 5 月,中办、国办印发《关于推进实施国家文化数字化战略的意见》,指出其核心关键问题是打通、共享与集成,不改变现有文化数据的物理存在与分布现状,通过常用网络协议、API 标准与数据编码规范,实现开放共享和虚拟连接[4]。2022 年 8 月《“十四五”文化发展规划》发布,提出要以国家文化大数据体系建设为契机,提升公共文化数字化水平;同时,强调运用大数据、人工智能等新技术,促进内容生产和传播手段现代化。

欧美国家重视文化遗产数据资源的可持续建设,积极引入 AI 技术应用,投入大量资源,在宏观政策、中观组织、微观项目三个层面都有诸多实质性进展。2023 年 9 月,美国自然科学基金会(NSF)宣布投资 2 670 万美元构建首个开放知识网络(Open Knowledge Network,OKN)原型,18 个多学科、跨部门团队将构建知识图谱,建立连接和创建教育资料,打造值得信赖的开放知识网络,为实现 AI 驱动的未来而构建必要的公共数据基础设施,保障数据的开放、多样、包容[5]。2023 年 11 月,欧盟委员会拨款 440 万欧元建设欧洲文化遗产的数据共享空间,运用 AI 提高数据的质量、存续、利用与重用,开展了一系列 AI 赋能文化遗产项目,如 EUreka3D、5Dculture、DE – BIAS、AI4Europeana 等[6]。

在 AI 赋能下,文化遗产智慧数据的生成正进入效率提升、价值凸显的新阶段,数据驱动的文化遗产活化利用创新再上新台阶。图书馆长期以来在资源分类、编目、标引等方面具备专业优势[7],能够有力推动数据、模型、应用三者之间的粘合,助推数据治理能力体系构建。在图书馆应对 AI 的战略制定方面,国际图联 AI 特别兴趣小组召集人 Andrew Cox 牵头拟定了指导性文件,强调“描述性 AI”是 AI 在图书馆的最重要应用[8],即指各类馆藏资料的机器可读与检索可用,并指出图书馆虽然在技术能力上有劣势,但可以利用 AI 增强数据能力。

本文立足图书馆、档案馆与博物馆(以下简称图档博)等文化记忆机构的智慧数据建设场景,围绕智慧数据赋能文化遗产活化利用,尤其是对数字人文研究的支持,介绍全球最新实证研究发现、案例研究和最佳实践,结合中国本土实际,探讨 AI 新技术工具赋能智慧数据生成的可能性,帮助图档博馆员理解数据资源建设与 AI 技术应用相结合的创新思路。

1 文化遗产智慧数据相关认识

在中国式现代化进程中,数据成为重要的生产要素。国家数据战略顶层设计与数据基础制度建设步伐加快。2022 年 12 月“数据二十条”发布,2023 年 10 月 25 日国家数据局成立,2023 年 12 月 31 日,国家数据局牵头,联合 16 部门发布《“数据要素 ×”三年行动计划(2024 – 2026 年)》(以下简称《行动计划》)。《行动计划》以数据要素为核心,明确数据的基础资源和创新引擎两大作用,强化数据的乘数效应,即放大、叠加、倍增作用,统筹数据资源的整合、共享、开发与利用,这与此前发布的数字中国建设规划、文化数字化战略等一脉相承。根据《行动计划》中有关“数据要素 × 文化旅游”融合场景的重点行动表述,文化遗产智慧数据是实现文化数据价值挖掘和各类文化记忆机构数据贯通、促成中华文化数据库关联、助推文化大模型开发的聚焦发力点。

《行动计划》以激活数据要素潜能与价值为导向,这与智慧数据的内涵高度一致。将智慧数据的

概念阐释与内涵丰富化置于中国国家数据战略与政策发展环境中,智慧数据是发挥数据作为基础资源与创新引擎两大作用的内核所在,以数据价值联结 AI 技术,赋能多元化创新应用。

1.1 智慧数据再认识

智慧数据不是一个简单的技术炒作概念,是对数据本质与价值的深入探究,围绕高质量与高价值,提出了一种洞悉数据的高阶认知思维,其内涵要素与现实中各种数据处理业务要求相呼应。

2018 年笔者率先在国内系统化地提出了支撑数字人文的图档博智慧数据这一概念[9],以大数据为基础参照,阐释了智慧数据的丰富内涵,讨论了图档博机构立足资源优势的未来新作为。换言之,智慧数据是数据资源体系建设的目标与愿景,是数据高质量与价值化体现,其理念业已得到广泛关注。

智慧数据是在大数据语境中产生的概念,是实现带有多个 V 的大数据特征中最后一个 V——价值(Value)的方法,即通过对任何规模的可信的、情境化的、相关切题的、可认知的、可预测的和可消费的数据的使用,来获得重大的见解和洞察力,揭示规律,给出结论和对策。智慧数据是基于大数据的方法,即为了揭示“未知—未知”(Unknown - Unknown)而采取的方法。

智慧数据与数字人文视野下图档博数据的崛起和发展体现了从数字化到数据化、再到数智化的发展进程[10]。从情境化、可认知、可预测的特点来看,智慧数据通常带有自描述机制,背后有领域本体作支撑,使得这些数据符合特定的逻辑结构和形式规范,而且支持推理,由此形成智慧的基础,产生可预测和可消费的数据。智慧数据是一种便于机器理解的数据,同时也是人和机器都能读懂的编码化知识,而非只有机器可用的、难于人为表达的概率性隐性知识。智慧数据具有较强的可解释性,支持逻辑推理,这使得它能够用于多种用途,支持多种互操作,并且具有很强的可追溯能力,能够满足人文研究范式的需要,远远超出文化遗产资源搜索与浏览的基础层次。

近些年,人们对智慧数据存在不同场域和角度的认识,其智慧内涵和何以智慧的实现路径还在不断探索中。关于智慧化的认识,张晓林等认为,知识服务的智慧化在于问题解决场景下 DIKW(数据—信息—知识—智慧)转化过程产生的“知识增量”,突破就是智慧,突破过程就是智能[11]。在文化遗产、科技情报等诸多领域出现了有关智慧数据的研究讨论与项目探索:王晓光等认为,智慧数据代表着数据资源更高级的组织形态,具有富语义性、可计算性、可追溯性、可信性等特征[12];钱力等提出,科技情报智慧数据是面向具体应用场景,具有质量高、维度丰富、多源融合及认知理解等特征的科技大数据资源[13]。

在 AI 新时代,数据资源与智慧数据在供需关系框架下辩证统一起来,通过 AI 新技术组合创新,不断加快“量”(广、全)与“质”(深、专)的同步提升。对智慧数据的再认识可以得到进一步理解:智慧数据是高价值、高回报的高质量数据要素,能够运用 AI 技术高效处理,并能为 AI 所用,产生多样化的智能应用。

1.2 文化遗产内涵及其数据资源

文化记忆机构一直是文化遗产资料的保管者,肩负着保护和传承文化遗产、开展数字人文专业研究和推动相关文化向社会大众普及的责任。切实理解文化遗产内涵的多元性与包容性,是做好文化遗产数据的前提,可以避免形式化地“为做而做”。

“做数据”首先要理解文化遗产的精神内核。中华文明起源多元一体,相互影响与融汇,具有连续性、创新性、统一性、包容性、和平性等突出特性,并在发展中不断筑牢中华民族共同体意识。国际上不同文化、不同民族之间也建立起包容和谐的沟通与分享渠道,达成跨文化交流互鉴与理解。《联合国土著人民权利宣言》申明,所有民族都对构成全人类共同遗产的文化和文明的多样性和丰富多彩做出了贡献。其中第 31 条指出,土著人民有权保持、掌管、保护和发展其文化遗产、传统知识和传统文化体现形式及其知识产权[14]。在实践方面,由加拿大图档博及宣传教育部门等机构协同联动的全国本土知识与语言联盟(NIKIA - ANCLA)指出,以图档博为代表的文化记忆机构在本土知识与语言文化的再生和代际传承中发挥重要作用,包括尊重本土世界观,理解口头传统的连续性与活跃性,

意识到本土知识的收集、管理与长期保存会受到殖民化、政权等人为因素干扰。目前该联盟正在建设一个动态开放的多语种规范术语在线平台,对文化记忆机构使用的过时和不适宜的术语进行替换和更新[15]。欧盟委员会于2023年11月启动检测和处理文化遗产馆藏中的有害语言项目DE-BIAS,致力于使文化遗产元数据更具包容性。该项目创建的工具,可以自动检测有偏见的术语,并提供相关问题的背景信息及适当的术语建议[16]。

在尊重文化遗产多元性与包容性前提下,文化遗产数据既要FAIR(公正),也要CARE(保护)。FAIR(可找寻、可访问、可互操作、可重用)是被全球认可并应用的开放数据原则;CARE原则为促进本土人士(又称土著、原住民)数据的集体利益而设立,旨在实现包容性发展、创新、改善治理和公民参与,并实现公平,具体包括集体利益(Collective Benefit)、控制权(Authority to Control)、责任(Responsibility)与伦理(Ethics)四个方面[17]。FAIR原则与CARE原则的结合更能体现对文化遗产多元性的理解与包容。

文化遗产种类丰富多样,数据类型与结构差异大,对智慧数据的生成提出更高要求。文化遗产数据是核心要素,从价值出发,围绕成本与利润,对应的数据资源与智慧数据是两个衍生概念。首先,文化遗产数据资源体系是基础设施,需要投入大量物力和人力,也需要可持续推进的资源建设机制,强调"量"与"质"的基础要求达成,是不断投入的成本中心;其次,文化遗产智慧数据是在数据资源基础上,重视"质"的提升,强调激活数据,析出数据价值点,为AI技术提供高质量的数据养料,实现智慧化场景应用,更多的是价值目标导向的高层要求,逐渐成为输出价值的利润中心。

在AI技术日新月异以及中华优秀传统文化焕活新生的背景下,2023年12月召开的第五届中国数字人文年会中,围绕"数实共生:预见数字人文未来图景"主题的主旨发言以及十个分论坛,深入研讨了中国数字人文研究的发展和趋势,其中"数字人文与智慧数据""数字人文与文化遗产数字化""数字人文与图档博资源建设"分论坛有众多主题报告和论文汇报直接分享了相关实例和研究发现,涵盖了各种类型的文化遗产[18]。

1.3 AI与多模态数据生成

由于文化遗产丰富多样,从数字化开始就面临数据处理的复杂性,涉及不同种类数据、不同载体数据和不同形态数据,除了文本、图像、音频、视频等常见的数据类型外,还包括测绘3D、传感器采集等其他专有数据类型,这些被统称为多模态数据(Multimodal Data)。

AI擅长处理多模态数据,具备以下功能:光学字符识别(OCR)、手写文字识别(HTR)、统计分析、自然语言处理、文本挖掘,机器学习、计算机视觉,提取实体、特征、颜色、组成和模式,进行相似性匹配、图像分类和标记等。生成式AI还可以生成多模态的内容,大致可以分为以下几类:①文本,例如ChatGPT根据对话提示词生成文字内容;②图像,例如DALL-E、Midjourney、Stable Diffusion等根据描述词生成图像;③音频,例如Meta发布的开源框架AudioCraft根据文本、音乐、音效以及原始音频信号训练后生成新音频;④视频,例如DeepBrain AI通过文本、语言(包括方言)、身体特征、手势等创建AI化身和合成创意视频。

除了以上提到的多模态生成内容,AI技术应用于文化遗产多模态数据分析,能更好地辅助生成情境化的结构化数据,实现异构数据融合、特征提取、时空关联等以前单一数据类型无法实现的综合性分析挖掘,使得智慧得以实现。在多模态数据集成层面,数据编织(Data Fabric,又译数据织物)是一种对多源异构数据进行链接(松耦合)的方法,可以实现数据松耦合的系统化集成,最终目的是打通数据通道,实现共享、管理与挖掘利用。随着Google Gemini的发布[19],对多模态数据融合处理、分析与理解能力增强,智慧数据的落地应用有了更丰富的想象空间和实现路径。大语言模型对多模态数据的集成处理能力不断增强,提升了识别、分析与生成效果,至少在技术层面上,对整合多种文化遗产数据展现出一定的可行性与预见性。

对AI的系统化认知和客观性判断是一个重要前提。AI不再是象牙塔里的专业研究,AI意识与认知在全球范围内快速普及,AI正在快速以技术使能(enabling)和应用赋能(empowering)影响各行各

业,惠及社会方方面面。总的来看,AI 有不同的任务场景,除了自然语言处理,还有逻辑推理、概念图表、数据统计以及提升系统评估的准确性。AI 目前的主要功能可见于人机对弈、模式识别、自动工程和知识工程[20]。本文特别关注知识工程,它以知识本身为处理对象,研究如何运用 AI 和软件技术,设计、构造和维护知识系统,用于专家系统、智能搜索引擎、计算机视觉和图像处理、机器翻译和自然语言理解、数据挖掘和知识发现等。

AI 技术的主要作用不是替代,而是辅助、支持和增强。许多 AI 工具通过背后的各种规范词表(如地理信息、名称规范等)和不断增强的本体知识库支持进行机器学习。在智慧数据价值体现方面,除了文本挖掘(带有时间参考、语义推理和推理实体),还能进行分类、语义标记、情感分析、事实挖掘,做写印和实体之间的提取关系,通过语义增强功能,进行知识管理和大数据分析,将数据转化为对情报分析人员有用的知识。GPT 除了用于其擅长的翻译领域之外,还可进一步升级,使用精确逻辑进行推理,从论文中抽取实体(节点),构建概念关系,形成初始语义关系网(Scaffolding Net)、初始本体(Proto - ontology)、路径推理(Path Inference)与摘要生成(Summary)等高级功能,运用比二维向量更高阶(三维及以上)的张量进行精确计算,以消除现有 GPT 不准确且不稳定的问题[21]。

Nature 近期发布了一项关于 AI 工具对科学研究影响的调研结果,对 1 600 位科研人员的调查分析显示,AI 在科学研究中得到普遍关注,科研人员首要关注的是机器学习[22]。AI 工具在科学研究中显现的优势是,提供更快的数据处理方法,实现以前做不到的计算速度,节省时间和金钱。与此同时,带来的劣势是,科研人员会更依赖 AI 工具的模式识别而缺乏深入理解,AI 会带来更深层次的数据偏见,不恰当的使用也会导致科学研究结果缺乏可信性和可验证性(不可再现性)。

1.4 AI 赋能文化遗产智慧数据

在数字人文研究的跨学科与多学科参与中,以图档博为代表的文化记忆机构,在文化记忆与知识保存方面,以独有的文化遗产资源整理与保存优势,为数字人文研究贡献了高质量的数据底座、随需应变的资源服务平台与多元应用的创新手段。

在 AI 新时代,文化记忆机构馆藏资源是重要的数据资源,是机器学习模型训练的高质量“数据养料”;馆藏资源为机器学习提供了训练数据,创造了最佳实践和案例,但也存在数据的版权、隐私、伦理与合理使用等挑战[23]。

AI 要能更好地赋能文化遗产智慧数据生成,需要以文化知识为基础,特别是对基础性古代文化知识的理解。陈力认为,数字人文研究需要基础条件,要让计算机理解古代文献,进行识别、转换与编码以实现古籍数字化;同时,也要借鉴古代类书等知识工具编纂方法,建设多维度、智能性的古典知识库[24-25]。在数字人文研究热潮中,国内涌现出各种类型的专题知识库与数字人文知识工具。以古籍数字化为例,中华古籍数字化成果显著,积累了大量影像资料,为进一步数据化与数智化打下了坚实基础,逐渐形成古籍智慧数据生成路径的中国做法。国家图书馆牵头建设中华古籍资源库,统一整合了国家图书馆(国家古籍保护中心)自建和征集的国内各级公共图书馆的古籍数字资源,包括数字古籍、数字方志、敦煌遗珍、中华古籍善本联合书目等,累积发布古籍影像资料数量达 10 万部(件)①,近些年新增开放种类与数量逐步扩大,面向社会大众提供公开免登录查阅。在数字化影像基础上,中华古籍智慧化服务平台启动建设,在知识深度加工、检索发现、传播交流、活化利用上,进一步提升知识服务的智慧化能力。值得一提的是,引入互联网公司(如百度、腾讯、字节跳动等),联合打造公益项目,为古籍数字化资源的深度开发与利用提供技术赋能。例如,国家图书馆与百度的战略合作,利用文心大模型助力古代方志和家谱数据资源活化,通过 AI 助力海外华人寻根;字节跳动设立古籍保护专项基金,参与国家图书馆的《永乐大典》影像数据库、东巴文汉文合璧《创世纪》知识库建设等。

在文化遗产数据资源持续积累与建设基础上,国外关于文化遗产数据项目的 AI 应用探索和相关研究,也值得跟踪借鉴。

(1)2017 年 IBM Watson 在巴西圣保罗美术馆推出“艺术之声”项目[26]。在参观过程中,根据参

观者定位信息,会触发“艺术之声”App 的 AI 助手,讲解特定绘画和雕塑信息及其背景故事,并能回答有关此作品的提问,大幅度提升参观的沉浸体验感。为实现问答,IBM 程序员花了六个月的时间与艺术策展人和专业学者合作,向 Watson 提供有关艺术作品的信息和答案,这些信息来自广泛的研究和数据,包括书籍、旧报纸、传记、采访和互联网等。对话的范围包括历史和技术事实,例如,作品与当代事件有什么关系,这幅画是用什么技术创作的等。这与当前大语言模型和知识库结合实现问答系统的技术路线类似。

(2)欧盟 Europeana 被誉为跨地区、开放共建共享的文化遗产数据资源平台建设的典范。该平台积极拥抱 AI 技术,值得学习与参考。2019 年,Europeana 技术工作组以欧洲八家文化遗产机构和 56 名文化遗产研究机构专业人士为对象,开展了一项针对 AI 和机器学习在文化遗产领域应用的全面调查,形成了《AI 与 GLAM 的关系》报告[27]。该报告涵盖的项目多元,主要聚焦于文化资产的数字化和可发现;报告指出,绝大多数受访者(91.8%)至少对一个 AI 主题感兴趣,而超过半数受访者(54%)在特定主题上具有专业知识;报告着重强调了跨部门合作、数据质量、工具的选择和实际应用、伦理问题、用户体验、展示 AI 的价值、项目成果及数据共享等关键主题;最后,该工作组提出了未来的行动计划,包括知识交流、数据共享、战略建议等,强调了在文化遗产领域提高对 AI 认识和应用的必要性,以及在实现这些目标过程中所面临的伦理、法律问题和社会挑战。目前 Europeana 启动了为期 27 个月(2023 - 2025 年)的 AI4Culture 项目,目标是建立一个 AI 能力建设平台,提供一系列与 AI 有关的资源,如开放标注的数据集、开源工具和培训资料等。利用 AI 技术,AI4Culture 项目计划实现以下功能:扫描图像并进行多语种文本识别、多语种字幕生成与验证、丰富图像中提取的信息并建立与其他文化内容的连接、文化遗产元数据的机器翻译[28]。

(3)2022 年 5 月,欧盟议会发布《人工智能在文化遗产与博物馆场景中的复杂挑战和新机遇》报告[29],指出欧盟成员国拥有丰富的文化遗产,众多知名博物馆及其珍贵藏品是文化创意产业的宝贵资源,但这些资源目前还依赖美国的技术平台和亚洲的技术设备。报告提到 AI 赋能文化遗产的典型实例包括以下几个方面:①AI 生成内容,帮助重建失去的文化遗产,例如巴黎圣母院火灾后的数字化虚拟重建;②AI 生成内容,尝试完成未完成的艺术作品,例如续谱贝多芬第十交响曲;③AI 与机器学习技术相结合识别古代文字的作者,例如笔纹技术;④AI 与测绘技术相结合对遗址进行三维数字化,例如无人机技术与数字测绘;⑤AI 用于智慧博物馆,实时检测文物保护环境,防护艺术犯罪等;⑥AI 与 VR(虚拟现实)、AR(增强现实)相结合,可增强博物馆现场观众和线上访客的沉浸式体验;⑦AI 促进数字化保存文化创意资源。

(4)2020 年,意大利阿里安娜·特拉维利亚理工学院第九次文化、遗产与景观(CDCPP)指导委员会全体会议上,列举了文化遗产领域 AI 技术应用的主题,包括破译古代语言、使用深度学习恢复古代文本、解码铭文标记、从楔形文字板中提取布局、自动识别(如罗马硬币上的头像)、自动化 3D 数字化程序、AI 和机器人、化学物理分析、通过 AI 检测未知文化遗产、通过 AI 检测文化遗产与艺术犯罪侦查等[30]。

(5)微软文化遗产项目通过 AI 为个人和组织赋能,致力于文化遗产的数字化保存与创新性传播。微软承诺在五年内投入 1 000 万美元,借助 Azure 云平台、AI 与机器学习技术,聚焦历史重要人物、古迹遗址、语言与文物四类资源[31]。该项目官网展示了几个典型项目:与希腊文化与体育部合作的“体验 2 000 年前的奥林匹亚与古代奥运会”②、与印度艺术与摄影博物馆合作的利用微软 AI 发现其纺织品收藏③、传奇艺术家索尔·勒维特的一生与作品体验 App④、与努纳武特政府合作的因纽特语振兴(涉及语音识别与机器翻译技术)⑤等。除此之外,微软与美国大都会博物馆和麻省理工学院合作,利用 AI 技术实现大都会博物馆藏品信息在全球范围的开放获取;与其他技术公司合作,通过混合现实和 AI 技术,在巴黎浮雕博物馆创造了全新的博物馆体验;在墨西哥西南部,参与保护世界各种语言,利用 AI 记录和翻译尤卡坦玛雅语和克雷塔罗奥托米语[32]。

这些实例充分说明,以数据资源建设与服务为优势的图档博机构在数字人文研究中的支撑作用不可或缺。

2 为多元异构文化遗产资源生成智慧数据

文化遗产资源在数字化基础上向数据化深入。以文献为代表的文本型文化遗产,主要包括非结构化、半结构化、结构化三种数据类型,也对应不同的层次和方法。从结构化文本,特别是元数据、数据库与索引,到语义化增强,已经总结了大量成熟可靠的数据处理方法与经验[33]。对国内外实例的分析表明,生成式 AI 中自然语言处理技术的应用可以有效提高文化遗产文本型数据的结构化处理效率,增强实体识别与语义关系揭示,为高质量的文化遗产知识图谱构建提供语料。值得注意的是,在为文化遗产资源生成智慧数据的过程中,面对的原始资料有多种类型,除了各种类型的历史资料,还有不断出现的新闻报道和研究成果,以及由官方和大众提供的多载体的图像文件和音视频文件等。下文结合全球范围的典型案例,梳理总结不同类型的文化遗产资源从数字化逐渐深化为智慧数据的做法。

2.1 机器可读文字的数据资源

进入 21 世纪以来,通过计算机制作文献内容已经成为主流的出版发布方法之一。以学术文章生成结构化数据为例,根据美国国立医学图书馆报道,2021 年完全靠人工标引的文章的平均索引时间为 145 天,这不包括书目数据审查所需的时间。自 2022 年 4 月以来,MEDLINE 收录的所有期刊均自动生成索引,并酌情进行人工审查和结果管理,通过机器学习不断提高效率和准确率。现在 MEDLINE 索引系统会在收到文章后 24 小时内自动完成标引,第二天即在 PubMed 中显示为 MEDLINE 索引[34],标引效率明显提升。

近年来,在文化遗产领域,由非机器可读文字原件转换为机器可读文字的数据资源已经大量增加,这一转变主要是通过数字图档博项目来实现。这类项目对简牍、古籍、打印机制作的资料甚至手写注释文字等进行增强技术处理,产生机器可读文字。日本人文开放数据学术中心的 KuroNet Kuzushiji 字符识别服务,利用深度学习技术,使用 RURI(2022 年 10 月开始采用的新模型)为符合国际图像互操作框架(IIIF)的图像提供多字符 Kuzushiji OCR 功能。日本古典书籍 Kuzushiji 数据集也是应用该技术开发的,其中每个字符都可以打开有 AI 深度学习的字库支持的内容,新旧字符的整合也遵循该数据集的创建策略[35]。

将数据集中不常出现的字符(例如手写注释)转换为机器可读的文字是一种挑战,国外机构也在尝试构建更大的数据集以提高识别的准确率。美国加州大学洛杉矶分校数字图书馆馆藏实验室的图书馆员、学生和工作人员开发实验代码[36],采用 Zooniverse 图书注释分类方案,通过机器学习技术自动侦测并突出显示印刷书籍中的手写注释。该过程分三个阶段:①识别包含注释的页面;②用注释描绘页面区域;③对各种注释类型进行分类。Zooniverse 是专为图书注释分类产生训练数据,机器学习实验用于侦测和分类数字化印刷书籍中的注释,结果在 GitHub 上公开[37]。该实验最终是要开发出一种工具,帮助研究人员识别 IIIF 托管的数字特藏的注释页面。

文化遗产资料需要转化为机器既可读又可处理的数据,并通过数字化流程进行重建。数字化后的数据一般还是非结构化的,数字人文领域的智慧数据方法强调通过组织和整合,将非结构化数据转换为结构化和半结构化数据。笔者曾通过智慧数据阐释美国肯特州立大学液晶研究所(Liquid Crystal Institute,LCI)的创新史,这里以这一研究项目为例来说明非结构化机器可读数据的处理方法。LCI 是世界液晶技术研究的发源地,也是过去五十年美国研究液晶技术最成功的机构之一,拥有的专利数量在美国和加拿大领先。项目主要是通过智慧数据分析,阐述 LCI 的技术创新历史。研究过程中,面临的研究数据来源多样,且大部分是非结构化数据,包括液晶研究所年度报告(其中有 1960 年代起多年用传统打字机打印出来的文献)、口述史、新闻报道、周年庆会议文稿;少量结构化数据包括大型学术和专利数据库、美国国家自然科学基金会资助项目数据库等。项目运用 Cogito Intelligence API 等工具对 OCR 后的 50 年的年度报告等进行文本挖掘,结合分类法、领域本体等,对创新历史中涉及的

人物、组织以及相应的成就资料(出版物、专利、赠款、演示文稿等)、主要发明、研究和开放人际网络、地理位置等与正确的时间轴相匹配,并用知识图谱呈现所有这些事物对象之间的正确关系。项目揭示了数据的价值,展示了数据分析方法,包括科学计量、网络理论、语义分析模型、历史测量、事实挖掘、推理、图论、时空数据分析等[38-40]。

由图档博原始资料生成的结构化元数据多为机器可读,其中也包括可以进一步做语义增值处理的数据,特别是其中的半结构化文本数据(主要是经过人工整理的大型资料记录)。例如,对档案元数据记录(称为 Finding Aids,直译为查找辅助工具)使用机器学习的工具进行语义分析,能自动抽取并生成实体和主题标签。Finding Aids 是美国档案界采用的标准方法,有其特有的格式,是一种包含详细内容分析、索引、元数据以及有关档案中特定记录集合信息的文档。一个档案里(例如某一事件、某一国家公园之建设、某位人物的特藏等档案)可能有上百个不同类型的原件,因此编制一个 Finding Aid 往往需要数月甚至数年,其所含材料内容中通常包括文件清单以及对材料、来源和结构的描述,具体文献或文物编号、名称,及其在特藏中的位置(盒、箱、文件夹)等诸多信息,是图档博数据中宝贵的高质量的第一手资料。一个 Finding Aid 里的半结构化数据包含描述性信息,例如创建者历史、作用域和内容注释部分,以及编目员制成的摘要。在对来自美国 16 个机构的 43 个 Finding Aids 档案记录的批量处理中,作者团队使用了语义分析工具 OpenCalais 的免费版本(现纳入 PermID 的 Intelligent Tagging 平台),抽取了八千多个实体和三百多个建议的社交标签,基本上对每个档案记录的分析都产生了数十个甚至上百个潜在的实体和主题标签,可为档案记录提供更多的检索点。通过 OpenCalais 分析正确识别的实体包括人名、组织机构名(公司、设施、组织)、地名(城市、大陆、国家、自然地带、省或州、地区)、事件(节假日、政治事件)等,对每个已识别实体进行相关性打分,可以作为判断该实体对档案馆藏整体重要性的依据。同时也发现,要保证“社交标签”“行业术语”和“产品”分类的准确性,仍需要人工校对和修正[41]。

在图档博数据处理中,语义中介(包括合并和映射)、标注、分析是首先采用的语义增强方法,通过使用已知实体之间的显式语义关系来分析、搜索、呈现信息。依据智慧数据生成策略[33],数据服务提供者在数据结构化的过程中应注意以下几点:①创建机器可理解、可处理、可采取行动的数据;②在数据链接、引用、传输、授权管理、使用、重用的过程中提供准确的数据;③允许一次生产、多次使用的高效的数据处理方法,进而支持数字人文研究。

2.2 非文本类数据资源

除了前述最常见、也是数量最多的机器可读文字资源以外,图像数据、音频数据、气味数据、实物数据等都是文化遗产数据的表现类型。当前 AI 技术对之进行有效处理,也已经展示了快速发展的潜力。

(1)图像

文化遗产资料数字化之后,很多都是以数字图像形式存在的,但也有分辨率高低的差异。对图像的计算机视觉技术得到了长足发展,对图像的展示、调用与深度分析是 AI 的重要应用场景之一。

①2D 图像。文化遗产数字化初期阶段,数字化扫描项目是最常见的,产生的图像数据所占比重最大,也是数字人文研究人员最常使用的资料类型。一般而言,图像数据需要经过识别、标注、解构才能进行分析。对于有文本内容的古籍文献,图档博机构通过 OCR 识别技术,或对图像进行矢量化生成文本数据,在此基础上才能实现浏览与检索功能,如中国国家图书馆《永乐大典》系统、“汉典重光”项目等。

AI 增强了图像数据的处理能力,对文化遗产图像进一步分析挖掘具有重大意义。目前主流的大模型都能对图像进行一定程度的内容分析,识别图中要素,生成图像的描述内容,并进一步产生猜测、推理与联想。虽然内容准确性还有待提升,但已具备了一定的可用性。例如,将人脸识别技术用于文物修复,文物虚拟修复师对安岳石刻佛像面部进行照片采集,借助 AI 技术提取佛脸特征,尝试为佛脸做“身份证”,通过数据量化分析,洞悉古代工匠的雕刻技艺之美,并以面部数据为依据,生成佛像脸部的虚拟修复效果[42]。

采用 AI 工具生成图像已经成为热门。例如，一位创作者使用 DeepAI 生成了四个图像。按照提示词“显示一个棕榈树密布的热带岛屿，阿波罗 11 号指挥舱停靠在海滩上”，分别由以下生成器生成四张艺术作品：①文艺复兴绘画生成器；②DeepAI 抽象绘画生成器；③幻想世界生成器；④印象派绘画生成器。那么，怎样为这些艺术作品创建可信的描述性元数据？这个实例也说明需要新的指南[43]。

②3D 图像。文化遗产的 3D 扫描与拍照数据已成为文化遗产多模态数据管理的重要组成部分。这一过程是将实物进行高精度扫描，通过 3D 建模技术，形成三维立体数据（以下简称 3D 数据）。文化遗产 3D 数据不仅对考古研究、文物保护与修复工作具有重要意义，而且通过与 3D 打印、VR、AR 以及 AI 技术相结合，在提升公众对文化遗产认知和参与方面发挥了显著作用。

对实体文物而言，3D 数据提供了近距离、多维度的观赏体验，弥补了现场观看实物时无法看到的更多角度与细节。例如，故宫博物院的“数字多宝阁”开放了一大批文物的 3D 浏览，配合简单文字介绍，为用户提供了便利的使用体验⑥。中国国家博物馆的“数说犀尊”专题展览⑦，是 AI 相关技术的应用集成，包括 3D 扫描、物联网、云计算、区块链、文献知识图谱、交互式可视化、3D 打印文创等，其专题展览背后是围绕西汉错金银云纹铜犀尊的多元数据融合，如文献、史料、实物等，是典型的智慧数据应用。未来会有更多国宝文物实现多维度、深层次的数字化、数据化与数智化，实现文化遗产活化利用从“数据说”到“智慧说”。

对遗址场所、非遗活动等而言，基于 3D 数据，可以通过模拟仿真等多技术集成，创新性地打造沉浸式临场体验、虚实结合的演绎剧场、增强型游戏和在线游览，以及开发实物文创衍生品等。例如，AI 图像处理在法国巴黎圣母院失火后的数字重建中起到重要作用[44]，还有数字敦煌项目的洞窟实景在线和数字藏经洞⑧、故宫博物院的“全景故宫”线上游览⑨、三星堆博物馆的考古方舱虚拟复原展示，以及“数说犀尊”的 3D 打印文创礼品等，都是代表性的应用。

与 2D 图像数据相比，3D 数据的采集、建模及制作工作量更大，难度更高，其数据质量与专业性程度要求也更高。通常这些数据由专业机构独家享有，不予以公开共享。在高精度专业级别，重要文物和重要遗址通常会有计划地开展 3D 数字化；而在低精度业余级别，可以通过手机拍摄实物不同角度的照片，结合前端 App 和后端云计算建模服务，低成本快速生成实物的 3D 数据，并与 3D 打印、教学展示等多应用场景联动。

3D 数据在视觉展示上具有冲击力，能够达到吸引眼球的先发优势。但是，目前 3D 类文化遗产项目在视觉与内容的深度融合方面仍有待加强，在带来视觉震撼的同时，也应将文本、声音等多种类型的数据集成起来，以实现“有表有里”的综合传达，这也是当下文化遗产活化利用的组合创新发展趋势。

随着技术的发展和应用，国内外已有不少采用 3D 制作并重现文化遗产的项目。2023 年中国数字人文年会的“数字人文与文化遗产数字化”专题论坛上，学者们分享了丰富多样的专题研究成果。欧盟委员会于 2023 年 11 月宣布拨款资助的 EUreka3D 重建项目，通过建设智能技术基础设施的知识中心，提升文化遗产机构的智慧化能力，它将拥有一个包含整体信息的知识库，以加强该行业 3D 内容和 3D 技术的创建和重用；另一个项目是 5Dculture，主要部署和展示 3D 文化遗产空间，提供和管理欧洲共同数据空间中有关时尚、建筑和考古等主题的 3D 数字文化遗产资产，该项目将探索教育、旅游和更广泛的文化创意领域的再利用方案。

以上项目案例正说明情境化数据的重要性，特别是在多维数字立体空间中，智慧数据能揭示同一地点和文物与不同时代的关系，以及同一类文物在不同空间下的意义和价值。

图像数据的处理与分析思路在不同应用场景中是相通的，也是可以相互借鉴的。美国审计署对医疗诊断中机器学习技术及其新兴用途进行评估，针对选定疾病（如某些癌症、糖尿病及并发症、阿兹海默症、心脏病、COVID－19）确定各种机器学习技术，其中大多数技术依赖 X 光或磁共振成像等影像数据，由此带来的好处包括及早发现疾病，医疗数据分析更一致，增加弱势群体获得照护的机会[45]。

该评估报告讨论的要点是目前针对五种疾病应用的机器学习医疗诊断技术,以及该技术开发和应用面临的挑战,并分析了应对挑战的政策选择。这些以非文字影像数据为基础的技术开发、应用挑战、政策选择等也是文化遗产图像数据可以侧重参考的内容。

(2)音频

从音频资料中也可提取智慧数据,辅助开展人文研究和文化遗产活化利用。

美国辛辛那提大学数字人文研究"关联阅读"(Linked Reading)项目使用音频机器学习技术,研究人员能够直接查询、分析和可视化来自多个独立数据集的音频数据,包括辛辛那提大学的 Elliston 诗歌档案和亚利桑那大学诗词中心的 Vico 诗歌档案收藏的资源(包含 450 多位诗人的 700 多首诗歌及诗歌相关内容录音的音频档案,录音跨越 70 多年)。通过对印刷文本进行语义分析,结合音频档案的声音分析,构建了关联数据平台。该项目可以帮助学者研究诗歌流派的阅读风格,比较口语诗歌与印刷文本不同的形式维度。其试点研究项目比较了两个档案中对同一作者作品的多次阅读,研究诗人如何背诵韵律,在表演中如何使用行列、标点符号和呼吸,以及一首诗前后的序言评论和典故如何为诗歌添加背景等。歌词的形式维度,正如在手稿或印刷文本中的旁注一样,创造了补充意义层。该平台能够发挥创意写作和电子媒体方面的优势,促进新的诗人、数字人文学者和声音学者开展一系列研究项目及合作[46-47]。

免费电子书平台古登堡项目(Project Gutenberg)经过与来自微软、谷歌和麻省理工学院的研究人员合作,利用基于神经网络的文本转语音技术,制作了 5 000 本免费开放的有声读物特藏,包含大约 35 000 小时的音频,通过 AI 合成人声大声朗读,让世界各地的有声书籍爱好者更容易接触到文学作品,使高质量有声读物的获取更加大众化、平民化[48-49]。

(3)气味

气味是文化遗产活化的重要感官维度之一。气味无形,以嗅觉为线索,需要根据文本、图像等相关资料,用 AI 进行集成分析。值得关注的一个特殊研究是英国和欧洲的科学家、历史学家和 AI 专家正在联手开发 Odeuropa 项目[50],旨在识别甚至重现 16-20 世纪初欧洲扑鼻而来的香气。该项目首先用 AI 筛选出七种语言的历史文本,以描述独特的气味及其背景,并识别图像(如绘画)中的芳香物品。在项目进行期间(2021-2023 年),Odeuropa 团队组织了各种嗅觉活动来测试其方法的可行性和有效性,例如,乌尔姆博物馆的导游气味之旅、柏林的 Malodour 研讨会以及阿姆斯特丹城市嗅探器城市之旅。2023 年底发布的产品为:①气味探索器:一个为每个对气味的历史和文化遗产感兴趣的人提供的独特搜索引擎;②气味历史与遗产百科全书;③在博物馆和其他遗产机构中处理气味的"操作方法"指南;④气味遗产图书馆:根据历史研究得出的气味成分,对特定欧洲文化具有或曾经具有重要意义的气味提供传统气味的嗅觉表征。

除此之外,Odeuropa 项目还带来了其他一些落地成果:①设计了各种开源工具和演示器,可以帮助导航该项目从数字遗产收藏中捕获的大量与气味相关的数据,这些数据资源均可免费使用,并且可在其 GitHub 空间及其开放获取语料库、词汇表和基准中找到;②开发用于分析嗅觉遗产信息的开放访问 Odeuropa 数据模型,提供用于描述气味及其相关体验的语义模型;③Odeuropa 团队与《阿姆斯特丹历史评论》合作设计了一个在线开放获取的教学模块:通过感知了解如何教授气味的历史,通过视频和作业,介绍在课堂上教授感官历史的最佳实践技术;④在线气味追踪器(仍在开发中):有助于识别气味研究不同方面的历史和当代信息[51]。

(4)文化实物

罗马帝国在线硬币(Online Coins of the Roman Empire,OCRE)项目是美国钱币学会和纽约大学古代世界研究所共同发起的,旨在帮助用户识别、分辨、研究丰富多样的古罗马硬币[52]。在 2012 年末发布第一版以后,该项目取得了巨大的成功与实质性长足发展,是智慧数据支撑数字人文研究的一个示范。全世界几乎所有收藏古罗马硬币的博物馆和历史研究中心都加入了该项目,并共享馆藏硬币数据。该项目设计了统一的本体知识库,汇聚几十个不同数据集的成千上万种硬币数据,以关联数据

形式发布,提供专业分析挖掘功能。除了数字化后机器可读的硬币全貌外,很多内容(如人物形象、符号)通过 IIIF 给予特定 URI 并纳入知识库,可按照类型、属性、关系等机器可处理的关联数据进行浏览、查寻和深度学习,在“定量分析”部分还显示硬币类型的平均测量值,按用户所选参数在后端进行类型分析或测量分析后直接生成图表。例如,通过比较硬币含金量、含银量、含铜量与帝王兴衰成败周期之间的关系,研究者不用亲身前往各个实体收藏机构,便能通过这些数据开展过去无法想象的研究,其功能远远超过了传统网站。

正如 Gruber[53] 指出的和我们直接尝试的经历那样,钱币学研究一直被视为历史学和考古学中一个深奥的子学科。通过成功引入直观的在线用户界面,采用以本体知识库为基础的语义增强的智慧数据,OCRE 可以作为揭示高度专业化知识的桥梁,为更多的考古学家和古典主义学者提供信息和定量分析研究结果,揭示出“未知—未知”。

2.3 大众提供的混合型数据

在文化遗产数字化保护中,往往涉及多来源、多类型的数据集成,大众提供的混合型数据也是重要的数据之一。下文将以一个日本地震资料数字化保护实例来展开论述。

2011 年 3 月东日本发生大地震和海啸后,日本国立国会图书馆建立了“东日本大地震档案库”,用于记录和报告地震灾害[54],收集的大部分资源由全国各地的志愿者提供。这个数字存储库中的资源多是非文字型的照片及其他类型的资料,比如有多张地震前“釜石大观音”照片。日本各地的地震档案资源都通过这个网站聚合,例如,青森地震档案库主要收录东日本大地震受灾地青森县八户市、三泽市、奥入濑町、阶上町的档案,档案类型既有官方文件,也有由市民提供的照片、视频和访谈记录[55]。日本筑波大学数字化存档项目为上述数字档案以及非物质文化遗产、表演艺术、灾难等非物质和体验实体的数字档案构建数据模型,曾召开两个专题研讨会对该模型展开讨论,并根据亚洲不同国家的非遗数字化研究成果完善该模型[56]。

在数据收集过程中,有形实体的(物理的、模拟的或数字的)存档资料容易获取。对于无形实体而言,通常只能收集到特定事件(如一个传统宗教仪式)的记录,以及与无形实体(如发生的自然灾害)相关的有形物件的资料,如近年来的非遗数字档案,包括以数字方式管理非遗保护的记录、围绕非遗事件的相关文档(文本和视听的)等。但无形实体与数字档案对象之间的关系,并不像有形实体数字档案那样明确且清晰,其数字档案元数据模式也不一定能像现有文化记忆机构数字馆藏元数据标准那样以实体为中心[57]。实际上,以事件为中心的元数据模式已为越来越多的文化遗产存储、研究、服务工作所采用。

2.4 专题知识整合与关联

在建设高质量知识库的基础上,如何与外部的高质量数据进行整合或形成关联是智慧数据价值实现的重要保障。以下以德国国家经济学图书馆(ZBW)的 20 世纪新闻剪报档案数据集(PM20)数据捐赠项目为例进行说明。该数据集收集了 1908 – 2005 年来自德国及世界各地的 1 500 多种不同来源的新闻报道,按照人物、机构与主题等进行分类组织,形成了高质量的规范化结构数据,并提供包括 25 000 个文件夹超过 200 万页的内容免费在线访问[58]。在 2019 年 Wikidata 七周年庆祝活动时,ZBW 将 PM20 数据集作为生日礼物捐赠给 Wikidata,以 CC0 许可协议发布,通过元数据匹配、映射、链接与补充,集成到 Wikidata,借助 Wikidata 平台促进数据共享、访问和利用[59]。

该项目在文献资源长期保存、数据开放理念认识、数据价值激活、技术平台借力等方面,值得文化记忆机构参考借鉴。PM20 提供查询提问、查询表单、查询语句 SPARQL 模板等多样检索方式,用户查询个人文件夹时,提供查询语句,Wikidata 根据人物出生地分布图,能提供多种分析结果。举例来说,按出生地生成经济学家地图,可以在 PM20 网页[⑩]上选取对应的语句模板,直接在 Wikidata 查询并获取相关的原始报纸页面和人物数据,也能进行交互式查询限定与可视化展示,如下页图 1 所示。这种交互方式远超维基百科普通网页浏览体验。PM20 带来的启示是:①从名称规范的角度制作地点和民族的知识图谱;②借助 Wikidata 背后的开源系统 Wikibase 开展文化遗产智慧数据开放平台建设。

在实现数字化之后,迈向数据化的过程中,不论是结构化、半结构化,还是非结构化数据,都需要考虑语义增强,对于机器可读文字,由知识组织系统和本体知识库所支持的机器学习工具可对其进行分析。在大数据与 AI 的共同作用下,人文学者逐渐适应数据驱动的数字学术环境,跨学科综合集成型的研究范式越来越普遍。

2.5 国家与区域级知识网络——智慧数据底座

国家和区域级知识网络构成智慧数据的底座。一个典型案例是芬兰语义计算研究组(Semantic Computing Research Group,SeCo,包括阿尔托大学理学院计算机科学系和赫尔辛基大学数字人文中心及多个跨学科合作单位)创建了环型数据发布 Sampo 模式以及若干服务端口。这是一套应用于文化遗产领域的通用数据模型,该研究组基于共享本体和知识库开发了关联数据创建与发布平台[60]。依据 Sampo 模型,SeCo 陆续建立了一系列专题知识库[61],并在语义网和 AI 持续加持下增进数字人文的深度研究。2001 年至今,在语义网的发展进程中,Sampo 模型不断引入本体、关联数据、知识图谱和 AI 新技术,已经得到充分发展,目前已有跨越各专业领域和部门的二十多个子项目,涉及文化、政府、地理、医疗、生物学、学习教育以及商务等领域,能够利用关联数据对异构资源进行数据整合和增值,并创建大规模情境化的集成数据集和知识图谱,包括开放共享的通用本体知识库(如时间、地理、人物、事件等)和针对不同领域的本体知识库,统一通过结构相似的 Sampo 端口进行服务。

Sampo 模型的结构是一个共享的内聚式本体基础架构[62],包括依据语义网标准建模的共享领域本体及其对应的元数据格式。不同来源的元数据描述向内部关联,形成语义增强的互联网络结构,最终形成一张巨大的知识图谱。

通常文化遗产数据 V 的内在相互关联性很强,但现实是这些数据发布在异质、分散的本地资源孤岛,很难在全球范围内得到充分利用,而且这些内容通常仅供人们阅读,较难作为数字人文专业分析和应用程序开发的数据。因此,SeCo 提出文化遗产关联资源的协作出版模型,为数字人文研究和应用创建共享资源服务和语义入口网站,根据六项设计原则,采用递进方式解决具体应用问题。六项原则

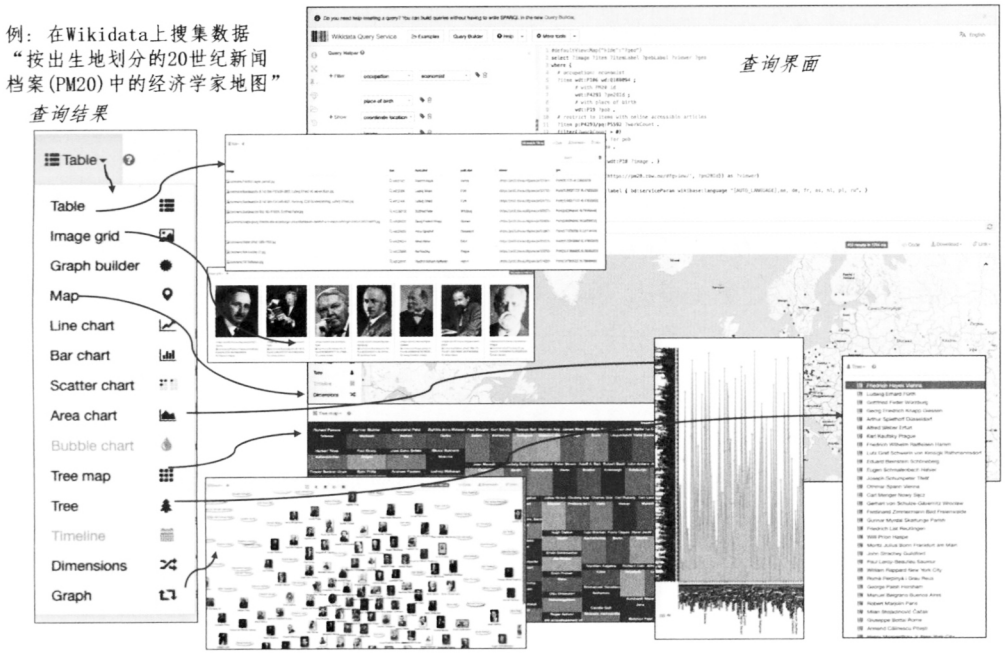

图 1 Wikidata 上 PM20 数据集的多样查询分析

注:采用已编好的查询提问(query)“Map of economists in PM20 by place of birth”可以直接在此页 * 获取,用于 Wikidata 搜索。

* 根据专家研究提供的各种查询提问页:https://www. Wikidataorg/wiki/wikiduta:WikiProiect20thcenturyPressArchivesUsecases。

依次为:①支持协作数据创建和发布;②使用一个共享的开放本体基础架构;③明确区分开放关联数据服务和用户界面;④为同一数据提供多种视角;⑤在系统门户中采用简单的两步循环(首先筛选出具有某些共同特征的目标人群,然后进行更详细的分析)的标准化做法;⑥除了数据探索之外,还支持数据分析和知识发现。

Sampo 项目展示了三代数字人文语义门户研究中的焦点转移。第一代的研究重心是语义门户的开发,围绕数据的协调、聚合、搜索和浏览;第二代的研究重心转移到为用户提供解决问题的交互式综合工具;第三代的研究重心是 AI 增强。未来研究人员可借助此门户找到并聚焦研究问题,在研究人员设定的约束条件下,尝试自动解决问题,并解释结果。

Sampo 的未来工作目标是开发基于 AI 的数字人文工具。这些工具不仅能为研究人员提供更有效的数据获取方式,还能帮助研究人员解读数据,尝试解决数字人文研究问题。AI 技术对于创作和丰富语义门户背后的知识图谱来说意义重大[63],Sampo 随着语义网技术、知识图谱、AI 技术等发展加持下的项目建设历程如图 2 所示。

芬兰 Sampo 模型和项目带来的启示是:图档博智慧数据的生成不是一蹴而就的,需要持续积累与渐进发展的实践路径,从数据的结构化开始,到本体、关联数据驱动的语义关联与丰富化,再到 AI 赋能的语义计算,为数字人文研究构筑坚实的数据基础设施与可持续发展的数据资源保障机制。

2.6 归纳:异构多模态文化遗产智慧数据生成

根据分析,本文归纳异构多模态文化遗产生成智慧数据的主要项目与技术要点(见下页表 1)。该表提供了一个文化遗产智慧数据生成的框架引导,从异构多模态数据到综合集成多元化数据,可进一步改编与扩充。

3 为非遗活态文化生成智慧数据

3.1 理解非遗活态文化内涵

在文化遗产活化利用中,非遗最贴近当下社会实践,是连接过去、现在、未来的活态文化。非遗之所以是活态文化,是因其将优秀传统文化的历史传承与当下的社会大众现实生活结合起来,以人为本,形成了可持续发展的文化生态。

物化(有形的)藏品和活态(无形的)非遗应统一在文化遗产活化利用的整体格局下。在理解文化遗产及构建数据资源体系时,应明确物质与非物质并无严格界限,两者的文化内涵是相通、关联的。在数据层面,不能将非遗项目以狭义的静物来看待,需要抓住活态的内核本质——“人”这一主体。物质文化遗产以物为载体,非物质文化遗产以人为载体。这里的“物”泛指文物、遗存,“人”指非遗传

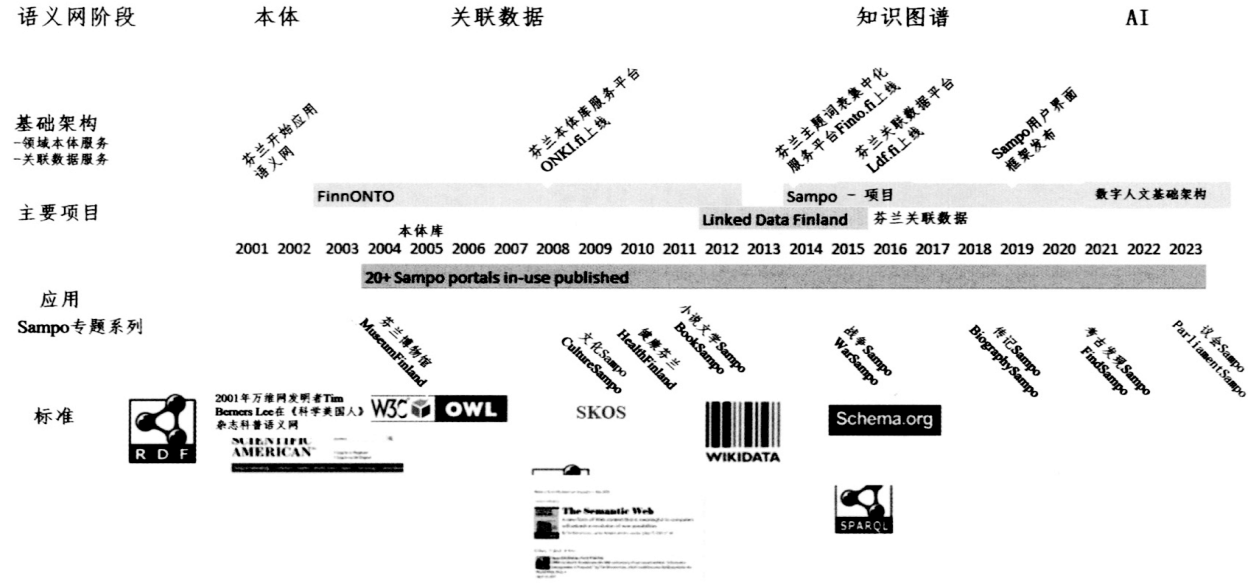

图 2 语义网技术发展和芬兰 Sampo 系列项目[64]

表 1 文化遗产智慧数据生成

数据类型			项目举例	技术赋能
异构多模态	文本类	机器可读文字数据（结构化程度）	美国 MEDLINE 索引编制	结构化数据自动生成
			日本人文开放数据学术中心项目	OCR 识别、IIIF、深度学习
			Zooniverse 图书注释分类项目	机器学习识别
			美国肯特州立大学液晶研究所（LCD）创新史项目	非结构化数据解析、文本挖掘
			美国档案特藏元数据 Finding Aids 项目	半结构化文本分析、语义增值
	非文本类	图像（2D/3D）	中华古籍资源库	OCR 识别、机器学习
			北京故宫博物院“数字多宝阁”项目	3D 建模、IIIF
			敦煌研究院“数字藏经洞”项目	虚拟现实、元宇宙
			美国审计署医疗诊断影像	机器学习识别
			DeepAI 生成图像	提示词工程、文生图
		音频	美国辛辛那提大学关联阅读项目	语音识别与文本语义分析相结合
			古登堡有声读物特藏项目	文本转语音、声音合成
		气味	重现欧洲香气 Odeuropa 项目	基于文本与图像等的嗅觉分析、多语种语料库
		实物	罗马帝国在线硬币（OCRE）项目	关联数据、本体、IIIF
综合集成		大众提供的混合型	日本筑波大学数字化存档项目	以事件为中心的数据建模、本体
		专题知识整合与关联	德国国家经济图书馆（ZBW）的 PM20 数据捐赠项目	Wikidata/Wikibase、关联数据
		国家与区域级知识网络	芬兰 Sampo 模型及系列项目	关联数据、本体、知识图谱、自动分析与推理

承人及相关传承、管理与体验的群体。非遗数据资源建设从静态平面的项目名录出发，应更多关注传承人及相关群体的活动，从关注物质的外在形态发展到挖掘非物质的内涵价值。

在历史发展长河中，人类创造的丰富的物质文化，逐渐与创造者分离，成为跨越时间的历史遗存。而非遗始终以人为主体，因人的存在而存在。非遗的丰富内涵是由人所拥有的知识、经验、技能、行为等构成的。当非遗传承人去世或传承断代导致失传，非遗项目也将不复存在。在 AI 新时代，AI 技术能全方位赋能非遗的保护、传承、发展和创新。

3.2 以人为中心开展非遗数据模型设计

非遗是以人为中心的活态体验型文化，具有多样性、丰富性、本土性及实践性等特征，其蕴含的技艺、民俗、知识体系等都与传承人的智力活动紧密相关，更多表现为口耳相传的人传人活动。由于传承人携带非遗基因，随着去世或无传承人，许多非遗项目种类，如传统音乐、传统技艺等濒临消失，造成“人去艺绝”局面。因此，以非遗代表性传承人为中心，开展活态文化的抢救性记录是长期且迫切的文化工程，是面向非遗活态文化的智慧数据建设的重要场景，也是弘扬多元包容的中华民族优秀传统文化最有力的抓手之一。

（1）非遗数据资源的现实基础

国家级非物质文化遗产代表性传承人抢救性记录工程（以下简称“记录工程”）于 2013 年开始试点，2015 年全面开展，面向各省市年岁已高（70 岁以上）的国家级非遗传承人，利用先进的多媒体技术手段，开展全方位数字化采集记录，主要形式包括口述史、传承教学、项目实践以及记录宣传片拍

摄等,相应的基础性数据工作包括文献资料搜集与整理、保护工作建档与建库等。记录工程是一个典型的多学科参与应用场景,涉及历史学、社会学、艺术学、传播学、信息资源管理学以及计算机科学与技术等。围绕记录活动,以图书馆为代表的信息资源管理学科和专业机构,充分体现出在文化遗产资源建设中的重要作用,“藏、建、用”并举,特别是围绕“大文献观”发挥资料搜集、整理、揭示与深度加工的专业特长,拓展了多元化社会教育与文化传播服务手段。

国家图书馆于2012年启动“中国记忆”项目,主要围绕重大历史事件与重要人物、非物质文化遗产两大类别,以口述史形式开展资料整理,涉及各种文献类型,以影像记录为特色,取得了丰富的资源成果。截至2020年,自建50多个专题库,累计数据量达到80TB级别[11],这些为机器学习训练、开展深度分析与挖掘提供了良好的数据基础。最近启动的图书馆共建项目“人口较少民族口头传统典藏计划”,拟对28个人口较少民族口头传统进行系统性记录,同时开展典藏专题资源库建设,这与记录工程的保护愿景和坚定文化自信的使命是一致的。

国家图书馆“中国记忆”项目中心负责起草了《国家级非物质文化遗产代表性传承人抢救性记录工程操作指南》(2016年),在具体操作与经验分享层面,各省市在该指南基础上,也相继出台各地适用的指南版本,国家图书馆出版社也出版了《国家级非物质文化遗产代表性传承人抢救性记录十讲》等相关资料。

记录工程形成的基础资料具备一定的数字化基础,是最宝贵的第一手活态文化数据。如何提高记录的质量与效率,AI能够提供帮助。要发挥作为数据要素的文化遗产的价值,需要进一步在数据化与数智化的递进上深挖提升。如何借助AI技术,识别、描述这些多模态数据,将其整合与集成为数据资源体系,并逐步上升为智慧数据,是当前重要的工作任务。在AI新时代,文化遗产数据资源的数据建模仍需要人来设计主导。AI技术能在遵守数据标准规范和指定规则下,较为高效地生成和处理数据,但仍需要人来审查其结果的可信性和可解释性。

(2)非遗数据模型要素

从数据角度认识非遗活态文化,需要打开格局和视野,关联多维要素。围绕非遗进行本体设计与知识库构建,已有的思路多是从非遗项目名录出发,关联传承人和文献。非遗保护记录工程是以传承人为中心开展的,传承人作为承载非遗项目及相关文化元素的生动鲜活个体,将事与物联系起来。非遗数据模型在“人—事—物”基本框架下深挖三者之间及各实体内部的语义关联,集成多模态数据类型,形成生动立体的数据视界,而非仅揭示项目与传承人之间的基本联系。

参考以事件为中心和以物件为中心的数据建模方法,结合非遗保护现实,笔者提出非遗数据建模的五个基本要素:项目—人—事(活动)—物—地。①非遗保护工作的起点是项目认定与名录建设,将非遗项目作为主要实体,其属性包括项目的基本文化信息、管理信息与保护信息等。②非遗传承人是非遗项目存在的根本所在,将非遗传承人(包括各级代表性传承人)作为“人”这一实体的主要组成部分,形成传承谱系表,其中还应包括学徒、当地民众、管理者等其他人物角色和团体组织。③非遗作为活态文化,事(活动)是区别于静态文化的重要方面,主要反映在民俗、习俗、社会活动方面,关联出其背后的历史文化背景,也是当下传承文化最有显示度的方面。④非遗虽然名称上体现为非物质性,但其存在依赖于现实各种实物形式。这里的“物”是广义的,涵盖非遗有关的各种实物,如非遗传统技艺的产品、民俗活动的服装、传统音乐的乐器、传统医学的药材和用具等,也包括记录非遗相关的文献和实物。⑤非遗项目都有特定的地域属性和民族特征,两者紧密相关。在地域方面,同一类非遗项目在不同地域发展出各自特色分支,如2011年入选UNESCO人类非物质文化遗产代表作名录的中国皮影戏包含唐山皮影戏、四川皮影戏、沙河皮影戏等37个地方流派。在民族方面,中国少数民族众多,少数民族非遗项目大多具有强烈的地域文化特色,如羌族的羌年、藏族的格萨尔等。

数据模型是数据资源体系的核心组件,奠定了数据要素、结构与语义基础,AI可以在效率提升、数据生成辅助、数据价值挖掘与服务应用开发等方面

赋能助力。

3.3 AI 赋能非遗智慧数据生成

非物质文化遗产的特点是多元化、多维度、多载体以及高度包容,与物质遗存、物件以及人物、地点、事件等有着直接的联系。基于以上对非遗数据基础的分析,AI 技术帮助非物质文化遗产保护的主要作用体现在以下方面。

(1)AI 提高非遗项目抢救性记录工作的效率与深度。主要体现在以下两个方面:①对传承人所承载的技艺、知识、记忆进行记录与保存,主要采用口述史、访谈等方式,将资料搜集与影像记录相结合,通过数字多媒体进行全面、真实记录,并对资料进行深度挖掘与利用。②利用 AI 技术转录和翻译口述历史、歌曲和故事,使研究人员和公众更容易理解它们,机器学习算法还可以分析这些内容背后的历史和文化,为不同社群的文化习俗和信仰提供有价值的见解和参考。

(2)AI 助力非遗数字档案保护与知识图谱生成。非遗主要通过口头或实践传承,随时间推移而面临很大的消失风险。AI 可以通过创建这些文化元素的数字档案来帮助减轻这一风险,确保这些文化习俗和传统不被遗忘;依据五维数据概念模型,通过本体、词表生成知识图谱,形成高质量的非遗智慧数据底座。

(3)AI 推动非遗创造性转化与创新性发展。开发交互式 AI 应用程序,帮助公众了解不同的文化习俗和传统,并提供个性化的学习体验。通过 AI 技术打造非遗传承人数字分身,模仿非遗传承人行为,例如为羌年代表性传承人释比制作数字人。运用 AR 与 VR 等多感官交互技术,在数字虚拟空间里复现非遗传统技艺工序,表演传统戏剧等,社会公众可以进行沉浸式体验,从而激发出参与非物质文化遗产保护的兴趣和行动。

虽然 AI 带来了令人激动的非遗保护与活化利用的畅想可能,但仍需警惕技术带来的负面问题[65]。AI 在这一领域的使用引发了一些伦理和实践方面的考虑,如要确保 AI 的使用尊重文化遗产记录和保存的社区的权利和隐私,此外也要考虑 AI 的局限性,不能依赖它作为保护和研究非物质文化遗产的唯一解决方案。

3.4 羌年智慧数据设计思路

羌年作为羌族文化的代表,是流行于四川省理县、茂县、汶川县、北川羌族自治县等多个羌族聚居地(乡镇)的传统节日,是中国国家级非物质文化遗产代表性项目,也入选了 UNESCO 急需保护的非物质文化遗产名录,其类别为社会实践、仪式、节庆活动[66]。

由于羌族没有文字,民族文化主要通过人与人的口口相传而得以延续。当前羌族年轻人大多说汉语,越来越少的羌族年轻人能够用羌语交流。20 世纪 90 年代,国家民族事务委员会开始推动羌语拼音方案,在“中国语言资源保护工程”采录展示平台上⑫,也对羌族语言进行了对应的记录保存。

现有的羌族文字资料都是清代以来用汉字记录的,如清道光时期的《石泉县志》。北川县档案馆在 2008 年地震前存有 8.5 万卷档案,主要是近现代档案与地方志。国家图书馆“中华古籍资源库”收藏有北川(古名石泉)县志的全文影像,可以利用 AI 与机器学习技术对这些数字化的图像(含矢量图)进行文本识别与挖掘。

2008 年 5 月 12 日发生的汶川地震对羌族文化造成重大损失。从非遗保护的角度来看,四川省非物质文化遗产保护中心及时开展抢救性记录保护工作,通过多媒体音像手段记录以羌族传承人为代表的一批非遗项目,目前非遗档案资料整理已有一定工作积累。以羌年为例,除了保护建档基本要求,还形成了《百转千音话羌年,消失边缘的羌族“释比”》等高质量纪录片。四川省非物质文化遗产保护中心编制的《四川省非物质文化遗产名录图典》(1—5 卷)主要以图文形式编排了各个批次的非遗项目内容。电子科技大学数字文化与传媒研究中心聚焦数字人文,采用档案记录方法,编制藏羌彝非遗档案系列资料,其中羌族卷分上下两卷,囊括羌族非遗项目 81 项(国家级和省级)[67-68]。这些都是羌族文化遗产的基础性图文资料,需要使用 AI 技术来对这些文本记录和音像资料进行分析与挖掘,以支持进一步创新活化利用,而不是以存放在库房的形式来进行保护。

在数字人文研究视野中,相比于物质文化遗产的数字化和数据化,羌族本土文化研究和促进多民

族文化交流的智慧数据生成面临很大挑战。从羌族文化遗产角度看,羌年作为传统节日,集成了锅庄舞、音乐、服装、人物(例如释比)等多种要素,不仅关乎羌族文化历史,还与当下羌族人民的日常生活紧密结合在一起。从羌族语言出发,丰富的民族活动和生活艺术表达形式,如歌舞、宗教仪式、口述故事、传统技艺(刺绣、头饰、服饰)等,共同构成了羌年非物质文化遗产的丰富内涵。

羌族文化涉及多个非遗项目类别,包括民俗、民间文学、传统戏剧、传统音乐、传统舞蹈、传统美术、传统技艺等。对不同类别非遗项目进行数字化与数据化时有不同的处理思路和特点,将各类别非遗项目关联集成,方能全面体现羌族文化景观,如图 3 所示。该图主要从数据建模扩展开来,下方的异构多元、多模态数据是 AI 赋能发挥作用的着力点,笔者从元数据生成的文化与语言角度,探讨羌族非遗文化资源的数据建模及与 AI 的结合。

4 总结与展望

文化遗产活化利用是历史文化传承与当代社会发展的共同价值取向。AI 技术带来了令人激动和兴奋的新机遇与新挑战,为文化遗产数据加工、处理与分析增添了新能力,数据量横向扩展和细粒度纵向深入,有力助推了文化遗产智慧数据的生成和价值实现。

文化遗产数据资源建设与智慧数据的生成不是一蹴而就的,不能单纯依靠技术来突破和完成。文化遗产数据资源建设是一盘棋,需要深入持续地理解文化的多元性、时代性以及文化遗产的多样性、丰富性,注重对文物、文献、音视频等不同载体资料的纵向挖掘与横向关联。面向文化遗产活化利用的智慧数据生成,需要满足“专—精—特—新”要求,即数据资源的专业化管理、精细化加工、特色化打造、创新性转化。

以图档博为代表的文化记忆机构应时刻守住文化遗产数据资源根基,以积极开放的心态拥抱 AI,尝试探索运用各种 AI 技术,面向文化遗产活化利用,更快、更好地生成文化遗产智慧数据。芬兰国家级数据基础设施的建设,有力地印证了 AI 赋能文化遗产智慧数据的生成需要持续积累和不断努力。如何做到与时俱进、厚积薄发,可为我国文化数字化战略中提到的“关联形成中华文化数据库”提供路径和实现方案,具有参考意义。

本文提出文化遗产智慧数据生成路径的四点策略。

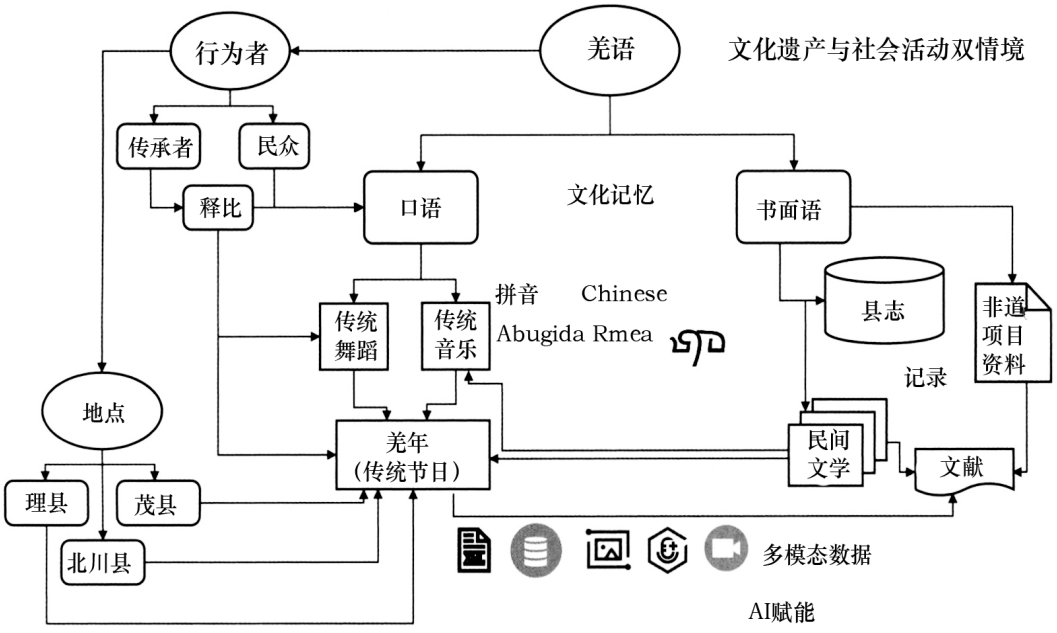

图 3 羌族非遗智慧数据设计思路

注:翻译、改编自 FAN Wei 在第 20 届都柏林核心和元数据应用国际会议(DCMI 2022)上的报告,详见 https://www.dublincore.org/conferences/2022/sessions/panel-cultural-and-linguistic-challenges/#rec4CBiatJy0bIryG。

(1)主动拥抱 AI,抓住新一轮 AI 赋能机遇,及时补齐数据基础设施短板,加强数据资源体系建设。

虽然图档博机构在文化遗产保护与传承上有着共同使命,但因其各自馆藏资源特点,具体业务侧重各有不同,需要在 AI 赋能文化遗产的多学科交叉方向下进行定位。以图书馆为例,专业优势体现在大文献观下的资源识别、描述、揭示、检索与发现,以智慧数据生成为目标,适合在多模态数据及其加工场景中深度参与。

值得一提的是,AI 需要足够的数据量(未必是海量)和高质量的数据,这对文化记忆机构的信息资源管理业务水平提出了进阶要求,例如更高效的辅助与自动化编目。在数据量要求方面,要为 AI 应用不断积累数据,进一步促进更大规模的数字化;在数据高质量要求方面,要求数据资源具备跨系统的互操作能力,并采用恰当的元数据标准、领域本体与词表。

与以前人工主导、机器套录相比,当前的生成式 AI 为结构化数据的加工与转换提供了加速器,在提示词工程与规则训练指导下,更高效地进行数据加工,让图档博的智慧数据基础更加牢固,这样才能支持更多的智能化应用,实现智慧数据的价值。

遵循文化遗产领域元数据标准规范,利用 AI 探索文化遗产结构化数据的自动化(辅助)生成,以及自动链接发现等。参考国外文化遗产领域的数据模型与元数据标准规范,结合中国本土文化遗产多元内涵与特质,制定应用纲要和 AI 使用规则,建设可持续发展的文化遗产数据资源体系。

强化各类知识组织系统与词表资源的对齐,利用 AI 提高自动标引和数据整合的能力。在 AI 技术赋能文化遗产智慧数据建设场景中,分类法、叙词表、本体等各类知识组织系统和受控词表不会被替代,依然会起到关键性中介与辅助作用,是高质量(结构化与语义化)、高价值(挖掘)的智慧数据生成的重要保障。

(2)尽快开展馆藏数据资源的“大语言模型 + 知识库”结合工作,实现智能分析与计算增强。

从文化遗产数据资源体系的数据建模出发,以实体为基础,强化以事件为中心的文化遗产活化和活态特点。高质量的结构化、关联化数据资源与大语言模型是互惠的结合。一方面,利用大语言模型开源框架自动辅助开发垂直领域的专业知识库;另一方面,知识图谱的高质量可信知识内容可以促进大语言模型应用优化,提高语义准确性。利用大模型与机器学习技术,提升文化遗产多模态数据的识别、处理、整合、分析与挖掘能力。同时,智慧数据也为大模型学习与训练提供数据养料,进一步优化 AI 技术的准确性与专业性,为数字人文深度研究赋能。

文化遗产数据资源高质量发展,需要数据与 AI 技术充分结合,发挥各自作用并相互增益,以数据驱动模型优化,用大模型改进数据质量,推动两者协同发展。

(3)鼓励更广泛的文化遗产数据开放与共享,支持活化利用的创新。

从文化遗产数据资源体系到智慧数据的实现,要积极主动利用各类 AI 工具,不只是简单编辑加工素材,更多的是要开展自身特有资源的创新活化利用,如数字游戏体验、文创产品等。根据实际情况,文化记忆机构考虑适度、有序开放自身馆藏文化遗产数据,让更多的社会公众有机会通过数字化手段进行观赏与体验。

(4)确保可信的智慧数据。

可信是 AI 技术应用的通用挑战之一,文化遗产领域也不例外。不论是人工采集还是机器生成的数据,都需要严谨地进行记录,形成文档,以便后续验证与复现。要对数据准确归档,并进行关联、分析与挖掘,同时警惕唯 AI 论、AI 幻觉与伦理等问题。由于文化固有的多元性,应避免特定文化时空语境下的敏感性和争议性问题,时刻保持文化包容性理解与处理,鼓励创新性数字文化叙事。

用好 AI 是综合性系统工程,存在复杂的挑战,涉及数据要素(数据量不足、数据质量差、价值挖掘浅等)、高计算资源消耗、内容版权、社会伦理、人才与专业技能以及资金不足等诸多问题。

最后,让 AI 回归技术,让文化回归文化。保持

文化底色更加鲜亮和突出,从数据中生成文化遗产活化利用的智慧,让文化遗产在当下“活起来”“火起来”。

注释:

①http://read.nlc.cn/thematDataSearch/toGujiIndex

②https://olympiacommongrounds.gr/explore

③https://interwoven.map-india.org

④https://www.commarts.com/exhibit/sol-lewitt-app

⑤https://news.microsoft.com/en-ca/2022/02/01/government-of-nunavut-preserving-endangered-inuit-languages-and-culture-with-the-help-of-artificial-intelligence-and-microsoft

⑥https://www.dpm.org.cn/shuziduobaoge

⑦https://www.chnmuseum.cn/zl/ztcl/202305/t20230529_258570.shtml

⑧https://dlc.e-dunhuang.com

⑨https://pano.dpm.org.cn

⑩https://www.wikidata.org/wiki/Wikidata:WikiProject_20th_Century_Press_Archives/Use_cases

⑪https://www.nlc.cn/cmptest

⑫https://www.zhongguoyuyan.cn/index

参考文献:

[1]UNESCO. Recommendation on the ethics of artificial intelligence[R/OL]. (2021-11-23)[2023-12-15]. https://unesdoc.unesco.org/ark:/48223/pf0000381137.

[2]习近平对非物质文化遗产保护工作作出重要指示[EB/OL]. (2022-12-12)[2023-12-15]. https://www.gov.cn/xinwen/2022-12/12/content_5731508.htm.

[3]习近平. 把中国文明历史研究引向深入 增强历史自觉坚定文化自信[J]. 求是,2022(14):4-9.

[4]崔海教. 提升数字文化建设水平[N]. 人民日报,2022-08-03(11).

[5]National Science Foundation. Open knowledge network roadmap:powering the next data revolution[R/OL]. (2022-09-15)[2023-12-15]. https://nsf-gov-resources.nsf.gov/2022-09/OKN%20Roadmap%20-%20Report_v03.pdf.

[6]Over € 4.4million granted to four new projects to enhance the common European data space for cultural heritage[EB/OL]. (2022-12-12)[2023-12-15]. https://pro.europeana.eu/post/over-4-4-million-granted-to-four-new-projects-to-enhance-the-common-european-data-space-for-cultural-heritage.

[7]熊远明,白雪华,吴建中,等. 国家文化数字化战略:图书馆的专业阐释与使命践行[J]. 中国图书馆学报,2022,48(4):4.

[8]Developing a library strategic response to artificial intelligence[EB/OL]. (2023-11-20)[2023-12-15]. https://www.ifla.org/developing-a-library-strategic-response-to-artificial-intelligence/.

[9]曾蕾,王晓光,范炜. 图档博领域的智慧数据及其在数字人文研究中的角色[J]. 中国图书馆学报,2018,44(1):17-34.

[10]曾蕾.《中国GLAM公开课》第九十六课:智慧数据与数字人文视野下的图档博数据崛起[EB/OL]. (2022-04-19)[2023-12-15]. https://lib.shu.edu.cn/info/1022/3555.htm.

[11]张晓林,梁娜. 知识的智慧化、智慧的场景化、智能的泛在化——探索智慧知识服务的逻辑框架[J]. 中国图书馆学报,2023,49(3):4-18.

[12]王晓光,侯西龙. 面向活化利用的文化遗产智慧数据建设论纲[J]. 信息资源管理学报,2023,13(5):4-14,43.

[13]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.

[14]联合国土著人民权利宣言[R/OL]. (2007-09-13)[2023-12-15]. https://www.un.org/development/desa/indigenouspeoples/wp-content/uploads/sites/19/2019/06/UN-Declaration-Rights-of-Indigenous-Peoples-DGC-WEB-CH.pdf.

[15]NIKLA-ANCIA. Respectful terminology platform project[EB/OL]. [2023-12-15]. https://www.nikla-ancla.com/respectful-terminology.

[16]DE-BIAS-Detecting and cur(at)ing harmful language in cultural heritage collections[EB/OL]. (2023-01-01)[2023-12-15]. https://pro.europeana.eu/project/de-bias.

[17]CARROLL S R, GARBA I, FIGUEROA-RODRíGUEZ O L, et al. The CARE principles for indigenous data governance[J/OL]. Data Science Journal, 2020(19)[2023-12-15]. https://datascience.codata.org/articles/10.5334/dsj-2020-043.

[18]第五届中国数字人文年会举行[EB/OL]. (2023-12-11)[2023-12-15]. https://news.whu.edu.cn/info/1015/448417.htm.

[19]Gemini Team, Google. Gemini: a family of highly capable multimodal models[R/OL]. (2023-12-19)[2023-12-25]. https://arxiv.org/abs/2312.11805.

[20]人工智能[EB/OL]. [2023-12-15]. https://upimg.baike.so.com/doc/2952526-3114987.html.

[21]Permion. Permion platform[EB/OL]. [2023-12-15]. https://permion.ai/permion_platform.html.

[22]NOORDEN R V, PERKEL J M. AI and science: what 1 600 researchers think[J]. Nature, 2023, 621(7980): 672-675.

[23]BENJAMIN C G L. The "collections as ML data" checklist for machine learning & cultural heritage[J/OL]. Journal of the Association for Information Science and Technology, 2023[2023-12-15]. https://doi.org/10.1002/asi.24765.

[24]陈力. 数字人文视域下的古籍数字化与古典知识库建设问题[J]. 中国图书馆学报, 2022, 48(2): 36-46.

[25]陈力. 中国古代的知识工具与古典知识库的建设[J]. 中国图书馆学报, 2023, 49(3): 19-40.

[26]Giving art a voice with Watson[EB/OL]. (2017-05-10)[2023-12-15]. https://travelblonde.medium.com/giving-art-a-voice-with-watson-1c1a235cb63a.

[27]AI in relation to GLAMs[R/OL]. (2019-12-12)[2023-12-15]. https://pro.europeana.eu/project/ai-in-relation-to-glams.

[28]AI4Culture—an AI platform for the cultural heritage data space[EB/OL]. (2022-11-14)[2023-12-15]. https://pro.europeana.eu/project/ai4culture-an-ai-platform-for-the-cultural-heritage-data-space.

[29]European Parliament. Artificial intelligence in the context of cultural heritage and museums: complex challenges and new opportunities[R/OL]. (2023-04-28)[2023-12-15]. https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2023)747120.

[30]ARIANNA T. Artificial intelligence applications to cultural heritage[EB/OL]. (2020-12-12)[2023-12-15]. https://rm.coe.int/artifical-intelligence-applications-to-cultural-heritage-by-arianna-tr/1680a096b8.

[31]Microsoft. AI for cultural heritage[EB/OL]. [2023-12-15]. https://www.microsoft.com/en-us/ai/ai-for-cultural-heritage.

[32]As technology like AI propels us into the future, it can also play an important role in preserving our past[EB/OL]. (2019-07-11)[2023-12-15]. https://blogs.microsoft.com/on-the-issues/2019/07/11/as-technology-like-aipropels-us-into-the-future-it-can-also-play-an-important-role-in-preserving-our-past/.

[33]曾蕾, 谭旭. 数据的语义增强——解读图档博支持数字人文的新动向[J]. 数字人文研究, 2021, 1(1): 65-86.

[34]MEDLINE 2022 initiative: transition to automated indexing[EB/OL]. (2021-12-01)[2023-12-15]. https://www.nlm.nih.gov/pubs/techbull/nd21/nd21_medline_2022.html.

[35]人文学オープンデータ共同利用センター. KuroNet くずし字認識サービス(AI OCR)[EB/OL]. [2023-12-15]. http://codh.rois.ac.jp/kuronet/.

[36]IIIF and AI/machine learning[EB/OL]. [2023-12-15]. https://training.iiif.io/iiif-5-day-workshop/day-four/iiif-and-ai.html.

[37]Omniscribe[EB/OL]. [2023-12-15]. https://github.com/collectionslab/omniscribe.

[38]ZENG M L, 之 UMER M, ZHANG Y. Revealing innovation history by using smart data—a conceptual and methodological exploration and demonstration[EB/OL]. [2023-12-15]. https://www.kent.edu/ischool/revealing-innovation-history-using-smart-data.

[39]HU T, ZENG M L, ZHANG X, et al. Spatial-temporal variation based innovation history visualization: a case study of the Liquid Crystal Institute at Kent State University[C/OL]//Digital Humanities 2017. Montreal, Canada, 2017[2023-12-15]. https://dh2017.adho.org/abstracts/571/571.pdf.

[40]LI H, ZENG M L, ZHANG Y, et al. Tackling innovation networks with smart data: a case study of the Liquid Crystal Institute at Kent State University[C/OL]//Digital Humanities 2017. Montreal, Canada, 2017[2023-12-15]. https://dh2017.adho.org/abstracts/334/334.pdf.

[41]ZENG M L, GRACY K F, 之 UMER M. Using a semantic analysis tool to generate subject access points: a study using Panofsky's theory and two research samples[J]. Knowledge Organization, 2014, 41(6): 440-451.

[42]CCTV 纪录片《码农的异想世界》[Z/OL]. (2022-05-07)[2023-12-15]. https://tv.cctv.com/2022/05/07/VIDEJiCOrwguCbd1ZjaQfuY4220507.shtml.

[43]SPENNEMANN D H R. Generative artificial intelligence, human agency and the future of cultural heritage[J/OL]. SSRN Electronic Journal, 2023[2023-12-15]. http://dx.doi.org/10.2139/ssrn.4583327.

[44]Arts Management & Technology Laboratory. A digital future for cultural heritage[EB/OL]. (2020-04-02)[2023-12-

15]. https://amt-lab.org/blog/2020/3/a-digital-future-for-cultural-heritage.

[45] Artificial intelligence in health care: benefits and challenges of machine learning technologies for medical diagnostics[R/OL]. (2022-09-29)[2023-12-15]. https://www.gao.gov/products/gao-22-104629.

[46] ZENG M L, LEE J. Smart data approaches to exploring independent datasets across disciplines, media, and perspectives for research in the humanities[C/OL]//Digital Humanities 2017. Montreal, Canada, 2017[2023-12-15]. https://dh2017.adho.org/abstracts/269/269.pdf.

[47] LEE J, BASNET A, CLARK H V, et al. Close listening and synesthetic reading across multiple poetry archives: tracking the performative afterlives of a poem[J]. Interférences Litteraires/Literaire Interferenties, 2021(25): 68-92.

[48] The Project Gutenberg open audiobook collection[EB/OL]. [2023-12-15]. https://marhamilresearch4.blob.core.windows.net/gutenberg-public/Website/index.html.

[49] Microsoft AI records 5 000 audiobooks for Project Gutenberg[EB/OL]. [2023-12-15]. https://thenewstack.io/microsoft-ai-records-5000-audiobooks-for-project-gutenberg.

[50] Odeuropa: smell heritage—sensory mining[EB/OL]. [2024-01-07]. https://odeuropa.eu/.

[51] LISENA P, SCHWABE D, VAN ERP M, et al. Capturing the semantics of smell: the Odeuropa data model for olfactory heritage information[C]//The Semantic Web(ESWC 2022). Cham: Springer, 2022: 387-405.

[52] OCRE[EB/OL]. [2023-12-15]. https://numismatics.org/ocre/.

[53] GRUBER E. Final report to the NEH for Online Coins of the Roman Empire[R/OL]. (2017-07-28)[2023-12-15]. https://archaeologydataservice.ac.uk/catalogue/adsdata/arch-3016-1/dissemination/pdf/2017/doa_post28093.pdf.

[54] National Diet Library Great East Japan Earthquake Archive. About the National Diet Library Great East Japan Earthquake Archive[EB/OL]. [2023-12-15]. https://kn.ndl.go.jp/static/about?language=en.

[55] National Diet Library Great East Japan Earthquake Archive. Aomori disaster archive[EB/OL]. [2023-12-15]. https://kn.ndl.go.jp/static/2014/04/1?language=en.

[56] Core Cultural Metadata Mode(CCMM) workshop[EB/OL]. [2023-12-15]. https://www.dublincore.org/conferences/2022/sessions/workshop-core-cultural-metadata-model/.

[57] SUGIMOTO S, KIRYADOS S, WIJESUNDARA C, et al. Metadata models for organizing digital archives on the web: metadata-centric projects at Tsukuba and lessons learned[C/OL]//2018 International Conference on Dublin Core and Metadata Applications. Porto, Portugal, 2018[2023-12-15]. https://dcpapers.dublincore.org/pubs/article/view/3968/2166.

[58] NUBERT J. 20th Century Press Archives: data donation to Wikidata[EB/OL]. (2019-10-24)[2023-12-15]. http://zbw.eu/labs/en/blog/20th-century-press-archives-data-donation-to-wikidata.

[59] WikiProject 20th century press archives[EB/OL]. [2024-01-07]. https://www.wikidata.org/wiki/Wikidata:WikiProject_20th_Century_Press_Archives.

[60] Semantic Computing Research Group[EB/OL]. [2023-12-15]. https://seco.cs.aalto.fi/.

[61] SeCo. Sampo model, data services, and series of semantic portals[EB/OL]. [2023-12-15]. https://seco.cs.aalto.fi/applications/sampo/.

[62] HYVÖNEN E. Using the semantic web in digital humanities: shift from data publishing to data-analysis and serendipitous knowledge discovery[J]. Semantic Web, 2020, 11(1): 187-193.

[63] HYVOÖNEN E. Digital humanities on the semantic web: sampo model and portal series[J]. Semantic Web, 2023, 14(4): 729-744.

[64] HYVOÖNEN E. How to create and use a national cross-domain ontology and data infrastructure on the semantic web[J/OL]. Semantic Web-Interoperability, Usability, Applicability, 2022[2023-12-15]. https://semantic-web-journal.net/content/how-create-and-use-national-cross-domain-ontology-and-data-infrastructure-semantic-web.

[65] FRĄCKIEWICZ M. Artificial intelligence and the study of intangible cultural heritage[EB/OL]. (2023-07-27)[2023-12-15]. https://ts2.space/en/artificial-intelligence-and-the-study-of-intangible-cultural-heritage.

[66] Qiang New Year Festival[EB/OL]. [2023-12-15]. https://ich.unesco.org/en/USL/qiang-new-year-festival-00305.

[67] 谢梅, 汪静泉. 藏羌彝非遗档案—羌族卷(上)[M]. 成都: 电子科技大学出版社, 2020.

[68] 谢梅, 汪静泉. 藏羌彝非遗档案—羌族卷(下)[M]. 成都: 电子科技大学出版社, 2020.

Exploring the Generation Path of Smart Data for the Activation and Utilization of Cultural Heritage in the New Era of AI

Fan Wei Zeng Marcia Lei

Abstract: With innovative technologies emerging and undergoing rapid iterative updates, the landscape of AI is constantly evolving, encompassing a multifaceted ecosystem that is beyond natural language – based processing and data – generating.

Based on the smart data construction scenarios of cultural memory institutions such as galleries, libraries, archives, and museums (GLAMs), this article explores the possibilities, challenges, and paths of the activation, utilization, and innovative inheritance of cultural heritage through generating smart data which has now entered a new phase characterized by high efficiency, in – depth analysis, and multi – modal integration in the new era of AI.

The paper first revisits the concept of smart data for cultural heritage. To consider the diverse features of cultural heritage data, the paper introduces the CARE principles (Collective benefit, Authority to control, Responsibility, and Ethics) which complement the existing FAIR principles (Findable, Accessible, Interoperable, and Reusable) by encouraging open, sustainable, and ethical data movements. Smart data can be efficiently processed using AI technology while also being used by AI to generate intelligent applications and produce high – quality data. Such advancement makes applications for promoting cultural heritage more abundant and practical.

The diversity of cultural heritage presents complexities in data processing. This paper sorts five generative patterns for diverse and heterogeneous cultural heritage resources: 1) machine – readable text data, including born – digital or turned – to – digital, unstructured and semi – structured; 2) non – textual data, such as images, sound, scents, objects, etc.; 3) hybrid data contributed by the community; 4) specialized knowledge integration and association; and 5) national level knowledge networks embracing all types of resources. The generation of smart data for GLAMs involves the processes with data structuring, ontology generating, and linked data – driven semantic association and enrichment, culminating in AI – enabled semantic computing.

The paper also emphasizes that human intelligence remains crucial in the data modeling and encoding of cultural heritage resources. Human oversight is essential to ensure the credibility, trustiness, and interpretability of the results.

Intangible cultural heritage reflects the passion and commitment of this paper. AI has the potential to enhance the efficiency and depth of salvage record – keeping while assisting in the protection of digital archives, the creation of knowledge graphs, and the innovative dissemination and creative transformation of intangible cultural heritage.

Balancing technological advancements with cultural sensitivity will be the key to shaping a future where technology and heritage coexist harmoniously. To overcome the challenges and unlock the full potential of AI in cultural heritage preservation and dissemination, generate trustworthy smart data, and advance innovative AI – powered applications, continuous investment in cross – disciplinary research in digital humanities will be essential.

Key words: Artificial intelligence; Cultural heritage; Activation and utilization; Smart data; Generation path