数据标注：文化数字化的战略支点

【要报要点】数据和算力、算法是人工智能三个基础要素。文化数字化进入人工智能时代，文化数据具有特殊战略价值，应加倍重视。本期要报建议，推进实施国家文化数字化战略，应立足文化资源及文化数字化工作基础，完善文化资源数据治理体系。在文化行业生成式人工智能大模型研发过程中，聚焦数据标注这一关键节点，取得文化数据资源竞争优势。

以大模型为代表的生成式人工智能，是数字化发展水平的重要标志，是数字化竞争的制高点。随着生成式人工智能在许多应用领域从辅助走向主导，文化数字化的人工智能时代也正在到来。8 月 28 日，国家文化大数据标识基地等二十多家相关机构，在北京召开文化大模型评测工作座谈会，就文化大模型如何赋能文化数字化建设进行深入研讨。在发展数字经济、建设数字中国和推进实施国家文化数字化战略、建成文化强国背景下，人工智能时代的文化数据资源具有重要战略价值。

一、文化数据是文化高质量发展的战略资源

（一） 文化数据是文化生产要素

数据是数字经济的生产要素，占有数据就是取得生产资料支配权。数据和算力、算法是人工智能的三个基础要素，数据集的数量和质量决定人工智能训练水平。据专业统计，全球网站56%是英文,只有1.5%是中文。北京智源人工智能研究院2021年就推出模型“悟道 2.0”，算法可能比 GPT-4 更复杂，但功能和影响力难以比拟，一个主要原因就是缺乏足够多的高质量数据。我国人工智能时代文化生产的数据资源基础，总量和质

量现状不容乐观。

（二） 文化数据具有意识形态属性

通用大模型和行业大模型只是模拟人类思维和价值观进行人机对话，其所有文化知识和文化艺术作品生成能力都源于训练语料中的文化数据，是数据库经过人为标记后强化学习的结果，实质上还是人类的思维和价值观。文化数据不同于公共数据、商业数据、企业数据、社交媒体数据，还蕴含丰富的人文情感、精神思想、主体意志等文化基因，以及叠加以后构成数字空间的集体意识、社群观念、公序良俗，乃至政治立场、民族认同、价值取向，具有天然的意识形态属性。虚实交织，关系到文化自信、文化安全的根基。

（三） 文化数据具有重要战略价值

数据化的文化资源，具有相对稀缺性。有研究认为，没有任何一项技术比今天的人工智能更依赖大规模高质量数据，“未来一个模型的好坏，20%由算法决定，80%由数据质量决定。” 从现实世界到虚拟世界，衍生出数字疆域及其主权和管治权，更是完全基于数据而存在。文化数据就是文化竞争乃至文化战争的战略资源。 在中、美、欧盟三大经济体人工智能国际竞赛中，管好用好优秀传统文化数字化存量资源和文化创造过程中文化数据增量资源，建立起以中文为核心的数据要素战略壁垒，是必然的战略选择。

二、数据标注是文化数字化战略支点

如果进行战略对标，美国人工智能已经在算力、算法、数据上取得先发优势。从行业应用角度，中美两国都处在发展初期。在文化领域，我国全面赶超，首先应该聚焦关键节点——抢占文化数据资源支配权。

（一）数据标注作为战略支点的可行性分析

数据标注，是基于训练人工智能模型的需要，对文本、图像、音频、视频等原始数据添加标签的过程。标注后的数据成为人工智能的训练数据，可以根据不同的训练任务创建为不同类型的训练数据集。目前衡量大模型训练水平的一个重要指标就是训练数据集的数量和质量。OpenAI 的 GPT3 就是以 45TB 的数据数量领先。关于数据质量，涉及特殊领域专业知识，涉及的文本、图像、音频、视频，都需要经过数据标注，大模型才能“读懂学会”。OpenAI 训练 GPT 的语料雇佣了大量肯尼亚劳工进行数据标注。百度的智能驾驶，就有数千人从事交通信息数据标注。文化数据标注的专业知识要求，背后是高昂的人力成本，极大限制了文化行业大模型研发。伦敦大学的一个人工智能研究团队做爵士乐音符数据标注，因为招募职业音乐家成本太高，只能限于小规模学术性研究，这是这一领域全球性普遍存在问题。我国不同，大量文化机构从业人员及高等教育持续不断输出文化专业人才，具有得天独厚的人资资源优势，完全可以支撑文化行业大模型训练所需海量高质量文化数据标注。2020 年，国家职业分类目录中增加了人工智能训练师这一新职业，其中包含的两个工种之一就是数据标注员。因此，建立大规模高水平文化资源数据标注工作体系，并在文化行业应用大模型国际竞争中取得领先，不仅有必要性也有可行性。

（二）完善的文化数据治理体系是重要的制度保障

关于数据治理，国家层面政策法规陆续出台，从“四梁八柱”到重点领域，治理体系持续完善。国家数据局近期组建，“协调推进数据基础制度建设，统筹数据资源整合共享和开发利用”，数据治理能力和治理效能有了组织保障。关于文化数据，在国家文化数字化战略布局中，以建设国家文化大数据体系为核心进行了全链条顶层设计。随着去年年底人工智能突然

爆发，文化数据的战略价值又提升到新的高度。不同于商业等其他行业数据，文化数据更多集中于公共文化机构，缺少投资驱动力和盈利吸引力，需要行政力量主导、社会力量参与、文化机构主动作为。长期看，更迫切需要市场参与，形成有为政府与有效市场协同能力。应对这一重大变化，目前还只有企业数据治理反应比较敏捷。国家战略层面上，急需更新理念，尽快采取针对性治理措施，完善面向人工智能的文化数据治理体系。

采用情况：本文于2023年10月被《文化和旅游智库报》采用
供稿单位：山东省艺术研究院（文化和旅游行业智库建设试点单位）、国家社科基金艺术学重大项目“科技赋能艺术生产与演出、演播研究”课题组
作者：林凡军 夏源 孟傲