跨学科视角下方言数字化体系的构建：语言学、计算科学与文化遗产保护的融合综述

何正锋　邓倩妍　刘星星　赵天欣　杨小梅*

广西职业师范学院 广西 南宁 530007

【摘　要】：方言作为语言多样性与地域文化的重要载体，在全球化背景下正面临严峻的生存危机。本文基于语言学、计算科学及文化遗产保护的跨学科视角，系统构建以人工智能为核心的方言数字化保护与传承体系。论文深入探讨了多模态语料库建设、人机协同标注分析、低资源方言语音处理与双向翻译等关键技术路径，并提出从协同共建到可持续运营的实践框架。本文识别了当前面临的技术适配、理论整合与社群参与等挑战，指出唯有通过持续的跨学科对话与伦理考量，方能构建技术赋能、人文关怀与生态可持续的方言数字化新范式。

【关键词】：方言数字化；跨学科融合；语言资源保护；人工智能

DOI:10.12417/2982-3846.26.01.004

1 引言：方言保护的数字化转向与跨学科必然性

方言作为语言多样性的核心组成部分与地域文化认同的关键载体，正面临着前所未有的传承危机。在全球化和标准化语言推广的双重影响下，许多富有特色的地方方言使用人群持续萎缩，代际传递链出现断裂，使其陷入功能退化与快速消亡的困境。传统的保护方法，如田野调查的纸质记录与静态音像存档，虽具历史价值，但受限于规模、时效性与互动性，难以应对动态的语言演变与社会需求，更无法有效激发新一代使用者的兴趣[1]。因此，依托数字技术，特别是人工智能，实现方言资源的系统化保存与创新性活化，已成为一项紧迫且必要的学术与社会议题。

方言数字化并非单纯的技术移植，而是一项深度的跨学科系统工程。它要求语言学提供对语言结构与社会功能的深刻解析，计算科学贡献高效的数据处理与智能分析模型，而文化遗产保护学则须确保这一过程的伦理合规性与文化可持续性。当前，相关研究已在语料库建设、低资源语言技术及数字人文项目等方面取得进展，但往往侧重于单一维度，缺乏一个将理论、技术与人文实践有机整合的协同框架。本文正是在此背景下，试图构建一个以跨学科融合为基础、以智能技术为驱动、以生态化传承为目标的方言数字化体系综述，旨在系统梳理关键问题、整合可行路径，并为“方的言”等实践平台提供理论参照，从而探索在数字时代延续乡土声音与文化记忆的可行范式。

2 三维学科视野的交叉与融合

2.1 语言学的结构化描述与社会性阐释

方言数字化体系的构建首先需要语言学理论提供对语言本体的精确描述框架。结构语言学为方言的音系、词汇及语法系统确立了形式化的分析方法，这是将鲜活的语言现象转化为可计算数据模型的前提。例如，方言中特殊的声韵调特征、特有的词汇构式及句法规则，都需要转化为机器可识别与处理的规范化表征体系[2]。与此同时，社会语言学视角的引入至关重要，它揭示了语言变异与社会结构、身份认同及代际传承之间的复杂关联。方言并非均质的系统，其使用模式、演变动力深受社区网络、语言态度及社会变迁的影响。因此，数字化体系不仅需记录语言的静态结构，更应具备捕捉并呈现其动态社会生态的能力。

2.2 计算科学的适应性建模与技术挑战

面对方言作为“低资源语言”的客观现实，计算科学的核心任务在于发展适应性强的智能化处理方法。当前自然语言处理与语音技术主要建立在资源丰富的标准语数据基础上，直接

作者简介：何正锋（2002-），男，汉族，籍贯：广西全州，本科，单位：广西职业师范学院，研究方向：物联网工程。

邓倩妍（2005-），女，汉族，籍贯：江西赣州，本科，单位：广西职业师范学院，研究方向：物联网工程。

刘星星（2003-），女，汉族，籍贯：广西玉林，本科，单位：广西职业师范学院，研究方向：工商管理。

赵天欣（2003-），女，汉族，籍贯：广西玉林，本科，单位：广西职业师范学院，研究方向：物联网工程。

通讯作者：杨小梅（1981 年-），女，汉族，籍贯：山东青岛，博士研究生，单位：广西职业师范学院，研究方向：计算语言学，人工智能。

本项目由国家级大学生创新创业训练计划项目资助，项目名称：广西职业师范学院 2025 年大学生创新创业训练计划项目《方的言-基于人工智能的地方方言传承与双向翻译平台》，项目级别：国家级，项目类别：一般项目，项目编号：202514684013X。

迁移至方言场景面临显著性能落差。因此，针对低资源场景的机器学习范式——如小样本学习、跨语言迁移学习以及自监督学习——成为关键技术突破口。这些方法旨在利用普通话或其他相关语言的知识，通过模型架构或算法层面的创新，实现对特定方言的高效适配。在语音层面，需解决方言连续语音识别中声学模型鲁棒性、口语化发音变异建模等问题；在文本层面，则需应对方言书面语料稀缺、语法非标准化带来的分析困难。

2.3 文化遗产保护的伦理框架与价值导向

方言数字化本质上是一项文化遗产的抢救与传承实践，必须遵循文化遗产保护的基本伦理与价值导向。这要求将技术过程置于“以社区为中心”和“活态传承”的伦理框架之下[3]。首先，必须充分尊重语言持有者的主体权利，在数据采集、使用与传播的全过程中贯彻知情同意原则，并建立完善的个人隐私与文化产权保护机制。其次，数字化工作的目标超越单纯的语言数据存档，更在于辅助构建与延续社群的文化记忆与身份认同。因此，技术系统的设计需具备文化敏感性，避免因技术简化或商业化应用导致文化意义的曲解或剥离。最后，可持续性原则要求数字化项目能与地方社区的发展需求相结合，不仅“取之于社群”，更应“用之于社群”，通过数字赋能激发社区内部的文化传承活力，形成外源性保护与内源性发展的良性互动。

2.4 学科融合：构建“数据-知识-文化”复合模型

上述三个维度的理论并非彼此独立，而是共同指向对方言这一复杂对象的整体性理解。语言学提供了被数字化“描述什么”的内容框架，计算科学解决了“如何数字化”的方法工具，而文化遗产保护学则确立了“为何数字化”的价值规范。三者的深度融合，旨在构建一个将方言视为“数据-知识-文化”三维复合体的整合模型。在此模型中，经过结构化标注的方言数据是基础，通过计算模型从中挖掘出的语言规律与文化模式构成可复用的知识，而这些知识最终服务于文化价值的阐释、传承与发展这一根本目的。

3 技术架构：核心模块与系统性实现路径

3.1 多模态语料库：标准化建设与伦理化采集

构建方言数字化体系的首要基础，在于建设一个高质量、可扩展的多模态方言资源库。这要求超越传统的单一录音模式，系统性地采集包含高清音频、同步视频、场景图像及文本转写在内的多维度数据，以完整记录方言使用的真实文化语境。为实现数据的可互操作与长期可用性，必须建立一套标准化的采集协议与元数据规范。元数据设计需涵盖语言特征、发音人社会属性、采集时空信息及文化场景描述等多个维度[4]。尤为关键的是，整个过程需嵌入严格的伦理框架。必须在采集前获取知情同意，明确界定数据的使用范围、归属权利及开放共享等级，并建立有效的隐私保护机制，确保技术过程始终尊重并保护语言社群的主体权益。

3.2 人机协同标注：智能化工具与专家知识融合

原始语料的价值需要通过系统化的标注与分析才能充分释放。面对海量数据，纯粹依赖语言学专家手工标注效率低下且难以规模化。因此，一个高效的人机协同标注平台不可或缺。该平台应集成先进的预处理工具，例如基于深度学习的自动语音切分与音素识别、以及针对方言文本的初部分词与命名实体识别。这些自动化结果将导入一个设计友好的协同工作界面，供语言学家和经过培训的社区成员进行校对、修正与深化标注[5]。这种“机器初标、人工精校”的模式，不仅能大幅提升标注效率与一致性，更能将语言学的专家知识与社区成员的内隐文化知识相结合，在标注过程中逐步构建起一个结构化的方言知识网络。

3.3 智能处理引擎：低资源环境下的关键技术突破

智能处理是赋予方言数字化体系核心能力的关键。在低资源约束下，方言的语音识别与合成面临独特挑战。语音识别模型需通过迁移学习、多任务学习或方言自适应技术，提升其对不同口音和变异现象的鲁棒性。语音合成则需在保证自然度的基础上，精准捕捉并再现特定方言的韵律、语调等“乡土韵味”，先进的神经语音合成与风格迁移技术在此领域具有应用潜力。在机器翻译方面，由于平行语料极度稀缺，需探索利用单语语料的无监督或半监督训练方法，构建跨语言语义表示，并在解码过程中引入方言的语法约束。此外，构建融合语言学知识的方言计算模型，以及将词汇、俗语、文化概念进行关联的方言知识图谱，能为上层应用提供深度的语义支撑。

3.4 沉浸式传播界面：构建数字时代的文化体验场景

技术的最终目的是服务于人的体验与文化的传承。因此，设计沉浸式、交互式的文化传播界面至关重要。基于虚拟现实（VR）与增强现实（AR）技术，可以构建高度仿真的方言文化场景，用户通过自然语音与虚拟角色进行互动，在情境中习得语言。交互式数字方言地图则能整合地理信息、实地录音、民间故事与历史影像，让用户直观探索方言的空间分布及其背后的社会历史脉络。这些界面设计的目标，是超越工具性的语言学习，通过营造情感共鸣与文化认同，使方言在数字空间中转化为一种可感知、可参与、可传承的活态文化经验，从而激发，特别是年轻一代，主动传承的内在动力。

4 实践框架：协同创新与可持续生态的构建

4.1 多方协同的工作机制：从专业分工到价值共创

方言数字化体系的落地实施，有赖于构建一个贯通“产学

研用”的协同创新网络。该网络的核心在于明确界定并有机整合多元主体的角色与贡献：语言学家与方言研究者提供本体分析的学术规范与质量控制；计算机科学家负责核心算法的持续优化与技术平台的稳健运行；文化遗产保护机构与地方文化工作者确保项目符合文化伦理并扎根社区语境；而作为文化持有者的方言社群成员，则应超越单纯的“数据提供者”角色，成为全过程的关键参与者、成果的共同阐释者与首要受益者。这种深度协同机制要求建立常态化的跨领域对话平台，制定共同认可的工作协议与数据标准，并通过参与式设计方法，确保技术开发始终回应真实的文化需求与社会价值，最终形成专业知识、技术能力与地方性知识相互赋能的价值共创格局。

4.2 生态化运营：保障体系长效发展的多维策略

为确保项目的可持续性，必须建立一套兼顾公益属性与内在活力的生态化运营策略。在资源层面，倡导“有限开放”原则，在严格遵循伦理协议的前提下，分级开放脱敏后的语料数据与基础工具包，吸引全球学术界共同攻克技术难题，同时通过 API 服务等形式为合规的商业创新提供支持。在社群参与层面，设计数字徽章、文化贡献积分、线下工作坊等复合激励体系，将线上参与与线下社区文化活动相连接，培育稳定的核心贡献者与活跃用户群体。在长效机制层面，积极探索“公共文化服务采购+专项研究基金+公益基金会支持+可持续商业增值服务”的混合资金模式。推动将方言数字化成果纳入地方公共文化服务体系，同时探索与数字出版、沉浸式文旅、语言科技等产业结合的轻度商业化路径，形成社会效益与项目自我造血能力之间的良性平衡，最终推动整个体系从项目制运营向可持续社会创新生态的平稳演进。

5 总结

尽管方言数字化体系构建前景广阔，但其发展仍面临多维度的严峻挑战。总之，方言数字化是一项融合技术理性与人文价值的长期事业。唯有通过持续的跨学科协作、迭代性的技术优化以及深度的社区参与，方能构建一个既具备技术先进性、又富有文化生命力的可持续传承体系，真正实现数字化时代语言文化遗产的存续、活化与重生。

参考文献：

[1] 解天仪.数字赋能下天津方言的保护与传承[J].中原文学,2025,(20):72-74.

[2] 原伟,邓耀臣.中国传统语言学双语知识图谱的建设与应用[J].外语与外语教学,2024,(03):111-124+149.

[3] 李传欢,刘托.“非遗文化”误用对非遗术语体系的影响及其规范路径[J].中国科技术语,2025,27(06):27-32.

[4] 周若彤.语料库语言学对古文字“字料库”构建的启示[J].今古文创,2025,(22):124-127.

[5] 张绍阳,张子卓,柳永利,等.基于小样本学习的方言语音识别方法[J].江苏大学学报(自然科学版),2025,46(06):692-698.