文 献 与 数 据 学报
第 6 卷 第 2 期 2024 年 6 月

# 水书信息化建设研究进展与挑战 *

武 帅[1] 杨秀璋[2]
（1. 南京农业大学信息管理学院，南京 210003；
2. 武汉大学国家网络安全学院，武汉 430072）

**摘 要：**[**目的 / 意义**] 水族文字作为濒危文字的代表，在国家政策的帮扶下已基本完成对其数字化保护，但在信息化建设上仍与其他象形文字存在差距。针对水书与其他象形文字间的差距，发现水书信息化建设的不足，是促进水书研究向智慧化应用发展转变的关键。[**方法 / 过程**] 通过对水书文献信息化的田野调查和文献梳理，回溯我国水书文献信息化建设进程及取得的成果，探究水书文献信息化建设的难点。[**结果 / 结论**] 水书文献的字法研究最为扎实，基本满足民族语言信息化建设需求。但词法的研究较为匮乏，且模型准确率有待提升，原因在于缺乏高质量的水书文献语料库。

**关键词：**民族地方志 水书文献 信息化建设 民族文献
**分类号：**G275
**DOI：**10.31193/SSAP.J.ISSN.2096-6695.2024.02.13

## 0 引 言

水书是水族古文字及其著编典籍的汉译统称。作为世界上“活着”的象形文字，收集、整理、研究、解读水书对了解水族历史文化、研究少数民族原始宗教和社会信仰、窥探中国古文字奥秘等有着重要意义。由于水族古籍和各类文献资料并未得到有效梳理和信息共享，一定程度制约了水族文化研究的深度和广度，间接导致国内外学术界对于中国水学研究认识不够清晰。同时，水书的记录载体多为刺绣、碑刻、木刻、纸张等，且仅靠数量稀少的水书先生代代手抄传承，随时间推移，大量水书古籍文献资料将不断损毁和流失，成为永远不可再生的宝贵资源。因此，水书作为水族文化发展衍生出的智慧结晶和宝贵财富，梳理其数字化抢救的研究进程，探讨如何使其“活”起来，是历史所赋予的时代使命，也是传承和发扬水族文化的主要措施。

我国国家档案局于 2000 年正式开启“中国档案文献遗产工程”[1]，采用科学高效的方法对逐渐老化、破损、失传的濒危水族文献进行数字化抢救、保护。中央民族工作会议于 2014 年提

* 本文系贵州省科技计划项目“基于大数据及图像识别的水族文献及濒危水书抢救性整理研究”（项目编号：黔科合基础[2020]1Y279）的研究成果之一。
[作者简介] 武帅（ORCID：0000-0002-1162-4308），男，博士生，研究方向为数字人文、信息资源管理，Email：2022214003@stu.njau.edu.cn；杨秀璋（ORCID：0000-0001-9648-9506），男，讲师，博士，研究方向为数字人文、民族文献，Email：yangxiuzhang@whu.edu.cn（通讯作者）。

出“积极培育中华民族共同体意识”，2019 年再次提出“不断推进中华民族共同体建设”，可看出国家对少数民族文化的重视程度，如何较好保护与研究民间民族文化已然成为现阶段较为热门的研究主题[2]。水书于 2003 年列入首批档案文献遗产名录中，随后国家档案馆将其列为收藏的重点民族古籍名录[3]。国家民委组织编写的民族问题丛书《中国少数民族》将水族列为第一种。2021 年 5 月中共中央办公厅、国务院办公厅联合印发《“十四五”全国档案事业发展规划》，明确将“国家电子档案战略备份中心建设项目”作为重点部署项目[4]。国家重视程度的提升一定程度上提高了对水书信息数字化的质量要求。

## 1　水书研究现状

传承至今的水书主要流传于贵州南部水族地区，创制年代相对久远[5]。我国对水族研究最早的记载为 1860 年莫友芝在《红崖古刻歌》的序中记述水书。民国期间的方志整理也部分涉及水书相关信息。值得庆幸的是，虽然相比于壮语、维吾尔语，水族文字并未引起学者重点关注，且在传承发展过程中遭受过多次劫难，存在断代现象，但水书至今仍以活态的形式在水族人民日常生活中得以延续。

### 1.1　水书整理保存的局限性

由于产权意识和实用需要，众多散落民间的水书无法被收集进行专业修复和妥善保存，现藏于图书馆或档案馆的水书不到总量的十分之一。更令人担忧的是，全国目前仅有不足千名水书先生，且年龄多在 60 岁以上，部分水书先生甚至未找到继承者就已过世。传统水书收集整理和研究的方式在人力、物力和技术上有一定的局限性，对大量散落民间的水书进行系统采集整理和编撰难度巨大，已无法满足信息化时代对水书等濒危民族古籍进行抢救的新要求。

### 1.2　水书研究的四个阶段

作为目前世界上稀有、完整、活态地保存并运用至今的水族古籍，水书被专家学者誉为世界象形文字的“活化石”，并于 2006 年纳入首批国家级非物质文化遗产名录。本文系统梳理水书抢救保护历程（详见表 1），并将其按时间线划分为四个阶段。20 世纪 90 年代前（阶段一），侧重水族文化源流研究，探究基础理论来源；20 世纪 90 年代（阶段二），侧重水书古文字考释，并开始水书破译研究；21 世纪 00 年代（阶段三），从国家记忆层面进行濒危水书抢救，初步尝试水族文字信息化建设；21 世纪 10 年代至今（阶段四），水族古籍信息化建设步伐加快，珍贵古籍得以保存。

我国对水族文献和水书的整理研究始于 20 世纪 50 年代。党的十一届三中全会后，在党和各级政府的关心下，水书抢救保护工作全面展开，该阶段在水书源流考据、文字考释、研究概况及水族文化变迁等方面积累了大量学术研究成果。21 世纪以来，国内外学者开始高度重视对濒危水族古籍的抢救工作，发掘、收集、整理大批水书文献资料。这一时期除了传统的对水书自身的研究和挖掘，也开始思考水书抢救工作中存在的问题和困难，探索出多种解决方案。随着计算机技术的发展和普及，水书研究正式进入信息化时代，以计算机辅助水书输入、整理、研究的方法走入学术视野，为后期进行水族知识图谱构建和水书数字化采集研究、构建水族本体并建立水书电子数据库提供了大量学术依据和理论基础。

**表 1　水书抢救保护历程大事表**

| 阶段 | 时间 | 事件 |
|---|---|---|
| 阶段一 | 1949 | 岑家梧定义了水书的种类、用途、内容、结构及来源，并指出水书抢救难度大的原因是除水家鬼师外多未认识水书，拉开抢救水书的序幕。 |
| | 1957 | 国务院设立三都水族自治县，族称定为“水族”。 |
| | 1963 | 中国科学院民族研究所探究水族迁徙原因及水语与其他语言关系。 |
| | 1980 | 三都水族自治县民族文史研究组探究水族历史和水书源流。 |
| | 1981 | 潘一志汇集正史、方志、碑碣、档案文书、民间传说及私人著作，探究水书与水族社会源流。 |
| | 1987 | 潘朝霖探究水族族源、迁徙原因，对水书与水族社会源流进一步考察。 |
| | | 王国宇就水书性质，将其分为：读本、阅览本、阴阳本、时象本、方位本、星宿本。 |
| | 1989 | 雷广正揭示了水族宗教的仪式、礼规和巫术等演变为人的意识形态观念，并解释自然现象的崇拜文化、人和物种来源的图腾文化，反映水族祖先崇拜和神鬼观念的巫术文化。 |
| 阶段二 | 1990 | 贵州省水家学会成立，奠定了水书研究和破译工作的开端。 |
| | | 雷广正归纳总结了水书基本情况、作用、字形结构特征。 |
| | 1991 | 王品魁对水书起源、用途、流传等方面进行探究。 |
| | | 陈昌槐列举 444 个水族与汉族文字对照表，公开首份《水族文字汉译一览表》。 |
| | | 国家民族事务委员会发布《关于进一步做好少数民族语言文字工作的报告》，明确指明民族语言应增强相关基础理论、应用理论、文字信息处理的科学研究。 |
| | 1992 | 韦忠仕系统梳理水书阶段研究成果，标志着民间水书整理及翻译工作正式展开。 |
| | 1993 | 王国宇列举水字与汉义、译音对照表，发布首份公开水族文字译音一览表。 |
| | | 石尚昭探究水族文字的性质、结构及读音，将水族文字分为：象形字、表意字和谐音字。 |
| | 1994 | 王品魁首次破译出版的《水书·正七卷壬辰卷》，为水书研究及抢救做出了开拓性的工作。 |
| | 1999 | 韦宗林发现水族文字反写现象是由文字本身、民族压迫、社会文化造成。 |
| 阶段三 | 2000 | 韦宗林创制水族古文字计算机输入法，并公开《水文与汉字对照表》《水文输入法基本字根总表》和《水族文字编码》。 |
| | 2002 | 石冬梅界定了水书的来源、分类、日常应用及历法。 |
| | | 三都县档案局从民间收集 180 多卷水书。 |
| | | 韦宗林对水书释义、字形、字音研究，发现水书是先秦时期从古文字分化出来的一种古字，并破译大量水书古籍。 |
| | | 3 月，荔波县获批首批“中国档案文献遗产名录”。 |
| | | 7 月，荔波县全面开展大规模水书抢救工作，征集民间各类水书数千册。 |
| | 2004 | 国家语言文字工作委员会发布《关于进一步做好语言文字信息化工作的若干意见》，强调做好濒危少数民族文字数字化整理和记录及保存工作，重视少数民族文字信息化建设，构筑民族语言文化高地。 |

续表

| 阶段 | 时间 | 事件 |
|---|---|---|
| 阶段三 | 2005 | 组建“水书文化研究所”，甄别、查重已有水书。 |
| | | 制定《水书文字字符总集》《水书文字形体规范标准》。 |
| | | 搭建水族文字信息化平台、字符总集字库。 |
| | | 初步建成水书文字字库、水书文字资料库系列。 |
| | | 为 444 个水文部件编码，设计 Win 系统的水书文字输入法。 |
| | | 韦宗林从时间与地域、书写字体形态、书写文化三个维度论证水族古文字是古代神本文化的活化石。 |
| | | 罗春寒系统梳理濒危水书在抢救环节所面临的主客观问题，确保水书抢救工作达到预期目标。 |
| | | 国务院首批国家级非物质文化遗产名录，将水书定名为“水书习俗”。 |
| | 2008 | 国务院将六册水书文献典籍档案列入中国第一批国家珍贵古籍名录（万年经镜、六十龙备要、吉星、庚甲、渺金、金银）。 |
| | | 欧阳大霖认为水书先生是“水书习俗”传承的桥梁和关键。 |
| | | 杨建玲首次从档案视角对水书研究，界定水书的“档案文献”与“文献档案”的概念区别。 |
| | 2009 | 邓章应界定了水文、水字、水书三者间的区别。 |
| 阶段四 | 2010 | 潘朝霖认为水书消亡的原因是社会变革、政策失误、外来文化、自身缺陷、功能弱化等。 |
| | 2011 | 戴建国指出水书中仪式等非文字形式在社会记忆传承中的重要性。 |
| | | 戴丹使用哈希表和哈希函数加快水书水字可视化输入匹配速度。 |
| | 2014 | 英国学者康蔼德和中国学者潘兴文联合开展水语调查研究。 |
| | 2016 | 三都水族自治县水书文化研究院成立首个“水书习俗少儿弟子班”招收 8 人（女 4 人），打破“传男不传女”戒律，降低水书习俗传承门槛，对水书保护与传承有积极的借鉴作用，缓解传承断层现象。 |
| | 2018 | 杨胜昭被聘为贵州民族大学特聘研究员，是中国首位被高校特聘的水书先生。 |
| | 2019 | 三都水族自治县水书文化研究院编纂的《中国水书》共收入 1453 种水书，共计 160 卷本，其中 1–100 卷为三都县、101–150 卷为荔波县、151–160 卷为潘藏卷。 |
| | 2020 | 国务院办公厅发布《关于全面加强新时代语言文字工作的意见》，提出现阶段迫切需要提升民族文字的信息化处理能力，并尝试推荐相关融媒体应用。 |
| | 2022 | 贵州省水书文献成功入选《世界记忆亚太地区名录》，是世界上仅存为数不多仍在使用的象形文字之一。 |
| | | 国务院将 79 册水书文献典籍档案列入中国国家珍贵古籍名录。 |

## 2　水族文字信息化研究进展

自 20 世纪 80 年代起，计算机技术已然成为档案文献保护的核心技术。然而，传统民族文字除拉丁字母外均未解决文字输入法实现的问题，一定程度上阻碍了少数民族研究的发展。

为了考察水族文字信息化研究进程，本文选取我国具有代表性的三种象形文字甲骨文、藏

文、东巴文，以甲骨档案（世界记忆国际名录 [2017]）、藏文典籍（国家珍贵古籍名录 [2020]）和纳西东巴古籍（世界记忆国际名录 [2003]）为考察依据，探究这三种象形文字研究所取得的阶段性成果，与水族文字（以贵州水书文献为依据）的信息化研究进程进行对比，分别从文字研究、文字数字化以及数字人文视角进行梳理，详见表 2。

**表 2　四种象形文字典籍信息化研究进程**

| 名称 | 甲骨档案 | 藏文典籍 | 纳西东巴古籍 | 贵州水书文献 |
|---|---|---|---|---|
| 文字 | 甲骨文 | 藏文 | 东巴文 | 水文 |
| 遗产项目 | 世界记忆国际名录［2017］ | 国家珍贵古籍名录［2020］ | 世界记忆国际名录［2003］ | 世界记忆亚太地区名录［2022］ |
| 文字研究 | 将甲骨文按书体风格划分为宾组、出组、何组、历组和黄组 | 按字形结构将藏文分为：位点（像素）、笔划、部件（构件）、单字四级 | 东巴文是目前世界上少有的还流行在民间的象形文字，尚未脱离图画阶段 | 水书基本情况、作用、字形结构特征 |
| | 卜辞包含前辞、命辞、占辞、验辞四个部分 | 《出行择日吉凶法》考释弥补敦煌汉文文献的缺陷和不足 | 东巴字尚处于繁杂的图画文字阶段，音节多，存在“几读并存”现象 | 列举水族文字汉译一览表 |
| | 最早的汉语双宾语句在花东卜辞中 | 以藏语句中动词对藏文句型进行分类 | 制作东巴文异体字总表《纳西象形文字谱》 | 列举水族文字译音一览表 |
| | “卜蕴诗心”标志中国文学诗性智慧的开源 | 将藏文特殊文化词按奈达分类法分为：生态、物质、社会、宗教、语言五类 | 制作《意符构字频度表》和《声符构字频度表》 | 探究水族文字的性质、结构及读音 |
| | 从古代社会、文献互证、跨学科综合、字形、辞例、汉字演变对甲骨文考释 | 构建藏文寨文中藏汉对音词汇表 | 从构件数量、结构、位置、功能、读音方面探究东巴文合文 | 首次破译出版水书《水书·正七卷壬辰卷》 |
| 数字化 | 消除甲骨文采集过程中的图像失真现象 | 采用深度学习模型对藏文手写数字识别 | 构建东巴文特征提取和识别方法 | 水文与汉文对照表、水文编码、输入法基本字根表 |
| | 实现未识别甲骨文字的输入 | 结合藏文文字与古籍版式特征合成藏文古籍模拟数据 | 统计东巴文字形的块数、孔数、端点数、三叉和四叉点的基本拓扑特征 | 初步建成水书文字字库、水书文字资料库系列 |
| | 利用有向笔段优化未识别甲骨文字的输入效果 | 构建藏文中“丁”字印刷体数据集并对其进行精确标注 | 实现不同情况下东巴文识别 | 以《水书常用字典》收录水族文字的特征，设计水族文字的编码规则 |
| | 以矢量描述的方式构建甲骨文字形描述库 | 以音节为单元对藏文图像文本的分割 | 实现多字数的东巴象形文字的高精度识别 | 提取水族文字轮廓，实现水族文字识别 |
| | 运用仿射变换自动生成甲骨文 | 实现多种字体的手写藏文的跨库文本检测 | 实现东巴象形文字文档图像的自动分割算法 | 实现复杂环境下的水族文字提取 |

续表

| 名称 | 甲骨档案 | 藏文典籍 | 纳西东巴古籍 | 贵州水书文献 |
|---|---|---|---|---|
| 数字人文 | 甲骨文字的“取象”之美是中国美学和艺术的真正源头 | 对藏文微博信息的情感分类 | 设计东巴文的元数据规范，并构建东巴经典古籍编目管理系统 | 设计水书文字识别原型系统 |
| | 集甲骨文字形库、著录库、文献库、文献资源数字化平台——殷契文渊 | 对藏文新闻文本的主客观句子分类 | 对东巴文分词处理 | 结合水书档案文献特征构建水书本体模型 |
| | 以人类学角度印证“■”为触发词，追溯摔牛运动至晚商社会 | 以图采样的方式对藏文的实体关系抽取 | 设计东巴画深层特征提取模块 | 构建大规模水书手写文字数据集 |
| | 按照等级将商代的简册文书分为“册”“典” | 结合句法结构对藏文进行高质量分句处理 | 实现东巴图像的情感分类 | 对水族古文字的高质量数字化提取 |

梳理可知，文字研究视角：甲骨档案已发展至对文章结构、特殊词法以及不同语言对比的探究论证阶段；藏文典籍已发展至对字形结构、词性分类、文献资料融合的探索考证阶段；纳西东巴文献已发展至对字形、字义、结构、字音等的初级研究阶段；贵州省水书文献已发展至水文汉译、水书释义、文字分类、字形字音归纳的入门研究阶段。

数字化视角：甲骨档案已发展至具有精准、系统、全面的甲骨文字图像数据库，实现未识别甲骨文字的计算机自动生成的智能数据（Intelligent Data）阶段；藏文典籍已发展至运用数据增强的方法构建特殊类型藏族文字图像数据库，通过标注实现对特殊手写藏文的精准识别的关联数据（Linked Data）阶段；纳西东巴文献已发展至通过特征提取及融合实现对单个东巴文字的精准提取，并能实现对篇幅级图像文本行的自动分割的语义数据（Semantic Data）阶段；贵州省水书文献发展至水书文字输入法构建，运用图像增强技术能够实现对复杂环境下的水书古籍文字提取的原生数据（Original Data）阶段。

数字人文视角：甲骨档案已实现从不同学科视角对典籍的潜在历史意义、文化价值进行更深层次探究，开展甲骨文文创开发、通过展览演绎实现宣传推广的传承性保护阶段；藏文典籍已实现对已数字化的藏文典籍结合深度学习模型对其进行实体关系抽取，结合藏文句法结构扩充藏文典籍语料规模的智慧性保护阶段；纳西东巴古籍已实现运用深度学习模型对东巴画进行情感特征提取、构建东巴古籍编码管理系统实现对古籍资料的存储和检索、尝试运用词向量转换的方式增强东巴文本的表示能力的资源储备、数字人文计算的再生性保护阶段；贵州省水书文献则还停留在运用图像识别技术提升水族文字的数字化提取，初步尝试结合水书档案文献特征构建水书本体模型的原生性保护阶段。

从上述视角对比四种象形文字典籍的信息化研究进程，初步发现贵州水书文献缺乏较为全面、准确、公开的图像文字数据集，整体研究进程滞后于其他三类象形文字典籍。贵州水书文献被纳入世界记忆亚太地区名录，一定程度上引起了学者的研究重视。因此，构建便于学者研究的水族文字图像数据集和贵州水书文献数字化档案是提升贵州省水书文献研究的首要任务。结合上述四种象形文字的研究发展情况，归纳四种象形文字信息化建设情况对比，见表 3。

**表 3　四种象形文字典籍信息化建设情况**

| | 甲骨档案 | 藏文典籍 | 纳西东巴古籍 | 贵州水书文献 |
|---|---|---|---|---|
| 数据类型 | 智能数据 | 关联数据 | 语义数据 | 原生数据 |
| 数据态 | 智能态 | 关联态 | 语义态 | 原生态 |
| 实现功能 | 语义推理<br>逻辑校验<br>知识图谱构建及补齐<br>…… | 关联数据挖掘<br>知识重组及融合<br>实体及本体对齐<br>相似度计算<br>…… | 手写体识别<br>本体及语义网构建<br>实体识别及关系抽取<br>自然语言处理<br>…… | 机器扫描<br>图像分割及文字提取<br>著录<br>…… |
| 功能效果 | 实现典籍隐性知识的自动生产 | 实现典籍显性知识的多模态融合 | 实现典籍显性知识的语义重组及编码构建 | 实现典籍元数据及内容的呈现 |
| 阶段特征 | 可推理<br>可自动<br>可追溯 | 易开源<br>广关联<br>强交互 | 自描述<br>可解释<br>可机读 | 易传播<br>可利用<br>便保存 |
| 所处的保护阶段 | 传承性保护 | 智慧性保护 | 再生性保护 | 原生性保护 |
| 阶段保护内容 | 宣传推广<br>展览演绎<br>文创开发 | 元数据及本体数据构建<br>语义数据及关联数据构建 | 全文本数字化<br>专题数据库构建<br>数字化图像 | 典籍修复及加固<br>典籍影印及出版<br>缩微复制及收藏 |

## 3　水族文字的字法处理研究

由于水族古文字发展严重滞后于其他文字的发展，使得水书并不能像其他地方志一样通过抄本的形式进行研究。因此，水族文字字法处理是实现水书自然语言处理的前提条件，是保障水书语义挖掘的重要前提。由于水族文字尚未形成规范的字符编码，方法主要分为基于光电扫描的电子输入和基于字符识别的文字输入。

### 3.1　基于光电扫描的电子输入

基于光电扫描的电子输入是指通过对水书光电扫描实现电子化处理，并通过文字识别技术实现水书字符编译，根据文本识别方式分为显式切分和隐式切分。

显式切分是指通过单字分类器构建“候选切分－识别”网络路径评价和最优路径搜索，实现对水书单体字符的形状特征的获取。现阶段，常用切分技术的精准率均达到 97% 以上[6]，为进一步提升切分精准性，在传统切分技术的基础上引入置信度转化机制[7]进一步提升模型字符效果，Wang 等[8]对该方法的设计原理进行了详细的阐述。Kimura 等[9]对切分块提取几何特征之后，使用修正后的二次判别函数（Modified Quadratic Discriminant Function，简称 MQDF）的非线性分类器实现对字符的分类。但由于 MQDF 模型复杂度相对较高，后期 Liu 等[10]在保证 MQDF 模型准确性的基础上结合最近邻分类算法（Nearest Prototype Classifier，简称 NPC）提升了模型分类速度。但由于切分问题一直没有得到较好的解决，一定程度上制约了显式切分文本分类的发展。随着深度学习技术的快速发展，逐步替代机器学习算法，单字识别和单句识别效果均得到不错的提升。

隐式切分是指通过事先设置候选字符实现对文本的识别，运用滑动窗口实现文本与序列之间的解码获得字符串的识别结果。Wang 等[11]通过使用滑动窗口实现文字编码，结合隐马尔可夫模型（Hidden Markov Model，简称 HMM）模型的先验概率对解码文字进行解码，最后通过维特比算法（Viterbi）筛选出后验概率最佳字符串。隐式切分的字符识别的训练过程不需要进行字符级的标注，大量节省标注时间。

### 3.2 基于字符识别的文字输入

光学字符识别（Optical Character Recognition, 简称 OCR）技术[12]的出现，使水族文献信息化建设更为高效、便捷，提升了古籍文字[13]、镌刻文字[14]、自然环境[15]等不同载体下民族文字字符特征提取的效果。随着 TH-OCR 2007[16]民族文字识别系统的实现，推动水族文字识别进入实用。但由于水族文字多存在于刺绣、碑刻、木刻等载体上，多数民间水书未能实现数字化。王晓娟等[17]提出一种基于 BP 神经网络的图像识别方法，对水书中的手写体图像部分进行归一化处理。杨秀璋等[18]提出一种基于自适应图像增强和区域检测的水族文字提取与分割算法，提升水书文字提取效果。丁琼[19]率先尝试卷积神经网络（Convolutional Neural Networks, 简称 CNN）算法实现对水书字符识别，为实现水书字符识别系统打下基础。汤敏丽等[20]尝试运用 Faster-RCNN 算法实现对页面级古籍水族文字识别，为规模化数字识别奠定基础。杨秀璋等[21]提出一种改进卷积神经网络的古文字图像识别方法，解决字形变化对水书识别效果的影响。

总体而言，现阶段水族文字字法处理研究主要集中于文字提取和水书文献的数字化，并未形成水书文献的数据库构建。

## 4 水族文字的词法分析研究

词法分析指开展以词为单元的文本信息处理，是语言智能化研究的重要标志。由于水书文献的词语间有着显性标识，其词法分析多为通用词的词性标注（Part of Speech Tagging）和特殊词的命名实体识别（Named Entity Recognition，简称 NER）研究。

### 4.1 水书的词性标注研究

词性标注旨在确认单位词的词性，但水族文字的自然语言处理基础较为薄弱，且尚未形成规范化的标注语料。因此，需在少量语料的前提下进行词性标注。杨蓓[22]利用少量语料训练 HMM 模型，结合 Viterbi 解码实现词性标注。王兴金等[23]在 HMM 模型基础上融合水族语言规则，进一步提升词性标注效果。考虑到稀疏词的存在，王兴金等[24]运用深度学习方法学习语言的构词特征，进行词性标注，但存在并行能力差、长远信息丢失和特征提取不充分等问题。

### 4.2 水书的命名实体识别研究

命名实体识别是自然语言处理中较为特殊的词性标注，旨在对目标语料中的专有名词进行标注，便于后期领域知识图谱的构建，是水书信息化建设向智能化应用转变的关键。本文按照模型的分类细粒度将水书 NER 模型分为基于字符、词汇和字符 – 词汇相结合的模型，进行相关研究梳理。

（1）基于词汇的命名实体识别模型

基于词汇的水书命名实体识别模型原理如图 1 所示，先通过《水族文字汉译一览表》将水书译为汉文后，再对中文分词处理后的词汇进行命名实体识别。该类方法通过改进中文分词效果，提升水书命名实体识别效果。Collobert 和 Weston[25] 以单词嵌入的方法代替传统手工制作的特征，运用 CNN 提取水书特征并结合条件随机场（Conditional Random Field，CRF）模型，实现对实体类别预测。Ma 等[26] 融合双向长短期记忆网络（Bi-directional Long Short-Term Memory，简称 Bi-LSTM）和 CNN，构建 Bi-LSTM-CNN 模型来提升语义特征提取效果。Chen 等[27] 以改进分词效果的方式，提升 Bi-LSTM-CRF 模型对实体词边界的识别效果。

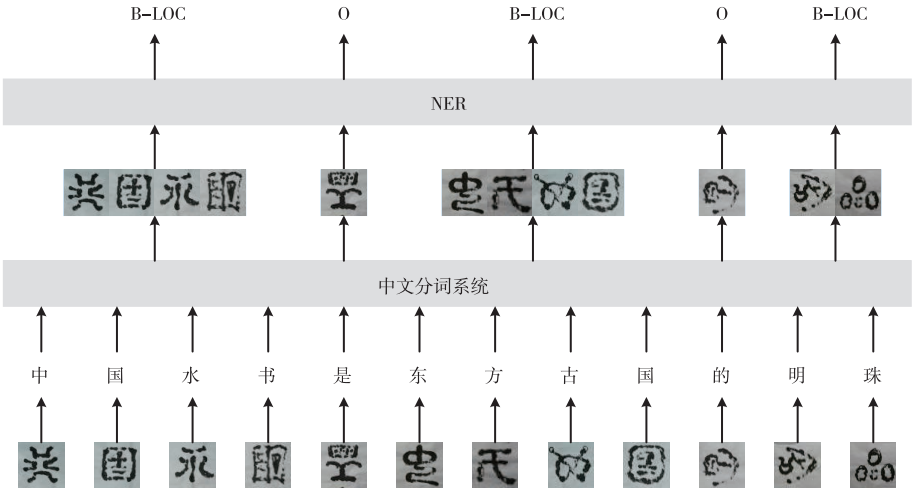


**图 1　基于词汇的水书 NER 模型原理展示图**

（2）基于字符的命名实体识别模型

基于字符的水书命名实体识别模型原理如图 2 所示，先按照《水族文字汉译一览表》将水书译为汉文，再对汉字字符进行命名实体识别，有效规避实体边界难以识别所造成的错误传播问题。Dong 等[28] 作为国内首支运用部首集实现基于字符的命名实体识别团队，开拓水书命名实体识别研究的途径。Zhu 和 Wang[29] 考虑字符间的依赖性，尝试使用双向门控循环网络（Bidirectional Gated Recurrent Unit，简称 Bi-GRU）实现句子级全局信息获取，并融入 Att 机制实现局部特征加权，初步实现语义融合。Gu 等[30] 创新性尝试两个规律性模块捕获字符的规则特征，构建正交空间融合所捕获的特征，实现局部语义增强。

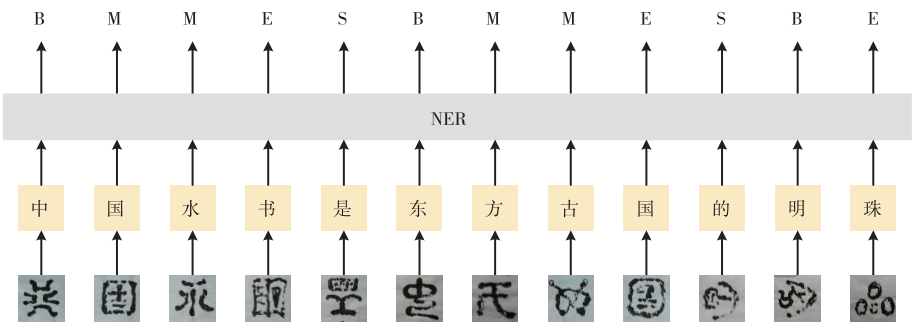


**图 2　基于字符的 NER 模型原理展示图**

（3）基于字符—词汇的命名实体识别模型

基于字符—词汇的命名实体识别模型是在基于字符的命名实体识别模型基础上，融入词汇的语义理解，以提升向量的语义表达能力。Ghaddar 等[31]率先提出一种基于格的长短记忆网络（Long Short–Term Memory，简称 LSTM）模型，通过字典匹配的方式，在 LSTM 模型基础上实现字符的语义增强，具体模型原理见图 3 所示。Sui 等[32]构建词—字符的包含图、转换图、字符格图的协作图网络模型，有效解决格结构的信息损失问题。Gui 等[33]在图神经网络的基础上融入词典特征和 Att 机制，缓解词汇冲突，解决语言歧义问题。

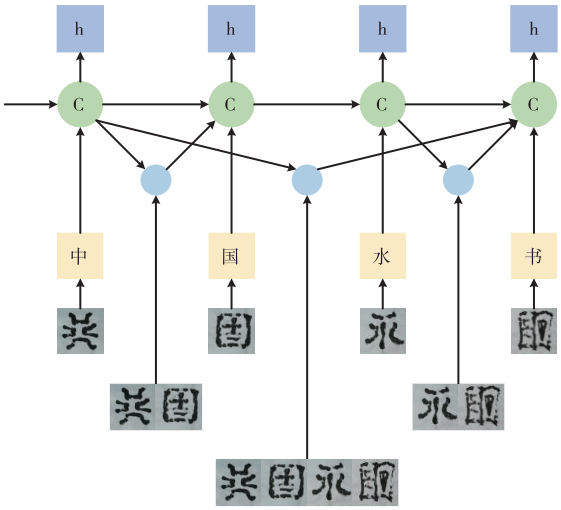


**图 3　Lattice LSTM 模型结构**

随着人工智能技术的进一步发展，Transformer 框架进入历史舞台，推动着图像识别和自然语言处理的发展。Xue 等[34]在 Transformer 框架基础上，增设相对位置编码的注意机制，提升字符与词之间的依赖关系。Li 等[35]在 Lattice LSTM 框架的基础上融入 Transformer 框架，实现字符与潜在词汇的信息交互，高效解决词汇冲突问题。

Transformer 发展越发成熟，以预训练模型为主在命名实体识别任务中的效果远优于基于词汇 – 字符等的静态嵌入模型的效果。Liu 等[36]使用特殊编码标记识别句中词的边界后，将其输入预训练模型进行编码，有效改善识别效果。

## 5　水书信息化建设的挑战及建议

通过文献梳理发现，现阶段水族文字的信息化建设远落后于其他象形文字的信息化建设，其建设核心仍停留在数字化存储层面，在信息化建设及智能化应用方面仍存在诸多挑战。

### 5.1　水书信息化建设所面临的挑战

参照已取得的阶段性成果，笔者认为水书信息化建设及智能化应用研究主要面临如下挑战。

（1）严重匮乏的语料库

水书语料库构建涉及语料的选取、收集、加工和分析。我国水书语料库的建设始于 20 世纪

90 年代，但受限于当时匮乏的基础理论研究和低效的档案数字化水平，语料库建设的规模和质量受到严重影响。且水书尚未设立健全且规范的语料标注方案，仅能依靠水书先生完成标注，严重制约水书信息化建设向智能化应用发展的进程。

（2）非正式文本占主体

水书的典籍材料相对较少，多为日常生活的记录，官方正式性文件更为稀缺，语料数据的类型分布不均，不利于模型的准确识别。且水书中占比最大的是水书先生实时记录内容和部分手抄本，该类数据存在复杂多样、文本语义混乱的特点，增加了文本信息化处理的难度。

（3）水书载体损坏严重

由于水族文字的特殊性，水书仅由为数不多的 60 岁以上的水书先生手抄得以传承。除水书先生手抄版外，记录水书的载体多为典籍、刺绣、碑刻、木雕、金雕等原生材质，随着时间推移，氧化腐蚀严重，严重阻碍水书的数字化提取。

### 5.2 对开展水书信息化建设的建议

针对上述挑战，笔者认为可从如下五个方面开展水书信息化建设。

（1）水书的自动化标注

水书的自动化标注是水书信息化建设向智能化应用发展的基础，能有效缓解研究者数据标注的压力。但预训练模型的训练数据并未包含水书文献数据，增加预训练模型中的水书标注数据，实现水书的自动化标注，是构建水书数据库迫切需要解决的问题。

（2）水书的细粒度挖掘

提升文本语料的细粒度，是利用文本挖掘技术准确获取知识的关键。由于水书包含天文、婚嫁、祭祀、地理等十类细粒度的实体类型，不同实体类型间专业知识的差异较大，需进行深层次的细粒度划分，才能为智能化应用提供较好的数据基础。

（3）构建水书的特征表示方式

作为象形文字的水族文字，其字形存在偏旁部首，字音存在元音辅音，这些特征有利于较好实现水书实体词边界的划分，因此，可以结合水族文字特征构建符合其构词规则的特征表示方式，达成符合其独特领域的自然语言处理任务。

（4）构建规范化的标注体系

迁移学习模型能实现同类型不同语言、同语言不同类型的自然语言处理研究，已在中文自然语言处理研究中的部分数据集中得以实现，这为水书技术研究提供了理论基础。但跨语言迁移学习存在标签匹配性和域间差异性的问题，因此，将迁移学习模型运用于水书领域，需尽早构建规范化的标注体系。

（5）预训练模型的轻量化

Transformer 架构已在大规模语料中得到较好的应用，预训练模型也较好地实现了跨领域、跨语言的迁移学习。然而，水书由于与预训练模型的语言特征存在极大差异，在微调过程中需要花费大量的算力，才能实现对文本语义的迭代学习，这就造成一定的资源浪费。如何在模型识别精度和轻量化之间达成平衡，是水书智能化应用的必经之路。

## 6 结 语

通过文献梳理发现，水族文字在国家政策的帮扶下，已实现基于统一编码的网络传输，这为水族文字的文本信息化和濒危水书的数字化保护、传承奠定了基础，但仍与其他象形文字研究存在较大差距。目前，我国水书的“字法”研究最为扎实，基本满足民族语言信息化建设需求。但“词法”研究较为匮乏，其主要原因是严重匮乏的语料库、非正式文本占主体和水书载体损害严重。考虑到水书信息化建设的市场份额小，无法获得业界青睐，同时，由于缺乏规范统一的信息化建设标准，学界在收集、整理和分析濒危水书时，本体构建水平参差不齐，难以提升下游文本挖掘任务的准确性，一定程度上阻碍了水书智能化应用的发展。

笔者认为，如何将无监督、多任务、小样本、零样本学习的低资源信息处理技术运用到水书信息化建设，以及后续的智能化应用，是攻克水书信息化建设难题的关键。

## 【参考文献】

［1］刘晓程，吴丽燕. 中华民族共同体视阈下的民族政策传播：内涵、功能与框架［J］. 新闻春秋，2022（5）：75–82.

［2］王玉英，张志杰，李念峰. 大数据时代少数民族传统文化信息处理模型构建与对策研究［J］. 情报科学，2022，40（7）：154–160，168.

［3］瞿智琳. 水书档案编纂现状探析［J］. 兰台世界，2016（1）：25–27.

［4］李明华. 关于建立国家电子档案战略备份中心的提案［J］. 中国档案，2022（3）：20.

［5］蒙耀远. 水族水书抢救保护十年工作回顾与思考［J］. 文史博览（理论），2016（1）：23–26.

［6］Yang W, Jin L, Tao D, et al. Drop Sample: A new training method to enhance deep convolutional neural networks for large–scale unconstrained handwritten Chinese character recognition［J］. Pattern Recognition, 2016, 58: 190–203.

［7］Liu C L. Classifier combination based on confidence transformation［J］. Pattern Recognition, 2005, 38(1): 11–28.

［8］Wang D H, Liu C L. Learning confidence transformation for handwritten Chinese text recognition［J］. International Journal on Document Analysis and Recognition, 2014, 17(3): 205–219.

［9］Kimura F, Takashina K, Tsuruoka S, et al. Modified quadratic discriminant functions and the application to Chinese character recognition［J］. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1987 (1): 149–153.

［10］Liu C L, Nakagawa M. Evaluation of prototype learning algorithms for nearest–neighbor classifier in application to handwritten character recognition［J］. Pattern Recognition, 2001, 34(3): 601–615.

［11］Wang Z R, Du J, Wang J M. Writer–aware CNN for parsimonious HMM–based offline handwritten Chinese text recognition［J］. Pattern Recognition, 2020, 100: 107102.

［12］Mori S, Nishida H, Yamada H. Optical character recognition［M］. John Wiley & Sons, Inc., 1999.

［13］薛春寒，金小峰. 基于迁移学习的少样本朝鲜语古籍文字的识别方法［J］.延边大学学报（自然科学版），2021，47（4）：350–355.

［14］仁青东主. 基于深度学习的藏文古籍木刻本文字识别研究［D］. 拉萨：西藏大学，2021.

［15］洪松，高定国，三排才让，等. 自然场景下乌金体藏文的检测与识别［J］.计算机系统应用，2021，30（12）：332–338.

［16］钱丽花. 统一平台的多民族文字文档识别系统研制成功［N］. 中国民族报，2007–01–30（001）.

［17］王晓娟，白艳萍. 基于BP神经网络的手写体数字的识别方法研究［J］. 数学的实践与认识，2014，44

（7）：112–116.

［18］杨秀璋，武帅，夏换，等. 基于自适应图像增强技术的水族文字提取与识别研究［J］. 计算机科学，2021，48（S1）：74–79.

［19］丁琼. Matlab平台下水书文字特征提取与分类方法实现研究［J］. 电子技术与软件工程，2020（14）：155–157.

［20］汤敏丽，谢少敏，刘向荣. 基于Faster–RCNN的水书古籍手写文字的检测与识别［J］. 厦门大学学报（自然科学版），2022，61（2）：272–277.

［21］杨秀璋，施奕，李娜，等. 一种改进卷积神经网络的阿拉伯文字图像识别方法［J］. 信息技术与信息化，2021（9）：6–11.

［22］杨蓓. 老挝语分词和词性标注方法研究［D］. 昆明：昆明理工大学，2016.

［23］王兴金，周兰江，张金鹏，等. 融合词预测的半监督老挝语词性标注研究［J］. 小型微型计算机系统，2019，40（12）：2500–2505.

［24］王兴金，周兰江，张建安，等. 融合词结构特征的多任务老挝语词性标注方法［J］. 中文信息学报，2019，33（11）：39–45.

［25］Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning［C］//Proceedings of the 25th International Conference on Machine Learning. 2008: 160–167.

［26］Ma X, Hovy E. End–to–end Sequence Labeling via Bi–directional LSTM–CNNs–CRF［C］//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016: 1064–1074.

［27］Chen X, Qiu X, Zhu C, et al. Long short–term memory neural networks for Chinese word segmentation［C］// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 1197–1206.

［28］Dong C, Zhang J, Zong C, et al. Character–based LSTM–CRF with radical–level features for Chinese named entity recognition［M］//Natural Language Understanding and Intelligent Applications. Springer, Cham, 2016: 239–250.

［29］Zhu Y, Wang G. CAN–NER: Convolutional attention network for Chinese named entity recognition［C］// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 3384–3393.

［30］Gu Y, Qu X, Wang Z, et al. Delving deep into regularity: A simple but effective method for Chinese named entity recognition［C］// Proceedings of the NAACL–HLT (Findings). Seattle: Association for Computational Linguistics, 2022: 1863–1873.

［31］Ghaddar A, Langlais P, Rashid A, et al. Context–aware adversarial training for name regularity bias in named entity recognition［J］. Transactions of the Association for Computational Linguistics, 2021, 9: 586–604.

［32］Sui D, Chen Y, Liu K, et al. Leverage lexical knowledge for Chinese named entity recognition via collaborative graph network［C］//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP–IJCNLP). 2019: 3830–3840.

［33］Gui T, Ma R, Zhang Q, et al. CNN–based Chinese NER with lexicon rethinking［C］// Proceedings of the Twenty–Eighth International Joint Conference on Artificial Intelligence. Macao: IJCAI, 2019: 4982–4988.

［34］Xue M, Yu B, Liu T, et al. Porous lattice transformer encoder for Chinese NER［C］//Proceedings of the 28th International Conference on Computation Linguistics. Barcelona: International Committee on Computational Linguistics. 2020: 3831–3841.

［35］Li X, Yan H, Qiu X, et al. FLAT: Chinese NER using flat–lattice transformer［C］//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics. 2020: 6836–6842.

［36］Liu W, Fu X, Zhang Y, et al. Lexicon enhanced Chinese sequence labeling using BERT adapter［C］//

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021: 5847–5858.

# Research Progress and Challenges of Informatization Construction of the Literature of Shui Nationality

Wu Shuai[1]　Yang Xiuzhang[2]

(1. College of Information Management, Nanjing Agricultural University, Nanjing 210003, China;
2. School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China)

**Abstract:**［**Purpose/Significance**］As a representative of endangered scripts, Shui script has basically completed its digital protection with the help of national policies, but there is still a gap with other hieroglyphic scripts in the informatization of Shui script. In view of the gap between Shui script and other hieroglyphic scripts, the discovery of deficiencies in the informatization of the Literature of Shui Nationality is the key to promoting the transformation of the study of Shui script towards the development of intelligent applications.［**Method/Process**］Through the field survey and literature combing of Shui script informatization, we retrace the process of Shui script informatization construction in China and the results achieved, and explore the difficulties of Shui script informatization construction.［**Result/Conclusion**］The study shows that the research on the "character syntax" of the literature of Shui Nationality is the most solid, which basically meets the needs of the national language informatization construction. However, there is a lack of lexical research, and the accuracy of the model needs to be improved because of the lack of high-quality corpus of Shui script.

**Keywords:** Minority local chronicles; Literature of Shui Nationality; Information construction; National literature

（本文责编：王秀玲）