

FamilySearch家谱资源数字化开发利用 实践及启示*

王春迎¹ 王舒芳¹ 姬恒¹ 温芳芳^{2,3}

(1. 郑州大学信息管理学院, 郑州 450001; 2. 中国科学技术信息研究所, 北京 100038; 3. 浙江财经大学图书馆, 杭州 310018)

摘要: 家谱资源是重要的历史文化遗产, 文化数字化战略背景下促进家谱资源数字化开发与利用成为必由之路。采用网络调查法对全球最大的家谱组织FamilySearch展开调研, 发现FamilySearch基于社会需求多途径收集家谱资源, 针对实践需要多维度开发数字家谱资源, 面向不同人群多样化提供数字家谱服务。结合我国家谱资源建设与利用现状, 应拓宽资源收集渠道, 丰富家谱资源类型; 综合采取多项措施, 促进家谱资源深入开发; 面向不同用户群体, 提供针对性家谱服务。

关键词: 数字资源; FamilySearch; 家谱; 资源建设; 开发利用

中图分类号: G271 **DOI:** 10.3772/j.issn.1673-2286.2023.08.004

引文格式: 王春迎, 王舒芳, 姬恒, 等. FamilySearch家谱资源数字化开发利用实践及启示[J]. 数字图书馆论坛, 2023(8): 34-43.

家谱是记载同宗共祖的血亲集团世系、人物、事迹的历史图籍, 它与正史、方志构成中华历史大厦三大支柱, 是中华民族悠久历史的重要组成部分, 是极为珍贵的历史文化遗产^[1]。《关于推动数字文化产业创新发展的指导意见》指出要促进优秀文化资源数字化, 实施数字内容创新发展工程^[2]。《关于推进实施国家文化数字化战略的意见》强调到“十四五”时期末, 要基本建成文化数字化基础设施和服务平台^[3]。加强家谱资源数字化开发与利用对国家文化数字化战略具有重要意义。当前, 国内外学者从家谱资源开发技术^[4-6]、家谱数据管理方案^[7-8]、家谱众包平台的设计与实现^[9-11]等方面开展了系列研究, 但是缺乏对家谱资源数字化项目从多元资源收集、多技术手段开发到最终提供多样化服务全流程整体方案的论述。因此有必要对大规模、高成熟度、长期运行的家谱资源数字化项目进行调研分析。

FamilySearch是1894年在美国犹他州成立的一个非营利性家谱组织(原名犹他家谱学会), 目前已成为全球最大的家谱组织, 其开发整理了133亿条可按姓名检索的家谱记录、50亿张家谱图像、50多万本在线家谱图书^[12](截至2023年4月), 面向全球用户免费提供家谱查询、在线修谱等数十种家谱服务。无论是在规模、技术还是服务方面, FamilySearch都具有一定的代表性, 因而本研究采用网络调查法, 从家谱资源收集、家谱资源数字化开发、数字家谱服务提供3个维度对FamilySearch进行调研分析, 借鉴其发展经验, 以期推动我国家谱资源的数字化开发与利用。

1 基于社会需求多途径收集家谱资源

FamilySearch通过自行收集、建立合作、用户提供

收稿日期: 2023-06-10

*本研究得到河南省教育厅人文社会科学研究项目“黄河文化遗产大数据知识图谱建设研究”(编号: 2022-ZZJH-417)、2022年河南兴文工程文化研究专项项目“河南家谱研究”(编号: 2022XWH047)资助。

等方式收集了大量各类型的家谱资源, 并对其开展了数字化工作。

1.1 家谱资源的类型

FamilySearch的家谱资源按类型可以分为印刷型家谱资源、缩微型家谱资源、口述家谱资源和数字化家谱资源。

1.1.1 印刷型家谱资源

印刷型家谱资源包括家谱书籍、家谱期刊、户籍登记记录、人物志、地方志、人口普查记录、兵役记录、教堂记录、遗嘱、出生证明、婚姻登记记录、讣告等具有家谱价值的历史文献。这些资源中蕴含着丰富的家谱信息, 以讣告为例, 其中一般包含死者的父母、配偶、子女等家庭成员的姓名、性别和人物关系信息, 记录详实, 可以作为家谱资源开发与研究的第一手资料。印刷型家谱资源可以直接提供给读者阅读, 用户通过遍布全球的FamilySearch家庭历史中心即可免费阅览该类资源。

1.1.2 缩微型家谱资源

缩微型家谱指的是利用缩微摄影技术将手写或印刷型家谱文献缩摄在感光材料上所形成的家谱。1938年起, 犹他家谱学会开始建设缩微型家谱资源, 并将其存储在专门用于保管家谱缩微品的花岗岩山记录库中。目前该记录库收藏了来自120多个国家的240多万卷缩微胶卷, 积累了数十亿份家谱记录。为了更好地保存家谱和提供家谱服务, 20世纪末, FamilySearch开始对其收藏的缩微型家谱进行数字化转换。在数字化转换之前, 用户可以在FamilySearch各家庭历史中心利用缩微胶片阅读机查看缩微型家谱; 在数字化转换后, 用户可以通过FamilySearch网站直接在线访问阅览。

1.1.3 口述家谱资源

口述家谱也称口传家谱, 指的是通过口传心授流传下来的反映家族世系的家谱^[13]。当拥有口述家谱记忆的长者去世后, 相关家谱信息也会随之丢失, 因此有必要搜集整理口述家谱资源。FamilySearch从20世纪60年代开始在汤加、斐济等太平洋岛国收集口述家谱, 后

又扩展到印度尼西亚和马来西亚等东南亚国家以及非洲各国。在口述家谱的收集过程中, FamilySearch会聘请和培训当地人用当地语言进行访谈, 使用摄像机和录音机记录被访谈者提供的家谱信息, 保存收集到的音频文件, 为每个人生成照片、录音和姓名记录, 并根据访谈内容创建家谱树。目前, FamilySearch已经收集数百万条口述家谱记录, 构建了100多万棵口述家谱树, 可供用户在官网进行检索和收听。

1.1.4 数字化家谱资源

FamilySearch的数字化家谱资源主要由两部分组成: 一是转化型数字资源, 即由印刷型、缩微型家谱转换而成的资源; 二是原生数字资源, 即用户直接通过FamilySearch官网上传、在线编修的家谱资源。迄今为止, FamilySearch已经收集数以亿计的家谱图像和在线家谱记录, 用户可以通过其官网的各个模块访问这些数字化资源。

1.2 家谱资源的收集方式

FamilySearch的家谱资源主要来自于专业团队的搜集、与其他机构的合作以及用户捐赠。

1.2.1 组建专业团队, 收集家谱资源

FamilySearch由专门的家谱收集团队到世界各地收集家谱资源, 保存人口普查、遗嘱、公民身份信息等各种类型的家谱记录。自1938年起, 犹他家谱学会雇佣摄影师在世界各地使用缩微技术拍摄各国公民的出生、死亡等原始记录。2019年, FamilySearch向全球45个国家和地区派遣了300多个摄制组, 拍摄了1.7亿张数字化家谱图像。此外, FamilySearch还招募世界各地的志愿者拍摄照片, 收集遗嘱、讣告等家谱资源。

1.2.2 多元机构合作, 拓展家谱资源

FamilySearch在成立初期仅是一个本地组织, 为了搜集更多家谱资源, FamilySearch与国内外多家机构开展了项目合作。

(1) 与美国的图书馆、档案馆等机构合作收集家谱资源。例如, FamilySearch与美国公共数字图书馆

(Digital Public Library of America, DPLA) 合作, 获得来自与DPLA合作的2 000余家图书馆、档案馆、博物馆的1 300多万件文化遗产资源^[14]。FamilySearch与美国奥农达加县公共图书馆(Onondaga County Public Library)进行合作, 以数字方式保存奥农达加县公共图书馆收藏的40 000多份家族史、地方史、墓地记录等历史文献^[15]。

(2) 与多家海外机构合作拓展家谱资源。除了与美国的机构合作外, FamilySearch还与100多个国家的10 000多家档案馆、图书馆和互联网供应商等签订了合作协议。例如, FamilySearch与意大利政府、意大利国家档案馆以及各地档案馆合作, 保存了意大利各地区的出生、婚姻、死亡等民事记录的数字化图像。与加纳共和国的Smave公司合作, 对2 500多万条2010年加纳人口普查记录进行扫描和数字化, 并将其添加到FamilySearch数据库中^[16]。

1.2.3 接受用户捐赠, 丰富家谱资源

用户捐赠也是FamilySearch家谱资源的重要来源, 可以通过线下和线上两种途径实现。在线下捐赠方面, 用户可以到美国犹他州家谱图书馆及世界各地的家庭历史中心捐赠家谱书籍、家谱期刊、出生证、结婚证、遗嘱等资源; 在线上捐赠方面, 用户可以通过FamilySearch网站在线编修自己的家谱树, 上传与自己家族相关的照片、文档、音频等资源。目前, 用户通过FamilySearch网站贡献了数十亿条家谱记录。

2 针对实践需要多维度开发数字家谱资源

为提升资源开发效率、保证资源开发质量, FamilySearch围绕技术工具使用、家谱标准制定、数字人文团队建设、众包平台搭建4个方面进行了数字家谱资源开发与利用实践。

2.1 从数字化到数据化, 实现家谱资源多粒度开发

FamilySearch数字家谱资源开发过程可被划分为“数字化转换”和“数据化开发”两个阶段: 前者强调将家谱资源转化为数字形式, 后者则强调从家谱内容中

提取出结构化数据。开发流程不同, 最终产出的家谱数据粒度也不同。通过汇总梳理相关资料, 绘制了FamilySearch数字家谱资源开发技术流程图(见图1)。

2.1.1 家谱资源的数字化转换

家谱资源数字化转换指的是将各类非数字化家谱资源转换为数字化家谱资源。对于口述家谱, 其数字化转换方式是使用录音设备对被采访者讲述的家谱信息进行录音, 并通过数码相机拍摄相关照片。对于印刷型家谱, 其数字化转换方式有两种: 一是先使用缩微相机进行缩微化处理, 然后使用缩微胶卷扫描仪对缩微型家谱进行数字化转换。为了提升数字化转换的效率, FamilySearch开发了专用的软硬件系统Scanstone用于自动化处理图像裁切和对比度调整等任务, 减少人工处理导致的图像遗漏现象^[17]。二是直接对印刷型家谱进行数字化转换, 为此FamilySearch引入了300多台数码相机用于拍摄数字化家谱图像, 还研发了专用软件dCamX^[18], 用于辅助家谱图像的加工处理, 保证采集质量。

针对现有商业化光学字符识别(Optical Character Recognition, OCR)方案识别家谱文献准确率不高这一难点, FamilySearch训练了专用的OCR和手写文本识别(Handwritten Text Recognition, HTR)模型, 以提高家谱文献识别的准确率。对于缺乏训练数据的语言, FamilySearch则通过迁移学习技术改善识别效果, 例如FamilySearch在训练葡萄牙语HTR模型时采用了大量的西班牙语训练数据, 最终训练出来的葡萄牙语HTR模型也具有很高的准确度。

2.1.2 家谱资源的数据化开发

数据化以数字化为基础。FamilySearch综合应用机器学习方法和基于规则的信息提取方法, 深入经过数字化的家谱内容提取结构化的家谱数据。

在机器学习方面, 通过命名实体识别、关系抽取等自然语言处理技术, FamilySearch从非结构化的家谱文本中提取出姓名、日期、人物关系等信息, 并将其组织为结构化数据。此外FamilySearch还训练了专用的机器学习模型来从表格布局类的家谱文献中提取信息。对于一些更复杂的家谱布局结构, 则需要编写一些基于规则的代码来辅助或取代机器学习模型进行信息提取。

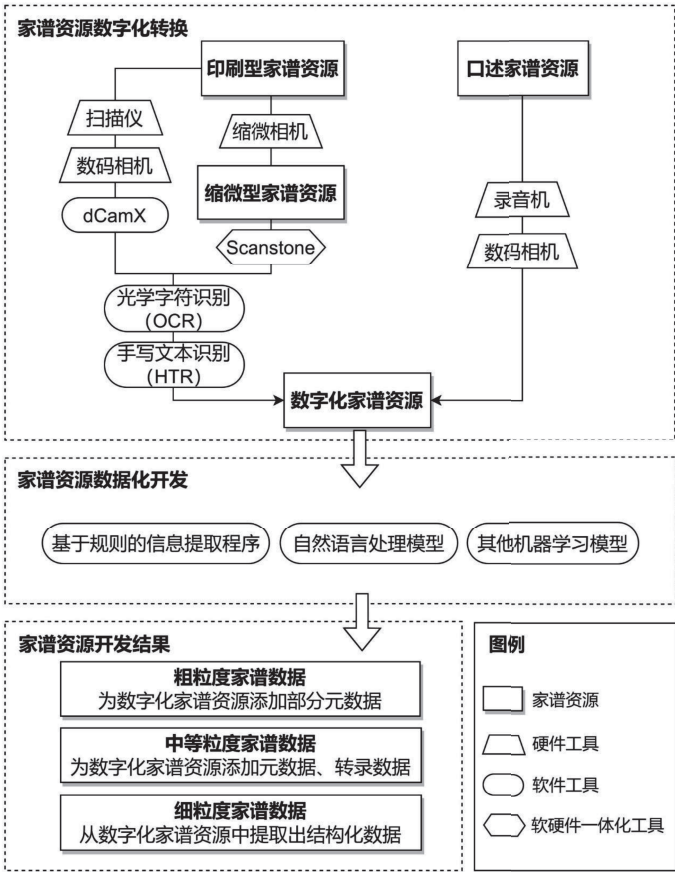


图1 FamilySearch数字家谱资源开发技术流程

2.1.3 家谱资源多粒度开发结果

家谱资源开发流程的终点是产出多粒度的家谱数据，不同粒度的开发成果可为用户提供不同类型的家谱服务。

家谱资源的粗粒度加工主要是指将家谱资源转换为数字格式并添加部分元数据，以方便在线访问并提供基础检索功能。FamilySearch网站的“家谱图像”（Images）模块就提供海量粗粒度家谱图像数据，其中的大部分图像仅支持按拍摄地点、时间、记录类型3种元数据进行检索，不支持更深入的姓名检索。

中等粒度加工成果在粗粒度加工成果的基础上附加一些转录数据，例如通过OCR或HTR从原始家谱文献中提取出的全文数据。这些数据并未经过严格校验，可能包含一定比例的识别错误，但仍能在一定程度上满足用户的全文检索需求。FamilySearch的“家谱书籍”（Books）模块就提供中等粒度的家谱数据，该模块除支持按书名、作者进行检索外，还利用转录数据提供全文检索服务。

细粒度加工成果主要指的是从非结构化家谱文献中提取出的结构化数据，例如从家谱中提取出的人物姓名、人生事件、世系关系等数据。这些数据经过校验以增强准确性，并支持以多种方式进行检索，是家谱资源数字化开发的理想成果。一些粗粒度和中等粒度的数据也会被渐进式地加工为细粒度数据。

2.2 引入多种标准, 保障家谱数据规范化处理

数字家谱资源开发过程环节众多且需要大量员工协同参与，为此FamilySearch引入了一系列标准来规范开发流程。

（1）图像和元数据标准。该标准是FamilySearch内部制定、内部使用的资源采集加工标准。其中：图像标准主要用于保证采集到的家谱图像内容的可读性，包含对图像色彩空间、对比度、文件格式等方面的要求；元数据标准则用于保证为图像添加的元数据信息的规范性与统一性^[19]。

（2）家谱数据标准GEDCOM。该标准是Family-

Search内部制定、内部使用并同时提供给外界共同使用的家谱内容描述标准,其定义了通用的数据表示语法和专用于家谱数据的世系链接语法,可以详细记录人、时、地、事等细粒度家谱信息。此外,GEDCOM还是一个家谱数据交换标准,可用于保障不同家谱软件之间以统一格式交换家谱数据。

(3) 家谱数据标准GEDCOM X。相较于GEDCOM, GEDCOM X支持不同序列化方法以增强家谱数据的通用性和可交换性,例如通过该标准记录的家谱信息可被序列化为XML、JSON等格式。GEDCOM X还添加了语义网标准以加深家谱数据的语义化程度和关联程度,例如为了描述家谱记录中常见的出生、死亡、结婚等事实信息, GEDCOM X定义了一组家谱领域专用的受控词表,表中每个词汇都对应一个独一无二的统一资源标识符(Uniform Resource Identifier, URI),用于提供该词汇的语义信息。

(4) 系谱证明标准(Genealogical Proof Standard)。系谱证明标准包含详尽的研究、完整且准确的来源引用、深入的分析和关联、解决相互冲突的证据、基于最有力的证据得出可靠的书面结论^[20]等5个部分, FamilySearch引入该标准以增强收集到的家谱信息的可信性。

2.3 组建数字人文团队,开展家谱跨学科研究

数字家谱资源开发是一个综合型任务, FamilySearch打造了人文和信息资源管理学科以及跨学科和数字化方向的数字人文团队,负责专业化处理开发过程中的文献研究、资源管理、技术研发等方面的任务。

(1) 文献研究任务。该方面的任务要求工作人员结合具体的历史地理背景对家谱文献内容做出专业的研究解读,辅助数字家谱的整理构建。该任务主要由系谱学家、家族史研究人员负责,这些工作人员往往拥有家谱、历史、语言等方面的专业技能,熟悉各国历史和家族史,在对家谱文献进行研究时可以提供专业的知识与经验。

(2) 资源管理任务。该方面的任务着眼于对FamilySearch收集的各类家谱进行数字化管理,主要由信息资源管理学科的元数据专家负责。元数据专家需要拥有图书管理、图书编目、元数据管理等方面的专业知识,能够完成家谱书籍目录管理、地理名称数据库管理、受控词表构建、元数据维护与改进等方面的任务。

(3) 技术研发任务。该方面的任务包括打造家谱

资源开发工具、编写相应的应用程序等,主要由数字化方向的工作人员负责。其中:机器学习工程师主要负责训练家谱领域的机器学习模型、从家谱文献中提取家谱数据,软件开发工程师则主要负责建设家谱网站、开发移动端应用程序。

除针对特定任务引入特定方向的专业化工作人员外, FamilySearch还会招纳跨学科方向的工作人员以处理特殊类型的开发任务,促进各部门之间的分工协作。

2.4 建设家谱众包平台,促进家谱资源合作式开发

众包的核心理念是利用公众智慧来解决较复杂的问题^[10], FamilySearch在众包方面进行了长期探索实践。在早期实践阶段, FamilySearch主要招募志愿者对海量家谱资源进行处理,例如1988年其招募志愿者从寺庙记录中提取了数千万条家谱记录。2006年, FamilySearch推出了FamilySearch Indexing众包平台,主要用于从扫描版的家谱中转录出可供搜索的数字索引。数十万名全球志愿者通过该平台提供的专用程序与网页参与了家谱转录工作,形成了超过10亿条可供检索的家谱记录。2022年, FamilySearch又推出了Get Involved众包平台,志愿者可以通过该平台与人工智能程序协同工作,对程序自动转录出的家谱信息进行审核修正,提高转录效率与准确性。此外FamilySearch还开发了专门的Get Involved移动应用程序,方便志愿者通过移动设备,利用碎片化时间参与众包任务。

3 面向不同人群多样化提供数字家谱服务

3.1 面向普通用户提供数字家谱查询与编修服务

(1) 家谱查询服务。FamilySearch收集整理了海量家谱树、家谱历史记录、家谱图像等资源,用以提供各具特色的查询服务,满足用户多样化需求。①家谱树查询服务。用户可以通过官网的“家谱树”(Family Tree)模块,按姓名、时间、人生事件等检索家谱数据,检索到的世系信息支持以扇形视图、水平视图、垂直视图等方式进行可视化呈现,方便为用户提供寻根问祖、查找亲属等服务。②家谱历史记录查询服务。用户

可以通过“家谱记录”(Records)模块检索从结婚证明、出生证明、人口普查记录等官方资料中提取出的人物信息。相比其他资源,历史记录具有更高的可信度和历史价值。③特色家谱信息查询服务。用户可以通过“活动”(Activities)模块的“姓氏起源”(Surname Origins)服务查询某个姓氏的含义、起源、人口分布等信息;通过“关于我的一切”(All About Me)服务查询用户出生年份的头条新闻、流行歌曲、足球冠军等。

(2)家谱编修服务。FamilySearch为用户提供在线修谱、上传家庭回忆记录、链接家谱资料等服务。用户可以通过这些服务构建出可信的、相互关联的高质量家谱树。①在线修谱服务。用户可以通过官网“家谱树”模块可视化地编辑自己家族的世系关系。在编辑过程中,FamilySearch基于资料匹配与逻辑判断给出一些智能提示,例如提醒用户检查填写的信息是否有误、检查其家谱树中的人物是否和其他家谱树中的人物是同一个人。这些提示可以提升用户编写家谱信息的质量,帮助用户把个人家谱树和全球共享的大家谱树连接起来。②上传家庭回忆服务。用户可以通过官网“家庭回忆”(Memories)模块上传文档、照片、音频等家庭回忆记录。上传后的家庭回忆既可以和用户的家谱树链接在一起,增强家谱信息的可信度,也可以共享给其他亲属,共同构建关于某个祖先的回忆。③家谱资料链接服务。用户可以检索FamilySearch采集到的海量家谱数据,并将其中有价值的信息直接链接到自己的家谱树当中,增强自己家谱树的可信性。

3.2 面向开发人员提供家谱数字技术支持服务

为了促进数字家谱资源的创新利用,FamilySearch面向开发者提供RESTful应用程序接口(Application Programming Interface, API)、开发文档、开发工具包等技术支持服务,开发者借此创建了大量图表类、管理类和游戏类家谱应用程序。

在注册成为FamilySearch的解决方案供应商并创建自己的应用程序之后,开发者就能通过FamilySearch提供的API获取家谱资源,进行程序开发。在开发过程中,开发者还可以接入FamilySearch提供的多种技术支持服务。①API文档。针对每个API,该文档都提供了详细的讲解和示例,帮助开发者快速掌握使用方法。②软件开发工具包(Software Development Kits,

SDK)。SDK中封装了大量可重复利用的工具类和工具函数,可以极大地提升开发效率,并且支持Java、JavaScript、Python等编程语言,方便各个平台的开发者下载使用。③应用程序示例。FamilySearch还提供了大量开源的应用程序示例,开发者可以仿照这些示例快速编写自己的应用程序。

3.3 面向科研人员提供数字家谱资料与研究百科服务

(1)家谱资料服务。FamilySearch向研究人员免费提供高质量的家谱书籍、家谱树、家谱历史记录等资料,辅助家谱研究工作的开展。在家谱书籍方面,研究人员可以通过FamilySearch网站在线访问其收集的50多万本家谱书籍,开展家谱研究。例如,王日根等^[21]使用从FamilySearch获取的《云阳毛氏三修族谱》《茹氏四修族谱》等书籍,研究了明清时期湘中、湘南地区家族人口增长的整体趋势及差异性。家谱树、家谱历史记录中则包含了大量的人物姓名、性别、世系关系等信息,同样具有宝贵的研究价值。例如,Lleras-Muney等^[22]使用1940年美国全面人口普查数据并结合FamilySearch家谱树数据研究了受教育程度与寿命之间的关系。

(2)研究百科服务。FamilySearch面向研究人员提供由专家编写的、包含十万多篇家谱文章的专业化研究百科网站,许多与家谱研究相关的主题都可以在该百科中找到。例如,该百科按地域编制了家谱研究所需的资源索引,研究人员可以直接通过该索引查看汇总好的出生、婚姻和死亡记录。

4 FamilySearch家谱资源数字化开发利用实践对我国的启示

中国家谱资源数字化工作兴起于20世纪80年代^[23],先后经历了家谱书目数据库、家谱全文数据库^[24]、家谱知识服务平台3个建设阶段,产生了台湾地区家谱联合目录数据库、中国谱牒库、中华寻根网、上海图书馆家谱知识服务平台(以下简称“上图家谱知识服务平台”)等典型项目。与国内项目相比,FamilySearch规模更大、技术更先进,拥有诸多可供借鉴的亮点与特色。从资源收集、开发、服务3个方面归纳FamilySearch实践方案的核心优势,面向我国家谱资源开发利用的困境与难题提出针对性建议。

4.1 拓宽资源收集渠道, 丰富数字家谱资源类型

收集家谱资源是对其进行数字化开发和利用的基础。目前国内家谱机构在收集过程中普遍存在如下问题: 一是专业团队缺乏。国内大多数家谱机构并未成立专门的资源收集团队, 收集工作多由其他岗位的员工兼职负责, 这在一定程度上影响了收集效果。二是机构合作较少。国内各地区、各部门在家谱资源的收集和保存上相互独立, “孤岛效应”严重^[25]。三是收集渠道和类型单一。我国家谱机构多侧重于通过线下渠道收集家谱资源, 忽视了利用线上渠道收集家谱资源的便捷性。此外, 其收集的资源类型也多为纸质版的家谱书籍, 对其他类型的资源不够重视。上图家谱知识服务平台虽建设有线上收集渠道, 但主要收集扫描版家谱书籍, 对口述家谱、数字化家谱、家谱相关资料缺乏关注。

相比之下, FamilySearch的资源收集方案更为完善。借鉴其经验, 国内家谱机构首先应加强人才队伍建设, 组建具有历史、地理知识, 熟悉地方风土人情与方言, 同时熟练掌握现代化采集技术的专业化团队。其次, 应加强与不同国家、不同地区图书馆、档案馆等机构的合作。FamilySearch主要通过联合众多机构, 采用资源互换共建模式而非被动接受捐赠模式进行共建共享。国内家谱收藏机构众多, 通过建设家谱资源互换共建网络可以有效促进资源共享, 避免重复建设。最后, 应加强线上家谱平台构建, 鼓励用户通过网站上传不同类型的家谱, 同时根据我国的实际情况, 加大对记载在甲骨、青铜器、竹简、石碑、布帛等载体上的家谱资源的收集力度。

4.2 综合采取多项措施, 促进数字家谱深入开发

4.2.1 训练领域专属模型, 整合系统开发方案

数字化转型与数据化加工是家谱资源开发的两个重要环节, 但国内家谱平台一直难以完成从数字化到数据化的跨越。虽然上图家谱知识服务平台和中华寻根网提供部分结构化家谱数据, 但主要为书目元数据或小规模的人工抽取数据。深入家谱内容, 提取海量结构化家谱数据并提供知识服务的愿景一直难以实现, 这构成了国内家谱平台转型升级的核心困境。

FamilySearch从数字化到数据化的发展方案具有借鉴意义。参考其经验, 国内平台首先应构建家谱领域专用的人工智能模型。虽然从书籍内容中提取结构化数据的方法在不断改进, 但具体到家谱领域, 仍会面临版面复杂、图像破损等诸多特有难题, 通用模型难以实际应用, 因此, 训练领域专属的OCR、命名实体识别、关系抽取模型成为必要举措。其次, 应整合基于人工、基于规则和基于深度学习的方法, 形成系统性开发方案。虽然深度学习方法在单项任务中性能较好, 但家谱开发是一个体量庞大、情况多变的巨型工程。面对复杂家谱布局, 基于规则的方法仍有用武之地, 并且更特殊的情况仍需人工介入处理, 只有三者协同才能保障数据化转型的真正实现。

4.2.2 制定家谱内容标准, 进行长期维护更新

建立家谱标准是统一家谱信息描述格式, 促进资源共建共享的重要途径。相较FamilySearch, 国内家谱标准相对陈旧, 需要从如下两个方面改进升级。首先, 应尽快制定细粒度的家谱内容描述标准, 为提供家谱知识服务做准备。目前国内广泛使用《家谱元数据著录规则》等元数据标准, 但这些标准无法对家谱内容中的人、时、地、事及其相互关系进行精确描述, 难以满足深入开发需求。借鉴GEDCOM, 制定国内家谱内容描述标准的核心要务是开发一套成熟且强大的描述方案, 使其可以充分表达家谱内容中的出生、死亡、婚姻、家庭等细粒度信息。此外应注重保持数据模型的灵活性, GEDCOM因数据模型固化而受到批评^[26], 国内方案可在优先适配中国传统宗族结构与现代家庭结构的基础上, 增强对不同国家家庭结构的兼容性, 为国际化做准备。其次, 应对家谱标准进行详细说明与长期维护。上图家谱知识服务平台开放了基于关联数据的家谱本体, 但其说明文档相对简陋, 上手难度较高, 这在一定程度上影响了推广效果。参考FamilySearch相关经验, 应为标准建立专门网站, 进行独立维护; 提供详细说明文档, 方便其他机构接入; 通过迭代不断修复旧缺陷、引入新技术, 保持标准与时俱进。

4.2.3 组建数字人文团队, 协作开发家谱资源

在团队建设方面, FamilySearch与国内平台的核心差异在于其形成了分工明确的数字人文团队。数字家谱

资源的开发是一项综合型任务,需要不同专业人员的协同参与。在实践中,国内的上图家谱知识服务平台起初没有明确分工,随着开发推进,其也逐渐形成了不同的工作团队,这证明了分工的必要性。

国内平台在建设团队时,可以借鉴FamilySearch的成熟经验,采用文献研究、技术研发、资源管理这一分工方案,实现技术团队与人文团队的分工协作,同时由信息资源管理工作对家谱书籍以及开发过程中产生的海量资源进行系统管理。此外,应注重跨学科研究人员的招聘,充分发挥其在促进团队协作、攻克复杂跨学科问题中的作用。

4.2.4 开发移动众包平台,引入人工智能协助转录

家谱资源规模庞大、体例多样、排版复杂,仅依靠家谱机构有限的工作人员难以在短期内完成大规模数字化开发任务,因此,借助众包平台提升开发效率成为必要选择。国内,上海图书馆建设有网页版的历史文献众包平台,但其产出相较于FamilySearch众包平台的数十亿条记录仍有数量级差距,如何提升众包效率成为亟待解决的问题。

审视FamilySearch众包平台十几年的发展历程,可以总结出如下两点针对性建议:一是国内家谱机构应注重开发移动端众包平台。40多亿移动互联网用户是众包平台不可忽视的重要参与力量,FamilySearch的Get Involved众包平台仅在Google Play平台便拥有十余万下载量,成为重要众包入口。针对国内环境,开发众包App或者微信小程序成为提升众包参与度的重要途径。二是应引入人工智能程序,与志愿者协同工作。赵宇翔等^[27]指出,在文化遗产众包方面,除Web建站等通用技术外,人工智能、OCR等领域专用技术的应用会给资源数字化带来巨大突破。FamilySearch的实践印证了该观点,从“人工转录”到“人工智能转录+人工审核”的模式转变推进了其众包平台的重大代际升级,这也是国内众包平台下一阶段的重要迭代方向。

4.3 面向不同用户群体,提供针对性数字家谱服务

对国内家谱平台而言,针对不同类型用户提供特定服务的思想尚在萌芽状态。例如,中国谱牒库更多作为专业数据库面向研究人员提供服务,上图家谱知识服

务平台则主要面向普通用户提供服务,面向开发者提供的服务仍处于早期建设状态。相比之下,FamilySearch面向普通用户、开发人员、研究人员提供定向服务,其服务体系具有重要参考意义。

4.3.1 面向普通用户:深挖家谱内容,丰富呈现方式

国内家谱网站主要面向用户提供粗粒度的姓氏百科、家谱书目检索等服务,但普通用户往往对深入了解家谱内容、可视化呈现家谱信息更感兴趣。借鉴FamilySearch相关经验,国内家谱平台可从如下方面加以改进:第一,注重利用细粒度家谱数据提供新型服务,例如利用结构化家谱树提供家谱寻亲服务、利用转录数据提供族人历史文献检索服务等,以充分挖掘家谱这一独特数据类型社会价值;第二,提供多种家谱可视化方式,应尽可能提供水平、垂直、扇形等视图,支持放缩、拖动等多样化操作,多维度呈现家族信息;第三,通过地理信息系统(Geographic Information System, GIS)、时间轴等可视化方式提供迁徙地图等特色服务,满足用户个性化需求。

4.3.2 面向开发人员:完善开发者服务,实现数据增值

开发者是对家谱数据进行编程加工,将其转换为多样服务,实现人文数据增值的重要力量。国内,除上图家谱知识服务平台为开发者提供家谱数据接口外,其余平台并不重视开发者服务。相比之下,FamilySearch的开发者服务构建成熟,服务完善、组织清晰是其核心特征。国内机构可以借鉴FamilySearch代码、文档、应用库三模块结构。在代码方面,除通过API对外提供数据外,还应注重多语言SDK开发,在平台方一次性封装重复逻辑,避免开发者重复劳动,提升数据增值效率。此外还应注重应用库建设,促进开发成果的展示与分发。在资源组织方面,国内机构应建设“开发者中心”专用网站,对代码、文档等开发者所需的特色资源进行整合,提供一站式服务。

4.3.3 面向研究人员:提供文献数据,强化科研参与

在人文学科向数字人文研究新范式转型的浪潮中,

升级成为数字人文研究平台、提供全面科研服务是数字时代文化遗产机构必须把握的重要发展机遇^[28],但国内家谱平台较少参与科研。FamilySearch广泛参与科研,仅中国知网中便有不下数十篇论文使用其家谱资源展开研究。

对FamilySearch的科研实践进行梳理,可以总结出两条核心路径:一是提供全面、可获取的家谱书籍,进行“文献参与”。相较于资源有限或需付费使用的国内家谱平台,FamilySearch收集了50余万本家谱书籍供用户免费在线访问,这使得其广受人口史、民族史、社会学等领域研究人员的青睐。例如,吴铮强^[29]指出,相较于温州市图书馆的残本,FamilySearch所藏《苍坡方巷李氏宗谱》为完帙且可免费在线访问,为科研提供了极大便利。二是提供海量、细粒度的结构化家谱数据,进行“数据参与”。数据参与的特点是量级巨大,例如在Lleras-Muney等有关教育与寿命关系的研究^[22]中,FamilySearch的500多万条出生与死亡记录提供了良好的数据支撑。在后续建设中,粗粒度的文献参与和细粒度的数据参与可以成为国内家谱平台开辟科研服务版块的重要参考路径。

5 结语

家谱资源是家族历史和文化遗产的珍贵载体,文化数字化战略对家谱资源的深入开发利用提出了更高的要求。本文以全球最大的家谱组织FamilySearch的数字化实践为例,从资源收集、开发、利用3个方面,全流程梳理其整体实践方案,而后对中华寻根网、上图家谱知识服务平台等典型平台展开调研,定位国内家谱平台发展方向。国内家谱平台在后续建设中应着重完善线上收集渠道、促进数字家谱深入开发、面向不同用户提供定向服务,以此实现从家谱数据库到家谱知识服务平台的转型升级,促进家谱资源价值的充分发挥。

参考文献

- [1] 王鹤鸣. 史界瑰宝 不朽盛业: 近百年来开发中国谱牒资源简述[J]. 图书馆杂志, 2001, 20 (4): 9-12.
- [2] 文化部关于推动数字文化产业创新发展的指导意见[EB/OL]. [2022-10-26]. http://www.gov.cn/gongbao/content/2017/content_5230291.htm.
- [3] 中共中央办公厅 国务院办公厅印发《关于推进实施国家文化数字化战略的意见》[EB/OL]. [2022-07-22]. http://www.gov.cn/zhengce/2022-05/22/content_5691759.htm.
- [4] CAMPANYA A J, CONESA C J, MAYOL S E. Modeling genealogical domain-an open problem[C]//Proceedings of the International Conference on Knowledge Engineering and Ontology Development, 2012: 202-207.
- [5] 张磊, 胡新. 以现代技术重现传统文献的探索: 上海图书馆的古籍数字化工作[J]. 数字图书馆论坛, 2006 (12): 5-11.
- [6] FOLKMAN T, FURNER R, PEARSON D. GenERes: a genealogical entity resolution system[C]//2018 IEEE International Conference on Data Mining Workshops (ICDMW), 2019: 495-501.
- [7] 夏翠娟, 刘炜, 张磊, 等. 基于书目框架(BIBFRAME)的家谱本体设计[J]. 图书馆论坛, 2014, 34 (11): 5-19.
- [8] SCHULTZ E S, ABBOTT P D. Metadata: key to high-volume access to records[J]. Archiving Conference, 2011, 8 (1): 44-46.
- [9] FORNÉS A, LLADÓS J, MAS J, et al. A bimodal crowdsourcing platform for demographic historical manuscripts[C]//Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, 2014: 103-108.
- [10] 刘倩倩, 夏翠娟. 家谱知识服务平台众包模式的设计与实现[J]. 图书馆论坛, 2020, 40 (5): 10-15.
- [11] HANSEN D L, SCHONE P J, COREY D, et al. Quality control mechanisms for crowdsourcing: peer review, arbitration, & expertise at familysearch indexing[C]//Proceedings of the 2013 Conference on Computer Supported Cooperative Work, 2013: 649-660.
- [12] FamilySearch.org facts[EB/OL]. [2022-07-26]. <https://www.familysearch.org/en/newsroom/company-facts>.
- [13] 王鹤鸣. 中国少数民族原始形态家谱调研报告[J]. 安徽史学, 2021 (5): 151-159.
- [14] 国家图书馆研究院. 美国公共数字图书馆与美国犹他家谱学会开展合作[J]. 国家图书馆学刊, 2016, 25 (4): 28.
- [15] Onondaga County Central Library Project will make over 40, 000 historic records freely accessible[EB/OL]. [2022-10-26]. <https://www.familysearch.org/en/newsroom/onondaga-county-central-library-project-will-make-over-40000-historic-records-freely-accessible-online>.
- [16] At risk records preserved in Ghana[EB/OL]. [2022-07-26]. <http://news-africa.churchofjesuschrist.org/at-risk-records->

- preserved-in-ghana-?country=ghana.
- [17] The new face of family history[EB/OL]. [2022-08-06]. <https://www.genealogytoday.com/articles/reader.mv?ID=1188>.
- [18] Technology in action-technology helps FamilySearch hit major milestone[EB/OL]. [2022-08-06]. <https://www.nextscan.com/technology-in-action-technology-helps-familysearch-hit-major-milestone/>.
- [19] LAXMAN R J. FamilySearch digital image standard: a key to quality[J]. Archiving Conference, 2010, 7 (1): 81-89.
- [20] Board for Certification of Genealogists. Ethics and standards[EB/OL]. [2022-07-19]. <https://bcgcertification.org/ethics-standards/>.
- [21] 王日根, 叶再兴. 明清湖南家族人口增长的趋势及差异: 以族谱资料为中心的考察[J]. 南方人口, 2016, 31 (5): 70-80.
- [22] LLERAS-MUNEY A, PRICE J, YUE D H. The association between educational attainment and longevity using individual level data from the 1940 census[J]. Journal of Health Economics, 2022, 84: 102649.
- [23] 黄显功. 上海图书馆的家谱数字化资源服务: 古籍保护效果的一个实例[J]. 图书馆学刊, 2008, 30 (1): 4-6.
- [24] 毛建军. 中国家谱数字化资源的开发与建设[J]. 档案与建设, 2007 (1): 22-24.
- [25] 加小双. 国外“寻根问祖”网站兴盛探讨: 以Ancestry.com为例[J]. 寻根, 2014 (1): 19-25.
- [26] 肖禹. GEDCOM数据模型述略[J]. 图书情报导刊, 2016, 1 (9): 127-130.
- [27] 赵宇翔, 练靖雯. 数字人文视域下文化遗产众包研究综述[J]. 数据分析与知识发现, 2021, 5 (1): 36-55.
- [28] 刘炜, 谢蓉, 张磊, 等. 面向人文研究的国家数据基础设施建设[J]. 中国图书馆学报, 2016, 42 (5): 29-39.
- [29] 吴铮强. 宋元明时期苍坡李氏家族变迁史: 兼论所谓的宋元明转型问题[J]. 浙江社会科学, 2017 (11): 122-132, 159.

作者简介

王春迎, 女, 博士, 讲师, 研究方向: 信息检索与服务、数据治理、信息资源建设。

王舒芳, 女, 本科生, 研究方向: 信息资源建设。

姬恒, 男, 硕士研究生, 研究方向: 信息资源建设。

温芳芳, 女, 博士, 副研究馆员, 通信作者, 研究方向: 开放数据、信息素养, E-mail: wenfangf@126.com。

Practices on the Digital Development and Utilization of Genealogical Resources by FamilySearch and Implications

WANG ChunYing¹ WANG ShuFang¹ JI Heng¹ WEN FangFang^{2,3}

(1. School of Information Management, Zhengzhou University, Zhengzhou 450001, P. R. China; 2. Institute of Scientific and Technical Information of China, Beijing 100038, P. R. China; 3. Zhejiang University of Finance and Economics Library, Hangzhou 310018, P. R. China)

Abstract: Genealogical resources are important historical and cultural heritage. Promoting digital development and utilization of genealogical resources has become a necessary path in the context of cultural digitization strategy. The study uses a web-based survey method to investigate FamilySearch, the world's largest genealogy organization, and finds that FamilySearch has collected genealogy resources in multiple ways based on social needs, developed digital genealogy resources in multiple dimensions to meet practical needs, and provided digital genealogy services for different groups of people in various ways. Under the current situation of construction and utilization of genealogy resources in China, it is necessary to broaden the channels of resource collection and enrich the types of genealogical resources; take multiple measures to promote the in-depth development of genealogical resources; provide targeted genealogical services for different user groups.

Keywords: Digital Resource; FamilySearch; Genealogy; Resource Construction; Development and Utilization

(责任编辑: 王玮)