

AI 赋能档案自动化整理的实践探索 ——以采集工程为例

郑悦萍

(北京应用物理与计算数学研究所)

DOI: 10.65285/RWYSHKX.V2I4.019

摘要: 本文从中国科协采集工程档案资料收集整理编研实务场景中的问题出发,探索以 AI 辅助 VBA 代码开发赋能科学家档案自动化整理的实践路径,构建元数据著录、学术年表生成、资料长编整合三者的数据关联与联动机制,实现档案一次著录多业务复用,为档案资源的自动化整理提供实践范式。

关键词: 档案; 自动化; 元数据; 流程联动; AI

Practical Exploration of AI-Driven Automation in Archival Organization: A Case Study of Data Collection Projects

Yueping Zheng

Beijing Institute of Applied Physics and Computational Mathematics

Abstract: This paper based on the collection engineering practice scenarios and archives management work specifications, this paper uses AI to assist VBA code development, builds an automated linkage system with standardized metadata tables as the core center, and establishes a metadata description, academic chronology generation, and data compilation integration. The strong data association and process linkage mechanism realizes one-time description, data reuse, and synchronous derivation, providing a practical paradigm for the automatic sorting of archival resources.

Keywords: archive; automation; metadata; Process linkage; AI

前言:

中国科协老科学家学术成长资料采集工程持续纵深推进,海量涵盖手稿、照片、录音录像、证书证件、科研实物的科学家档案完成数字化转化,形成类型丰富,价值突出的科学家学术档案资源体系。采集范围涵盖手稿、信件、证书、著作、论文、照片、口述资料整理稿、音视频等 15 大类原始资料,是记录老科学家学术历程与科研精神的重要载体。按照工程技术规范,除实物类资料以高清拍照方式存档外,其余 14 大类非实物资料均应尽量通过扫描、转录、格式转换等手段完成数字化处理,为后续系统化整理、长期保存与开发利用奠定数据基础^[1]。

当前采集工程使用的官方离线采集工具基于 VBA 宏开发,主要用于资料基础信息登记、元数据著录、格式规范校验等基础操作,未实现目录数据与年表、长编的自动化关联与成果派生,三大核心工作环节呈现“串行滞后、数据割裂”的低效工作模式。采集实务工作中,档案目录数据著录完成后,年表、长编需人工逐一调取多格式资料、手动排版整合,不仅造成前期目录数据闲置浪费,更导致年表、长编汇编工作长期滞后于著录环节,出现内容不一致、时序混乱、重复工作量大等问题。针对这一核心痛点,本文通过定制化 VBA 工程开发,打通三大环节数据流转壁垒,构建“数字化归集—元数据著录—年表生成—长编整合”一体化联动体系,在数据采集处理与成果输出等环节应用

自动化技术,全面提升采集工程整体工作效率与成果规范性。

一、VBA 与 AI 融合的自动化档案整理体系设计

(一) 体系设计核心原则

为破解采集工程档案整理的现实瓶颈,本文构建以 VBA 为执行引擎、以 Excel 元数据为中枢、以 AI 为辅助开发手段的档案自动化整理体系,体系设计严格遵循四项核心原则:第一,唯一数据源原则,以 Excel 元数据表为全部业务环节的权威数据来源,年表、长编均从元数据自动生成,确保数据同源、口径统一、逻辑一致。第二,Excel-Word 深度联动原则,充分发挥 Excel 在结构化数据存储、计算、排序、路径管理方面的优势与 Word 在长文档排版、格式规范、成果输出方面的优势,实现目录与资料相关成果汇编数据互通。第三,离线安全与本地化运行原则,全部程序依托 Office 原生组件运行,不上传数据、不联网处理、不依赖外部接口,保障档案资料安全可控。第四,AI 功能边界严格限定原则,人工智能仅用于辅助 VBA 代码编写、逻辑优化、错误调试与批量代码生成,不介入档案内容识别、文本抽取、史料处理、语义理解与内容修改,确保原始档案不受算法干预、保持原始性与权威性。

(二) 业务逻辑思路

以 15 大类资料数字化为前提,以采集工程及标准化元数据表为唯一核心数据源,通过 VBA 代码开发,搭建“数



图4 VBA 导出年表代码界面

(二) 资料长编一体化汇编

资料长编是采集工程中相对复杂、工作量较大的成果形态,要求按时间顺序完整汇集原始资料、图文混排、题名清晰、注释规范、可直接归档利用。VBA 汇编长编以元数据中的存储路径为核心索引,跨模态编研档案资源,按时序自动抽取调用文本、图片类电子文件,自动插入对应资料并添加规范题名标识。PDF 文档需经过 OCR 文字识别后转成 WORD 后调用, JPG 照片与 TIFF 高清扫描件, VBA 可以直接调用,以嵌入式方式插入,保证版式稳定、不漂移、不跨页错乱。通过 VBA 自动化汇编,人工复制、粘贴、插入、排版、校对等重复劳动基本被消除,长编编制周期从数月缩短至几天以内,且能够稳定处理大批量、大体积、多类型资料,同一事件不同档案资源通过时间、摘要线索快速完成规范化整合,确保采集资源之间的关联构建严谨、准确,打破传统档案编研工作模式固化、单维、线性、封闭的局限,^[4]满足采集工程规模化整理需求。

四、应用成效

以元数据为核心的自动化体系在采集工程资料整理中取得了明显实效。项目先把 15 类不同形式的资料统一完成数字化处理,为后续自动化整理搭建了稳定的数据基础。在实际工作中,自动化提取元数据大大缩短了基础数据整理用时,使得与档案整理相关的三项工作能够同步开展,有效解决了以往年表与长编制作长期滞后的问题,整体工

作效率显著提高。实践中的自动化路径除代码编写外,其余操作都在本地离线环境完成,能够确保科学家档案资料的安全。系统基于 Office 平台运行,不用额外购置软硬件,部署简单易推广。同时,借助 AI 辅助编写程序,也降低了工具开发的门槛,普通档案人员不用掌握复杂编程技术,就能快速搭建出贴合业务、稳定好用的自动化处理工具。

结语:

针对中国科协老科学家学术成长资料采集工程涉及的 15 大类资料高效规范整理需求,本文提出在规范资料数字化转化、形成电子文件集合的基础上,基于 VBA 二次开发搭建“数字化归集-元数据著录-年表-长编”联动自动化体系,以唯一数据源实现采集工程三大业务数据互通、流程联动,一定程度上解决了多源异构资料整理低效、年表长编等汇编工作滞后于资料著录的难题,大幅提升了采集工程工作效率,具有一定的实务应用性与可操作性。本文作为档案 AI+ 实践的探索,将 AI 的使用严格限定于代码开发环节,既充分发挥人工智能提升自动化工具研发效率的作用,又严格恪守档案原始性、真实性与业务安全可控的基本要求,为档案 AI+ 提供合规、安全、可落地的实践范式,为老科学家学术成长资料标准采集、深度利用提供了实践思路,对推动科技档案智能化转型具有重要的现实意义与推广价值。

参考文献:

- [1] 中国科协.老科学家学术成长资料采集工程采集工作规范(暂行)[J].中国科技史杂志,2011(02).
- [2] 贾君枝,张绪阳,张贵香.档案数据编目:架构、业务流程与功能[J].档案学通讯,2024(1).
- [3] 林慕婵,蔡颖.基于 Excel VBA 的数码照片著录模板的设计与实现[J].计算机与现代化,2011(9).
- [4] 席姗姗.档案智能编研的基础建设、技术应用与实践策略[J].山西档案,2025(11).

作者简介: 郑悦萍(1991.06-),女,汉族、山西人、硕士研究生、馆员、研究方向:档案数字化、信息化,口述与实物档案管理。