AI生成元数据赋能图书馆资源建设的实践与启示*

——基于国内外案例调查

张谙宁　叶兰　周文琦　张倩　张欢庆

摘要　分析国内外图书馆在资源建设中应用AI生成元数据的实践经验，为国内图书馆提升AI应用能力和资源建设水平提供参考。通过调研国内外20个实践案例，围绕实施主体、资源对象、应用技术、应用场景和赋能成效总结实践现状和经验。研究从两个方面提出实践建议与启示：一是制定AI生成元数据的技术应用方案，包括基于大语言模型的提示词工程、基于机器学习的领域数据训练和基于知识图谱的检索增强生成；二从战略规划、质量监控和人才能力等方面保障AI生成元数据赋能成效。

关键词　人工智能　元数据　资源建设　图书馆

分类号　G254

DOI　10.16603/j.issn1002－1027.2025.04.010

引用本文格式　张谙宁，叶兰，周文琦，等.AI生成元数据赋能图书馆资源建设的实践与启示——基于国内外案例调查[J].大学图书馆学报，2025，43(4)：90－104.

1　引言

元数据是关于数据的数据。元数据是图书馆资源建设的基石，它不仅关系到资源的描述与组织，而且是实现资源可发现、可访问、可互操作及可重用的基础。AI生成元数据指利用人工智能(Artificial intelligence，AI)技术自动提取或创建描述图书、文件、图像、音频、视频、数据集等资源的元数据。随着AI技术的发展，AI生成元数据正在重塑资源描述、组织、发现与利用的方式，给图书馆资源采访、编目、资源开发与利用等业务工作带来变革与创新，同时为用户利用图书馆资源带来新体验。北美研究型图书馆协会(Association of Research Libraries，ARL)的一项内部调查显示[1]，AI技术可被应用于图书馆工作的多个方面，包括自动编目和元数据生成、增强信息检索发现以及馆藏资源的保护和保存等领域。国际图联(International Federation of Library Associations and Institutions，IFLA)人工智能特别兴趣小组发布的《图书馆对人工智能的战略响应》[2]报告指出，人工智能在图书馆领域最强大的应用之一是“描述性AI应用”，包括馆藏资源的规模化描述、AI增强或创建元数据等方向。由此可见，AI生成元数据不仅是技术发展的必然结果，更是图书馆资源建设创新与发展的重要驱动力。为此，本研究调研与分析国内外图书馆利用AI生成元数据赋能资源建设的典型案例，总结AI生成元数据赋能资源建设的应用场景和应用成效，为国内图书馆在资源建设中应用AI生成元数据提供实践案例参考与启示。

2　研究现状

目前，关于AI生成元数据在图书馆资源建设中的应用主要集中在以下四个方向：

(1)自动化元数据采集。林家华(Chia-Hua Lin)等[3]对机器人流程自动化(Robotic Process Automation，RPA)采集与处理元数据方面的潜力进行了研究，提及新加坡国家图书馆利用RPA从电子书平台自动提取电子书使用数据，为采购工作提供数据支持；日本枚方市中央图书馆通过RPA自动抓取馆藏信息用于采购流程中的馆藏检查；美国国

* 广东省图书馆科研课题“AI生成元数据赋能图书馆资源建设的实践研究”(编号：GDTK24025)的研究成果之一。
通讯作者：叶兰，邮箱：yel@szu.edu.cn。

会法律图书馆借助RPA转录政府文件元数据并补充国会议员身份信息，简化人工检索与比对流程。

(2)自动分类与主题标引。自动分类与自动主题标引常用于处理海量资源的标引场景[4]。随着AI技术的发展，自动分类和主题标引的准确度显著提升。科拉尔卡·戈卢布(Koraljka Golub)等[5]利用芬兰国家图书馆开发的自动主题工具Annif对瑞典国家图书馆书目数据进行杜威十进制分类号的自动分配测试，发现在三级分类任务中达到了66.82%的准确率。国内学者戎璐[6]利用大语言模型ChatGPT进行中图法自动分类的实验，发现三级类目分类任务准确率突破70%。

(3)自动生成书目记录。以ChatGPT为代表的生成式AI在自动生成书目记录(包括描述与标引)方面取得了显著进展。理查德·布鲁斯托维茨(Richard Brzustowicz)[7]展示了使用ChatGPT准确生成MARC记录的能力。珍妮·博登哈默(Jenny Bodenhamer)[8]进一步评估了生成式AI在生成书目记录过程中的优势和不足，提出了在实际应用中应注意的改进方向。格伦·格林利(Glen Greenly)[9]通过引入检索增强生成技术(Retrieval-Augmented Generation，RAG)，允许访问未嵌入在大语言模型中的编目标准，开发了“CatalogerGPT”自动编目工具，进一步提高了生成MARC记录的准确性。

(4)元数据语义增强。AI技术可以通过深度标引和实体识别等功能增强图书馆元数据，从而提升资源在检索和发现方面的效果。迪拉瓦尔·阿里(Dilawar Ali)等[10]针对历史报纸文献资源提出一套融合计算机视觉与机器学习的处理流程，首先通过版面分析将报纸图像分割为单篇文章，实现文章级访问；在此基础上，进一步识别文章中的人物、事件等命名实体，并与Wikidata中的实体进行链接，为报纸资源建立语义级元数据并实现跨资源的关联利用。

总体而言，AI生成元数据应用于图书馆资源建设，能够有效提升效率、精确性和智能化水平，为用户提供更便捷和丰富的资源访问与利用途径，同时支持图书馆在数字化时代更好地管理和优化馆藏资源。

但是，目前关于这一主题的研究仍处于起步阶段，研究内容多集中于AI生成元数据的技术可行性和初步应用，且多是个案介绍，较少形成系统全面的实践案例研究，较少对应用场景的梳理和总结。基于此，本研究通过调研国内外图书馆AI生成元数据赋能资源建设的实践案例，总结不同的应用场景和成效，并据此提出进一步的实践建议，以期为图书馆资源数字化与智慧化转型提供参考。

3 案例调查与基本情况分析

本研究通过系统调研国内外多个案例，深入分析了AI生成元数据在图书馆资源建设中的实际应用。研究采用网络调查法与文献分析法相结合的方式，搜集已有研究、相关文献以及网络资源，以获取案例的详尽信息。案例的选择基于以下标准：(1)选取已有一定成果且资料较为详实的案例，以确保研究数据的可靠性和代表性。(2)注重选择不同地区、不同规模和不同类型的图书馆，以提升案例的多样性和广泛性。在此基础上，最终筛选了18个国外案例和2个国内案例，涵盖了不同实践类型的图书馆，确保研究的覆盖面。调研时间为2024年6月至10月，具体案例基本情况详见表1。

在对案例基本情况进行分析时，本研究借鉴美国国会图书馆实验室提出的人工智能规划框架(Artificial Intelligence Planning Framework)，该框架指出，AI项目实施的三个核心要素为：人员(People)、数据(Data)和模型(Model)[11]。其中，“人员”对应图书馆AI项目中的实施主体；“数据”即资源，指AI项目中图书馆所处理的各类馆藏资源对象；“模型”对应于AI技术与工具。

3.1 实施主体

在实施主体方面，本研究涉及图书馆类型包括国家图书馆、高校图书馆和省市公共图书馆三类。不同类型图书馆在技术与资源条件方面存在明显差异，直接影响其技术路径选择。

(1)国家图书馆。国家图书馆在资源描述与组织领域的AI技术应用处于领先地位，主要原因有：①馆藏资源丰富。国家图书馆得益于悠久的发展历史和出版物法定呈缴制度，通常拥有丰富的馆藏资源，如德国国家图书馆依据《德国国家图书馆法》和法定缴存条例，拥有自1913年以来出版的所有印刷资源。自2006年起，缴存范围扩展到数字出版物，数字资源的激增直接推动其探索机器自动分类的实

表1 AI生成元数据在图书馆资源建设领域的应用案例

序号	图书馆	项目名称	项目简介	应用技术	应用场景	实施步骤
1	瑞典国家图书馆	瑞典语大模型训练项目[12]	基于谷歌的BERT模型，结合瑞典语馆藏资源训练瑞典语文本识别模型KB-BERT及语音识别模型VoxRex	LLM；ASR	生成描述性元数据	瑞典语文本识别模型应用实例：提取政府文档主题[13]；识别数字化报纸结构特征（如标题、时间等）[14]；识别明信片内容（如人物、建筑、事件等）[15] 瑞典语语音识别模型应用实例：转录议会辩论音频并提取内容描述信息[16]
2	丹麦皇家图书馆	声音检索（Sound Search）项目[17]	通过语音识别模型Whisper将丹麦语广播和电视档案音视频资源转为文本，实现精准检索	ASR；NLP	生成描述性元数据	将1987以来的广播和电视等视听资源转换为可检索文本（2005年以前的资源为物理载体形式）；提取电视/广播的年代、频道等元数据；实现视听资源检索
3	俄克拉荷马州立大学图书馆	学位论文自动编目项目[18]	使用生成式AI为学位论文自动创建摘要和关键词字段	LLM	生成描述性元数据	评估生成式AI工具，包括Copilot、Claude、ChatGPT和Gemini的免费版及订阅版；通过提示词工程请求摘要和关键词；结果评估，Copilot订阅版模型效果最优
4	圣地亚哥州立大学图书馆	电子政府文档自动编目项目[19]	使用ChatGPT为政府文件生成摘要、非控制主题词字段	LLM	生成描述性元数据	利用Python提取政府文档PDF元数据，包括头标区、005－008、040、245、264、3XX、856字段；使用ChatGPT为PDF生成520、653字段；后续计划使用ChatGPT生成分面应用主题词（FAST）
5	芬兰国家图书馆	Annif项目[20]	创建Annif自动主题和分类标引的开源工具包，专注于文本文档的分类	NLP	主题标引与分类标引	集成多种文本分类算法，包括Maui、fastText、Omikuji和基于TensorFlow的神经网络模型；选择文档语料库和训练模型；为新文档生成主题与分类建议
6	德国国家图书馆	自动编目系统（Automatic Cataloguing System）项目[21]	利用AI技术对出版物进行自动主题与分类标引，进一步提高机器自动标引的质量	前期采用NLP，后期探索使用LLM	主题标引与分类标引	前期实践：采用文本分类算法实现资源的自动分类，生成杜威十进制分类号（DDC）和德国国家图书馆主题词规范档（GND） 现阶段：计划引入新的LLM模型，进一步提升自动主题分类的质量，并开发灵活、可重用的开源工具集成到编目流程
7	佛罗里达大学图书馆	机器辅助索引（Machine Aided Indexing，MAI）项目[22]	自动提取资源中的地名信息，为学位论文和特藏期刊添加地理位置主题词	NLP	主题标引与分类标引	与商业公司合作构建命名实体识别（Named Entity Recognition，NER）流程，从数字化馆藏中提取地名信息，创建佛罗里达地名分类词典（Thesaurus of Florida Place Names，FLGEO）；将识别出的地理位置主题词自动添加到学位论文和特藏期刊的主题字段元数据
8	中佛罗里达大学图书馆	FAST主题词自动分配项目[23]	通过OpenAI提供的大语言模型生成机构数据库中电子资源的FAST主题词	LLM	主题标引与分类标引	借助RAG技术，基于OpenAI的text-embedding-3-small模型将完整FAST主题词表构建为向量知识库；使用机构数据库中记录的标题和摘要在向量数据库中检索，获取语义最相近的前25个FAST主题词；利用ChatGPT综合标题、摘要和25个候选主题词，从中筛选出最相关的5个FAST主题词；将最终选定的FAST主题词输出至Excel文件供人工审核

续表

序号	图书馆	项目名称	项目简介	应用技术	应用场景	实施步骤
9	加州大学戴维斯分校图书馆	学位论文主题分类项目[24]	使用ChatGPT为学位论文分配国会图书馆主题词(Library of Congress Subject Headings, LCSH)	LLM	主题标引与分类标引	构建提示词，给出示例和要求；使用Python将提示词和学位论文的标题和摘要输入ChatGPT；对生成的LCSH主题词进行评估
10	香港浸会大学图书馆	AI辅助主题词分配实验[25]	探索和评估AI为图书分配LCSH主题词的应用效果可行性	LLM	主题标引与分类标引	选择2万条中医学类书目记录并导出，包括唯一标识符、题名和主题词，并转换为CSV格式；将数据上传至Poe.com平台，选择Claude模型构建机器人问答助手；测试模型生成主题词的效果，包括从现有记录中查找主题词以及为新记录生成主题词
11	美国国会图书馆	探索计算描述(Exploring Computational Description, ECD)项目[26]	测试使用多种AI技术创建电子图书MARC记录的有效性	NLP；LLM	创建书目记录	阶段一：测试5种NLP工具生成电子书元数据的效果 阶段二：基于开源LLM(如Llama2、Mistral、Gemma)，重点研究如何优化LLM提示词、微调LLM，并利用向量搜索匹配规范数据生成书目记录，并提供更多电子书用于训练和调整模型；输出BIBFRAME格式
12	比利时皇家图书馆	自动化编目流程项目[27]	利用AI模型和微软的Power Automate平台自动化图书编目流程	CV；NLP	创建书目记录	图书关键页面扫描；使用AI模型提取元数据；检索比利时、法国、德国等国家图书馆在线书目以及ISNI、VIAF等标识符数据库补充元数据；主题标引；地理名称识别；信息整合与元数据标准遵循；数据传输与输出；人工验证
13	广东省立中山图书馆	采编图灵项目[28]	采用基于神经网络的图书专用AI图像处理模型，实现印刷图书采编流程自动化	CV	创建书目记录	扫描图书题名页、版权页、封底、封面、书脊等关键信息页；应用图书专用AI模型识别ISBN、价格、题名等关键字段；自动套录联编数据
14	卡普兰诺大学图书馆	AI编目工具“CatalogerGPT”开发项目[29]	开发基于ChatGPT的CatalogerGPT编目工具，简化创建原始MARC记录的过程	LLM	创建书目记录	提供关键书页(封面、题名页、版权页和目录)的文本内容或图像；生成MARC记录；编目人员审查并提供反馈；模型根据反馈进行改进 版本更新：CatalogerGPT 2.0正在开发中，改进内容包括支持更多输入/输出格式，增加访问本地书目记录以及其他系统的功能
15	卢森堡国家图书馆	eLuxembourgensia项目[30]	基于数字化馆藏历史报纸内容，利用ChatGPT实现对馆藏报纸内容的语义对话检索功能	CV；LLM	元数据语义增强	对馆藏印刷报纸和期刊数字化，结合CV技术识别版面结构，从800万篇文章中提取出950万个段落，并为每个段落绑定如下基础元数据：标题、日期及URL(定位该段的唯一链接)等；使用OpenAI的Ada模型对每个段落生成语义向量，并与上述元数据一并存入向量数据库；实现用户以自然语言提问后，在语义向量库中进行相似度匹配，检索出最相关段落，并通过绑定的元数据返回段落的标题、日期和来源

续表

序号	图书馆	项目名称	项目简介	应用技术	应用场景	实施步骤
16	挪威国家图书馆	Maken项目[31]	通过AI驱动的相似性算法，基于图像内容特征，为用户提供馆藏图书封面和馆藏历史照片的相似度推荐	CV；LLM	元数据语义增强	利用馆藏资源训练挪威语大模型；使用挪威语大模型识别约50万张图书封面和50万张历史照片内容特征，如场景、颜色、人物分布等；将内容特征转化为向量嵌入Elastic-Search索引系统；系统计算向量之间的相似度，判断图书封面、历史照片之间的相似性；实现用户点击图书封面或历史照片时，可基于内容相似性自动推荐与之相近的其他资源
17	大英国家图书馆	MapReader项目[32]	利用CV技术解析数字化历史地图，开发开源的地图图像分析工具	CV	元数据语义增强	将地图图像内容转换为可处理的文本数据；集成地点实体链接功能，对地图图像进行位置信息标注；开发开源地图检索工具，实现地图数据的智能化检索与应用
18	新加坡国家图书馆	链接数据管理系统项目[33]	采用NLP技术提取实体并为实体添加URI，提供统一的资源检索与访问	NLP	元数据语义增强	利用NER工具从数字资源中提取人物、组织和地点等实体；开发链接数据管理系统，统一管理多来源元数据，并为实体添加URI，支持链接数据服务
19	斯坦福大学图书馆	物种发生（Species Occurrences，SPOC）项目[34]	通过解析学位论文PDF，提取物种属种、地点、时间等海洋生物观察数据	NLP	元数据语义增强	使用开源工具（GROBID）解析学位论文PDF，提取全文和元数据；利用NER工具（spaCy）提取论文中的物种属种、地点、时间等实体；通过权威机构数据库验证实体准确性；上传实体数据至全球生物多样性信息设施（Global Biodiversity Information Facility，GBIF）
20	埃默里大学图书馆	Wayfinder项目[35]	使用ChatGPT从《非裔美国人报纸和期刊：国家书目》自动识别及提取关键元数据	CV；LLM	元数据语义增强	将1998年出版的《非裔美国人报纸和期刊：国家书目》数字化；使用ChatGPT自动识别和提取图书内容中每一条书目的关键元数据，如标题、著者、出版日期和主题；将识别到的名称、地点等实体链接到维基数据，构建在线互动式书目网页

注：LLM（Large Language Model），即大语言模型；NLP（Natural Language Processing），即自然语言处理；CV（Computer Vision），即计算机视觉；ASR（Automatic Speech Recognition），即自动语音识别。

践[36]。②技术力量强大。许多国家图书馆具备强大的技术实力，设立了专门的AI实验室，如美国、瑞典和挪威等国家图书馆均已建立AI实验室，积极参与图书馆领域大语言模型的研发及原生AI应用的开发。③资源描述与组织标准经验丰富。国家图书馆在资源描述与组织标准方面积累了丰富的经验，通常作为资源描述格式、著录标准的主要制定者，且主导建设并持续维护主题词表、分类法等知识组织系统，因而具有在资源描述与组织上开展AI应用探索的优势。

（2）高校图书馆。高校图书馆更多依赖外部技术支持来提升个性化资源建设能力。具体表现在：①与外部技术公司合作。部分高校图书馆通过与专业技术公司合作，定制AI技术应用。如佛罗里达大学图书馆与信息服务公司Access Innovations合作，定制开发命名实体识别（Named Entity Recognition，NER）流程提取地名实体，实现资源的自动地理主题标引。②使用通用型AI工具。部分高校图书馆则采用通用AI工具生成元数据，如斯坦福大学、圣地亚哥州立大学、俄克拉荷马州立大学和加州大学戴维斯分校等高校的图书馆利用OpenAI、谷歌、微软等公司提供的生成式AI工具支持资源建设和元数据生成工作。

（3）省市公共图书馆。与国家图书馆类似，省市

公共图书馆整合其资源基础与技术优势，以满足区域资源的自动化加工与智能服务需求。如广东省立中山图书馆结合区域需求优化编目流程，实施“采编图灵”项目。

3.2 资源对象

本研究所涵盖的案例实践中，资源对象按来源可分为原生数字资源和实体资源的数字化形式两类。资源类型的差异对图书馆应用AI技术生成元数据的难易度有直接影响。

(1)原生数字资源。原生数字资源不需要繁琐的数字化过程(如扫描、转录等)来转换为可用格式，其技术实施的复杂性和成本较低，通常成为高校图书馆优先探索AI技术应用的对象。典型的原生数字资源包括学位论文、期刊论文和政府文件等。①以学位论文与期刊论文为对象的AI应用。典型案例如加州大学戴维斯分校图书馆，其利用ChatGPT对学位论文进行主题自动标引；中佛罗里达大学图书馆则为其机构知识库中的期刊论文自动生成FAST主题词。②以政府文件为对象的AI应用。典型案例如圣地亚哥州立大学图书馆，其利用ChatGPT为政府文件自动生成摘要字段和非控制主题词字段。

(2)实体资源的数字化形式。实体资源的数字化通常由大型公共图书馆主导，涉及书籍、报纸、地图及电视、广播等视听资源的物理保存载体等多种形式。这类资源通过扫描或转录等技术实现数字化后，再应用AI技术进行元数据提取和处理。①以印刷图书为对象的AI技术应用。典型案例如比利时皇家图书馆和广东省立中山图书馆，它们通过扫描印刷图书的关键页面信息，提取相关著录元数据。②以地图资源为对象的AI技术应用。典型案例如大英图书馆，其对历史地图进行扫描和数字化解析，将图像内容转换为机器可读数据。③以报纸资源为对象的AI技术应用。典型案例如瑞典国家图书馆，其对2014年至今所有瑞典出版的报纸进行扫描，利用瑞典语模型KB－BERT对数字化后的文本进行识别并提取相关字段。④以视听资源的物理保存载体为对象的AI技术应用。典型案例如丹麦皇家图书馆，其在2005年之前通过磁带保存广播和电视资源。在声音检索项目中，利用语音识别技术将这些磁带内容转录为数字文本，并提取年代、频道等元数据字段。

3.3 应用技术

人工智能是指通过多种技术，使机器能够从数据中学习并不断改进，从而模拟人类智能行为[37]。根据《国家人工智能产业综合标准化体系建设指南》[38]对AI技术的分类，结合图书馆元数据生成任务的特点，本研究将调研案例中应用的AI技术类型归纳为：大语言模型、自然语言处理、计算机视觉以及自动语音识别。其中，大语言模型作为自然语言处理的重要组成部分，由于其快速发展和广泛应用，在案例中被专门提及较多，因而在分类时将其单独列出。

(1)大语言模型(LLM)。凭借强大的语言理解与生成能力，LLM技术在调研案例中应用最为广泛，擅长处理各类文本资源内容，在元数据生成、摘要提取、主题标引等任务中表现出较高适应性。同时，其适用范围正逐步拓展至多模态资源的处理。如瑞典国家图书馆结合瑞典语馆藏资源训练瑞典语文本识别模型与语音识别模型，探索将LLM应用于图像、文本、音视频等多种资源的识别与处理。

(2)自然语言处理(NLP)。NLP技术能够深入分析和理解文本资源，主要应用于文本信息与实体提取以及文本分类两大方向。在文本信息与实体提取方面，可提取资源基础描述信息或实体(如人、地点、物品)，同时结合外部数据(如Wikidata)为实体添加URI。在文本分类方面，调研案例多采用芬兰国家图书馆开发的Annif工具。该工具集成多种文本分类算法，可为文本资源自动分配主题词或分类号，是NLP在文本分类中的典型应用。

(3)计算机视觉(CV)。CV技术主要用于资源的数字化加工与图像内容理解，应用方向包括通过OCR将资源数字化为可编辑文本；通过物体检测和图像分割等方式标注图像关键元素。如卢森堡国家图书馆在处理印刷报纸和期刊馆藏时，首先通过OCR实现内容数字化，然后结合版面结构解析自动提取文章区块、标题与日期等基础元数据，为后续语义检索与对话访问提供基础。

(4)自动语音识别(ASR)。自动语音识别技术主要用于将音视频等资源转录为文本，方便后续的文本分析、元数据提取等。如丹麦皇家图书馆通过语音识别模型Whisper转录1987年以来的广播和电视资源为可检索文本，并提取年代、频道等元数据，实现视听资源的内容检索与利用。

4 应用场景

根据文献分析和案例调研，AI生成元数据在图书馆资源建设环境中的应用场景主要有生成描述性元数据、主题标引与分类标引、创建书目记录以及元数据语义增强四种类型。其中，“生成描述性元数据”与“创建书目记录”在任务上有所交叉，“创建书目记录”侧重传统编目领域图书的MARC书目记录的生成，“生成描述性元数据”指为文本、图像、音频、视频等各类资源生成基本的描述性元数据（不一定是MARC元数据）。

4.1 生成描述性元数据

调研案例中，部分图书馆（见表2）开始应用AI技术为文本、图像、音频、视频等资源生成题名、年代、摘要、关键词等描述性元数据。针对文本资源，AI技术可自动生成多种描述性元数据内容，如瑞典国家图书馆为馆藏政府文件生成政府部门、时间和聚类主题元数据；圣地亚哥州立大学图书馆生成政府文件的摘要和非控制主题词；俄克拉荷马州立大学图书馆生成学位论文的摘要和关键词。对于图像资源，瑞典国家图书馆识别馆藏明信片内容，提取人物、建筑、事件等元素。对于音视频资源，可以利用智能语音技术将音视频资源转录为文本并提取元数据，如丹麦皇家图书馆提取音视频资源的年代、频道等元数据。

表2 部分图书馆描述性元数据生成的具体内容

图书馆	资源对象	描述性元数据内容
瑞典国家图书馆	政府文件	政府部门、时间、聚类主题
瑞典国家图书馆	政府议会音频	姓名、性别、年份和选区等元数据
瑞典国家图书馆	明信片	人物、建筑、事件等元素
丹麦皇家图书馆	电视和广播音视频	年代、电视/广播的频道
俄克拉荷马州立大学图书馆	学位论文	摘要、关键词
圣地亚哥州立大学图书馆	政府文件	摘要、非控制主题词

4.2 主题标引与分类标引

在图书馆资源管理中，资源标引工作通常包含主题标引与分类标引两部分。主题标引是指根据文献内容为其分配相关的主题词或主题概念，以便能够通过主题词检索相关资料。这一过程通常依托特定的主题词表或词典，如美国国会图书馆主题词（LCSH）、分面应用主题词（FAST）和汉语主题词表等。分类标引则根据资源内容及特定的分类体系，如杜威十进分类法（DDC）、美国国会图书馆分类法（LCC）和中国图书馆分类法，为资源分配分类号，以便将其系统地组织到特定知识领域或类别中。本研究调研的案例中，部分图书馆（见表3）已开始利用AI技术进行主题标引和分类标引的实践。

目前自动标引实践已形成专门的标引工具，如Annif，该工具集成多种文本分类方法，在图书馆领域内外得到广泛使用。在图书馆领域内部，如调研案例中的芬兰国家图书馆和德国国家图书馆已采用Annif进行自动分类；美国国会图书馆和比利时皇家图书馆也将其纳入书目记录自动生成系统，用于主题标引。此外，多个机构知识库，如于韦斯屈莱大学和莱布尼茨经济信息中心，也利用Annif协助论文主题索引。在图书馆领域外，芬兰广播公司Yle使用Annif为在线新闻文章分配标签。

除了利用专门的标引工具外，另一种方式是使用生成式AI工具（如ChatGPT和Claude）进行自动标引。生成式AI工具因其较高的易用性和灵活性，常作为高校图书馆探索自动标引的首选工具，如中佛罗里达大学图书馆选择ChatGPT生成论文的FAST主题词；加州大学戴维斯分校图书馆选择ChatGPT生成论文的LCSH主题词；香港浸会大学图书馆选择Claude生成图书的LCSH主题词。

表3 部分图书馆自动标引实践的具体内容

图书馆	资源对象	AI工具	标引类型	
			主题标引	分类标引
芬兰国家图书馆	所有文本型资源	Annif	LCSH	DDC
德国国家图书馆	所有文本型资源	Annif	GND	DDC
中佛罗里达大学图书馆	机构知识库论文	ChatGPT	FAST	—
加州大学戴维斯分校图书馆	学位论文	ChatGPT	LCSH	—
香港浸会大学图书馆	图书	Claude	LCSH	—

4.3 创建书目记录

自动创建符合编目标准的书目记录是AI技术在图书馆应用中的重要方向。该应用场景通过AI技术生成结构化、标准化的书目记录，实现对图书资源的自动编目，最接近图书馆传统编目工作内容。调研案例中，已有图书馆形成了较为完整的自动化编目流程以及专门的智能辅助编目工具。

(1)自动化编目流程。比利时皇家图书馆致力于开发和应用自动化编目流程，基于微软的Power Automate平台成功实施了基于AI技术的编目工作流程，关键步骤见表4。当前，该系统正计划将自动编目范围扩展至包括期刊、地图、乐谱等在内的其他资源类型。

表4 比利时皇家图书馆自动编目流程

步骤	具体内容
数据采集与扫描	使用带有自动边界检测功能的应用程序(如Microsoft Lens或Adobe Scan)扫描关键页面，包括题名页、版权页和封底
使用AI模型提取数据	题名页信息检测模型(AITitlePage)：识别题名、作者、出版商、年份和出版地点等元数据 版权页信息检测模型(AIColophon)：识别ISBN、出版商、年份、版权信息和原始题名等元数据 封底信息检测模型(GetText)：从封底提取与书籍分类相关的内容
书目记录检索	通过识别到的ISBN，执行三个HTTP请求，检索比利时皇家图书馆、法国国家图书馆和德国国家图书馆的书目记录
HTTP查询	国际标准名称标识符(ISNI)查询：使用作者姓名和题名向ISNI发起HTTP请求以获取名称唯一标识符 虚拟国际规范文档(VIAF)查询：向VIAF数据库发送HTTP请求，使用作者姓名和题名检索VIAF标识符和作品的永久链接
主题标引	基于封底文本或外部记录中的摘要等信息，通过Annif和预训练的微软模型进行文本分类
地理名称识别	从封底文本中提取地名实体，并通过GeoNames查询确认
信息整合与标准遵循	收集所有必要信息，包括自动检测结果、HTTP查询结果、实体提取和主题标引，生成符合MARC 21和BIBFRAME元数据标准的元数据
数据传输与记录输出	生成的文件通过API直接发送到图书馆目录，或传输到服务器并自动导入图书馆管理系统
人工验证	编目员对自动创建的记录进行验证和调整

(2)智能辅助编目工具的开发。卡普兰诺大学图书馆引入RAG，将编目标准转化为向量形式存储于向量知识库，开发基于ChatGPT的智能编目工具CatalogerGPT，简化编目员创建MARC记录的过程。CatalogerGPT工具提供了创建原始MARC记录、建议主题词和分类号、MARC记录错误检测及外语资料编目等功能(见表5)。目前，该工具正在卡普兰诺大学图书馆的小规模馆藏中进行测试，以优化其功能及实际应用效果。

表5 CatalogerGPT的主要功能

功能	说明
创建原始MARC记录	接受文本、图像、PDF输入，生成符合标准的MARC记录
主题词和分类号建议	提供LCC、LCSH以及DDC分类号建议
错误检测	检查MARC记录中的潜在错误并提供改进建议
外语资料编目	支持多种语言的书籍编目

4.4 元数据语义增强

语义增强是指采用各种技术和方法为资源添加语义元数据以有效增加数据价值的策略[39]。通过对调研案例的分析，当前AI生成元数据的语义增强主要集中在以下两个方向。

(1)实体提取与关联数据结合。基于NER技术从资源中提取人物、地名、组织等实体，并链接至外部数据源(如Wikidata)，为元数据添加实体的唯一标识，从而实现对具体概念的准确指向，并增强与其他系统之间的关联能力。如新加坡国家图书馆通过实体识别，并在数据字段中增加实体的URI链接，实现多来源元数据统一管理。类似地，大英图书馆、斯坦福大学图书馆和埃默里大学图书馆也通过实体提取和外部数据的关联，进一步强化了元数据的互联性。

(2)语义向量嵌入与相似度计算。通过将资源内容及其元数据信息转换为语义向量，在原有元数据基础上引入基于资源内容生成的向量表示，支持基于语义相似度的资源检索与推荐。如卢森堡国家图书馆将数字化报纸按段落划分，使用嵌入模型生成段落语义向量，并为每段绑定日期、来源、URL等元数据，用户以自然语言提问后，通过向量相似度匹配出最相关段落，并通过绑定的元数据返回段落标题、日期和来源。挪威国家图书馆的Maken项目对图像资源(图书封面和历史照片)采用类似的处理方式：通过提取图像内容特征并转化为向量嵌入索引系统，用户点击图书封面或历史照片时，可基于向量之间的相似度推荐与之相近的其他资源。

5 赋能成效

从案例实践的进展来看，AI生成元数据已在图书馆资源建设中展现出重要的赋能效应，主要体现在以下四个方面。

(1)非结构化与多模态资源的规模化揭示。AI技术在非结构化和多模态资源的规模化发现中发挥了重要作用。相较于结构化文本，音视频、图像、地图等资源因格式多样、结构复杂，长期难以实现有效的描述与组织。AI技术能够自动提取资源特征，使大量非结构化资源得以描述和发现。如丹麦皇家图书馆提取1987年以来的广播和电视资源的年代、频道等元数据；瑞典国家图书馆识别馆藏17409张明信片内容，为音视频、图像等特色资源的规模化揭示提供了现实可行的路径。

(2)优化编目工作流程。AI技术在优化图书馆元数据生成流程中展现出显著优势。传统编目的著录环节，编目员需要手工著录大量描述性信息，而AI技术的引入通过自动化流程减少人工编目工作量，使编目流程更加高效。如比利时皇家图书馆构建自动编目流程，由AI扫描图书题名、版权、封底等页面，生成符合MARC 21和BIBFRAME元数据标准的元数据。广东省立中山图书馆“采编图灵”项目应用图书专用AI模型识别来自图书题名页、版权页、封底、封面、书脊等关键信息页的ISBN、价格、题名等关键字段，通过ISBN自动套录联编数据，大幅提升编目效率，图书上架时效从传统模式的两三天压缩至十分钟[40]。此外，在传统编目的主题标引环节，编目员往往需要阅读图书内容、查阅受控词表，并结合专业判断分配合适主题词，整个过程耗时费力。AI技术的引入可在此过程中发挥辅助作用，通过快速分析图书封面、目录、前言等关键页面，提取核心内容信息，初步生成主题建议，供人工复核与修订。如德国国家图书馆借助Annif工具自动进行内容分析和主题标引，实现对德语、英语资源的批量处理，2023年为15万余种数字出版物(包括德语和英语)自动分配GND主题词[41]，大幅提升主题标引效率。

(3)支持科研与教学。AI生成元数据为学术研究提供了更精准的资源匹配与分析能力。传统馆藏组织方式通常基于固定分类体系，难以满足跨学科

研究和深度数据挖掘的需求。AI技术通过实体识别、知识关联等方式，使资源的组织更加灵活和智能，能够精准匹配研究人员的专业化需求，实现更高效的学术信息检索与利用。如大英图书馆通过识别地图中涉及的地理位置，支持历史地理学研究；斯坦福大学图书馆从学位论文中自动提取海洋生物观察数据，并实现可视化展示，为生物多样性研究提供支持；埃默里大学图书馆将馆藏资源精准匹配到课程内容，促进馆藏资源与教学融合。

(4)提升资源服务个性化与智能化。AI技术通过自然语言检索和智能推荐等手段，使图书馆资源服务更加个性化和智能化。传统的馆藏检索主要依赖规则化的字段查询，而AI技术以自然语言问答方式使检索更直观，增强用户与资源的交互体验。如卢森堡国家图书馆部署了聊天机器人，支持用户通过自然语言提问，如用户以对话形成提问“请总结1944年1月1日发生的事件”。系统将从数字化的八百万篇报纸文章中检索相关答案，并支持用户根据回复进一步开展对话式提问。瑞典国家图书馆建立支持语义搜索的馆藏明信片网站，用户可输入如“含教堂的图片”的提示，系统将返回相关明信片资源，实现更精准的内容检索。此外，AI技术还可实现基于用户浏览内容的智能推荐，增强个性化体验。如挪威国家图书馆在检索界面嵌入相似性推荐按钮，当用户点击一本图书封面或一张历史照片时，系统可根据图像内容相似度推荐相关资源，提升资源发现的效率与相关性。

需要说明的是，尽管AI生成元数据带来了上述赋能成效，但AI在元数据生成中仍然发挥的是辅助作用，目前仍作为辅助工具，主要体现在：①从应用的图书馆数量看，目前实际应用AI生成元数据的图书馆仅是少数，大多仍处于初步探索与实验阶段；②从资源的覆盖面看，图书馆大多以项目形式在小范围资源上开展AI生成元数据的探索与实验，并未拓展至本馆所有资源类型；③从AI生成元数据质量看，由于AI技术尚不成熟，其生成的元数据仍存在一定不准确性，元数据质量的控制依然离不开馆员，尤其是在主题和分类标引方面，更需依赖馆员的专业知识。因此，需要理性看待AI在元数据生成中的作用，既要充分发挥其优势，又要强化馆员与AI的协同合作。

6 建议及启示

随着AI技术的发展，一些公共图书馆和高校图书馆的先行者已积极探索将AI技术融入元数据生成流程，以增强资源建设的效能。他们在资源对象的选择、技术工具的应用以及应用场景的构建等方面，正勾勒出人机协同生成元数据的初步发展路径，但也面临生成的元数据准确性不足、技术适配难度高、人员能力有待提升等现实挑战。尽管如此，这些前沿实践仍然为图书馆界树立了新的标杆，也为其他图书馆在资源建设智能化升级的过程中提供了宝贵的参考。基于已有实践的调研分析，本研究从制定AI生成元数据技术应用方案和保障AI生成元数据赋能成效两方面提出建议。

6.1 制定AI生成元数据技术应用方案

通过案例调研可以发现，不同图书馆在资源类型、技术选择与元数据需求等方面呈现出多样化特征。因此，图书馆应结合自身条件，制定契合实际的AI生成元数据技术应用方案。结合案例实践，本研究提出以下三种常见技术应用方案。

(1)基于大语言模型的提示词工程

元数据的自动生成是提升图书馆资源管理效能的重要手段。图书馆通常借助大语言模型的自然语言处理能力，通过提示词工程自动生成题名、摘要、关键词、主题词等字段。由于具有技术门槛低、部署成本小等优势，该方案已成为当前AI生成元数据实践中最为常用的技术应用方案之一。在实际应用中，提示词的设计是关键。图书馆可参考一些通用的提示语框架，如DOAE提示语框架[42]，其将提示内容划分为任务描述(Description)、输出要求(Output)、注释(Annotation)和示例(Example)四个要素，其中任务描述与输出要求为必选要素，注释与示例为可选要素。以圣地亚哥州立大学图书馆为例，该馆使用ChatGPT处理政府文档，通过结构化提示词生成摘要(520字段)和非控制主题词(653字段)，在提示设计中包含任务描述、输出要求和注释三项内容(采用零样本方式，未提供示例)，其生成非控制主题词的提示内容见表6。对于资源结构复杂或专业性较强的任务，图书馆可视情况引入少量样本示例，通过示例数据提升模型生成元数据的准确性。

表6 DOAE提示语框架及参考样例

提示语要素	功能说明	圣地亚哥州立大学图书馆生成政府文档的非控主题词的提示词举例
任务描述	模型需完成的具体任务	基于提供的政府文档内容，生成非控主题词
输出要求	设定输出格式与表达规范	提取3－5个核心关键词，用“,”分割，仅回复关键词
注释	提供背景或资源内容	提供政府文档正文、标题或关键章节内容
示例	提供零样本或少样本示例	无

(2)基于机器学习的领域数据训练

图书馆业务场景对元数据的规范性与稳定性要求较高，通用模型的输出效果往往难以满足实际可用的质量标准。基于高质量的图书馆专业数据训练机器学习模型成为提升AI生成元数据质量的常见方法。模型训练在此过程中起着决定性作用。具体过程通常包括以下几个步骤：①数据准备：根据需求收集高质量已标注数据，如图书馆的馆藏数据，包括文本、图像、音频、视频等各类资源。实际应结合具体的任务需求选择适配的数据，如训练模型生成MARC记录，需准备图书资源内容(全文或关键页)，以及对应的完整MARC记录，包括题名、责任项、出版项、主题词、分类号等字段。②数据预处理：对原始数据进行清洗与格式标准化，满足模型训练的输入要求。如删除无效或缺失关键信息(如题名、责任者、分类号等字段)的记录，将原始标注数据(如MARC格式)转换为结构化格式(如CSV或JSON)。③模型选择：根据任务特性及模型性能，选择适用的模型。如文本字段生成(如题名、出版项等)，可优先选择预训练语言模型(如BERT、GPT系列)；主题或分类标引可采用通用文本分类模型或专用的主题标引工具(Annif)。④数据训练：常将数据集划分为训练集、测试集与验证集(如按8∶1∶1比例)，使用训练集对模型进行训练，测试集对模型性能进行测试，验证集用于评估模型表现。⑤评估与优化：常通过准确率、召回率、F1分数(精确度和召回率的调和平均值)等定量指标，以及人工定性评估的方式，对模型生成元数据进行质量评估。若测试结果未达预期，可进一步调整模型参数、优化数据结构，或增加训练数据以提升性能。以美国国会图书馆的ECD项目(阶段一)为例(见表7)，其以图书馆书目系统里的MARC数据为基础训练多种AI模型，评估模型生成书目元数据的效果。

(3)基于知识图谱的检索增强生成

在语义网与关联数据环境下，图书馆元数据工作正从传统的著录与标引任务延伸至语义增强与智能发现等更具深度的应用方向。为提升元数据在语义准确性、上下文理解与知识关联方面的能力，基于知识图谱的检索增强生成技术日益受到关注。该技术主要包括以下步骤(见图1)：①外部知识库构建。构建两个知识库：一是基于结构化数据构建知识图谱数据库，用于存储实体及其关系信息；二是将非结构化资源向量化，构建向量数据库，用于捕捉语义相似度。②检索。在接收到用户查询后，系统同时在知识图谱数据库和向量数据库检索。③语义增强。将检索到的知识图谱信息(如实体及关系)和向量数据库信息(如语义上最接近的文档段落)与用户查询组合，形成增强后的上下文。④大语言模型生成。在增强后的上下文基础上，调用大语言模型进行回答。

其中，向量数据库的建设是实现检索增强生成的重要基础。当前主要通过本地资源向量化构建语义库，将馆藏资源嵌入为向量，利用其语义相似性增强元数据的上下文关联。如卢森堡国家图书馆将数字化报纸按段落划分，使用嵌入模型生成段落语义向量，并为每段绑定日期、来源、URL等元数据，用户以自然语言提问后，通过向量相似度匹配出最相关段落，并通过绑定的元数据返回段落标题、日期和来源。然而，基于向量数据库的检索主要依赖语义相似度，往往忽视了资源之间的结构化关系，如不同资源之间的实体关联难以被准确识别和利用。为进一步增强元数据的语义，可以引入知识图谱作为外部知识来源，构建基于知识图谱的检索增强生成模型。知识图谱以结构化方式组织实体及其间的关系，具备良好的语义解释力和可扩展性。可选用通用知识图谱资源(如Wikidata、GeoNames等)作为支撑，也可以图书馆领域的主题词表(如LCSH、FAST)和名称规范数据(如VIAF)等为基础构建面向图书馆应用场景的专用知识图谱，增强AI生成元数据的语义背景与实体间的逻辑关系。

表7　美国国会图书馆ECD项目(阶段一)流程

步骤	具体内容
数据准备	数据集包含共计23130个电子书文件，格式为PDF和EPUB，总数据约250GB，主要语种为英语。数据分为四个子集，分别是：CIP子集(13802个条目)、开放获取子集(5835个条目)、电子存档电子书子集(403个条目)和法律报告子集(3750个条目)。每个子集的电子书文件存储在一个文件夹中，并包含一个单独的MarcXML文件，记录该子集内所有电子书的书目信息
数据预处理	使用多种工具(如Pandoc或Calibre)将PDF和EPUB格式转换为纯文本； 字符集编码的标准化，将未编码为unicode的文件转换为unicode； 并规范空白字符
模型/工具应用	GROBID模型，用于生成：标题、作者姓名、唯一标识符(DOI、ISSN、ISBN等)、关键词、版权信息
	Annif工具，用于生成：体裁/形式(使用LCFGT作为词表)、主题词(使用LCSH作为词表)
	NLP工具spaCy和BERT系列模型(包括DistilBERT、RoBERTa等)，用于生成：标题、作者姓名、唯一标识符、发行日期、创建日期、体裁/形式(使用LCFGT作为词表)、主题词(使用LCSH作为词表)
	带布局特征的NLP工具(从电子书布局特征，如文本大小、页面位置等提取信息)，用于生成：作者姓名、唯一标识符、发行日期、创建日期
评估与改进	将模型预测的元数据与原始MARC记录中的字段进行比较来评估模型的性能； 编目人员对模型生成的内容进行人工审核，提供质量评估反馈以改进模型
实验结果	部分字段——如唯一标识符(LCCN、ISBN)、作者和题名——提取准确率约为85%，而其他字段如主题词、体裁/形式和日期的生成准确率仅在50%或更低

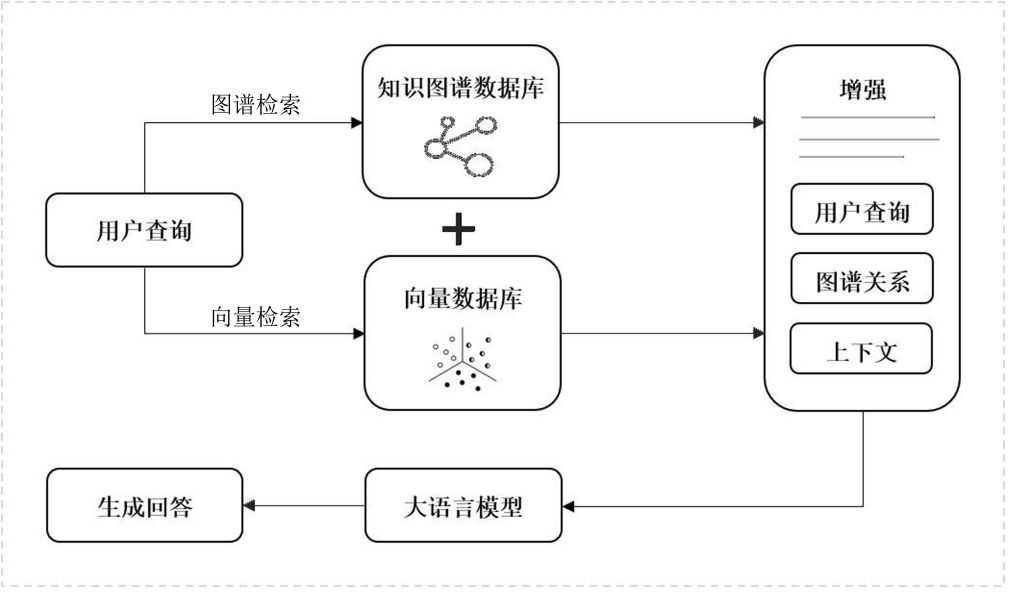

图1　基于知识图谱的检索增强生成方案

6.2　保障AI生成元数据赋能成效

AI生成元数据在图书馆资源建设中的应用，不仅可以促进非结构化与多模态资源的规模化揭示，优化编目工作流程，还增强了资源组织对科研与教学需求的适应能力，提升了资源服务的智能化与个性化水平。为持续发挥AI生成元数据在资源建设方面的效用，图书馆可从战略规划、质量监控、人才能力三方面构建保障机制，确保AI生成元数据赋能成效的长期可持续性。

(1)战略规划保障。AI生成元数据的应用不仅

是一项技术革新，更意味着图书馆元数据生产范式向智能化、协同化方向的系统性转型。图书馆应将AI生成元数据的应用纳入自身发展战略，以实现资源建设、管理与服务的持续优化。如德国国家图书馆于2021年启动“自动编目系统”项目即是作为德国国家人工智能战略组成部分[43]，体现出德国国家图书馆在数字化转型和智能化服务方面的前瞻性布局。当前，图书馆正处于“十五五”规划的关键阶段，智慧图书馆建设已成为推动行业高质量发展的重要方向。图书馆应结合数字化转型与智慧图书馆建设总体目标，制定AI技术在本地元数据管理中的发展方向，推动AI技术与元数据应用的有序融合。

(2)质量监控保障。AI生成元数据的质量直接影响图书馆信息组织与检索效果，因此需建立完善的质量监控和评估体系，确保元数据的准确性、一致性和可用性。图书馆可采用自动检测＋人工监督＋反馈优化的模式，建立动态质量监控机制，持续提升元数据质量。如比利时皇家图书馆在自动化编目流程中结合人工校验与机器检测，利用开源软件QA Catalogue检测数据结构，对质量问题进行人工修正，优化编目流程，形成持续改进的循环。美国国会图书馆通过构建人机协作编目模式“人在回路”(Human in the Loop，HITL)，建立机器与人工协同的工作流程，并开发系统原型，使AI生成的元数据经过人工复核与调整，数据质量符合编目标准。

(3)人才能力保障。注重提升资源建设馆员在AI技术领域的能力。生成式AI的快速发展正在从多个方面改变图书馆元数据工作的方式，包括自动生成元数据、智能辅助元数据质量控制等方面，理解与应用AI技术及大语言模型已成为元数据实践领域的核心能力之一[44]。图书馆需要重视资源建设馆员AI应用能力的培养和提升，可参考行业通用的人工智能素养框架，如面向图书馆员的人工智能素养框架[45]培养馆员AI素养。同时，可根据图书馆业务需求，鼓励馆员成立AI应用与学习小组，通过有组织的小组式、团队式学习，并结合具体业务实践的探索，快速提升馆员的AI应用技能[46]。

7 结语

AI生成元数据不仅是技术发展的必然结果，更是图书馆资源建设创新与发展的重要驱动力。本研究通过调研分析国内外20个图书馆应用AI生成元数据赋能资源建设的实践案例，为国内图书馆提升AI应用能力和资源管理水平提供参考。目前许多项目仍处于起步探索阶段，还未形成较为系统性的实施路径与模式，期待随着AI技术的发展及图书馆应用AI技术能力的提高，AI生成元数据在资源建设中的实践路径与模式更加完善。

参考文献

1 American Libraries Association. Evolving AI strategies in libraries: insights from two polls of ARL member representatives over nine months[EB/OL]. [2024－12－26]. https://www.arl.org/resources/evolving-ai-strategies-in-libraries-insights-from-two-polls-of-arl-member-representatives-over-nine-months/.

2 International Federation of Library Associations and Institutions. Developing a library strategic response to Artificial Intelligence [EB/OL]. [2024－12－26]. https://www.ifla.org/g/ai/developing-a-library-strategic-response-to-artificial-intelligence/.

3 Lin C H, Chiu D K W, Lam K T. Hong Kong academic librarians' attitudes toward robotic process automation[J]. Library Hi Tech, 2022, 42(3): 991－1014.

4 Kleppe M, Veldhoen S, Waal-Gentenaar M E T, et al. Exploration possibilities automated generation of metadata [EB/OL]. [2024－12－26]. https://doi.org/10.5281/zenodo.3375192.

5 Golub K, Suominen O, Mohammed A T, et al. Automated Dewey Decimal Classification of Swedish library metadata using Annif software [J]. Journal of Documentation, 2024, 80(5): 1057－1079.

6 戎璐. 面向图书自动分类的大语言模型提示学习研究[J]. 图书馆学研究, 2024(1): 86－103.

7 Brzustowicz R. From ChatGPT to CatGPT: the implications of Artificial Intelligence on library cataloging[J]. Information Technology and Libraries, 2023, 42(3) [2025－06－24]. https://doi.org/10.5860/ital.v42i3.16295.

8 Bodenhamer J. The Reliability and usability of ChatGPT for library metadata[EB/OL]. [2024－12－26]. https://hdl.handle.net/11244/339626.

9 Greenly G. CatalogerGPT[EB/OL]. [2024－07－07]. https://glengreenly.wixsite.com/catalogergpt.

10 Ali D, Milleville K, Verstockt S, et al. Computer vision and machine learning approaches for metadata enrichment to improve search ability of historical newspaper collections[J]. Journal of Documentation, 2024, 80(5): 1031－1056.

11 Library of Congress. Introducing the LC Labs Artificial Intelligence Planning Framework [EB/OL]. [2025－05－25]. https://blogs.loc.gov/thesignal/2023/11/introducing-the-lc-labs-artificial-intelligence-planning-framework/.

12 National Library of Sweden. Sweden library AI open source projects[EB/OL]. [2024－12－26]. https://blogs.nvidia.com/blog/sweden-library-ai-open-source/.

13 National Library of Sweden. Topic models for Sous[EB/OL]. [2024－12－26]. https://kb-labb.github.io/posts/2021－05－04-topic-models-for-sous/.

14 National Library of Sweden. Swedish newspapers collection [EB/OL]. [2024－12－26]. https://tidningar.kb.se/.
15 National Library of Sweden. Bildsok：a search tool for images [EB/OL]. [2024－12－26]. https://lab.kb.se/bildsok/.
16 National Library of Sweden. RixVox：a swedish speech corpus [EB/OL]. [2024－12－26]. https://kb-labb.github.io/posts/2023－03－09-rixvox-a-swedish-speech-corpus/.
17 Royal Danish Library. Sound search：audio search tool [EB/OL]. [2024－12－26]. https://labs.statsbiblioteket.dk/sound-search/.
18 Oklahoma State University. Teaching student workers to use Generative AI to create metadata [EB/OL]. [2024－12－26]. https://www.xcdsystem.com/tla/program/pAJExb8/index.cfm?pgid=806&sid=38182&abid=113425.
19 San Diego State University. Enhancing cataloging of electronic government documents with programming and OpenAI [EB/OL]. [2024－12－26]. https://osf.io/39bws.
20 National Library of Finland. Annif：automated metadata tagging [EB/OL]. [2024－12－26]. https://annif.org/.
21 German National Library. Automatic cataloguing system [EB/OL]. [2024－12－26]. https://www.dnb.de/EN/Professionell/ProjekteKooperationen/Projekte/KI/KI.html.
22 Ma X，Dinsmore C，Van Kleeck D，et al. Finding Florida—Implementing Machine Aided Indexing in an academic library [C]// Proceedings of the International Conference on Dublin Core and Metadata Applications. Dublin Core Metadata Initiative，2023.
23 Deng S，Piascik J，Chow E H C，et al. AI in action：reimagining metadata and cataloging with Chatbots and OpenAI [EB/OL]. [2025－06－03]. https://rex.libraries.wsu.edu/esploro/outputs/conferencePresentation/99901181238301842.
24 Chow E H C，Kao T J，Li X. An experiment with the use of ChatGPT for LCSH subject assignment on electronic theses and dissertations [J]. Cataloging & Classification Quarterly，2024，62(5)：574－588.
25 香港浸会大学图书馆. 初探利用人工智能提供国会图书馆主题词 [EB/OL]. [2024－12－26]. http://www.cccna.org/AI_LCSH_final.pdf.
26 Library of Congress. Exploring computational description [EB/OL]. [2024－12－26]. https://labs.loc.gov/work/experiments/ECD/.
27 Royal Library of Belgium. One automatic cataloging flow：tests and first results [EB/OL]. [2024－12－26]. https://repository.ifla.org/handle/123456789/2686.
28 肖燕. 图书采分编的自动化技术运用——以广东省立中山图书馆“采编图灵”为例 [J]. 图书馆论坛，2024，44(4)：20－28.
29 Greenly G. CatalogerGPT [EB/OL]. [2024－12－26]. https://glengreenly.wixsite.com/catalogergpt.
30 National Library of Luxembourg. Chatbot for digital archives [EB/OL]. [2024－12－26]. https://chat.eluxemburgensia.lu.
31 National Library of Norway. Maken [EB/OL]. [2024－12－26]. https://www.nb.no/maken/.
32 MapReader. Introduction to MapReader [EB/OL]. [2024－12－26]. https://mapreader.readthedocs.io/en/latest/introduction-to-mapreader/.
33 Dresel R. Designing a linked data service across borders and time zones：the National Library Board's experience [C]//Proceedings of the International Conference on Dublin Core and Metadata Applications. Dublin Core Metadata Initiative，2024.
34 Stanford Libraries. AI projects [EB/OL]. [2024－12－26]. https://sul-dlss-labs.github.io/ai-projects/.
35 Emory University Libraries. Wayfinder：AI research tool [EB/OL]. [2024－12－26]. https://wayfinder.libraries.emory.edu/.
36 German National Library. Collection mandate [EB/OL]. [2024－12－26]. https://www.dnb.de/EN/Professionell/Sammeln/sammeln_node.html.
37 Cox A. The impact of AI，machine learning，automation and robotics on the information professions：a report for CILIP [EB/OL]. [2024－12－26]. https://eprints.whiterose.ac.uk/174522/.
38 工业和信息化部 中央网络安全和信息化委员会办公室 国家发展和改革委员会 国家标准化管理委员会关于印发国家人工智能产业综合标准化体系建设指南(2024版)的通知 [EB/OL]. [2024－12－26]. https://www.gov.cn/zhengce/zhengceku/202407/content_6960720.htm.
39 曾蕾，谭旭. 数据的语义增强——解读图档博支持数字人文的新动向 [J]. 数字人文研究，2021，1(1)：65－86.
40 赵美花. 人工智能在图书采分编领域的实践与展望——以广东省立中山图书馆“采编图灵”为例 [J/OL]. 图书馆论坛，1－8 [2025－05－30]. http://kns.cnki.net/kcms/detail/44.1306.G2.20250418.1624.002.html.
41 Poley C，Uhlmann S，Busse F，et al. Automatic subject cataloguing at the German National Library [J]. LIBER Quarterly：The Journal of the Association of European Research Libraries，2025，35(1)：1－9.
42 赵晓伟，祝智庭，沈书生. 教育提示语工程：构建数智时代的认识论新话语 [J]. 中国远程教育，2023，43(11)：22－31.
43 German Federal Government. Artificial intelligence strategy [EB/OL]. [2024－12－26]. https://www.ki-strategie-deutschland.de/.
44 Association for Library Collections & Technical Services. Core competencies for cataloging and metadata professional librarians [EB/OL]. [2024－07－06]. http://hdl.handle.net/11213/20799.
45 CHOICE. Building an AI literacy framework：perspectives from instruction librarians and current information literacy tools [EB/OL]. [2024－12－26]. https://www.choice360.org/research/white-paper-building-an-ai-literacy-framework-perspectives-from-instruction-librarians-and-current-information-literacy-tools/.
46 四川大学图书馆. AI赋能专业馆员行动正式开启 [EB/OL]. [2024－12－26]. https://lib.scu.edu.cn/node/126703.

作者贡献说明：
张谙宁：论文撰写与修改
叶兰：设计研究框架，论文修改
周文琦、张倩、张欢庆：案例调研

作者单位：深圳大学图书馆，广东深圳，518060
收稿日期：2025年2月10日
修回日期：2025年6月26日
（责任编辑：支娟）

AI-Generated Metadata Enabling Library Collection Development: A Case Study

ZHANG Anning　YE Lan　ZHOU Wenqi　ZHANG Qian　ZHANG Huanqing

Abstract: AI-generated metadata is both an inevitable result of technological progress and a key driver of innovation in library collection development. This study examines practical cases in which AI-generated metadata has empowered collection development in libraries, both domestic and international. This study summarizes diverse application scenarios and outcomes, and proposes practical recommendations aimed at supporting the digital and intelligent transformation of library resources. In terms of methodology, the study adopts a combination of online surveys and literature analysis to collect detailed information from existing research, relevant publications, and online sources. Case selection follows two main criteria: (1) prioritizing cases with proven outcomes and well-documented data to ensure the reliability and representativeness of the study; (2) cases involving libraries of different regions, scales, and types to enhance breadth and comparability. A total of 20 cases are ultimately selected. The study conducts case analysis to outline the current state of practice across five dimensions: library types, resource types, technologies, application scenarios, and outcomes achieved. In terms of library types, the findings reveal that the selected cases involve national libraries, academic libraries, and provincial or municipal public libraries. These different types of institutions show varying levels of technical capacity and resource conditions, which directly influence their choice of AI application pathways. In terms of resource types, materials are categorized into born-digital resources and digitized physical resources. The distinction between these formats determines the technical complexity and approaches required for generating metadata with AI. Regarding technologies, the cases feature the use of large language models (LLMs), natural language processing (NLP), computer vision (CV), and automatic speech recognition (ASR). Typical application scenarios of AI-generated metadata include the creation of descriptive metadata, subject and classification indexing, bibliographic record generation, and semantic enrichment of metadata. In terms of outcomes achieved, the use of AI-generated metadata has contributed significantly to the large-scale disclosure of unstructured and multimodal resources, the optimization of cataloging workflows, support for teaching and research, and the enhancement of personalized and intelligent resource services. Based on the case analysis, the study offers two key recommendations: (1) Libraries should develop context-specific technical strategies for AI-generated metadata, including prompt engineering with LLMs, domain-specific training using machine learning, and retrieval-augmented generation based on knowledge graphs. (2)Libraries should establish support mechanisms covering strategic planning, quality control, and personnel development to ensure the sustainable impact of AI-generated metadata in library resource development.

Keywords: Artificial Intelligence; Metadata; Collection Development; Library