中南大學
CENTRAL SOUTH UNIVERSITY

面向生成式AI的高质量数据集构建和科研服务模式创新

——以中南大学图书馆为例

汇报人：中南大学图书馆 周芬

课题成员：崔永 王利君 蒋文梅 袁新辉

中南大学 CENTRAL SOUTH UNIVERSITY

目 录

contents

中南大學
CENTRAL SOUTH UNIVERSITY

01

研究现状和背景

一、研究背景和现状 | 研究背景

中南大學 CENTRAL SOUTH UNIVERSITY

生成式AI重塑科研与图书馆服务

政策牵引

“人工智能+教育”、数据要素、教育数字化等国家战略推动高校构建智能化知识基础设施。

- 教育部等五部门关于印发《“人工智能+教育”行动计划》的通知
- 国家数据局等部门关于印发《“数据要素×”三年行动计划（2024—2026年）》的通知
- 教育部等九部门关于加快推进教育数字化的意见

技术变革

大语言模型、多模态模型、RAG和智能体快速发展，科研服务从检索走向生成、推理与协作。

图书馆转型

高校图书馆不再只是文献收藏者，而是高质量数据资产管理者、智慧赋能者和开放生态构建者。

高质量数据集是生成式AI应用的“燃料”和“事实锚点”，决定科研服务的可靠性、专业性和可持续性。

一、研究背景和现状 | 问题提出

通用大模型在学术场景中的三类短板

专业深度不足

通用模型对垂直学科概念、术语体系、实验数据和学科知识谱系掌握不够深入。

事实幻觉风险

生成内容可能存在错误引用、知识过时、逻辑不严谨等问题，难以直接满足科研严谨性要求。

合规边界复杂

商业数据库、科研敏感数据、用户行为数据等难以直接用于训练和调用，必须建立可信治理机制。

解决路径：建设面向生成式AI的高校图书馆高质量数据集，使数据具备清洁可信、语义丰富、机器可读、可被模型检索/训练/推理调用等能力。

一、研究背景和现状 | 研究意义

从资源数字化到数据要素化、智能化

理论意义

重新定义图书馆馆藏对象：从“文献”拓展为“高质量、可计算、可标注的数据集”。

推动智慧图书馆理论从资源数字化走向数据要素化与智能化。

实践意义

为高校图书馆提供战略转型路线：建设可信数据基础设施，支撑智能文献综述、科研趋势预测、论文查证与写作辅助等服务。

最终目标：将图书馆打造为高校“人工智能+科研”的数据枢纽、知识基座与服务中台。

一、研究背景和现状 | 核心概念

生成式AI与高质量数据集

生成式人工智能

依托大模型和多模态学习，能够生成文本、图像、代码等内容。在科研服务中，重点体现为跨文献综合、知识推理、启发式生成与智能交互。

高质量数据集

以馆藏和科研数据为基础，经过版权清洗、隐私脱敏、多模态对齐、细粒度语义标注与质量评测，可被算法和智能体直接调用的垂直领域语料集合。

本课题关注的不是“数据量有多大”，而是“数据能否安全、准确、可持续地支撑生成式AI科研服务”。

一、研究背景和现状 | 研究现状

国内研究现状

Domestically

关注生成式AI在智能检索、参考咨询、阅读推广、学科服务、数字包容等场景中的应用；
但系统阐述“AI就绪数据集”构建流程的研究不多。

Abroad

国外研究现状

研究数据管理、开放科学、数据仓储、FAIR原则与数据出版较成熟；

但面向生成式AI训练/微调/RAG的图书馆数据集框架仍在探索。

以“高质量数据集”为核心，贯通数据治理、数据集构建、平台支撑和生成式AI科研服务创新。

中南大学
CENTRAL SOUTH UNIVERSITY

02

国内外数据集建设调研

二、国内外数据集建设调研 | 调研目标

- 梳理国内外高校在应对人工智能大模型数据需求时的前沿做法。
- 从“资源—系统—服务”的传统观察，转向“面向生成式AI的数据资源体系”观察。为后续提出高质量数据集建设框架、科研服务模式提供参考和依据。

二、国内外数据集建设调研 | 调研对象

中南大學 CENTRAL SOUTH UNIVERSITY

国外样本

选取U.S. News排名前20的美国高校，重点观察图书馆和相关数据服务部门在数据治理、数据建设、平台支撑和AI应用方面的实践。

国内样本

选取双一流高校和学术机构典型案例，关注哲学社会科学、法学、考古、科技文献、教育知识图谱等垂直领域数据集建设。

识别“哪些能力能够支撑生成式AI”，并据此推导高校图书馆的数据集建设路径。

二、国内外数据集建设调研 | 调研维度

面向生成式AI的四大调研维度

1、数据治理：合规与可信数据底座

开放授权、版权审查、隐私脱敏、伦理审查、分级访问

2、数据建设：语料重构与加工

机器可读化、OCR/解析、语义标注、知识图谱、多模态对齐

3、平台支撑：接口与系统环境

数据仓储、API、DOI/PURL、向量检索、受控计算空间

4、应用场景：服务对象与闭环

科研评价与推理科研推理、文本挖掘、学科建设、AI问答、趋势分析

维度逻辑：治理保证“能用” → 建设实现“好用” → 平台支撑“可调用” → 场景验证“有价值”

二、国内外数据集建设调研 | 国外调研总览

Top 20高校的共性趋势——从研究数据管理走向AI可调用的数据基础设施

开放与合规并重

重视开放授权、敏感数据分级、隐私脱敏与伦理流程，为AI语料使用奠定边界。

平台化能力突出

Dataverse、机构数据仓储、REST API、JSON/JSONL等接口提升数据可发现与可调用性。

数据建设重语义化

推动OCR、文本清理、实体抽取、关联数据与知识图谱建设，提高机器可理解性。

应用面向AI for Science

支撑文本挖掘、科研验证、医学/工程模型研发、自然语言检索与科研推理。

国外实践启示：图书馆不只是提供网页检索入口，而是在建设面向机器读取、程序调用和模型服务的数据基础设施。

二、国内外数据集建设调研 | 国外调研

1 数据治理 开放授权与可信管控

解决生成式AI语料的版权、隐私与伦理边界

哈佛 / 耶鲁

开放元数据与机器可读记录，推动公共领域资源用于文本分析与模型测试；

斯坦福/约翰霍普金斯

医学数据采用分级访问、去标识化、IRB审查等机制，支撑AI研发与训练。

杜克/加州伯克利

对人类受试者数据、商业授权文献建立去标识化与受控TDM环境。

高校图书馆建设AI数据集，必须把“可用”建立在“可信、合规、可追溯”的前提之上。

二、国内外数据集建设调研| 国外调研

2 数据建设 语料重构与知识化加工

将非结构化资源转化为模型可学习、可推理的数据

哈佛

- 公有领域著作语料库：OCR、文本清理、IIIF与REST接口，辅助语言模型训练和自然语言检索。

MIT

- DrivAerNet++：工程数据格式化与跨模态特征对齐，支持算法挖掘和物理模拟。

康奈尔

- arXiv机读语料：论文文本与元数据结构化，支持文本挖掘、趋势分析和算法开发。

布朗/范德堡

- 数字人文图谱/RDF数据：从历史文本提取人物、地点、时间和关系，支撑计算人文分析。

二、国内外数据集建设调研 | 国外调研

3 平台支撑 接口、仓储与安全计算环境

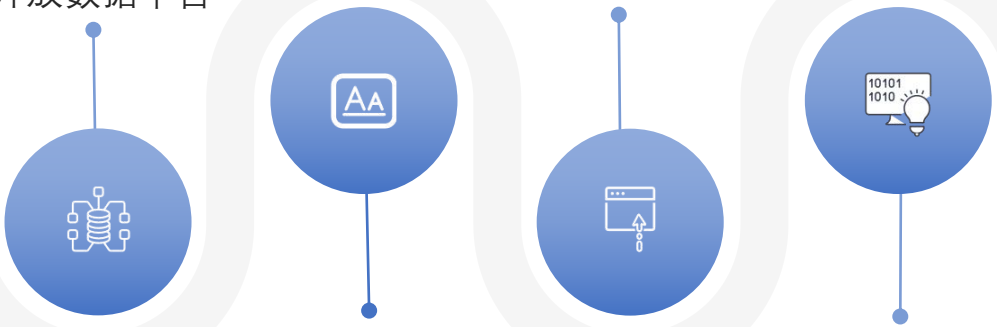
从读者网页检索转向程序化、高频、规模化调用

数据仓储

Dataverse、
机构知识库、
开放数据平台

接口调用

RESTful API、SPARQL、
OAI-PMH



标识与格式

DOI、PURL、JSON、
JSONL、RDF

计算环境

云端Jupyter、安全计算
环境、校内计算中心

平台演进方向：把图书馆数据从“人可以查到”升级为“机器可以读到、模型可以用到、服务可以嵌入到科研流程”。

二、国内外数据集建设调研 | 国外调研

4 应用场景 支撑AI科研与文本挖掘

- 语言模型训练与自然语言检索测试
- 医学/工程AI模型研发与多模态评测
- 计算社会科学与数字人文文本挖掘
- 科研实验数据重复验证与趋势发现

建设“面向机器可读与可调用”的数据基础设施，是生成式AI服务落地的前置条件。

二、国内外数据集建设调研 | 国内调研 | 总览

垂直深耕与场景赋能特征明显

建设路径：从特色资源建设走向学科知识基座与AI科研服务支撑

政策牵引

紧密对接“数据要素×”“人工智能+”战略

知识化加工

强调实体抽取、本体建模、知识图谱与结构化事实网络

特色切入

以特色馆藏、优势学科、行业需求为入口

场景落地

服务数字法学、考古、数字人文、航运治理等具体科研场景

国内实践更强调“领域知识基座”和“场景化应用牵引”，为高校图书馆开展垂直数据集建设提供了可借鉴路径。

二、国内外数据集建设调研 | 国内调研 | 数据治理

政策导向与多主体协同：从资源管理走向数据资产治理

1. **上海交大图书馆**：“面向法学人才培养的AI知识库和智能体构建”入选上海市高质量数据集先行先试项目，体现图书馆牵头、学院协同、应用牵引。
2. **上海海事大学**：“海上事故数据集”依托交通运输可信数据空间实验室，强调原始数据获取、清洗治理、标准化存储、质量评估和数据应用全链条。
3. **复旦/高校文科平台**：考古、古籍、数字人文等特色资源建设强调专家参与、学科规范、资源确权与知识结构化，支撑人文社科数据治理。

数据治理要从“资源归集”走向“权属清晰、标准统一、责任明确、可安全使用”的数据资产治理。

二、国内外数据集建设调研 | 国内调研 | 数据建设

垂直领域知识图谱：从文献集合走向领域知识基座

⇨ 复旦大学

考古资料高质量数据集：面向考古材料、遗址、器物、年代、关系等要素进行结构化组织，支撑人文社科开放协同与知识推理。

⇨ 上海交通大学

法学AI知识库与智能体：围绕法学人才培养与数字法学，构建规则、案例、知识点和问答任务数据。

⇨ 浙江大学

古籍特藏发布平台：目录、全文和专题多视角揭示，为古籍OCR、文本挖掘、传统文化知识发现奠定数据基础。

⇨ 上海海事大学

海上事故数据集：原始数据获取、清洗治理、标准化存储、分层构建、质量评估，服务航运领域智能化管理。

二、国内外数据集建设调研 | 国内调研 | 平台支撑

- 依托机构库、数字图书馆、专题数据库建设数据底座
- 建设实体关联、知识图谱、智能体调用和数据流转能力
- 探索可信数据空间、受控沙盒和跨部门数据协同
- 不足：统一接口、向量化检索、质量评测和持续更新机制仍需加强

二、国内外数据集建设调研 | 国内调研 | 应用场景

- 数字法学：教学、案例分析、法律知识问答与智能体服务
- 数字人文/考古：古籍识别、知识关联、文物/遗址信息组织
- 学科情报：主题发现、学者画像、机构评价与趋势研判
- 行业治理：航运事故分析、风险预警与精细化管理

国内实践启示：高质量数据集建设应从具体学科和真实场景出发，以应用倒逼数据标准、质量评测和平台能力建设。

二、国内外数据集建设调研 | 国内外调研综述

比较维度	国外	国内	共同问题/建设需求
数据治理	开放授权、隐私脱敏、分级访问较成熟	政策牵引强，多主体协同和特色资源治理突出	需建立覆盖全生命周期的标准与合规机制
数据建设	侧重通用语料、开放元数据、机器可读与关联数据	侧重垂直领域、实体抽取、知识图谱和深度知识化	需从“资源数字化”升级为“AI就绪数据集”
平台支撑	Dataverse、API、DOI/PURL、受控环境能力较强	机构库、智慧图书馆、专题库和智能体底座逐步发展	需实现标准接口、向量化、数据分发与受控调用
应用场景	AI for Science、NLP、医学/工程模型、文本挖掘	数字法学、考古、数字人文、行业治理等场景	需形成“数据—模型—服务—反馈”的闭环

国内外实践虽路径不同，但都指向同一目标——建设面向生成式AI、可治理、可调用、可持续优化的数据资源体系。

二、国内外数据集建设调研 | 调研启示

调研揭示的五类建设要求

统一汇聚多源异构数据

整合馆藏、科研成果、研究数据与开放资源，形成可治理的数据池；

建立全流程数据治理

围绕版权、隐私、清洗、标注、质量评测和更新形成标准体系

构建面向生成式AI的数据集

形成预训练、SFT、RAG、评测与Agent任务等复合型数据集

强化平台与治理保障

以API、向量检索、可信空间和算力环境支撑安全可调用

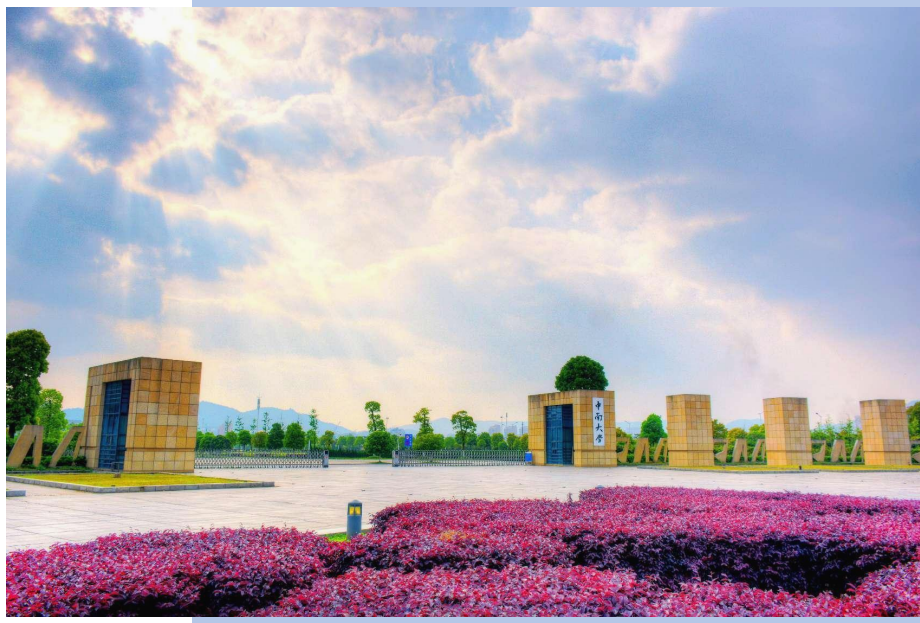
应用反馈闭环

以服务效果反向驱动数据持续优化

中南大学
CENTRAL SOUTH UNIVERSITY

03

高质量数据集建设框架



三、高质量数据集建设框架 | 建设思路

传统资源建设	面向生成式AI的高质量数据集建设
以文献收藏、整理、揭示为主	以AI可用数据资产构建为主
关注资源完整性与可检索性	关注机器可读、语义可理解、模型可调用
面向读者检索与获取	面向大模型训练、微调、RAG和智能服务
数据加工以元数据著录为核心	数据加工扩展到清洗、标注、对齐、评测
服务方式以资源提供为主	服务方式转向知识生成、智能问答与科研辅助

围绕“数据从哪里来—如何治理—如何构建AI数据集—如何保障调用—如何形成反馈闭环”展开框架设计。

三、高质量数据集建设框架 | 建设思路

整体思路：从“馆藏资源”到“AI就绪数据集”

馆藏/科研资源 → **数据治理加工** → **高质量数据集** → 生成式AI服务

文献、特藏、机构成果
科研数据

清洗、脱敏、标注、对齐、
质量评测

可信、可溯、机器可读、语
义丰富

RAG问答、智能咨询、情报
发现、沙盒计算

**“围绕生成式AI需求
重构数据资源体系”**

核心转化：把“可阅读的文献资源”转化为“可计算、可验证、可调用的知识数据资产”。

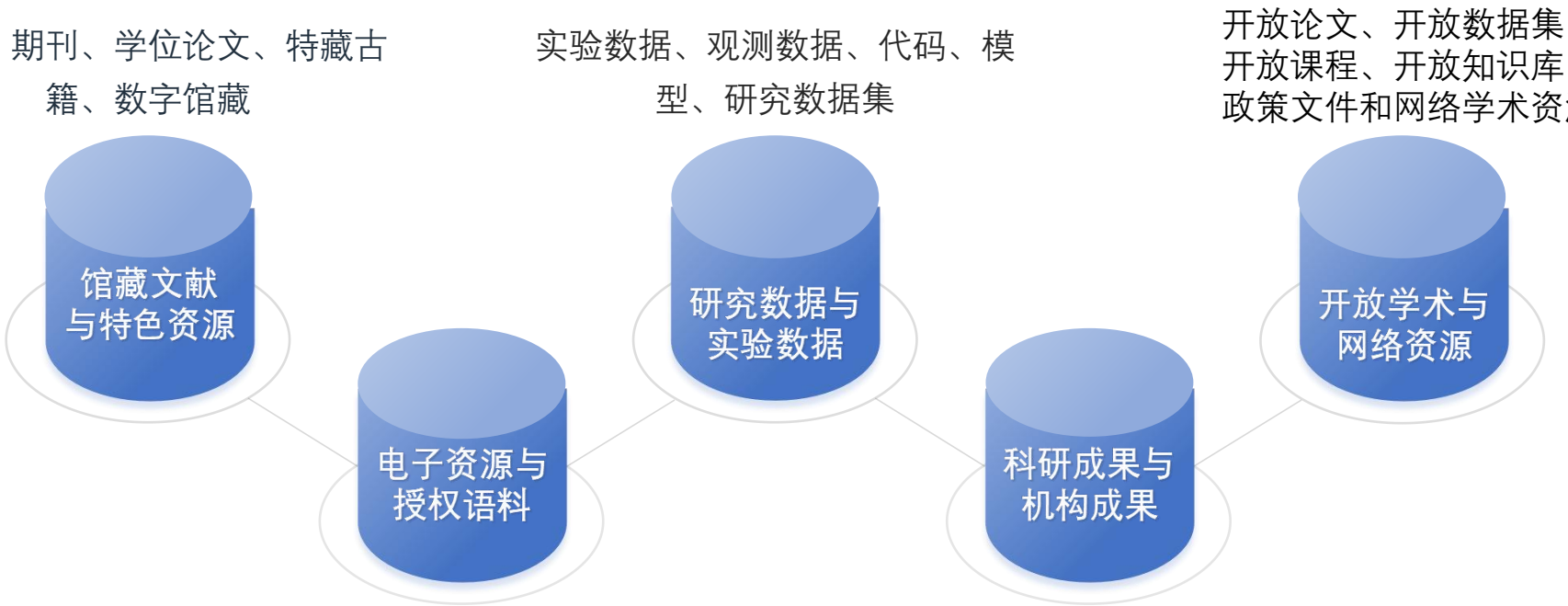
三、高质量数据集建设框架｜构建框架

搭建“数据来源—治理加工—AI数据集构建—平台治理—科研服务应用—反馈迭代”的闭环

框架强调：高质量数据集不是一般数据库，而是面向大模型训练、微调、检索增强、评测与智能体调用的AI-ready数据基础设施。

三、高质量数据集建设框架｜构建框架｜数据来源层

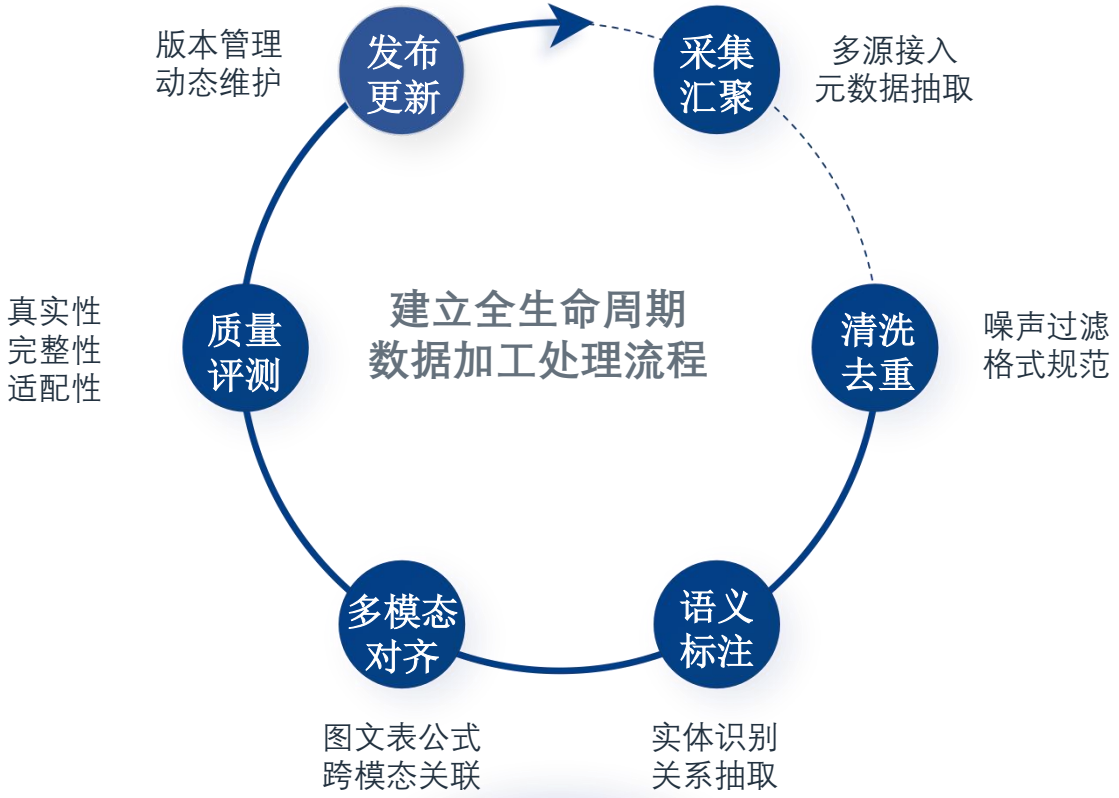
从馆藏资源、科研资源和开放资源中形成可治理数据池



数据来源特征：多源异构、多模态、跨学科、动态更新。
建设要点：源头数据必须同步记录来源、权属、许可、版本、更新时间和可用边界，为后续合规使用奠定基础。

三、高质量数据集建设框架 | 构建框架 | 数据治理层

建立全生命周期数据加工处理流程



从“粗放数字化”转向“AI可用加工”，特别关注图表、公式、非结构化文本和多模态数据的机器可读转换。

三、高质量数据集建设框架 | 构建框架 | 数据集构建层

中南大学 CENTRAL SOUTH UNIVERSITY

五类核心数据集

预训练语料库

面向领域大模型预训练或持续预训练，汇聚高质量学术文本、特色馆藏、科研成果和开放学术资源，提升模型对高校科研语境和学科知识的理解能力。

指令微调数据集

围绕高校科研服务任务，构建问答对、摘要生成、文献推荐、科研咨询、查新辅助、写作辅助等指令数据，提升模型任务响应能力。

检索增强知识库（RAG）

将权威文献、机构知识库、学科专题资源和科研成果转化为可检索、可引用、可溯源的知识库，为生成式AI提供事实支撑，降低幻觉风险。

评测基准集

围绕学术问答、文献综述、知识发现、情报分析、科研评价等任务构建评测样本，用于检验模型输出的准确性、可靠性和服务效果。

Agent工具调用/任务数据

围绕科研流程中的复杂任务，构建工具调用、任务拆解、流程编排和人机协同数据，支撑AI馆员、科研助理等智能体应用。

高质量数据集不是单一语料库，而是支撑“训练—微调—检索增强—评测—智能体服务”的复合型数据体系。

三、高质量数据集建设框架 | 构建框架 | 数据集构建层

核心构建动作：

语义标注

开展实体识别、关系抽取、主题标引、关键词规范化；
构建学者、机构、论文、项目、基金、专利、学科等核心实体；
形成可供模型理解和推理的语义数据。

知识组织

建立分类体系、主题词表、本体模型和知识图谱；
实现不同学科知识组织系统之间的映射与融合；
支撑跨学科知识发现、关联分析
和智能问答。

质量评测

从准确性、完整性、一致性、多样性、时效性、可追溯性等维度评价数据质量；
增加模型适配性评价，检验数据对生成式AI任务的支撑效果；
形成数据集质量控制和准入机制。

版本管理与持续更新

建立数据集版本控制、更新记录和引用标识；
根据资源变化、用户反馈和模型效果持续迭代；
形成动态维护的权威数据池。

关键路径：资源汇聚 → 语义加工 → 知识组织 → 质量评测 → 持续更新。

三、高质量数据集建设框架 | 构建框架 | 保障层

平台支撑与治理保障层：保障高质量数据集长期建设、稳定运行和持续赋能

数据标准与接口API

制定统一的数据格式、元数据规范、标注规范和接口标准，支持数据集被系统、模型和智能体高效调用。

算力与存储环境

建设稳定的数据存储、索引、向量化处理和计算环境，为大规模数据处理、模型调用和应用验证提供支撑。

数据可信空间/受控沙盒

针对受版权、隐私或安全限制的数据，建立受控访问、可计算不可搬移、安全审计的数据使用环境。

伦理审查与安全治理

完善版权合规、隐私保护、算法偏见、数据滥用和生成内容风险审查机制，确保数据集建设安全可信。

组织协同与人才保障

推动图书馆、信息化部门、科研管理部门、学院和技术团队协同，培养具备数据治理、AI应用和学科服务能力的复合型馆员。

治理保障的重点：让数据“可用不可乱用、可算不可外泄、可更新可追溯”

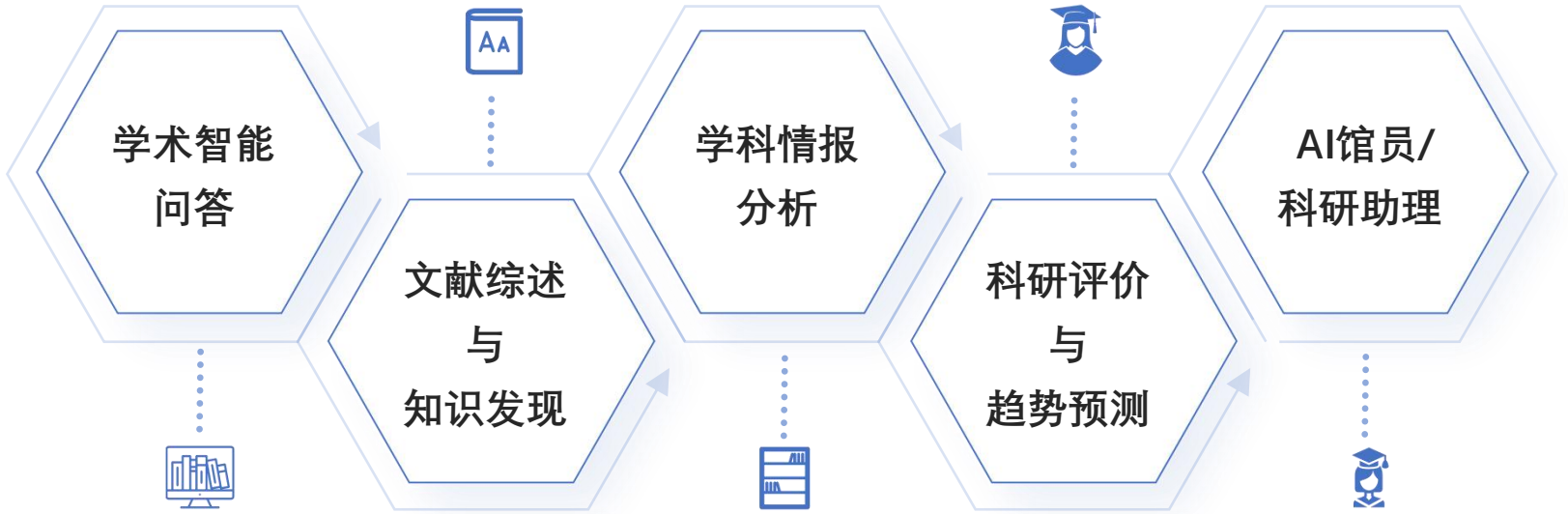
三、高质量数据集建设框架 | 构建框架

框架价值（一）：形成“数据集—模型/知识库—智能服务”的闭环

应用反馈 → 效果评估 → 数据迭代优化 → 服务能力提升

三、高质量数据集建设框架 | 构建框架

框架价值（二）： 沉淀可落地的科研智能服务能力



中南大學
CENTRAL SOUTH UNIVERSITY

04

科研服务创新模式

四、科研服务创新服务模式 | 总体设计

高质量数据集建设的最终目的，是将数据资产转化为可感知、可交互、可验证、可持续优化的科研服务能力。

由“资源供给”转向“数据驱动的科研协作”

1. **学科增强知识服务**
   面向垂直学科的RAG问答、文献综述、知识发现
2. **人机协同智能咨询**
   AI馆员+馆员审核，提供全天候、可追溯咨询服务
3. **科研评价情报发现**
   基于多源数据开展学者画像、机构分析和趋势研判
4. **受控沙盒数据计算**
   在安全环境中支持受版权/敏感数据的联合计算和文本挖掘

服务逻辑：数据集支撑模型 → 模型增强业务 → 业务反馈数据 → 数据持续优化

四、科研服务创新服务模式 | 模式1

面向学科增强的知识服务模式

以学科专题数据集、RAG知识库和知识图谱为基础，建设垂直学科知识服务能力。围绕文献综述、热点识别、研究脉络梳理、跨学科知识关联等场景提供智能支持。

服务对象：学科馆员、科研团队、研究生、重点实验室。

价值：提升学科服务的专业深度、事实可靠性和知识发现能力。

中南大学实践：地球科学AI知识库由图书馆与二级学院共建，汇聚1000余种相关图书、3万余篇中文期刊论文和近6000篇外文期刊论文。

通过细粒度知识抽取、多模态语义化处理和质量监督机制，形成垂直领域知识库；再以RAG实现“先检索、后生成”，有效缓解专业问答中的幻觉问题。

学科专题数据集+RAG知识库，是高校图书馆开展深度学科服务的可行路径。

四、科研服务创新服务模式 | 模式2

基于人机协同的智能咨询服务模式

分析构建AI馆员/科研助理，承担常见咨询、资源导航、检索策略生成和基础科研辅助。采用“AI初答—馆员审核—知识库回写”的闭环，提升服务效率并控制生成风险。

服务对象：全校师生、科研人员、学科馆员。

价值：实现全天候咨询、个性化响应与服务知识沉淀。

中南大学实践：AI馆员以馆务数据和FAQ问答对为初始知识来源，覆盖开放时间、借阅规则、空间预约、业务办理等高频咨询。通过“AI初答—馆员审核—知识库回写”的闭环，依托门户网站、微信公众号和线下大屏提供7×24小时咨询与语义化书目检索服务。

自2025年3月上线以来，已服务师生2万余人次，知识库命中率提升90%。

AI馆员不是替代馆员，而是通过人机协同实现服务增效、知识沉淀和持续优化。

四、科研服务创新服务模式 | 模式3、4

面向科研评价的情报发现服务模式

- 依托论文、专利、项目、基金、机构成果等多源数据，构建学者画像、机构画像和学科态势分析能力。结合知识图谱与生成式AI，可支持主题演化分析、竞争格局识别、合作网络挖掘和科研趋势预测。

  服务对象：科研管理部门、学科建设部门、学院负责人、科研团队。

  价值：推动图书馆从文献支持走向数据驱动的学科情报服务与决策支持。

基于受控沙盒的数据计算服务模式

- 针对受版权、隐私或安全限制的数据，建设“可计算不可搬移”的受控计算环境，支持文本挖掘、语义分析、模型调用与联合计算。通过可信空间、权限控制、过程留痕和审计机制，在保障合规的前提下释放数据价值。

  服务对象：重点实验室、学科团队、科技查新与知识产权服务场景。

  价值：在安全边界内实现高价值数据利用，为复杂科研任务和AI应用提供可持续支撑。

中南大學
CENTRAL SOUTH UNIVERSITY

05

总结与展望

五、总结与展望

研究总结

梳理国内外高校高质量数据集建设实践，提炼四个调研维度；
构建面向生成式AI的高校图书馆高质量数据集建设框架；
提出四类科研服务创新模式，推动数据资产转化为服务能力。

落地保障

组织转型：建设跨部门数据资源协同创新团队。
人才建设：培养数据策展人、算法训练师、提示词工程师。
经费算力：设立数据与智能建设专项经费。

治理保障

合规审查：强化版权、隐私、伦理与安全边界。
开放协同：构建校内外共建共享生态。
持续评估：以服务效果反向优化数据集质量。

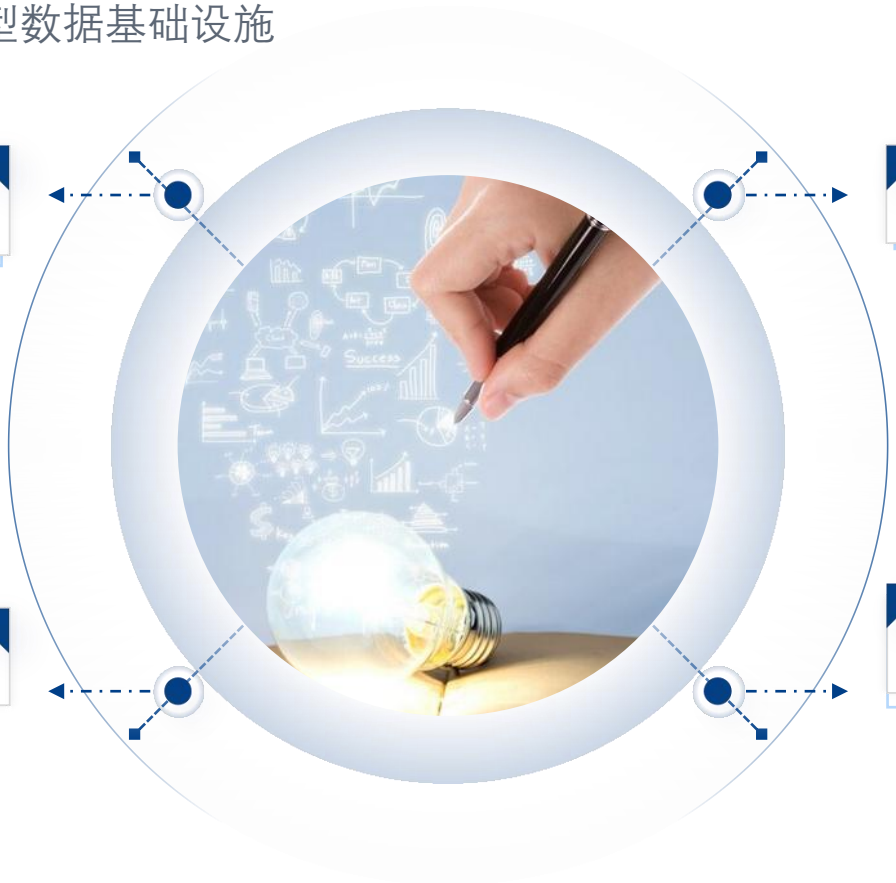
高质量数据集是高校图书馆融入“人工智能+科研”的关键抓手，也是连接资源建设、数据治理、模型应用和科研服务创新的核心枢纽。

五、总结与展望

面向生成式AI，建设高校图书馆新型数据基础设施

1、聚焦重点学科和高价值场景

优先在优势学科、科技查新、知识产权、科研评价等场景持续深化应用。



2、融入学校AI整体布局

形成“学校建平台、图书馆供数据、师生享服务”的协同模式，补齐算力和平台短板。

3、推动数据共建共享联盟

探索区域性、行业性数据联盟，共建元数据、特色资源和语义标注数据。

4、完善双环驱动机制

以治理闭环保障数据质量，以价值闭环优化服务效果，形成可持续迭代能力。

中南大學
CENTRAL SOUTH UNIVERSITY

感谢聆听，
敬请批评指正！