Java公共文化资源项目

系统分析与设计

目录

1. 项目概述
2. 角色与场景
3. 需求分析
4. 数据来源与更新策略
5. 基础资源与算法
6. 系统总体架构
7. 数据模型设计
8. 可视化设计
9. 验证方案与评测
10. 技术栈与部署
11. 里程碑与工作流
12. 交付清单

1 项目概述

目标

本项目的核心目标是：

构建一个基于AIGC技术的公共文化资源建设系统。实现从原始文化素材到最终知识服务的全流程、智能化能力，具体实现以下三点：

- **全流程能力整合**：实现从多模态资源的自动化采集、深度解析与智能标注，到文化内容的AIGC生成，再到知识图谱的结构化存储，最终通过多模态检索与智能问答提供服务的闭环管理能力。
- **知识化体系构建**：通过实体识别、关系抽取等手段 ，将非结构化的文化信息转化为结构化的知识图谱。

> *例如，系统能够自动识别出绘画作品中的作者、年代、流派等实体信息，并构建它们之间的关联，形成一张庞大的文化知识网络。*

- **服务能力创新**：基于构建的知识体系，赋能创新的文化服务场景。

愿景

我们的愿景是：将本系统打造成为公共文化服务领域的“智能资源生成平台 + 知识服务引擎 + 交互门户”。

- **智能资源生成平台**：系统能够利用AIGC技术，根据需求自动生成多样化的文化内容，成为文化内容生产的加速器。
- **知识服务引擎**：通过将分散的资源整合为结构化的知识图谱，本系统将成为一个强大的后台知识引擎。系统能精准响应复杂的查询需求，并以结构化的形式返回答案，为上层应用提供知识动力。
- **交互门户**：为用户提供一个直观、友好、智能的文化探索门户，从而能降低公众接触和理解优秀文化的门槛，提升文化服务的互动性和体验感。

范围

- 资源类型：文本、图像等文化资源。
- 内容领域：音乐、舞蹈、书法、绘画、戏曲、非遗等。
- 功能覆盖：采集、标注、生成、检索、问答、数据集管理。

非公开范围

- **非公开版权资源的非法抓取：**本项目严格遵守国家著作权法及相关法律法规。所有数据采集行为都将限定在官方明确授权或公开协议允许的公共数据源范围内。系统将建立版权合规审查机制，确保不侵犯任何第三方的合法权益。
- **高精度AIGC模型的从零训练：**考虑到巨大的算力成本和时间周期，本项目不会进行大规模AIGC模型的预训练。我们将使用**开源的预训练模型**，并通过模型微调和提示工程等技术，使其适应公共文化领域的特定需求。

---

2 角色与场景

角色（用户）

- **管理者：** 系统的日常运维人员，负责资源审核、数据质量监控与用户管理。
- 普通用户：
  - **普通民众/学生**：对文化艺术感兴趣的社会公众或在校学生，以学习和探索为主要目的。
  - **文化研究员：**高校或研究机构的学者，需要进行深度资料挖掘与关联分析。
  - **策展人/内容创作者：**博物馆、文化馆的策展人员或新媒体编辑，需要快速生成宣传材料和寻找创作灵感。

典型场景

- **S1：专题内容生成：**如输入主题“XX市木版年画的传承与创新”，系统自动生成包含核心知识、代表作品图片及传承人介绍的结构化初稿，辅助文化宣传任务的完成。
- **S2：多模态检索：** 一名服装设计师在看到一幅古代绘画作品中的纹样，将其截图上传至系统。系统通过多模态检索，不仅找到了该画作的详细信息，还能推荐同时期具有相似纹样元素的瓷器、服饰及建筑。
- **S3：智能问答：** 一名正在撰写论文的学生，通过自然语言提问：“查询京剧在清代发展的主要特点及代表人物？”系统迅速返回包含关键特点、代表人物、经典剧目等信息的结构化答案，并清晰列出原始文献或资料出处。
- **S4：文化实体图谱可视化：**如要探究苏轼的生平及其影响时，系统直观能地展示苏轼与同时代的文人（如黄庭坚）、政治事件、其创作的书法与文学作品之间的复杂关系网络，极大地提升研究效率。
- **S5：AIGC资源质检与标注修正：** 管理员能批量修正错误标签，并将修正后的数据加入验证集，以优化后续模型的识别准确率，实现系统的持续迭代。

---

3 需求分析

功能需求

- **多模态数据采集**：支持文本、图像的爬取与导入；
- **智能分析与标注**：文化实体识别、属性提取、关系构建；
- **AIGC资源生成**：基于提示词生成文化内容（文本、图像等）；
- **知识图谱构建**：结构化存储文化实体与关系；
- **多模态检索**：支持图文互搜、语义检索；
- **智能问答（RAG）**：基于检索增强生成的问答系统；
- **后台管理**：资源管理、模型配置、标注审核、数据集导出。

非功能需求

- 性能：~~万级资源检索响应<500ms~~；生成任务支持异步队列；
- ~~可用性：SLA 99%；~~
- 安全性：访问控制、数据脱敏、版权合规；
- 可维护性：模块化设计、模型热更新、日志与监控。

---

4 数据来源与更新策略

数据来源

- 国家公共文化云、博物馆、非遗中心、文化馆等官网；
- 公开文化数据集（如CNKI文化类论文、公开非遗名录）；
- 用户上传内容（需审核）。

更新策略

- 增量采集：定时任务 + 变更检测；
- 频率：日更，重点资源实时监控；
- 版本化管理：资源版本留痕，支持回滚与diff。

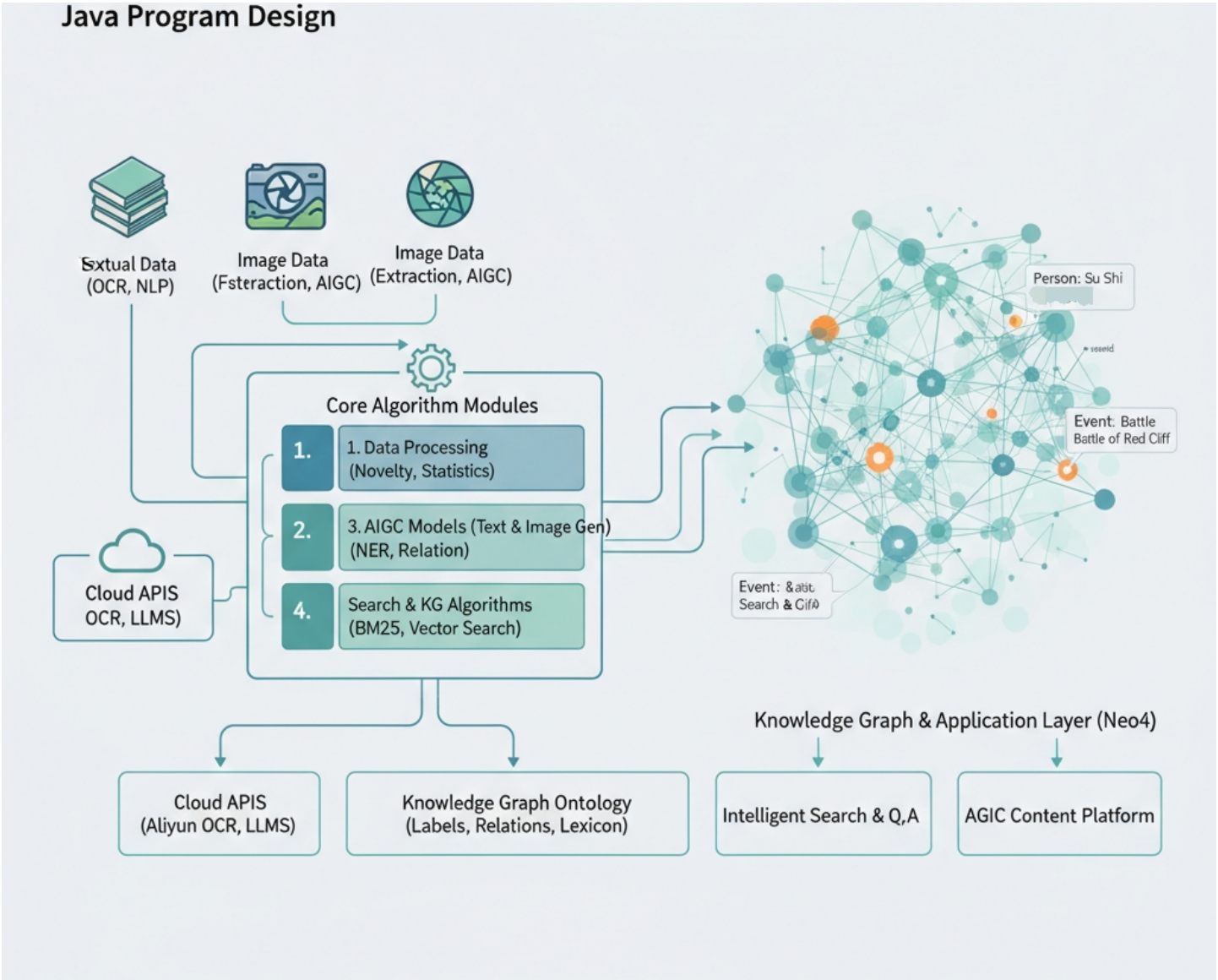
---

5 基础资源与算法

以下简要阐述支撑该项目所需的核心算法与基础模型。我们的选型原则是优先采用成熟的预训练模型与开源库，结合项目需求进行适配与应用，以确保项目的可行性与有效性。

5.1多模态数据处理技术

我们的目标是让系统能够深度理解并评估海量的文化资源。在数据接入初期，我们会运用多种统计分析算法，对资源的分布、热点及内容特征进行宏观分析。同时，为保证数据的高质量和信息增量，我们将引入新颖性计算算法，用于评估新采集的内容是否提供了新的知识，以此避免数据冗余，提升系统价值。对于图像中的文字，我们会调用阿里云等平台的API进行识别，确保高精度。



5.2AIGC生成模型

本项目的核心亮点之一是具备内容辅助创作的能力。我们将集成AIGC模型（如大语言模型），实现文化介绍文本和艺术图像的自动化生成。这样做的好处是，能够丰富平台内容生态，并为文化宣传提供多样化的创意素材。这些模型将通过独立的Python服务提供，由Java主应用进行调用和管理。

5.3自然语言处理(NLP)核心算法

为了从复杂的文本中精准地抽取出结构化知识，我们将构建一套文化领域标签体系与本体。我们的抽取方案会结合NLP模型与规则抽取两种方式。一方面，利用NLP模型进行泛化实体识别；另一方面，基于我们构建的领域词典和词间关系规则，进行高精度的信息抽取。这种混合模式的好处是，既能保证对海量数据的广泛覆盖，又能确保核心文化知识（如特定人物、流派、事件）抽取的准确性，为构建高质量知识图谱打下基础。

5.4检索引擎算法

为提供精准、高效的检索体验，我们设计了结合传统关键词与现代语义理解的混合搜索方案。该方案融合了BM25等经典算法与向量检索技术。其优势在于，不仅能响应用户的精确查找需求，更能理解其模糊的意图，甚至支持跨模态的“以图搜文”，从而显著提升信息检索的召回率和用户满意度。

5.5知识图谱相关算法

为了网络化、可视化地呈现文化知识，我们将运用图构建算法，把通过NLP技术抽取的知识实体与关系，加载到图数据库中，形成文化本体的实例化。此外，我们还将引入关系强度计算等关键算法，用于量化实体间的关联紧密程度（例如，两位画家影响力的强弱）。这样做的好处是，不仅能展示知识的全貌，还能揭示知识间的内在联系与主次关系，为用户提供更有深度的探索视角和更智能的问答服务。

---

6 系统总体架构

系统共有六个核心层级：采集层、解析层、生成层、存储层、应用层、运维层

采集层

作为系统入口，采集层主要负责从多样化来源获取公共文化资源。技术实现上，包括多模态爬虫、API接入、用户上传等多种方法

多模态爬虫利用Java、Python等语言实现的网络爬取工具，从国家公共文化云、文化馆、博物馆、非遗中心等平台自动化抓取资源，确保数据准确全面

API接入通过标准接口与外部公共文化机构对接，实现高效的数据拉取，避免重复建设

用户上传功能允许管理员或授权用户通过Web界面上传本地资源，支持文件验证和初步元数据获取。

解析层

本层主要对采集到的原始数据进行规范化处理和知识提取，包括多模态解析、实体识别和关系抽取等实现方法。

多模态解析模块针对不同类型的资源使用专用解析器，确保数据最终转换为统一格式

实体识别则通过规则驱动或现成的自然语言处理工具（如Standard NLP）识别文化实体，如任务、事件、文物等

关系抽取通过构建语义三元组（如 实体，属性，值），利用图算法提取实体间的关联，形成结构化知识

生成层

作为系统的创新核心，生成层依托AIGC技术自动创建新公共文化资源，包括AIGC模型服务和提示工程等方法。

AIGC模型服务集成外部AI API或本地虚拟模型，根据现有资源特征和服务需求生成新内容，例如合成新图像、扩展文本描述或创作虚拟非遗视频

提示工程通过优化AI输入提示，使用模板化和动态调整策略，确保生成资源的准确性和文化相关性。

存储层

提供可靠的数据持久化支持，分为MySQL（结构化存储）、向量数据库（检索优化）和文件存储（资源管理）。

MySQL数据库存储元数据、三元组和数据集，使用Java数据库连接实现高效的数据库基本操作，确保事务性和索引优化

向量数据库（如Milvus，目前最主流的开源向量数据库之一）存储嵌入向量，支持相似度检索，用于快速匹配文化资源

文件存储则采用本地文件系统或云存储（如阿里云OSS），保存原始多媒体文件，并通过哈希索引与数据库关联

应用层

本层面向用户提供实际服务，包括检索、问答、图谱可视化以及后台管理。

检索模块基于RAG（Retrieval-Augmented Generation）框架，从存储层拉取相关数据，实现自然语言搜索

问答应用支持多轮对话，使用会话管理上下文，并通过指标（如召回率、相似度）评估质量

图谱可视化利用前端库（如Echarts，JS的前端可视化库）渲染知识图谱，展示实体关系

后台管理提供资源审核、用户权限和日志查看功能

运维层

本层保障系统的稳定运行，包括任务调度、监控和模型管理。

任务调度使用Quartz框架（开源的Java任务调度框架）定时执行采集和生成任务

监控模块集成日志工具和性能指标收集，实现实时警报

模型管理支持AIGC模型的版本控制、更新和性能调优

- 总体而言，该架构以Java等为核心技术栈，遵循MVC（Model-View-Controller）模式和面向对象编程原则，实现模块化开发，便于团队协作和迭代。通过这一分层设计，我们不仅满足了项目对AIGC公共文化资源建设的功能需求，还为未来扩展（如集成更多AI模型或多模态支持）预留了接口。

---

7 数据模型设计

7.1文化资源表设计

核心字段的结构主要包括基础标识字段（ID、标题、创建时间）、内容类型字段（资源类型、文件格式）、来源信息字段（数据来源、原始URL）、内容特征字段（文本内容、图像特征向量、音频转录

文本）、生成标记字段（是否AIGC生成）、版本管理字段（版本号、变更说明）。通过数据库MySQL按照文化领域进行水平分表，对文化资源表进行向量化存储，使用FAISS存储图像和文本特征向量，MySQL存储向量ID，还可以运用MySQL存储文件元数据。通过这些分类与存储方式，可以构建统一的多模态资源存储体系，支持高效的资源检索、版本管理和质量监控。

7.2 文化实体表设计

实体表可以分为几类体系：人物实体、作品实体、事件实体、概念实体和地点实体，并且对于他们各自的属性也进行了相关的设计，设计要点为：1.核心属性：实体名称、类型、描述、来源 2.时空属性：时期年代、地理坐标、文化区域 3.特征属性：风格特征、文化价值 4.扩展属性：相关图像、数字资源链接。通过MySQL JSON字段进行动态属性的存储，还可以基于BERT的实体链接技术，解决同名的实体歧义问题；并且该文化资源系统面向范围为全球，所以支持多语言的国际化检索，需存储实体名称的多语言版本。通过一系列的技术实现，可以建立标准化的文化实体知识体系，为知识图谱构建和智能问答提供结构化基础。

7.3关系表设计

关系表的设计分为关系类型体系和关系属性设计两部分，关系类型体系：创作关系、影响关系、时空关系、相似关系、组成关系，关系属性设计为关系强度（可基于共现频率和语义相关度的量化评分）、关系证据（支撑关系的图像、来源）、时空约束、置信度评分（关系可靠的概率评估）。可以通过使用Neo4j存储实体关系网络，并且基于BERT的关系分类模型，从文本中自动提取实体关系。

7.4用户表设计

用户主要分为两类，一是普通用户，具备浏览、检索、问答的基本权限，二是系统管理员，具有用户管理、系统配置、数据维护的管理权限。还要对于用户的行为进行追踪，包括检索行为、交互行为、生成行为、标注行为。可以使用Elasticsearch存储用户的行为日志，对行为进行相关分析。通过用户表的设计，可以为设计者提供直观的数据，进一步优化用户体验。

7.5问答会话表设计

主要针对系统中的问答交互功能，可以对会话进行上下文管理，对多轮对话进行追踪，收集用户反馈。通过向量数据库存储对话上下文，支持对话的长期记忆，完整记录检索-增强-生成的环节，进行问答效果的分析，进一步优化用户体验。

7.6 标注记录表设计

标注任务分为几个体系：实体标注任务（文化实体识别、关系建立）、质量标注任务（相关性评分）、语义标注任务（主题分类、情感倾向）。在标注任务进行的过程中，需要对于标注质量进行控制，可以采用多人标注机制：通过多人独立标注，计算标注的一致性，并且还需要引入专家审核流程，构建机器标注——多人标注——专家审核的保障。可以通过引入标注工具的数据接口，集成标注平台，并且需要实时监控标注任务的进度和质量，进行相关工作的动态调整。

---

8 可视化设计

文化资源画廊（多模态展示）

页面采用整齐排列的文化资源卡片网络，可考虑在用户将鼠标移动到某个卡片时播放该卡片资源的简介朗读音频

每个资源卡片包含缩略图、标题、简短描述和分类标签（能否实现卡片平移交替？中部是否可以放入一个视频框，无声播放视频，可选用2-3个较长的视频循环播放）

顶部提供分类筛选栏和搜索框（可增加AI检索按钮）tkinter

点击任意卡片可以弹出详细层，完整地展示资源内容和相关信息

通过以上方式来提供直观地文化资源浏览体验，让用户能够快速发现和欣赏各类文化内容，增强系统的可访问性和用户参与度

【顶部右上角或左上角提供用户上传按钮，用户可点击通过文件或者文字方式上传文化资源，上传后进入审核区域，审核要求参考国家相关标准，由AI进行初审，通过AI审核后进入人工审核，人工审核通过方可加入文化实体数据库】

知识图谱可视化

（数据库中的文化资源即文化实体）

用交互式的关系网络图来展示，不同颜色节点对应不同的文化实体，节点之间的连线表示实体关系，用粗细反映关系强度。可以实现缩略功能，点击任一节点时高亮显示其**直接**关联节点。

（文化实体之间的关系考虑采用大模型生成json格式直接传入数据库，可能会存在许多问题，可预先通过人工标注校准，优化模型生成代码参数或提示词，提高召准率）

用图形化的方式来展现文化实体之间的复杂关系，帮助用户理解文化知识的关联结构，支持探索性的学习和研究。

生成效果对比界面

在AIGC界面，采用分屏对比，一侧显示原始文化资源，另一侧显示AIGC结果（采用RAG方案），可以滑动对比、缩放某一侧所占的空间

页面下方显示生成质量评分和用户反馈的评论区，最下方显示AI交互框和文件上传按钮，用户可以继续添加要求实现AIGC内容的修改

问答交互界面

页面底部固定为输入区域，支持多轮对话（可以考虑采用memory agent）。高亮显示AI回答中的文化实体，并插入超链接。支持滑动查看对话历史，并提供清除历史功能（为避免大幅占用空间，点击清除后即彻底删除。）

【能否实现词向量关联，通过选中AI生成的某一段内容就可以搜索出相关的文化实体】

后台数据看板

使用BI工具自动化生成，需要修改时仅需调整BI参数，整体布局以多图表组合的仪表盘形式为主，仪表盘中包含资源数量增长的趋势图、各分类文化资源占比、用户访问量（历史、年、月、周、日、实时，突出实时访问量）。可以通过BI工具实现时间范围或文化实体范围的筛选，所有图表能够联动更新。从而为管理员提供系统运行视图，支持决策和运维管理，保障系统的稳定运行，能够在出现BUG或访问量即将超出限制前作出判断。

---

9 验证方案与评测（侧重用户使用感受）

一、评测目标

根据系统用户需求及系统功能，从生成质量、检索性能、回答准确率、标注一致性和用户体验五个维度进行系统测评，实现主观评价与客观评测相结合、大模型测评与人工测评相结合，同时形成一套可服用的测评机制，为后期迭代优化提供支持。

评测维度	具体指标	说明
生成质量	人工评分（1-5分）	从 准确性、可读性、美观度、创新性等维度进行人工打分
	CLIPScore	使用 CLIP 模型计算生成图像与输入的文本提示词的一致性，作为客观的评价指标
检索性能	检全率	统计 “检索结果中相关资源数量 / 测试集标注的全部相关资源数量”
	检准率	统计”检索结果中相关资源数量/检索结果中的全部资源数量“
	F1值	计算检全率和检准率的加权平均
	响应时间	记录用户发起检索到获得结果的平均时间
问答准确率	问答匹配度	使用huggingface/bert-base-chinese预训练模型计算系统回答与标准答案的 BERTScore，对自动化评分存疑的结果进行人工判断。
标注一致性	F1 值	通过 Python 计算 Precision（标注正确数量 / 总标注数量）、Recall（标注正确数量 / 应标注数量），

		最终得 F1=2×(Precision×Recall)/(Precision+Recall)
用户体验	系统易用性（SUS 量表）	使用系统可用性量表（SUS）收集用户主观评价
	任务完成率	设计典型用户任务，确认任务完成率

10 技术栈与部署

一、后端技术

Java（Servlet + Tomcat）

- **实现方案**：使用Java Servlet构建系统主后端，负责请求转发、会话管理及与前端的接口通信；通过Tomcat部署Web应用。
- **意义**：确保系统结构清晰、扩展性强，并与课程教学内容（Java程序设计）相契合。

Python（AIGC模型服务）

- **实现方案**：独立运行Python微服务，用于加载和调用AIGC模型（文本生成与图像生成），与Java后端通过HTTP接口交互。
- **意义**：实现Java主系统与AI模型的解耦，提高灵活性和模型可替换性。

二、前端技术

Thymeleaf（改为python、HTML、css等）

- **实现方案**：作为模板引擎，用于动态生成HTML页面，实现服务器端渲染。
- **意义**：保证页面内容与数据库数据同步更新，便于开发与维护。

ECharts + JavaScript

- **实现方案**：ECharts用于绘制文化资源趋势图、访问量图及知识图谱关系图；JavaScript负责页面交互逻辑与数据可视化动态更新。
- **意义**：增强用户视觉体验，使文化知识的展示更直观生动。

三、数据库技术

MySQL

- **实现方案**：用于存储结构化数据（如实体、关系、用户、日志等），通过JDBC接口与后端交互。
- **意义**：保证数据一致性与可靠性，适合结构化文化资源管理。

向量数据库（FAISS）

- **实现方案**：存储文本与图像的嵌入向量，实现语义检索与相似度匹配。
- **意义**：支持跨模态检索（以图搜文、以文搜图），显著提升检索体验。

四、部署架构

（本地部署或租用服务器）

- **意义**：提高系统访问性能与安全性，确保多用户同时访问时的稳定性。

五、AI模型技术

接入API

- **实现方案**：~~（可对大语言模型进行微调）~~用于文本生成与实体识别。
- **意义**：提升自然语言处理与问答模块的准确性与流畅度。

Stable Diffusion（需不需要换其他模型）

- **实现方案**：用于文化主题图像生成与图像增强，支持AIGC资源创作模块。
- **意义**：丰富系统的多模态生成能力，提升文化传播的艺术表现力。

六、部署意义

- **Java + Python 分层结构**：兼顾系统稳定性与AI创新能力。
- **前后端协同可视化**：实现“文化数据—知识图谱—用户体验”的完整闭环。
- **综合数据库方案**：既支持传统结构化查询，又兼容现代向量检索需求，为系统扩展奠定基础。

11、 里程碑与工作流

1. M1（第 1 月）： 多模态采集与解析链路搭建

核心目标：完成多模态数据采集通道建设、预处理与标注准备

关键任务：对接公开数据源与用户上传通道；调用阿里云API实现图像 OCR、文本标准化；引入新算法进行冗余过滤；制定标签体系，明确图像标注维度

2. M2（第 2 月）：AIGC 生成与基础检索开发

核心目标：实现 AIGC 生成与检索功能

关键任务：独立的Python服务提供集成AIGC模型，设计提示词模板；开发关键词与向量检索；联调采集 - 解析 - AIGC - 检索链路

3. M3（第 3 月）：知识图谱与 RAG 问答开发

核心目标：构建知识图谱与 RAG 问答

关键任务：提取实体关系并存储于 Neo4j；搭建 RAG 框架；开发图谱可视化；测试问答准确率以及编写问答测评报告

4. M4（第 4 月）：系统集成、测试与部署

核心目标：整合系统并上线

关键任务：打通前后端与模块接口，配置权限；开展功能 / 性能 / 安全性测试，撰写测试报告；部署环境并编写部署文档；编写用户手册

12、交付清单

1. **文档类**：需求规格说明书、系统设计文档（含架构图、数据模型）、用户操作手册、测试报告、部署文档
2. **代码与数据类**：源代码（前后端 / 数据库 / 工具脚本）、配置文件、测试数据集、文化领域资源
3. **运行展示类**：系统安装包、演示视频、图谱样本数据

更新：9.25

1. 在 3.需求分析的“功能需求”部分的“**智能分析与标注**：文化实体识别、属性提取、关系构建；”处，应该进一步说明对图像数据进行数据标注时，标注哪些方面的属性；

   > 可能主要包括人物标注、文化背景与场景标注、文化符号与图标标注等

2. 算力支持方面：可能会报销调用api tokens的费用
3. 可以使用其他语言，RAG部分建议用Python
4. 在 3.需求分析的“功能需求”部分的“**AIGC资源生成**：基于提示词生成文化内容（文本、图像等）；”处，可以*参考其他文化资源相关的单位* 提供的智能问答包含的生成功能，并在可能的情况下，做更多的功能。