MuseLink-寰枢 需求规格说明书

1. 引言

1.1 编写目的

本文档旨在明确描述MuseLink-寰枢 知识图谱构建子系统的功能需求、系统特性与接口规范，作为后续系统设计、开发、测试及验收的依据。该子系统主要用于海外中国文物知识图谱的自动构建与管理。

1.2 项目背景

随着中国文物在海外博物馆的大量收藏，构建一套结构化、可链接、可拓展的文物知识图谱系统有助于实现对中国文化遗产的数字化保护与传播。本系统为该目标提供底层支撑，聚焦于数据采集、建模、补充与存储四大功能模块。

1.3 预期读者

本软件产品需求分析报告预期读者：

- 博物馆相关负责人
- 开发人员
- 文档编写人员

2. 综合描述

本项目是为博物馆开发的知识图谱，知识图谱中会存储三个海外博物馆中所收藏的中国文物的各种信息，并开放给其他开发人员使用

2.1 产品功能

系统将实现以下功能模块：

1. 数据爬取模块：从海外博物馆网站抓取中国文物的文本与图像数据；
2. 数据建模模块：将原始数据转换为三元组形式并标准化；
3. 数据补充模块：对缺失数据字段进行自动或人工补充；
4. 数据存储模块：将三元组数据写入图数据库，并发布为开放链接数据。

2.2 用户角色与特点

1. 系统管理员：负责系统运行管理，查看日志与导入/导出数据；
2. 爬虫调度人员：发起爬取任务，查看与编辑抓取结果；
3. 图谱设计人员：定义实体关系类型，管理图数据库；
4. 普通用户（未来拓展）：通过界面查看构建后的知识图谱信息。

2.3 运行环境

环境	详细信息
服务端	操作系统：Windows11 处理器：AMD R7-6800H CPU 内存：16G

环境	详细信息
客户端	操作系统：Windows11 处理器：AMD R7-6800H CPU 内存：16G
MySQL服务器	
Neo4j图数据库服务器	操作系统：Alibaba Cloud Linux 3.2104 LTS 64位 CPU&内存： 2核(vCPU)2 GiB 公网带宽：1 Mbps

3. 系统功能需求

3.1 数据爬取模块

编号	功能点	说明
F-1	网站数据批量抓取	系统支持导入网站URL列表，自动抓取所有中国文物页面内容
F-2	文物信息提取	自动解析网页结构，提取名称、介绍、朝代、材质、尺寸等信息
F-3	图像下载与统一命名	抓取并保存文物图像，自动生成统一格式的图像文件名

3.2 数据建模模块

编号	功能点	说明
F-4	数据结构化转换	将原始数据转化为三元组形式：主语-谓语-宾语
F-5	字段标准化与映射	对不同博物馆字段进行映射，统一成内部标准属性
F-6	CSV输出与MySQL导入	支持以CSV格式导出三元组，导入MySQL数据库缓存备份

3.3 数据补充模块

编号	功能点	说明
F-7	基于关键词的数据补爬	对缺失字段内容（如创作者）可调用百科资源补全
F-8	人工补充与编辑支持	支持手动填写补充信息并提交入库
F-9	补充信息版本控制	对所有补充内容进行版本记录与追踪管理

3.4 数据存储模块

编号	功能点	说明
F-10	Neo4j图谱存储	将三元组导入Neo4j图数据库，创建节点与边
F-11	图谱可视化接口支持	提供图结构展示与查询API接口
F-12	LOD格式数据发布	支持开放链接数据格式输出，供外部系统对接使用

4. 其他非功能需求

1. 性能：支持日均处理不少于1000条文物记录
2. 安全：管理模块仅限授权用户访问，操作日志需完整记录
3. 可用性：系统应支持断点续爬、数据回滚机制
4. 可维护性：各模块代码应模块化封装，便于后期扩展与维护
5. 兼容性：系统支持主流浏览器和MySQL、Neo4j数据库平台

5. 外部接口需求

5.1 用户接口

提供网页形式的图形用户界面（GUI）

6. 数据需求

数据类型与字段如下：
- 文物基本信息：ID, 名称, 介绍, 朝代, 材质, 尺寸, 图片链接
- 三元组格式：文物id — 名称 — 描述；文物id — 属于 — 博物馆 文物id — 属于— 年代；文物id — 作者 — 作者名称
- 图谱节点类型：文物、朝代、作者、描述