MuseLink-寰枢：知识图谱构建子系统设计文档

1.引言

1.1 文档目的

本文档旨在详细描述"MuseLink-寰枢"知识图谱构建子系统的技术架构和实现方案，为开发团队提供清晰的实现指导，同时为后续维护和迭代提供参考依据。

1.2 系统功能概述

知识图谱构建子系统旨在面向海外博物馆中所藏中国文物的信息进行系统化采集、处理与存储，并以结构化方式构建面向开放链接数据的知识图谱。其核心功能模块包括以下四项：

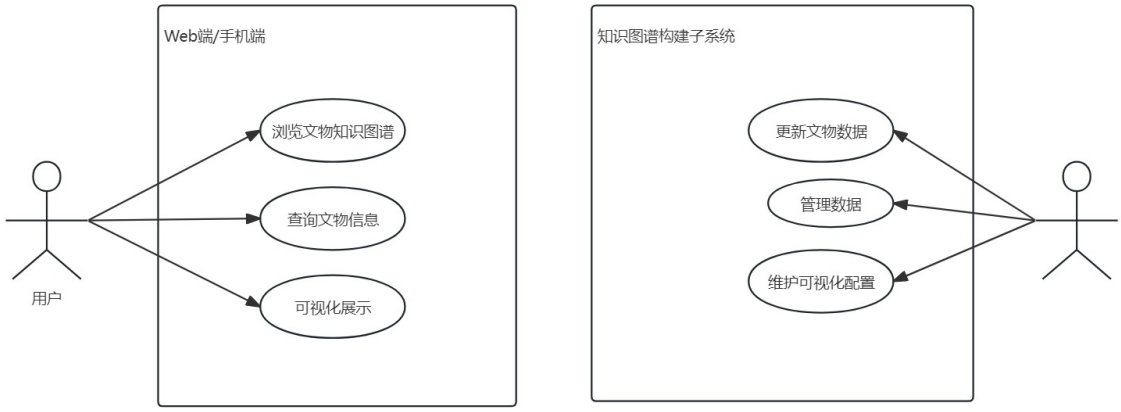
1. **数据爬取模块**
   负责自动化获取海外博物馆网站中的中国文物相关信息，包括文物名称、图片、年代、材质、介绍等内容，并按照预定格式对数据进行保存与组织，确保后续处理环节的数据规范性与完整性。
2. **数据建模模块**
   将爬取到的原始信息转化为知识图谱中通用的三元组结构（主语-谓语-宾语），并以CSV格式存储，导入MySQL数据库；图像资源将统一上传至图床进行管理与引用。
3. **数据补充模块**
   针对初步爬取数据中存在的缺失或不完整项（如文物创作者信息、背景知识等），该模块支持从第三方开放资源（如百度百科）进一步抓取补充数据，或通过人工标注方式完善现有数据体系，提升图谱的知识完整性。
4. **数据存储模块**
   将标准化后的三元组数据导入图数据库（如Neo4j或Virtuoso），用于支持后续的知识可视化、语义查询、图谱分析等多种下游任务。同时，系统支持将构建完成的知识图谱发布为开放链接数据资源，为其他模块或外部系统调用提供支持。

1.3 目标读者

- 开发团队：数据工程师
- 测试人员：功能测试、性能测试专员

2.系统架构设计

知识图谱构建子系统采用模块化设计，整体系统架构如图所示：



系统由数据采集、处理、建模、补充与存储五个主要子模块组成，模块间通过接口进行解耦协作，形成从原始数据到知识图谱构建的闭环流程。

---

3. 接口设计

1. 数据管理接口

- **数据爬取接口**
  - `museumCrawl(urlList)`：接收博物馆网址列表，批量爬取中国文物信息。
  - `supplementalCrawl(keywordList)`：根据缺失字段关键词，从指定开放资源平台（如百科类网站）补充数据。
- **数据处理接口**
  - `dataToTriple(rawData)`：将原始数据格式转化为标准三元组结构。
  - `entityEnhancement(entityList)`：针对特定实体，调用外部数据源补充属性信息。
- **数据存储接口**
  - `saveToGraphDB(tripleData)`：将三元组数据写入图数据库（Neo4j/Virtuoso）。
  - `publishAsLOD()`：对知识图谱进行开放链接数据封装，供外部系统访问与使用。

4.数据设计

1. MySQL 数据设计

本部分数据结构与后台管理子系统保持一致，用于原始数据缓存与管理。

2. Neo4j 图数据库设计

- 节点（Nodes）

类型	字段名称	数据类型	长度	能否为空	能否重复	说明
文物节点	文物编号	数值型	15	否	否	
	名称	字符型	100	否	是	可视化图中圆圈里面的字

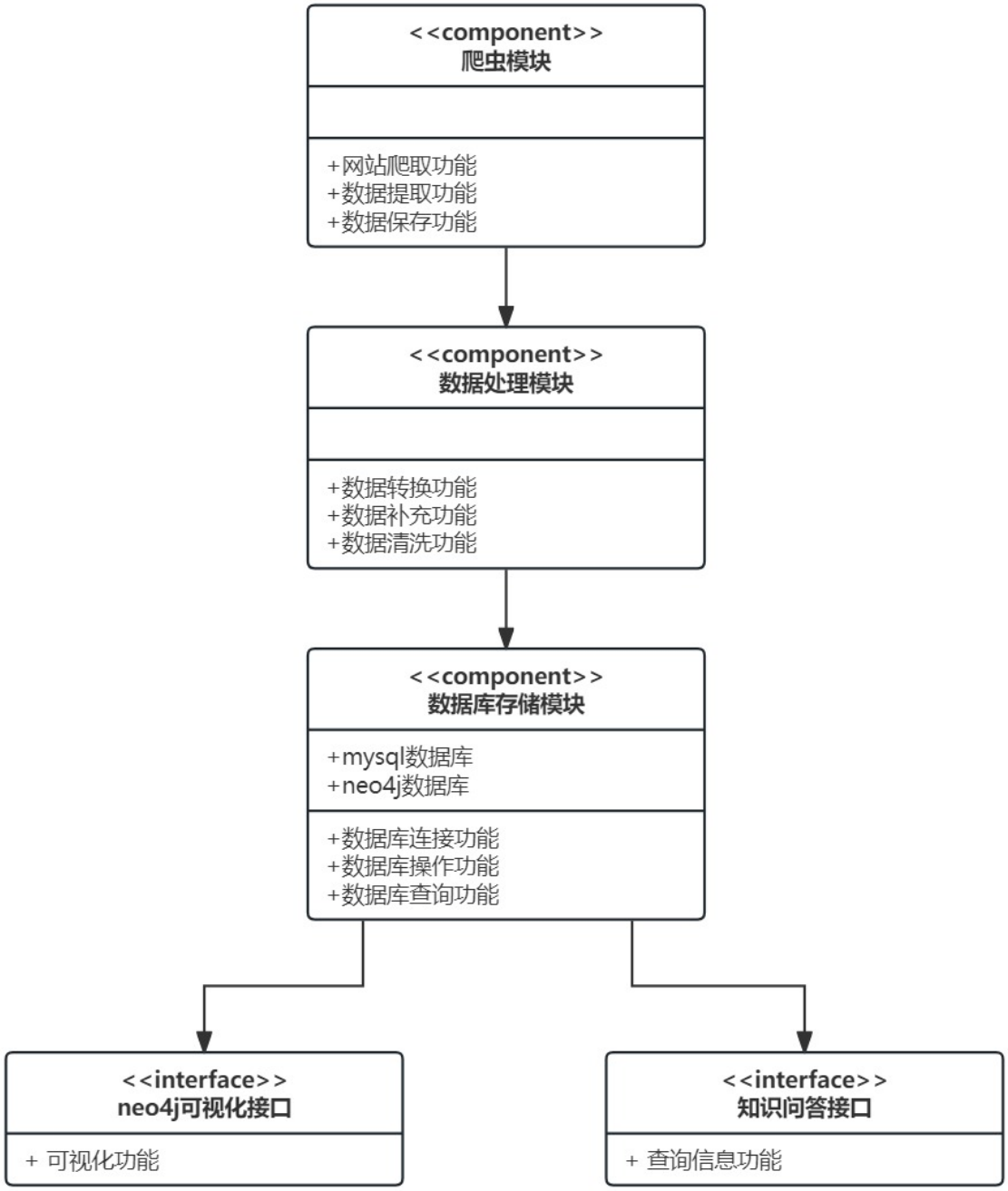
类型	字段名称	数据类型	长度	能否为空	能否重复	说明
	介绍	字符型	500	是	否	
朝代节点	朝代名称	字符型	50	否	否	可视化图中圆圈里面的字
作者节点	作者名称及介绍	字符型	100	否	否	可视化图中圆圈里面的字

- 边（Relationships）

关系名称	关系类型	方向	源节点	目标节点	属性	说明
制造于	制造于	文物 -> 朝代	文物节点	朝代节点	无	文物属于某个朝代
创作者	创作者	文物 <- 作者	文物节点	作者节点	无	文物的作者
藏品来源	藏品来源	文物 -> 藏品来源	文物节点	藏品来源节点	无	文物的来源

5.模块功能设计

各功能模块的逻辑结构如图所示：



- **数据采集模块**：负责网页结构解析、内容提取、图片下载与命名处理。
- **数据建模模块**：完成字段映射、关系抽取、三元组生成。
- **数据补充模块**：自动爬虫与人工校验结合，完善不完整信息。
- **图谱存储模块**：Neo4j驱动接入，实现节点与边的写入操作。

6.界面设计


Database Information
Use database
neo4j
Node labels
*(41,716)
Artifact
Artist
Museum
Period
_GraphConfig
_NsPrefDef
Relationship types
*(103,926)
作者
包含
年代
Property keys
_applyNeo4jNaming
_classLabel
_classNamePropName
_dataTypePropertyLabel
_domainRel
_handleMultival
_handleRDFTypes
_handleVocabUris
_keepCustomDataTypes
_keepLangTag
_objectPropertyLabel
_rangeRel
_relNamePropName
_subClassOfRel
_subPropertyOfRel
artifact_title
creation_date
descripe
ex
id
label
name
period
title
url
neo4j$
neo4j$ MATCH (m:Museum)-[r1:包含]→(a:Artifact) MATCH (a)-[r2:年代]→(p:Period) MATCH (a)-[r3:作者]→(artist:Artist) RETURN m, r1, a, r…
Graph
Table
Text
Code
Overview
Node labels
* (955)
Museum (1)
Artifact (500)
Period (220)
Artist (234)
Relationship types
* (1500)
包含 (500)
年代 (500)
作者 (500)
Displaying 955 nodes, 1,500 relationships.
$:server status
Connection status
This is your current connection information.
You are connected as user neo4j
to neo4j://123.56.94.39:7687
Connection credentials are stored in your web browser.