知识库管理界面基础核心功能（MVP 版）需求文档

一、核心功能模块
- 1. 系统管理（用户、日志等）
- 2. 服务状态监管仪表板（首页）
- 3. 组件及服务管理（按服务类型分组）
- 4. 备份恢复（独立功能模块）
- 5. 数据批量导入
- 6. 非功能性需求（部署方式、用户体验）
- 7. 开发优先级建议

1. 用户与权限管理（基础版）
功能需求：
- 用户登录与登出
- 两种角色：管理员（完整权限）和操作员（仅查看）
- 用户管理（创建、编辑、删除）
- 基础操作日志（记录关键操作）

技术方案建议：
- 推荐使用独立鉴权服务，基于 Token 的身份验证（JWT），支持后续对 anything-llm 及其他服务的用户管理及鉴权能力。
- 考虑轻量级数据库存储用户信息（如 SQLite、文件存储）
- 用户创建、更改、登入的简单日志记录到文件即可满足第一阶段需求

2. 服务状态监管
功能需求：
- 服务健康状态监控（在线/离线/异常）
- 关键资源使用情况：
  - 推理服务：状态、GPU/NPU 使用率、内存使用
  - 向量库：连接状态、存储空间使用量
  - 文档解析：服务状态、任务队列
- 基础服务控制：启动、停止、重启
- 服务日志查看功能

技术方案建议：
- 通过各服务的健康检查接口或管理 API 获取状态
- 可使用简单的 HTTP 轮询机制，间隔 15-30 秒
- 服务控制可通过调用各服务的控制接口实现
- 日志可通过代理查看或直接访问日志文件

3. 组件部署与管理

3.1 一键部署向导

功能需求：
- 部署模式选择：全新部署和对接现有服务
- 硬件环境选择：英伟达和昇腾
- 推理引擎选择：vLLM 和 MindIE
- 逐步配置向导：
  1. 管理界面配置，所有服务总览和导航
2. 推理引擎及模型服务配置
  3. 向量库配置
  4. 文档解析服务配置
- 能生成、导出、导入部署脚本或配置文件支持自动化配置

技术方案建议：
- 可提供不同部署方式的模板（如 Kubernetes manifests、Ansible playbooks、shell 脚本）
- 根据用户选择动态生成配置文件
- 支持导出配置包供用户离线部署

3.2 对接现有服务
功能需求：
- 手动配置服务地址和端口
- 连接测试功能
- 配置保存和管理

3.3 模型配置管理
功能需求：
- 已加载模型列表查看
- 模型配置参数管理：
  - 推理参数：并发数、输入 token 限制、输出 token 限制
  - 批处理参数：批处理大小、批处理延迟
  - 资源参数：GPU/NPU 分配策略
- 配置保存、导出、导入支持应用一点配置
- 模型部署、接入、配置更改、和模型升级及切换功能

技术方案建议：
- 通过推理服务的配置 API 或配置文件管理参数
- 支持配置模板，便于复用
- 配置变更时可能需要重启服务，需明确提示

4. 备份与恢复

4.1 备份功能
功能需求：
- 全量备份（手动触发）
- 增量备份（基于时间或变化触发）
- 备份进度显示
- 备份列表查看

- 备份验证（完整性检查）

4.2 恢复功能
功能需求：
- 从备份点恢复
- 恢复进度显示
- 恢复后验证

4.3 应用重刷
功能需求：
- 所有服务的镜像重新部署
- 基于备份的重置流程做恢复

5. 数据批量导入
批量文件导入功能应支持：
- 从文件系统路径（本地、NFS）及对象存储路径批量导入数据到知识库，并支持检索。
- 提供专用的批量导入面板，支持状态显示、可断点续传的传输机制，以及基于文件名与修改日期（校验和）的变化检测功能。

具体要求：
1. 状态显示 – 提供任务级别总览，显示文件总数与已处理数量。支持以下操作：
   - 全部任务的重新开始/暂停。
   - 单个任务的开始/暂停。
   - 失败任务的重新开始/清除。

2. 变化检测 – 支持启用/禁用扫描功能，并允许配置扫描间隔：
   - 可选模式：实时、每小时、每日、每周、每月。
   - 提供用于管理备份周期的控制面板。

二、技术方案建议（非强制）

6. 部署方案需求
5.1 容器化部署
- 优势：环境隔离、易于分发、扩展性好
- 实现方式：Docker-compose、Kubernetes
- 建议：使用容器编排工具简化管理

5.2 备份方案选择
基于应用原生能力及文件系统的混合备份方案
- 适合：所有数据都存储在文件系统中
- 数据库及向量库：使用相关组件的原生备份能力
- 文件数据：使用文件系统能力
- 采用简单的时间点增量备份，易于理解和维护

5.3 监控方案建议
轻量级方案：
 - 各服务提供 API 健康检查端点
 - 管理界面定期轮询
 - 关键指标记录到日志文件

5.4 安全性需求
- HTTP 和 HTTPS 支持
- 密码加密数据导出（配置及备份）存储（至少 SHA256）
- 操作审计日志

5.5 界面设计要点

导航结构建议
```
建议分组：
- 监控仪表板（首页）
- 部署配置（集中配置入口）
- 服务管理（按服务类型分组）
- 备份恢复（独立功能模块）
- 系统管理（用户、日志等）
```

关键交互流程

A. 部署向导流程：
```
1. 选择部署类型 → 2. 配置服务参数 → 3. 生成部署包 → 4. 执行部署/提供指引
```

B. 备份流程：
```
1. 选择备份类型 → 2. 配置备份参数 → 3. 执行备份 → 4. 查看结果
```

C. 恢复流程：
```
1. 选择备份点 → 2. 确认恢复选项 → 3. 执行恢复 → 4. 验证结果
```

5.5 非功能性要求

性能建议
- 管理界面响应时间：关键操作<3 秒，普通页面<1 秒
- 状态轮询间隔：15-30 秒（避免过载）

- 备份性能：增量备份应在 5 分钟内完成

可靠性
- 管理界面自身需要高可用
- 关键操作需要确认提示
- 操作失败需要有明确的错误信息和恢复指引

安全性
- 数据传输使用 HTTPS
- 敏感信息加密存储
- 操作需要权限验证

可维护性
- 清晰的日志记录
- 配置可导出/导入

6. 开发优先级建议

必须实现支持昇腾及英伟达生态
1. 用户登录和基础权限
2. 所有服务状态监控，启动/停止控制
3. 对接现有服务配置包括推理引擎、模型、基本配置管理
4. 一键式部署及配置包括推理引擎、模型、知识库方案的基本配置管理
5. 知识库方案部署及全量及增量的备份和恢复