上海双地信息系统有限公司

2019-8-2

上海图书馆  
全文OCR等检索功能项目

需求分析报告

目 录

[1. 功能设计 2](#_Toc24012944)

[1.1. 功能架构 2](#_Toc24012945)

[1.2. 全文OCR资源扩展 3](#_Toc24012946)

[1.3. 检索功能 3](#_Toc24012947)

[1.3.1. 普通检索 3](#_Toc24012948)

[1.3.2. 高级检索 3](#_Toc24012949)

[1.3.3. 专业检索 4](#_Toc24012950)

[1.3.4. 检索浏览 4](#_Toc24012951)

[1.3.5. 简繁检索一致性 4](#_Toc24012952)

[1.3.6. 文献导航 5](#_Toc24012953)

[1.4. 整本浏览 5](#_Toc24012954)

[1.4.1. 整本浏览 5](#_Toc24012955)

[1.4.2. 整本检索 6](#_Toc24012956)

[1.5. 镜像站服务 6](#_Toc24012957)

[1.5.1. 镜像站管理 6](#_Toc24012958)

[1.5.2. 镜像站前端 7](#_Toc24012959)

# 功能设计

## 功能架构

## 全文OCR资源扩展

对于全文OCR资源扩展，服务平台主要有以下两方面工作：

1. 数据框架扩展

数据框架是整个服务平台的基础组件，主要是用来维护各类型资源的元数据定义的。数据框架包含了文献类型管理、文献分类管理、文章类型管理、文献字段管理、文献文章字段管理、检索引擎字段管理、数据版本控制等功能。

为了扩展支持全文OCR资源，需要在文献、文章、检索引擎等各个层面进行相应的扩展，配置元数据支持精品资源的呈现、导航、检索等功能。

1. 数据转换扩展

数据转换发生在数据从生产平台到服务平台迁移的过程中，新的资源加入，我们需要为新资源配置相应的转换规则，保障数据从生产平台转换的数据，经过数据转换后，能够满足在服务平台数据框架中定义约束。

## 检索功能

### 普通检索

普通检索，为了实现对全文OCR资源的正文文本内容的检索支持，我们是通过将正文文本内容加入到检索引擎的All 字段中实现。All字段包括整个检索引擎支持的所有检索字段的内容，因此将正文文本加入到All中，即可在普通检索中检出。

### 高级检索

服务平台的高级检索的资源关联的字段是通过配置高级检索字段来实现，我们为全文OCR资源配置相应的支持高级检索的字段，分别按照正文、图片和广告进行了配置。

另外，在高级检索的产品树上，根据要求加上了全文OCR资源的节点，一旦用户选择了这个节点，那在检索下拉字段中自动添加全文OCR特定的高级检索字段，用户可以使用这些字段进行组合检索。

### 专业检索

专业检索中，为了支持全文OCR资源，我们为需要扩展的检索字段，配置特定代码，这些代码要求唯一性，和其他所有的字段编码都不一样，这样，用户在检索框中输入合法的检索语法，包含这些字段的时候，后台即可按照唯一编码，转换为实际的检索语句，能够支持全文OCR的特定字段的检索。

### 检索浏览

检索结果按照之前的需求分析要求，我们对摘要和全文都进行了高亮呈现，按照不同的呈现类型进行显示。全文高亮需要经过特殊的处理，主要是因为在加工过程中，这些数据包含了一些特殊的标记，需要对这些内容进行过滤。

### 简繁检索一致性

另外，我们将通过权限的Schema设计，从检索引擎的数据定义层面，将简繁检索结果一致性问题进行解决。具体思路是，用户输入的信息全部转换为简体内容，然后基于Schema中预先转换为简体的字段进行检索，显示的时候仍然显示原有的内容，这样可以保障简繁检索结果的一致性。

唯一需要注意的是，因为涉及简繁转换，我们检索的实际是简体字段，高亮的内容也是基于简体内容，需要将高亮信息同步到原始的字段内容，然后对用户呈现，这里基于的原则是按字转换，一个简体字和一个繁体字对应的原则。

### 文献导航

对于全文OCR资源，文献导航是归于近代期刊。为了实现这个目标，就要求将全文OCR资源在文献分类中归于近代期刊下，我们在数据框架上进行这样的定义。

另外，为了实现文献导航的OCR显示和隐藏的功能，通了将通过开关的方式是，来构造检索筛选语句，保障OCR资源按照需要进行显示。

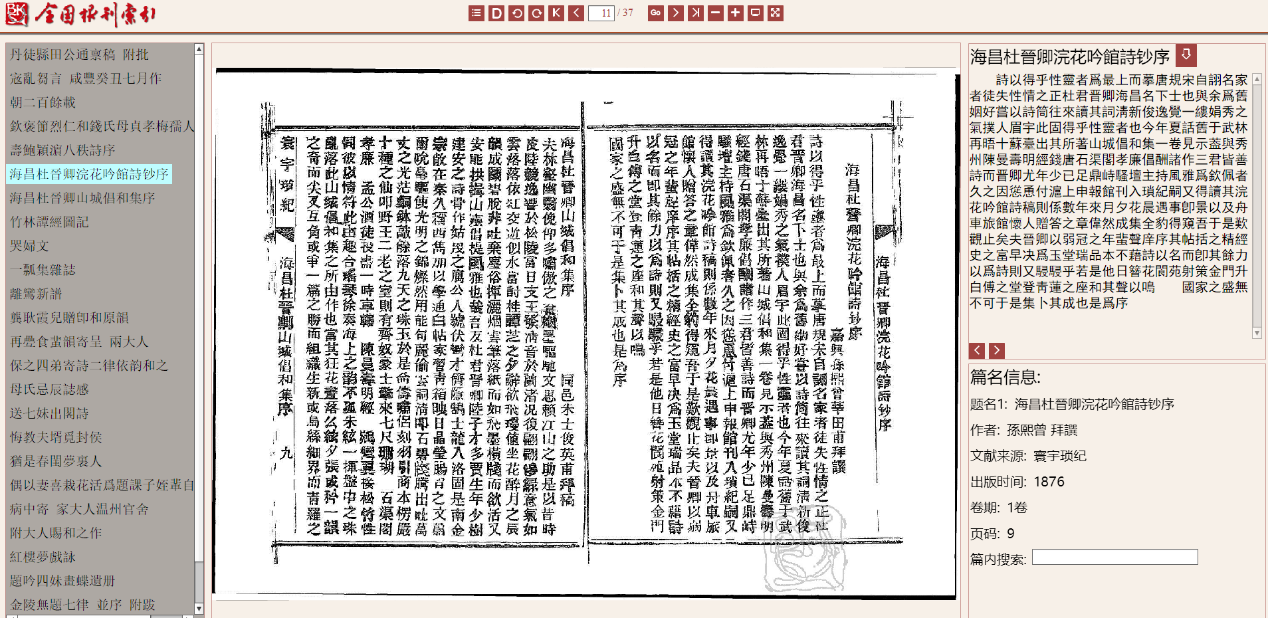
## 整本浏览

### 整本浏览

全文OCR资源的整本浏览，是比较复杂的，我们必须支持图文并排的呈现，能够按照目录结构进行组织。

目前，整个服务平台的整本浏览主要是基于Flex技术实现，我们同样继承了这个传统，仍然采用Flex技术，按照组件化的设计思想来实现这一功能。在数据的组织上，我们按照全文OCR的资源特点进行整理和整合，充分利用缓存技术，既可减少交互次数，又可减少交互内容，提高整体响应效率。

整个整本浏览的设计如下图所示：



1. 顶部是功能按钮区，提供了需求中定义的各种功能；
2. 左侧是目录去，可以呈现整本刊物的目录结构；
3. 中间是图片区，显示当前页面的图片；
4. 右侧上面是全文区，显示当前选中文章在当前页面的内容；
5. 右侧下面是篇名信息区，提供了篇名的相关信息。

### 整本检索

整本检索，是对单页文本内容提供检索功能，检索匹配的内容红色高亮显示，同时也支持简繁检索。为了实现这一目标，同时减少和服务端的交互次数，我们将文本内容的简体版本和原始版本全部传递到客户端，检索过程中，输入的内容按照简体版本进行检索，匹配项目按照位置同步到原始内容上，保障整个整本检索的正确性和高效性。

## 镜像站服务

### 镜像站管理

镜像站管理为了支持全文OCR资源，我们将进行如下改造：

1. 升级索引结构，支持全文OCR的索引数据，这个可以通过在服务平台进行数据框架扩展后，通过导出的方式进行Schema输出。
2. 升级缓存工具，全文OCR资源和原有的期刊不一样的地方在于包含了图片和广告数据，这些新的类型也需要加入到缓存中；
3. 将加密算法应用到新的索引和资源转换中，保障新资源的镜像安全。

### 镜像站前端

镜像站前端按照服务平台的功能调整进行代码同步调整，因为镜像站采用的是不同框架，因此只有部分核心逻辑可以重用，框架部分的调整仍然需要按需进行改造。