DI项目

化学和生物序列检索

接口说明书

知识产权出版社

长城计算机与系统有限公司

DI项目组

2015年3月

版本历史说明

|  |  |  |  |
| --- | --- | --- | --- |
| **版本** | **作者** | **日期** | **备注** |
| V1.0.1 |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

目 录

[版本历史说明 1](#_Toc416976782)

[目 录 i](#_Toc416976783)

[1. 概述 1](#_Toc416976784)

[1.1. 编写说明 1](#_Toc416976785)

[1.2. 范围说明 1](#_Toc416976786)

[1.3. 相关人员 1](#_Toc416976787)

[2. 化学结构检索 2](#_Toc416976788)

[2.1. 流程示意图 2](#_Toc416976789)

[2.2. 流程描述 2](#_Toc416976790)

[2.2.1. 化学检索流程描述 2](#_Toc416976791)

[2.2.2. 数据入库流程描述 3](#_Toc416976792)

[3. 生物序列检索 5](#_Toc416976793)

[3.1. 流程概览 5](#_Toc416976794)

[3.2. 流程描述 5](#_Toc416976795)

[3.2.1. 生物序列检索流程 5](#_Toc416976796)

[3.2.2. 数据入库流程描述 7](#_Toc416976797)

[4. 接口定义一览表 8](#_Toc416976798)

[4.1. 化学检索接口定义 8](#_Toc416976799)

[4.1.1. 化学式检索控件 8](#_Toc416976800)

[4.1.2. 化学式检索接口 8](#_Toc416976801)

[4.1.3. 查看化学物质接口 9](#_Toc416976802)

[4.1.4. 查看商家接口 9](#_Toc416976803)

[4.1.5. 批次数据入库进度查询接口 9](#_Toc416976804)

[4.1.6. 数据加载调用通知接口 9](#_Toc416976805)

[4.2. 化学数据更新文件定义 10](#_Toc416976806)

[4.2.1. 化学数据控制文件 10](#_Toc416976807)

[4.2.2. 化学数据索引文件 11](#_Toc416976808)

[4.2.3. 化学数据包文件 14](#_Toc416976809)

[4.3. 化学数据反馈文件定义 14](#_Toc416976810)

[4.3.1. 化学数据反馈控制文件 14](#_Toc416976811)

[4.3.2. 化学数据反馈数据文件 14](#_Toc416976812)

[4.4. 生物序列检索接口定义 18](#_Toc416976813)

[4.4.1. 生物序列检索页面 18](#_Toc416976814)

[4.4.2. 生物序列检索调用接口 24](#_Toc416976815)

[4.4.3. 生物序列检索任务查询接口 24](#_Toc416976816)

[4.4.4. 生物序列详情接口 25](#_Toc416976817)

[4.4.5. 批次数据入库进度查询接口 26](#_Toc416976818)

[4.4.6. 数据加载调用通知接口 26](#_Toc416976819)

[4.5. 生物序列数据更新文件定义 27](#_Toc416976820)

[4.5.1. 生物序列控制文件 27](#_Toc416976821)

[4.5.2. 生物序列数据索引文件 28](#_Toc416976822)

[4.5.3. 生物序列数据包文件 30](#_Toc416976823)

[4.6. 生物序列数据反馈文件定义 30](#_Toc416976824)

[4.6.1. 生物序列反馈控制文件 30](#_Toc416976825)

[4.6.2. 生物序列反馈数据文件 30](#_Toc416976826)

# 概述

## 编写说明

本文档主要用于定义化学结构检索和生物序列检索的整个流程，明确出版社与长软、长软与鹰谷在化学结构和生物序列检索流程中的职责，并对具体工作提出清晰定义。

## 范围说明

本文档是出版社与长软、长软与鹰谷之间的工作接口沟通文档，在此基础上将细化生成接口定义文档。

出版社和鹰谷之间化学结构和生物序列检索需求，不包含在本文档。

各方内部与接口无关的工作，诸如鹰谷化学式检索控件、ICS号注册机制、出版社生物序列ID生成机制等内容，不包含在本文档。

## 相关人员

本文档的预期读者是DI项目相关方管理人员、需求人员、设计人员、开发人员。

# 化学结构检索

## 流程示意图



## 流程描述

### 化学检索流程描述

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 流程概述 | 流程描述 | 责任方 | 相关方 |
| 1 | 化学式检索入口 | 鹰谷提供化学式检索控件，长软集成到检索页面；用户通过控件输入化学结构式检索时，长软调用化学式检索接口，访问鹰谷化学检索引擎获得对应化学物质检索结果； | 长软 | 鹰谷 |
| 2 | 化学结构概览页面 | 长软接受鹰谷化学物质检索结果反馈，在页面进行化学结构概览展示。 | 长软 | 鹰谷 |
| 3 | 3.1用户查看化学物质 | 鹰谷提供“查看化学物质”接口，供长软进行调用展示，使用户可以查看化学物质信息； | 鹰谷 | 长软 |
| 3.2用户查看商家 | 鹰谷提供“查看商家”接口，供长软进行调用展示，使用户可以查看化学物品供应商信息； | 鹰谷 | 长软 |
| 3.3用户进行相关专利检索 | 用户在化学结构概览页面，依据化学结构ICS，调用专利检索引擎进行专利检索； | 长软 | TRS |
| 4 | 专利概览 | 长软在概览子系统，对化学物质专利检索结果展示； | 长软 | TRS |
| 5 | 专利详览 | 用户可以选择专利文献，进行详览。 | 长软 | TRS |

### 数据入库流程描述

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 流程概述 | 流程描述 | 责任方 | 相关方 |
| 1 | 化学结构数据文件提供 | 出版社将化学结构数据（国内/国外数据文件），按照数据文件交换规范，以批次数据方式提交长软。  其中：1)控制文件、索引文件、数据包文件要求见数据交换规范；2）数据文件提供时，提供标准格式文献号和化学数据文件对应关系。 | 出版社 | 长软 |
| 2 | 数据包校验 | 数据运维平台对数据包进行校验（MD5校验）后，将化学结构数据上传到指定FTP目录（更新发送目录）； | 长软 |  |
| 3 | 数据更新入库 | 鹰谷定时扫描指定FTP目录（更新发送目录），获取文件进行数据入库； | 鹰谷 |  |
| 4 | 批次数据入库进度查询 | 鹰谷提供批次数据入库进度查询接口，供长软按控制文件名称，查询鹰谷数据入库处理进度； | 鹰谷 | 长软 |
| 5 | 映射关系反馈 | 鹰谷返回标准格式文献号和化学信息（ICS、化学物质名称）对应关系文件、错误记录及错误原因描述文件，上传到FTP目录（更新接受目录）； | 鹰谷 |  |
| 6 | 检索库更新 | 长软定时扫描指定FTP目录（更新接受目录），对反馈数据进行处理，其中映射关系更新到TRS检索库，异常记录进行后续处理。 | 长软 |  |
| 7 | 数据加载调用通知接口 | 鹰谷提供数据加载调用通知接口，供长软按照控制文件名称调用，通知鹰谷对特定批次数据，在指定时间加载入库；  主要用于异常数据更正后，需实时加载的更正批次数据。接口调用后，数据更新按照正常流程处理。 | 鹰谷 | 长软 |

# 生物序列检索

## 流程概览



## 流程描述

### 生物序列检索流程

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 流程概述 | 流程描述 | 责任方 | 相关方 |
| 1 | 生物序列检索页面 | 鹰谷提供生物序列检索页面，由长软对界面重新整合到DI项目； | 鹰谷 | 长软 |
| 2 | 生物序列检索接口 | 用户页面输入检索条件，确定检索后，由长软集成页面调用鹰谷生物序列检索接口，发起生物序列检索； | 长软 | 鹰谷 |
| 3 | 生物序列检索异步任务管理 | 长软调用生物检索接口后，建立异步任务管理机制，按照鹰谷接口返回的估计时间，进行异步任务结果查询处理； | 长软 | 鹰谷 |
| 4 | 检索结果查询服务接口 | 鹰谷提供异步检索结果查询接口，由长软异步管理任务调用，获得生物检索结果数据文件存放相对路径和名称。 | 鹰谷 | 长软 |
| 5 | 生物检索结果文件生成 | 鹰谷对于生物序列检索结果，在双方共享的检索结果交换目录下，生成两个XML格式的检索结果文件；其中一个为检索结果列表展示数据，一个文件为检索结果对应的生物序列比对结果。  检索结果列表展示数据要求如下：1）应按照匹配度从高到低排列；2）提供检索总记录数，以实现页面分页浏览；3）提供对应于每条数据记录，在详细数据文件中，生物序列对比结果数据的偏移索引。 | 鹰谷 | 长软 |
| 6 | 检索结果展示页面 | 长软依据检索返回文件，实现检索结果展示，包括分页显示、查看生物序列详情、查看比对结果详情。 | 长软 |  |
| 7 | 查看生物序列详情 | 长软调用生物序列详情查询接口，依据生物序列ID，获取生物序列详情信息； | 鹰谷 | 长软 |
| 8 | 查看专利文献 | 长软使用生物序列检索结果中的专利文献标准号，调用详览子系统查看专利文献。 | 长软 |  |

### 数据入库流程描述

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 流程概述 | 流程描述 | 责任方 | 相关方 |
| 1 | 生物序列数据文件提供 | 出版社将生物序列数据文件（含国内/国外数据文件），按照数据文件交换规范，以批次数据方式提交长软。  其中：1)控制文件、索引文件、数据包文件要求见数据交换规范；2)生物序列数据文件由出版社以专利文献为粒度加工成FASTA格式, 每个专利文献只有一个生物序列文件；3）一个生物序列文件可以包含多个生物序列；4）出版社针对每个生物序列生成唯一生物序列ID（只保证序列ID唯一，序列不去重）；5）对生物序列需要在详情页面展示的属性，在生物序列索引文件中提供。 | 出版社 | 长软 |
| 2 | 数据包校验 | 数据运维平台对数据包进行校验（MD5校验）后，将生物序列数据上传到“更新发送目录”FTP目录； | 长软 |  |
| 3 | 数据更新入库 | 鹰谷定时扫描指定FTP目录（更新发送目录），获取文件进行数据入库； | 鹰谷 |  |
| 4 | 批次数据入库进度查询 | 鹰谷提供批次数据入库进度查询接口，供长软按控制文件名称，查询鹰谷数据入库处理进度； | 鹰谷 | 长软 |
| 5 | 映射关系反馈 | 鹰谷返回标准格式文献号和生物序列ID对应关系文件、错误记录及错误原因描述文件，上传到“更新接受目录”FTP目录； | 鹰谷 | 长软 |
| 6 | 检索库更新 | 长软定时扫描指定FTP目录（更新接受目录），对反馈数据进行处理，其中映射关系更新到TRS检索库，异常记录进行后续处理。 | 长软 |  |
| 7 | 数据加载调用通知接口 | 鹰谷提供数据加载调用通知接口，供长软按照控制文件名称调用，通知鹰谷对特定批次数据，在指定时间加载入库；  主要用于异常数据更正后，需实时加载的更正批次数据；接口调用后，数据更新按照正常流程处理。 | 鹰谷 | 长软 |

# 接口定义一览表

本文档定义的接口均为REST接口。

## 化学检索接口定义

在目前环境下为保障项目顺利实施，化学检索接口采用现有提供的JDK接口进行开发，但双方系统间接口应基于共同的JDK等开发环境。双方系统间接口应采用JDK1.7版本。正式提交的接口为REST方式。

### 化学式检索控件

提供方：上海鹰谷

提供化学结构式输入控件；配合长软进行页面集成。

### 化学式检索接口

提供方：上海鹰谷

输入：通过鹰谷控件输入的化学式及其参数、请求最大查询记录数（预设为1000）、每页记录数、请求记录页数；

处理逻辑：在页面通知长软发起化学式检索；鹰谷模糊匹配检索化学物质库，按照相似度从高到底，提供当前页化学物质检索结果列表；

输出：总记录条数、当前页显示的化学物质ICS号的XML或JSON数据

说明：比如按照“最大查询记录数1000条、每页20条、第2页数据”请求时，鹰谷需按照相似度从高到底，提供第21-40条化学数据；

实现方式：REST接口

### 查看化学物质接口

输入：化学物质ICS号

处理逻辑：长软调用鹰谷查看化学接口，由鹰谷检索化学物质库，提供化学物质详细信息；

输出：化学物质详细信息XML或JSON数据

实现方式：REST接口

### 查看商家接口

输入：化学物质ICS号

处理逻辑：长软调用鹰谷查看商家URL接口，由鹰谷返回商家URL地址；

输出：商家信息URL的字符串

实现方式：REST接口

### 批次数据入库进度查询接口

输入：控制文件名称

处理逻辑：鹰谷查询控制文件数据入库进度，返回本批次数据入库进度；

输出：错误代码、错误描述、本批次数据入库状态（尚未入库、入库中、入库完成），并返回入库详细进度的XML或JSON数据（文献总数、文件总数、成功入库记录数、出错记录数，以及每个数据索引文件入库明细[数据索引文件名称、文献总数、文件总数、入库记录数、出错记录数]）；

实现方式：REST接口

### 数据加载调用通知接口

输入：控制文件名称、加载时间（14位YYYYMMDDHHMISS的字符串）

处理逻辑：鹰谷接收到控制文件名称后，判断是否已经启动本批次数据加载；如未启动控制文件对应的数据加载，则按照要求时间启动本批次数据加载；

输出：调用结果的返回代码、返回结果描述

实现方式：REST接口

## 化学数据更新文件定义

### 化学数据控制文件

数据控制文件用于表示数据交换批次。数据控制文件符合XML格式规范，文件编码为UTF-8。

* 数据控制文件定义

|  |  |  |
| --- | --- | --- |
| **节点** | **节点类型** | **说明** |
| control节点 | | |
| type | 属性 | 数据资源类型。  CHEMICAL：化学物质（指专利文献与化学物质对照数据） |
| dowork | 属性 | 是否可以启动数据装载。本字段表示数据文件是否生成完全，是否可以启动数据装载工作。  0：需等待数据文件完成  1：可以启动数据装载工作  **当本属性为1时，可以启动数据装载工作，且一个批次内数据，除删除数据包外可以并发处理。** |
| filelist\file节点 | | |
| filename | 属性 | 数据索引文件名称。 |
| status | 属性 | 数据更新方式（优先处理删除数据包）  C：新增  U：更新  D：删除  CU：新增或更新  对中国专利文献，涉及到需要创建新化学物质CID问题，需要明确区分新增、更新和删除。 |
| section | 属性 | 数据资源类型，扩展字段。 |
| sequence | 属性 | 数据加载顺序，扩展字段。 |
| md5 | 属性 | 对索引文件校验字段。 |

* 化学控制文件样例：



### 化学数据索引文件

#### 中国化学数据索引文件

中国化学数据索引文件符合XML格式规范，文件编码为UTF-8。中国化学数据索引文件无对应的数据包文件，在索引文件上提供化学对应的simles数据。

* 化学数据索引文件定义

|  |  |  |
| --- | --- | --- |
| **节点** | **节点类型** | **说明** |
| content节点 | | |
| contenttype | 属性 | 用于区分中国化学结构和国外化学结构数据。  CHEM\_CN：中国化学数据（中国专利数据无数据包文件）  其他情况或无此节点：非中国化学数据 |
| dateExchange | 属性 | 数据交付日期，8位，YYYYMMDD |
| dateProduced | 属性 | 数据生产日期，8位，YYYYMMDD |
| patcnt | 属性 | 数据包内专利文献数量 |
| status | 属性 | 数据更新方式  C：新增  D：删除 U：全字段更新 |
| content\doclist节点 | | |
| topic | 属性 | 数据主题 |
| country | 属性 | 国别，符合国别标准代码 |
| docNumber | 属性 | 案卷编号  对于专利著录项目数据包，“国别+案卷编号+文献类型”三字段关联是标准格式的文献号； |
| kind | 属性 | 文献类型 |
| PNO | 属性 | 原始格式的文献号 |
| PNS |  | 标准格式的文献号 |
| datePublication | 属性 | 公开日，8位，YYYYMMDD |
| amount | 属性 | 化学式个数 |
| status | 属性 | 数据更新方式  当content节点的status属性确定后，doclist的status属性与上面一致 C：新增  D：删除 U：全字段更新 |
| content\doclist\SMILES节点 | | |
| SMILES | 元素 | Mol文件转化成的字符串 |

* 中国化学数据索引文件样例



#### 国外化学数据索引文件

国外化学数据索引文件符合XML格式规范，文件编码为UTF-8。每个数据索引文件所对应的数据文件包，其大小不超过2G。

* 化学数据索引文件定义

|  |  |  |
| --- | --- | --- |
| **节点** | **节点类型** | **说明** |
| content节点 | | |
| file | 属性 | 数据压缩包的文件名称，ZIP格式压缩 |
| dateExchange | 属性 | 数据交付日期，8位，YYYYMMDD |
| dateProduced | 属性 | 数据生产日期，8位，YYYYMMDD |
| patcnt | 属性 | 数据包内专利文献数量 |
| filecnt | 属性 | 数据包内文件总件数 |
| size | 属性 | 数据包大小 |
| md5 | 属性 | 数据包MD5校验码（在数据加载前需验证数据包MD5校验码） |
| status | 属性 | 数据更新方式  C：新增  D：删除 U：全字段更新  CU：新增或更新 |
| content\doclist节点 | | |
| topic | 属性 | 数据主题 |
| country | 属性 | 国别，符合国别标准代码 |
| docNumber | 属性 | 标准格式的案卷编号  对于专利著录项目数据包，“国别+案卷编号+文献类型”三字段关联是标准格式的文献号； |
| kind | 属性 | 文献类型 |
| PNO | 属性 | 原始格式的文献号 |
| PNS | 属性 | 标准格式的文献号 |
| datePublication | 属性 | 公开日，8位，YYYYMMDD |
| path | 属性 | 数据文件所在的相对路径 |
| status | 属性 | 数据更新方式  当content节点的status属性确定后，doclist的status属性与上面一致 C：新增  D：删除 U：全字段更新  CU：新增或更新 |
| content\doclist\filelist节点 | | |
| amount | 属性 | 文件数量 |
| path | 属性 | 文件所在父目录名 |
| content\doclist\filelist\file节点 | | |
| filename | 属性 | 文件名称  数据包解压缩后，“doclist\path”属性+“doclist\filelist\path”属性+本节点filename属性，获得数据文件解压后的访问路径及名称 |
| filetype | 属性 | 文件类型，如csv等 |
| section | 属性 | 数据类型码，化学物质文件数据类型码为“CHE” |

* 国外化学数据索引文件样例



### 化学数据包文件

国外化学数据包文件是逗号分割的csv文件，在索引文件中存储了专利文献号和化学物质文件的对应关系。

* 国外化学数据包文件样例



## 化学数据反馈文件定义

鹰谷完成化学数据入库后，返回专利标准文献号、化学物质ID（ICS）、化学物质名称的对应关系数据文件。

### 化学数据反馈控制文件

化学数据反馈控制文件，参见“化学数据更新文件定义”部分控制文件定义。数据反馈时应按照原控制文件名称、索引文件名称进行反馈，以便于对数据更新过程进行对应记录。

### 化学数据反馈数据文件

#### 中国化学数据反馈文件

中国化学数据反馈文件符合XML格式规范，文件编码为UTF-8。

* 化学数据反馈文件定义

|  |  |  |
| --- | --- | --- |
| **节点** | **节点类型** | **说明** |
| content节点 | | |
| contenttype | 属性 | 用于区分中国化学结构和国外化学结构数据。  CHEM\_CN：中国化学数据（原数据返回） |
| dateExchange | 属性 | 数据交付日期，8位，YYYYMMDD（原数据返回） |
| dateProduced | 属性 | 数据生产日期，8位，YYYYMMDD（原数据返回） |
| patcnt | 属性 | 数据包内专利文献数量（原数据返回） |
| status | 属性 | 数据更新方式（原数据返回）  C：新增  D：删除 U：全字段更新 |
| content\doclist节点 | | |
| topic | 属性 | 数据主题（原数据返回） |
| country | 属性 | 国别，符合国别标准代码（原数据返回） |
| docNumber | 属性 | 标准格式的案卷编号（原数据返回） |
| kind | 属性 | 文献类型（原数据返回） |
| PNO | 属性 | 原始格式的文献号（原数据返回） |
| PNS | 属性 | 标准格式的文献号（原数据返回） |
| datePublication | 属性 | 公开日，8位，YYYYMMDD（原数据返回） |
| status | 属性 | 数据更新方式（原数据返回）  当content节点的status属性确定后，doclist的status属性与上面一致 C：新增  D：删除 U：全字段更新 |
| ICS | 属性 | 文献号对应的化学物质ID，多值间用半角分号分割 |
| icsnames | 属性 | 文献号对应的化学物质名称，多值间用半角分号分割 |
| content\doclist\SMILES节点 | | |
| simle | 属性 | 提交的smile值（原数据返回） |
| code | 属性 | 错误代码，“000000”表示处理无错误 |
| errmsg | 属性 | 错误描述，对错误代码的文字描述信息；当错误描述内容过长且有可能影响XML解析时，使用CDATA标签进行返回。 |

* 中国化学数据反馈文件样例



#### 国外化学数据反馈文件

国外化学数据反馈文件符合XML格式规范，文件编码为UTF-8。

* 化学数据反馈文件定义

|  |  |  |  |
| --- | --- | --- | --- |
| **节点** | | **节点类型** | **说明** |
| content节点 | | | |
| file | | 属性 | 化学数据索引文件名称，即提供给数据对应加工的索引文件名称；（原数据返回） |
| dateExchange | | 属性 | 数据交付日期，8位，YYYYMMDD（原数据返回） |
| dateProduced | | 属性 | 数据生产日期，8位，YYYYMMDD（原数据返回） |
| patcnt | | 属性 | 数据包内专利文献数量（原数据返回） |
| filecnt | | 属性 | 数据包内文件总件数（原数据返回） |
| status | | 属性 | 数据更新方式（原数据返回）  C：新增  D：删除 U：全字段更新  CU：新增或更新 |
| content\doclist节点 | | | |
| topic | | 属性 | 数据主题（原数据返回） |
| country | | 属性 | 国别，符合国别标准代码（原数据返回） |
| docNumber | | 属性 | 标准格式的案卷编号（原数据返回） |
| kind | | 属性 | 文献类型（原数据返回） |
| PNO | | 属性 | 原始格式的文献号（原数据返回） |
| PNS | | 属性 | 标准格式的文献号（原数据返回） |
| datePublication | | 属性 | 公开日，8位，YYYYMMDD（原数据返回） |
| path | | 属性 | 数据文件所在的相对路径（原数据返回） |
| status | | 属性 | 数据更新方式（原数据返回）  当content节点的status属性确定后，doclist的status属性与上面一致 C：新增  D：删除 U：全字段更新  CU：新增或更新 |
| ICS | | 属性 | 文献号对应的化学物质ID，多值间用半角分号分割 |
| icsnames | | 属性 | 文献号对应的化学物质名称，多值间用半角分号分割 |
| content\doclist\filelist节点 | | | |
| amount | 属性 | | 文件数量 |
| path | 属性 | | 文件所在父目录名 |
| content\doclist\filelist\file节点 | | | |
| datafile | | 属性 | 化学数据文件名称，即返回原数据包内的数据文件名称（原数据返回） |
| code | | 属性 | 错误代码，“000000”表示处理无错误 |
| errmsg | | 属性 | 错误描述，对错误代码的文字描述信息；当错误描述内容过长且有可能影响XML解析时，使用CDATA标签进行返回。 |

* 国外化学数据反馈文件样例



## 生物序列检索接口定义

### 生物序列检索页面

鹰谷应提供如下图所示宜的核酸序列检索和氨基酸序列检索页面，并配合长软进行页面集成。生物序列检索页面如下图：

#### 核酸序列检索页面



#### 核酸序列检索页面说明

| **序号** | **栏位名称** | **栏位说明** | **备注** |
| --- | --- | --- | --- |
| 1 | 国别 | 1. 数据来源于生物序列数据库来源国 2. 下拉列表方式，仅单选 3. 来源国要通过来源国字典进行维护。 4. 默认显示中国 | 选择项 |
| 2 | 选库别 | 1. 单选，选项仅包括核算序列和氨基酸序列 2. 与国别组合形成生物序列检索数据库类别 3. 默认无选中项 | 选择项 |
| 3 | 检索序列输入栏 | 1. 大文本框 2. 最多输入2000字符 3. 可以手工拷贝和粘贴生物序列信息 4. 默认为空 5. 需要支持用户打开本地ST.25和FASTA格式的生物序列文件，将生物序列数据粘贴在输入区 | 可编辑 |
| 4 | 表达式检索类型选择 | 1. 选择项，包括核酸和氨基酸 2. 单选 3. 默认无选中项 4. 用于指定表达式要在检索库内检索的生物序列类 |  |
| 5 | Blast检索算法 | 1. 单选框，每次只能选择一个； 2. MegaBlast、非连续性MegaBlast、Blastn三种选 3. 默认选择MegaBlast |  |
| 6 | 设定子序列 | 1. 编辑栏，可输入 2. 用于展示指定符合检索条件的生物序列的序列段数据 3. 如果两个栏位数据为空，则展示每个生物序列的全部序列数据 4. 如果起始框内数据为空而终止框内数据不为空（且不为零），则展示从1-终止框内数字的序列段数据；如果起始框内数据不为空而终止框内数据为空，则展示从起始框内数值到序列结束之间的序列段内数据；如果两个编辑框内数据都不为空，且起始框内数字小于终止框内数据，则展示两个数值之间的序列段数据。如果两个编辑框内数值相等，则展示该数值对应的序列数据。 5. 两个数值编辑框内只能填写大于0的数值型数据。 | 编辑 |
| 7 | 结果返回序列条数 | 1. 下拉框，单选 2. 具体参数由第三方产品提供商提供 | 选择框 |
| 8 | 启动短序列调整 | 1. 复选框，默认选中 2. 提供自动检测短序列相应调整参数选项 | 选择 |
| 9 | 相似性阈值 | 1. 下拉框，单选，默认为空 2. 2、具体参数由第三方产品提供商提供 | 选择 |
| 10 | 匹配次长度 | 1. 下拉框，单选 2. 具体参数由第三方产品提供商提供 | 选择 |
| 11 | 检索范围内最多允许匹配次数 | 1. 下拉框，单选 2. 具体参数由第三方产品提供商提供 | 选择 |
| 12 | 匹配/未匹配参数 | 1. 下拉框，单选 2. 具体参数由第三方产品提供商提供 |  |
| 13 | 空位罚分 | 1. 下拉框，单选 2. 具体参数由第三方产品提供商提供 |  |
| 14 | 过滤器 | 1. 复选框，默认选中 2. 低复杂度区域一个选项 |  |
| 15 | 屏蔽 | 1. 复选框，包含仅屏蔽查找表和屏蔽小写字母两种选项 2. 默认选择仅屏蔽查找表 |  |
| 16 | 恢复默认参数 | 1. 按照系统默认参数设置，对页面参数和输入表达式进行重新设置。 2. 具体默认参数包括范围和设置规则，由第三方指定。 3. 刷新页面数据 | 按钮 |
| 17 | 检索 | 通知页面调用生物序列检索接口，接受预计响应时间，启动异步任务管理 | 按钮 |

#### 氨基酸序列检索页面



#### 氨基酸序列检索页面说明

| **序号** | **栏位名称** | **栏位说明** | **备注** |
| --- | --- | --- | --- |
| 1 | 国别 | 1. 数据来源于生物序列数据库来源国 2. 下拉列表方式，仅单选 3. 来源国要通过来源国字典进行维护。 4. 默认显示中国 | 选择项 |
| 2 | 选库别 | 1. 单选，选项仅包括核算序列和氨基酸序列 2. 与国别组合形成生物序列检索数据库类别 3. 默认无选中项 | 选择项 |
| 3 | 表达式输入栏 | 1. 大文本框   最多输入2000字符   1. 可以手工拷贝和粘贴生物序列信息 2. 默认为空 3. 需要支持用户打开本地ST.25和FASTA格式的生物序列文件，将生物序列数据粘贴在输入区 | 可编辑 |
| 4 | 表达式检索类型选择 | 1. 选择项，包括核酸和氨基酸 2. 单选 3. 默认无选中项 4. 用于指定表达式要在检索库内检索的生物序列类型 |  |
| 5 | Blast检索算法 | 1. 单选框，每次只能选择一个； 2. Blastp、PSI-Blast、PHI-Blast、Delta-Blast四种选项 3. 默认选择Blastp |  |
| 6 | 生物序列编码格式选择 | 1. 复选框；用于选择一字码或三字码；默认无选中项 2. 系统默认是一字码；支持用户针对核酸或氨基酸编码格式进行检索，一字码是默认项，不需要再界面展示。 |  |
| 7 | 设定子序列 | 1. 编辑栏，可输入 2. 用于展示指定符合检索条件的生物序列的序列段数据 3. 如果两个栏位数据为空，则展示每个生物序列的全部序列数据 4. 如果起始框内数据为空而终止框内数据不为空（且不为零），则展示从1-终止框内数字的序列段数据；如果起始框内数据不为空而终止框内数据为空，则展示从起始框内数值到序列结束之间的序列段内数据；如果两个编辑框内数据都不为空，且起始框内数字小于终止框内数据，则展示两个数值之间的序列段数据。如果两个编辑框内数值相等，则展示该数值对应的序列数据。 5. 两个数值编辑框内只能填写大于0的数值型数据。 | 编辑 |
| 8 | 结果返回序列条数 | 1. 下拉框，单选 2. 具体参数由第三方产品提供商提供 | 选择框 |
| 9 | 启动短序列调整 | 1. 复选框，默认选中 2. 提供自动检测短序列相应调整参数选项 | 选择 |
| 10 | 相似性阈值 | 1. 下拉框，单选，默认为空 2. 具体参数由第三方产品提供商提供 | 选择 |
| 11 | 匹配次长度 | 1. 下拉框，单选 2. 具体参数由第三方产品提供商提供 | 选择 |
| 12 | 检索范围内最多允许匹配次数 | 1. 下拉框，单选 2. 具体参数由第三方产品提供商提供 | 选择 |
| 12 | 打分矩阵 | 1. 下拉框，单选 2. 具体参数由第三方产品提供商提供 |  |
| 13 | 空位罚分 | 1. 下拉框，单选 2. 具体参数由第三方产品提供商提供 |  |
| 14 | 补偿调整 | 1. 下拉框，单选 2. 具体参数由第三方产品提供商提供 |  |
| 15 | 过滤器 | 1. 复选框，默认选中 2. 低复杂度区域一个选项 |  |
| 16 | 屏蔽 | 1. 复选框，包含仅屏蔽查找表和屏蔽小写字母两种选项 2. 默认选择仅屏蔽查找表 |  |
| 17 | 恢复默认参数 | 1. 按照系统默认参数设置，对页面参数和输入表达式进行重新设置。 2. 具体默认参数包括范围和设置规则，由第三方指定 3. 刷新页面数据 | 按钮 |
| 18 | 检索 | 调通知页面调用生物序列检索接口，接受预计响应时间，启动异步任务管理 | 按钮 |

### 生物序列检索调用接口

提供方：上海鹰谷

输入：生物序列检索入口页面用户输入条件及参数

处理逻辑：按照用户界面参数在鹰谷生物库进行生物序列检索，返回任务ID、预计反馈时间；

输出：任务ID、预计反馈时间；

实现方式：REST接口；

### 生物序列检索任务查询接口

提供方：上海鹰谷

输入：任务ID

处理逻辑：鹰谷依据任务ID，判断生物检索任务是否完成，并返回当前状态、生物检索结果列表文件、生物序列详情文件；

输出：任务ID、当前状态（检索完成、排队中、检索中、鹰谷定义的其他异常）、生物检索结果列表文件名称、生物序列详情文件名称；

实现方式：REST接口；

备注：在异步数据反馈的方式下，鹰谷在双方共享的检索结果交换目录下，生成检索返回的两个结果文件；其中一个为检索结果列表展示数据，一个文件为列表对应的生物序列比对信息。检索结果列表数据要求如下：1）应按照匹配度从高到低排列；2）提供检索总记录数，以实现页面分页浏览；3）提供对应每条数据记录，在详细数据文件中，对应生物序列对比数据的偏移量。

#### 生物检索结果列表文件格式

文本文件格式。

第一行：“半角等号”+检索到的生物序列总数，如“=100”

第二行开始为生物序列列表数据，每条记录一行，以半角分号分割各字段：

原始文献号(PNO);标准文献号(PNS);来源(doclist节点的country);生物序列ID(seq节点的id);序列类型;分数;期望值;起始位置(生物序列比对结果详情文件中起始偏移量，字节为单位);长度(比对结果数据字节长度)

#### 生物序列比对结果详情文件

文本文件格式，以字节为单位计算偏移量和比对结果详情。

以“Length=”开始，对生物序列数据以下面图例样式反馈。下图效果的每行数据应以回车换行符换行。



### 生物序列详情接口

提供方：上海鹰谷

输入：生物序列ID

处理逻辑：鹰谷依据生物序列ID，检索返回生物序列详情；

输出：调用结果代码（“000000”表示调用成功）、错误信息描述、生物序列详情XML或JSON串；

生物序列详情数据包括：

<?xml version="1.0" encoding="UTF-8"?>

<seq id="CN301967775A\_000001">

<length>729</length>

<type>DNA</type>

<gn>蚓激酶F-Ⅱ</gn>

<og>线粒体</og>

<organism>Eisenia fetida</organism>

<features>

<feature sequence="1">

<location>(1)..(324)</location>

<keywords>CDS</keywords>

<other>AMIDATION</other>

</feature>

<feature sequence="2">

<location>(6)..(732)</location>

<keywords>CDS</keywords>

<other> 蚓激酶F-Ⅱ全基因DNA</other>

</feature>

</features>

<seqtext>

<![CDATA[

------------以常规生物序列显示格式存放生物序列详情----------

]]>

</seqtext>

</seq>

实现方式：REST接口

### 批次数据入库进度查询接口

输入：控制文件名称

处理逻辑：鹰谷查询控制文件数据入库进度，返回本批次数据入库进度；

输出：错误代码、错误描述、本批次数据入库状态（尚未入库、入库中、入库完成），并返回入库详细进度的XML或JSON数据（文献总数、文件总数、成功入库记录数、出错记录数，以及每个数据索引文件入库明细[数据索引文件名称、文献总数、文件总数、入库记录数、出错记录数]）；

实现方式：REST接口

### 数据加载调用通知接口

输入：控制文件名称、加载时间（14位YYYYMMDDHHMISS的字符串）

处理逻辑：鹰谷接收到控制文件名称后，判断是否已经启动本批次数据加载；如未启动控制文件对应的数据加载，则按照要求时间启动本批次数据加载；

输出：调用结果的返回代码、返回结果描述

实现方式：REST接口

## 生物序列数据更新文件定义

### 生物序列控制文件

数据控制文件用于表示数据交换批次。数据控制文件符合XML格式规范，文件编码为UTF-8。

* 数据控制文件定义

|  |  |  |
| --- | --- | --- |
| **节点** | **节点类型** | **说明** |
| control节点 | | |
| type | 属性 | 数据资源类型。  BIOLOGY：生物序列数据 |
| dowork | 属性 | 是否可以启动数据装载。本字段表示数据文件是否生成完全，是否可以启动数据装载工作。  0：需等待数据文件完成  1：可以启动数据装载工作  **当本属性为1时，可以启动数据装载工作，且一个批次内数据，除删除数据包外可以并发处理。** |
| filelist\file节点 | | |
| filename | 属性 | 数据索引文件名称。 |
| status | 属性 | 数据更新方式（一个控制文件内，优先处理删除的数据包）  C：新增  D：删除 U：全字段更新 |
| section | 属性 | 数据资源类型，扩展字段 |
| sequence | 属性 | 数据加载顺序，扩展字段 |
| md5 | 属性 | 对索引文件校验字段 |

* 生物序列控制文件样例：



### 生物序列数据索引文件

数据索引文件符合XML格式规范，文件编码为UTF-8。每个数据索引文件所对应的数据文件包，其大小不超过2G。

* 生物序列数据索引文件定义

|  |  |  |
| --- | --- | --- |
| **节点** | **节点类型** | **说明** |
| content节点 | | |
| file | 属性 | 数据压缩包的文件名称，ZIP格式压缩 |
| dateExchange | 属性 | 数据交付日期，8位，YYYYMMDD |
| dateProduced | 属性 | 数据生产日期，8位，YYYYMMDD |
| patcnt | 属性 | 数据包内专利总数量 |
| filecnt | 属性 | 数据包内文件总件数 |
| size | 属性 | 数据包大小 |
| md5 | 属性 | 数据包MD5校验码（在数据加载前需验证数据包MD5校验码） |
| status | 属性 | 数据更新方式（使用控制文件来控制加载索引文件的顺序） C：新增 D：删除 U：全字段更新 |
| content\doclist节点 | | |
| topic | 属性 | 数据主题 |
| country | 属性 | 国别，符合国别标准代码 |
| docNumber | 属性 | 标准格式的案卷编号  对于专利著录项目数据包，“国别+案卷编号+文献类型”三字段关联是标准格式的文献号； |
| kind | 属性 | 文献类型 |
| PNO | 属性 | 原始格式的文献号 |
| PNS | 属性 | 标准格式的文献号 |
| datePublication | 属性 | 公开日，8位，YYYYMMDD |
| Format | 属性 | 生物序列文件的格式  FASTA：FASTA生物序列文件规范 |
| path | 属性 | 数据文件所在的相对路径 |
| status | 属性 | 数据更新方式  当content节点的status属性确定后，doclist的status属性与上面一致 C：新增 D：删除（删除是对文献级别的删除） U：全字段更新 |
| content\doclist\file节点 | | |
| filename | 属性 | 文件名称  数据包解压缩后，“doclist\path”属性+本节点filename属性，获得数据文件解压后的访问路径及名称 |
| filetype | 属性 | 文件类型，可以为空，预留字段 |
| section | 属性 | 数据类型码，生物序列文件数据类型码为“SEQ |
| content\doclist\file\seqlist节点 | | |
| amount | 属性 | 序列数目 |
| content\doclist\file\seqlist\seq节点 | | |
| id | 属性 | 序列ID |
| length | 元素 | 序列长度 |
| type | 元素 | 序列类型 |
| gn | 元素 | 基因名称 |
| og | 元素 | 亚细胞定位 |
| organism | 元素 | 来源生物体 |
| content\doclist\file\seqlist\seq\features\feature节点 | | |
| sequence | 属性 | 特征序号 |
| location | 元素 | 序列起始、终止位置 |
| keywords | 元素 | 名称或关键词 |
| other | 元素 | 其他信息 |

* 生物序列数据索引文件样例



### 生物序列数据包文件

生物序列数据文件由出版社以专利文献为粒度加工成FASTA格式, 每个专利文献只有一个生物序列文件；一个生物序列文件可以包含多个生物序列；数据包文件中对每个生物序列生成唯一生物序列ID（只保证序列ID唯一，序列不去重）；对生物序列需要在详情页面展示的属性，在生物序列索引文件中提供。

* 生物序列数据文件样例



## 生物序列数据反馈文件定义

### 生物序列反馈控制文件

生物序列数据反馈控制文件，参加“生物序列数据更新文件定义”部分控制文件定义；在控制文件中，应能明确和原数据文件的对应关系。

### 生物序列反馈数据文件

数据反馈文件符合XML格式规范，文件编码为UTF-8。

* 生物序列映射反馈文件定义

|  |  |  |
| --- | --- | --- |
| **节点** | **节点类型** | **说明** |
| content节点 | | |
| file | 属性 | 生物序列数据索引文件名称，即提供给数据对应加工的索引文件名称； |
| dateExchange | 属性 | 数据交付日期，8位，YYYYMMDD（原数据返回） |
| dateProduced | 属性 | 数据生产日期，8位，YYYYMMDD（原数据返回） |
| status | 属性 | 数据更新方式  C：新增  D：删除 U：全字段更新 |
| content\doclist节点 | | |
| topic | 属性 | 数据主题（原数据返回） |
| country | 属性 | 国别，符合国别标准代码（原数据返回） |
| docNumber | 属性 | 标准格式的案卷编号（原数据返回） |
| kind | 属性 | 文献类型（原数据返回） |
| PNO | 属性 | 原始格式的文献号（原数据返回） |
| PNS | 属性 | 标准格式的文献号（原数据返回） |
| datePublication | 属性 | 公开日，8位，YYYYMMDD（原数据返回） |
| Format | 属性 | 生物序列文件的格式（原数据返回）  FASTA：FASTA生物序列文件规范 |
| path | 属性 | 数据文件所在的相对路径（原数据返回） |
| status | 属性 | 数据更新方式（原数据返回）  当content节点的status属性确定后，doclist的status属性与上面一致 C：新增 D：删除（删除是对文献级别的删除） U：全字段更新 |
| biologyid | 属性 | 对应的生物序列ID，多值间用半角分号分割 |
| content\doclist\file节点 | | |
| datafile | 属性 | 生物序列数据文件名称，即提供给数据对应加工的数据文件名称； |
| code | 属性 | 错误代码，数据出错返回 |
| errmsg | 属性 | 错误描述，对错误代码的文字描述信息；当错误描述内容过长且有可能影响XML解析时，使用CDATA标签进行返回。 |

* 生物序列映射反馈文件样例

