# 1 基于微服务架构的多源异构大数据管理平台GTL2系统

## 系统架构

## GTL2旨在基于微服务架构，实现一个多源异构大数据管理平台，架构如图所示。

## data%20service.jpg

## GTL2平台具备下列主要功能：

## 1）具备结构化数据管理功能，也就是能够管理关系数据库中的数据；并支持以服务的方式提供增删查改；

## 2）具备文件和目录管理功能，并支持以服务的方式方式提供增删查；

## 3）具备非结构化文本数据处理与管理功能，包括文本抽取、文本索引、文本存储、文本搜索、中文语言处理等；

## 4）具备非结构化图形数据处理与管理功能，包括空间对象抽取，空间索引、空间数据存储和空间查询与分析等；

## 5）具备非结构化栅格数据处理与管理功能，包括栅格数据划分、栅格数据索引、栅格数据存储、栅格数据查询与分析等；

## 6）可以方便地为各种应用定制为符合行业规则的大数据管理系统；

## 7）可以统一的服务接口，为其他系统提供数据支撑；

## 8）有方便的用户管理界面。

## GTL2中需要开发的功能采用单文档发布方式。编码规则为GTL2-TODO-xxx，其中xxx为功能编号。其对应的文档为GTL2-TODO-xxx.docx。

## 已经实现的功能点：

## 3-1 输入一个数据目录和一个索引目录，根据传入的文件过滤器和文档映射方式，构建Lucene的Document。

## cn.edu.cug.gtl.lucene.document.DocumentCreator

## 3-2 从Office文件中提取文本，并提供按照整个文件一个字符串、一个自然段一个字符串、一行一个字符串三种方式组织

## cn.edu.cug.gtl.lucene.text.TextExtractor

## 3-3 文件过滤功能，包括单个文件过滤，多个文件组合过滤，用户定义文件过滤，提供了接口

## cn.edu.cug.gtl.lucene.file.DocumentFileFilter

## 外部只能通过该接口访问lucene提供的所有文件过滤器，内部文件过滤器包括：

## cn.edu.cug.gtl.lucene.file.AllFileFilter

## cn.edu.cug.gtl.lucene.file.OfficesFileFilter Offices文件（doc,docx,xls,xlsx,ppt,pptx,pdf,rtf）过滤器

## cn.edu.cug.gtl.lucene.file.TextsFileFilter 文本文件（txt, csv,tsv,html,xml,json）过滤器

## cn.edu.cug.gtl.lucene.file.ShapesFileFilter 图形文件(shp,dwg,dxf)过滤器

## cn.edu.cug.gtl.lucene.file.ImagesFileFilter 图像文件(jpg,png,tif,bmp)过滤器

## cn.edu.cug.gtl.lucene.file.CompoundFileFilter 复合文件过滤器

## cn.edu.cug.gtl.lucene.file.UDFileFilter 用户根据文件后缀自定义的过滤器

## 3-4 目录中文件访问器，提供了基类

## cn.edu.cug.gtl.lucene.file.DocumentFileVisitor

## 对于每个文件的处理功能，可以重载此基类，重载visitFile函数，例如

## DocumentCreator.CreatorFileVisitor 将每个文件传化成一个或多个Document

## DocumentIndexer.IndexerFileVisitor 将一个文件添加到索引库中

## 3-5 输入一个数据目录和一个索引目录，根据传入的文件过滤器、文档映射方式和更新方式，构建Lucene索引。

## cn.edu.cug.gtl.lucene.document.DocumentIndexer

## 3-6 输入一个索引目录，根据传入的检索字符串，输出全文检索结果

## cn.edu.cug.gtl.lucene.document.DocumentSearcher

## 开发过程中，建议的目录设置为：

## 代码放在 ~/git/gtl-java中，

## 测试数据放在 ~/git/data中。

## ~表示当前用户主目录。

## 在当前用户目录下,创建git目录

## ZhenwendeMacBook-Pro:~ zhenwenhe$ mkdir git

## ZhenwendeMacBook-Pro:~ zhenwenhe$ cd git

## ZhenwendeMacBook-Pro:git zhenwenhe$ git clone https://github.com/ZhenwenHe/gtl-java.git

## Series说明：

## 对于~/git/gtl-java/Series 需要将UCRArchive\_2018整个目录放置在~/git/data下。此外，~/git/data/log下必须有一个series.db文件，里面记录了运行参数。可以将~/git/gtl-java/Series/series.db文件拷贝过去。生成的测试结果放在~/git/data/output/hax\_nn.xls中。

## 如果不采用上述默认设置，则需要采用，

## Config.setDataDirectory()设置数据根目录。

## 在飞测试代码中，获取数据目录请采用Config.getDataDirectory()。

## 在编写所有的测试代码过程中，

## 请使用Config.getTestInputDirectory()

## 函数获取测试数据的根目录。

## 输出结果目录存放位置请采用

## Config.getTestOutputDirectory()

## 函数获取。

## 编译，进入gtl-java目录，采用mvn命令编译

## root@namenode1:/gtl-java# mvn package -Dmaven.test.skip=true