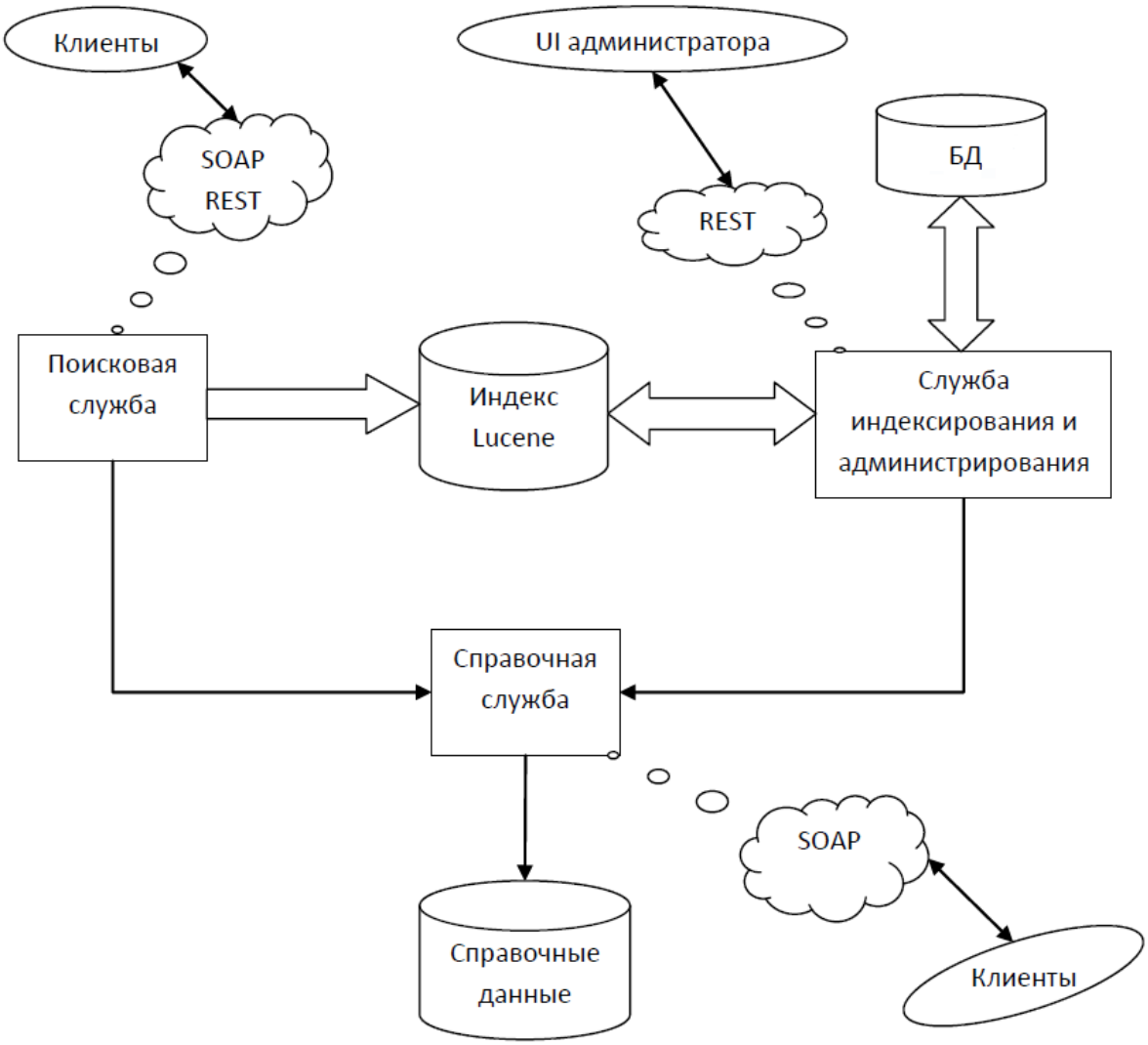
ДЕМЛИЗ
технология будущего - сегодня

2DL.Beagle

Описание системы

Общая схема



Описание поисковой системы

Система состоит из трёх основных модулей: индексатора, поискового сервиса и справочного сервиса.

Индексатор запускается в отдельном процессе с целью обеспечить его независимость от остальных модулей. Поисковый (search) и справочный (inquiry) сервисы запускаются как сервлет biblios.war на сервере приложений (tomcat). Кроме этих трёх основных модулей есть ещё дополнительный модуль - Администратор поисковой системы (biblioadmin.war). Модули используют единый файл конфигурации, в котором записаны основные параметры системы.

Индексатор

Индексатор представляет собой веб-сервис с интерфейсами SOAP и REST. Индексатор может создавать или обновлять поисковый индекс, необходимый для работы поискового сервиса. Индексатор получает исходные данные из т.н. источников данных, которые определены в файле конфигурации. Может быть определено несколько источников, данные из которых будут последовательно загружены и проиндексированы в один индекс. Основные параметры источника - имя файла с MARC записями, путь к хранилищу текстов документов и имя Java-класса источника. Экземпляры классов источников создаются индексатором по заданному имени класса перед началом индексирования.

Каждый источник данных загружает данные из своих входных файлов и преобразует в исходные данные для индексатора.

Разработаны источники данных для формата MARC21 и формата RUSMARC, кроме того, существует принципиальная возможность разработки источников и для других форматов, в т.ч. не MARC.

Индексатор. Краткое описание источников данных

Поисковая система использует индекс с т.н. общей схемой, т.е. существует единый для всех источников данных набор поисковых полей (author, title и т.д.), определённый свыше, каждый источник заполняет все или же часть этих полей, передавая таким образом информацию на вход построителю индекса(IndexWriter) из библиотеки Lucene. Перед тем как IndexWriter запишет данные в файлы индекса, он вызывает т.н. анализаторы. Для каждого поля может быть свой анализатор. Эти же анализаторы вызываются и поисковым сервисом при обработке некоторых типов поисковых запросов. Анализатор преобразует входные данные (выход источника данных) в поток лексем-токенов. Кроме анализаторов из библиотеки Lucene, используются разработанные специальные анализаторы для полей byAuthor, byTitle, storage_code, isbn, issn, title, note, theme, keyword, citation, series, card, content.

Индексатор. Краткое описание анализаторов

Как уже известно, анализатор преобразует входные данные, которыми могут быть строки, числа, текстовые потоки, в поток лексем-токенов (TokenStream). Преобразования могут быть любыми.

Анализатор для полей byAuthor, byTitle

Представляет входную строку как единый токен, преобразует её в нижний регистр, убирает повторяющиеся и крайние пробелы, заменяет 'ё' на 'е' и обрезает получившийся токен до заданной максимальной длины.

Анализатор для полей storage_code

Представляет входную строку как единый токен, убирает крайние пробелы и преобразует её в нижний регистр.

Анализатор для полей isbn

Представляет входную строку как единый токен и, в случае если это ISBN-10, преобразует её в ISBN-13. Если преобразование невозможно, то возвращает входную строку.

Анализатор для полей issn

Представляет входную строку как единый токен и, по возможности, преобразует к виду "\\d{4}\\-\\d{3}(\\d|X)". Если преобразование невозможно, то возвращает входную строку.

Анализатор для полей title, note, theme, keyword, citation, series, card

Разбивает входную строку на токены в соответствие с грамматикой Tokenizer1Impl.jflex, преобразует их в нижний регистр, выделяет словоформы для русского и английского языка и сохраняет исходную форму с префисом '$'.

Анализатор для полей content

Разбивает входную строку на токены в соответствие с грамматикой ContentTokenizerImpl.jflex с ограничением минимальной длины токена, преобразует их в нижний регистр, выделяет словоформы для русского и английского языка без сохранения исходной формы, применяет фильтрацию по т.н. стоп-словам, применяет специальную фильтрацию по особым образом составленному словарю.

Запуск индексации

Индексация может быть запущена с помощью API индексатора (в т.ч. через Администратор) или планировщиком по расписанию. Планировщик может запускать только индексирование изменений основного индекса. Через API возможен также запуск полного переиндексирования основного индекса, индексирования словарей и индексирования данных для фильтра анализатора поля content. По окончании индексирования индексатор вызывает функцию API поискового сервиса для перезагрузки индекса. Индекатор записывает информацию об индексировании в журнал, который можно просматривать в Администраторе, либо через API.

API индексатора

Кроме запуска индексирования, API позволяет контролировать текущее состояние индексатора и процесс индексирования, получать содержимое файла конфигурации, просматривать и изменять расписание запуска индексирования, просматривать журналы индексирования и поиска, статистику.

Поисковый сервис

Поисковый сервис представляет собой веб-сервис с интерфейсами SOAP и REST. Основная функция сервиса - поиск документов по индексу. Искать можно только по полям, проиндексированным заранее (см. выше).

Результатом поиска является список кратких библиографических описаний (БО) с подсветкой найденных токенов. Сервис выдаёт список страницами, размер и номер страницы задаются в запросе. Список может быть отсортирован по релевантности, автору, заглавию, году издания, дате изменения файла, дате добавления в индекс и ОКАТО (при заданном кодецентре), также, возможна добавление сортировки по любому другому полю.

Кроме БО список содержит дополнительную информацию, необходимую для приложений, использующих сервис. БО и дополнительная информация создаются поисковым сервисом в процессе обработки запроса на основе данных, сохранённых индексатором в индексе.

По запросу могут выводиться выдержки из текста документов с подсветкой и т.н. фасеты, т.е. различные классификационные категории с количеством найденных документов по ним.

Информацию о поисковых запросах сервис записывает в журнал поиска, который можно просматривать в Администраторе, либо через API.

Другие функции поискового сервиса

Кроме поиска сервис может выдавать информацию из справочников библиотек, каталогов, коллекций, баз данных, фондов, выдавать и изменять список документов, для которых имеются договора с авторами, выводить информацию по заданной записи (документу), поисковый образ, пояснение по расчёту релевантности.

Также сервис позволяет создавать подсказки по заранее проиндексированным словарям.

Справочный сервис

Справочный сервис представляет собой веб-сервис с интерфейсами SOAP и REST.

Функции сервиса:

- выдача наименований, цифровых и символьных кодов языков по ГОСТ 7.75-97
- выдача кодов и наименований специальностей по номенклатуре специальностей научных работников (в ред. Приказа Минобрнауки РФ от 11.08.2009 N 294)
- возможность создавать дополнительные справочники