Полнотекстовые базы данных ИРБИС

Материал из Wikipedia

Полнотекстовые базы данных ИРБИС – решение, обеспечивающее возможность ранжированного полнотекстового поиска по коллекциям текстовых документов (http://intranet.gpntb.ru/subscribe/?journal=ntb&year=2005&num=11&art=13) .

В основе данного решения:

- программные продукты *ИРБИС 64 для полнотекстовых баз данных*, в которых реализован Механизм полнотекстового поиска и
- *полнотекстовые базы данных ИРБИС* – как вид баз данных ИРБИС 64.

Содержание

Основные возможности ИРБИС 64 по работе с полнотекстовыми базами данных

Возможности для создателей электронных коллекций текстовых документов:

- Сформировать полнотекстовую базу данных ИРБИС – указать текстовые документы, по которым система обеспечит возможность полнотекстового поиска.
- С помощью программного обеспечения ИРБИС предоставить пользователям доступ к коллекции в локальной сети, Интернет или на CD/DVD.

Возможности для пользователей: получить доступ к коллекции текстовых документов для ранжированного полнотекстового поиска и просмотра найденных документов.

Полнотекстовые базы данных ИРБИС обеспечивают возможность работы с текстовыми документами в форматах: HTML, TXT, RTF, PDF, DJVU, DOC, XLS, PPT.

Отличие возможности *включения полнотекстовых документов в базу данных* от возможности *связывания документов базы данных с внешними объектами*

Не следует путать *возможность включения полнотекстовых документов* с *возможностью связывания документов базы данных с внешними объектами*.

Сходства возможностей:

- Обе эти возможности позволяют установить связь документов базы данных с внешними объектами.

Различия возможностей:

- *Возможность связывания документов базы данных с внешними объектами* ограничивается обеспечением простоты перехода пользователя от документа базы данных к внешнему объекту.
- *Возможность включения полнотекстовых документов в базу данных* обеспечивает полнотекстовый поиск, а также переход пользователя к найденным внешним объектам.

Программные продукты для работы с полнотекстовыми базами данных ИРБИС 64

Функциональность по работе с полнотекстовыми базами данных ИРБИС 64 обеспечивают следующие программные продукты:

- *АРМ Администратор полнотекстовых БД* – рабочее место специалиста, которое позволяет формировать полнотекстовые базы данных и обслуживать их.
- *АРМ Читатель для полнотекстовых БД* – рабочее место пользователя электронных коллекций.
- Веб-шлюз ИРБИС для полнотекстовых БД – обеспечивает доступ к коллекциям полнотекстовых документов пользователей Интернета (и/или локальной сети) с помощью веб-браузера.

Концепция полнотекстовых баз данных ИРБИС

Полнотекстовые базы данных ИРБИС отличаются возможностью индексирования текстов из внешних файлов. Что обеспечивает возможность организовать поиск этих текстов.

Эта концепция позволяет реализовать перечисленные выше возможности ИРБИС 64 по работе с полнотекстовыми базами данных.

Индексирование текстов из внешних файлов основывается на следующих идеях:

- Связывать *записи* полнотекстовой базы данных с текстами из внешних файлов, чтобы воспользоваться механизмом индексирования в ИРБИС.
- Реализовать специальный *метод индексирования 9*, позволяющий индексировать тексты из внешних файлов, связанные с *записями* полнотекстовой базы данных.

Для описания механизма связывания *записи* полнотекстовой базы данных с текстами из внешних файлов вводятся следующие понятия:

- Понятие *объекта полнотекстового поиска* – что может быть связано с *записью* полнотекстовой базы данных (и впоследствии проиндексировано).
- Понятие ссылки на *объекта полнотекстового поиска*, посредством которой *запись* полнотекстовой базы данных связывается с *объектом полнотекстового поиска*.

Формирование полнотекстовой базы данных

Формирование полнотекстовой базы данных предполагает добавление в полнотекстовую базу данных *текстов* (также принято называть *включение текстов в базу данных*) и их последующее индексирование.

Фактически, при добавлении *текстов*, они рассматриваются как объекты полнотекстового поиска, в соответствие каждому из которых в базе создаётся *запись*, содержащая ссылку на данный объект. Подробнее см. в статье *Схема полнотекстовой базы данных*.

Включение текстов в базу осуществляется с помощью АРМ Администратор, подробнее см. в подразделе *Включение текстов в полнотекстовую базу данных* статьи *АРМ Администратор полнотекстовых БД*.

Объекты полнотекстового поиска в ИРБИС

Концепция полнотекстовых баз данных ИРБИС предусматривает следующие виды *объектов полнотекстового поиска*:

Внешний текстовый файл

Файл допустимого типа. Список допустимых форматов приведён в подразделе *Основные возможности ИРБИС 64 по работе с полнотекстовыми базами данных*.

Отдельная страница многостраничного документа

Отдельная страница многостраничного документа (формата PDF или DJVU).

Файл, ассоциированный с текстом-"*подложкой*"

Файл допустимого типа, сопровождающийся текстовым файлом, содержащим текстовый слой. Список допустимых форматов приведён в подразделе *Основные возможности ИРБИС 64 по работе с полнотекстовыми базами данных*.

Понятие *объекта полнотекстового поиска* является одним из ключевых в концепции полнотекстовых баз данных:

- в процессе формирования в базе данных сохраняется ссылка на *объект полнотекстового поиска*;
- текст, связанный с *объектом полнотекстового поиска* подлежит индексированию;
- список *объектов полнотекстового поиска* (соответствующих поисковому запросу) будет выдан конечному пользователю в качестве результатов поиска.

Функциональные возможности ИРБИС 64 для полнотекстовых баз данных

Добавление файлов в полнотекстовую базу данных

При добавлении файла в полнотекстовой базе данных создаётся запись, в которой сохраняется ссылка на внешний файл.

Файлы для добавления могут быть указаны выборочно или может быть указана папка, из которой будут добавлены файлы.

Можно установить список расширений, чтобы были добавлены только файлы соответствующих типов.

Добавление многостраничных документов с разбиением на страницы

Файл PDF или DJVU может быть добавлен в полнотекстовую базу с разбиением на страницы (если выбрана соответствующая опция).

В этом случае объектом полнотекстового поиска является отдельная страница PDF или DJVU документа.

В полнотекстовой базе данных создаётся запись, соответствующая каждой странице добавляемого файла, а в каждой записи сохраняется ссылка с указанием номера страницы.

Если пользователь в результате поиска получил отдельную страницу текста, он имеет возможность перейти к другим страницам.

Примечание: с разбиением на страницы НЕ могут быть добавлены файлы PDF или DJVU, находящиеся в архиве.

Индексирование полнотекстовой базы данных

Индексирование полнотекстовой базы данных – процесс наполнения словаря базы данных словами из текстов из внешних файлов, добавленных в базу данных.

Индексирование документов, из которых невозможно извлечение текста

Если при включении полнотекстового документа найдена соответствующая ему подложка, то в соответствующем подполе базы данных сохраняется ссылка на файл подложки. Таким образом происходит *ассоциация* полнотекстового документа с соответствующей *подложкой*.

Если с полнотекстовым документом ассоциирована подложка, то при построении словаря будет использован содержащийся в подложке текст. Для показа пользователю будет использован сам включённый в базу документ.

Подробнее см. в подразделе *Ассоциация текста-подложки с включаемым в полнотекстовую базу документом* статьи *АРМ Администратор полнотекстовых БД*.

Индексирование файлов PDF, защищённых паролем

В ИРБИС возможно индексирование файлов PDF, защищённых паролем. Предусмотрена возможность использования только одного пароля. Подробнее см. в подразделе *Извлечение текстовых данных из PDF-файлов, защищённых паролем* статьи *АРМ Администратор полнотекстовых БД*.

Поисковые возможности

В ИРБИС для полнотекстовых БД предусмотрены следующие поисковые возможности:

- полнотекстовый поиск;
- поиск по элементам описания полных текстов (по умолчанию предусмотрен поиск по элементам Dublin Core);
- полнотекстовый поиск, дополненный ограничением по элементам описания;
- возможность уточнять полнотекстовый поиск при помощи *поиска в найденном* по элементам описания.

Устаревшие возможности ИРБИС 64 по работе с полнотекстовыми базами данных

Возможность размещения текстов в специально предназначенном архиве (не поддерживается с версии 2010.1)

Примечание: начиная с версии 2010.1 данная возможность не поддерживается, в связи с чем не рекомендуется к использованию.

Опции включения полнотекстовых документов, совместимые с данной:

- Возможно включение как указанных пользователем документов, так и всех документов из указанной папки.

Достоинства:

- Компактное хранение полнотекстовых документов (это относится к документам, хорошо поддающимся сжатию при архивации).
- При перемещении папки с базой данных не требуется изменений в ссылках на файлы, поскольку ссылки являются относительными (только имя документа в архиве).

Недостатки:

- Дополнительные затраты времени на разархивирование при обеспечении доступа к полнотекстовому документу.
- Тексты не доступны через Веб-ИРБИС.

Ссылки

См. также:

- АРМ Администратор полнотекстовых БД
- АРМ Читатель для полнотекстовых БД
- Установка и использование ИРБИС 64 для полнотекстовых баз данных
- Возможности АРМ Каталогизатор по работе с полнотекстовыми базами данных
- Известные проблемы и их решения, касающиеся работы с полнотекстовыми базами ИРБИС
- Рекомендации по обслуживанию баз данных ИРБИС
- Связывание документов базы данных ИРБИС с внешними объектами
- Механизм полнотекстового поиска
- Полнотекстовая база данных (вид баз данных ИРБИС)
- Схема полнотекстовой базы данных

Источники информации:

- Полнотекстовые базы данных в ИРБИС64 (http://irbis.gpntb.ru/read.php?48,17749)