АРМ Администратор полнотекстовых БД

Материал из Wikipedia

АРМ Администратор полнотекстовых БД – версия АРМ Администратор, предназначенная для работы с полнотекстовыми базами данных ИРБИС. АРМ Администратор представляет собой рабочее место специалиста, выполняющего операции над базами данных системы в целом в целях поддержания их актуального состояния и сохранности. *АРМ Администратор полнотекстовых БД* также служит для включения полных текстов в базу данных и их исключению из базы.

В этой статье описаны только специфические возможности *АРМ Администратор полнотекстовых БД*. Общие возможности АРМ Администратор описаны в статье АРМ Администратор.

Содержание

Создание новой полнотекстовой базы данных

Необходимые действия для создания баз данных, в том числе полнотекстовых, описаны в соответствующем разделе статьи АРМ Администратор.

Рекомендации по созданию полнотекстовых баз данных ИРБИС см. в соответствующем подразделе статьи *Установка и использование ИРБИС 64 для полнотекстовых баз данных*.

Включение текстов в полнотекстовую базу данных

Включение текстов в полнотекстовую базу осуществляется с помощью специальной формы, которая открывается при выборе пункта главного меню *Полнотекстовый сервис – Добавить / удалить полнотекстовые документы* (для версии 2009.1 *Полнотекстовый сервис – Добавить (удалить) тексты в БД*).

Предусмотрены следующие способы включения текстов в полнотекстовую базу данных:

- выбор текстов путём указания их местоположения на файловой системе (непосредственно, либо в архиве ZIP или RAR; вкладки: *Отдельные файлы*, *Папки* или *Архивы*);
- включение текстов из указанного электронного каталога (вкладка *Внешние объекты электронного каталога*).

Графический интерфейс этой формы представляет собой:

- набор вкладок и опций, относящихся к процессу включения текстов в базу данных ИРБИС;
- список включённых в полнотекстовую базу данных текстов;
- кнопки *Добавить*, *Удалить*, *Обновить*.

Процесс включения текстов начинается по нажатию кнопки *Добавить* и осуществляется в соответствии с выбранными опциями.

При нажатии кнопки *Удалить* происходит удаление из базы данных выбранных текстов.

Ниже описаны возможности, относящиеся к процессу включения текстов в базу данных ИРБИС.

Подробнее см. Включение текстов в полнотекстовую базу данных.

Виды ссылок на тексты (относительные или абсолютные)

В зависимости от выбранных опций в базе данных ИРБИС могут быть сохранены относительные или абсолютные пути к файлам полных текстов.

Использование относительных путей возможно в случае хранения полных текстов в папке базы данных.

Включение текстов из электронного каталога

Возможно включение текстов, являющихся внешними объектами электронного каталога. Если говорить точнее, возможно включение текстов, ссылки на которые содержатся в любой базе данных ИРБИС (начиная с версии 2014.1).

При этом в полнотекстовую базу добавляются все ссылки на тексты из выбранной базы данных.

Чтобы воспользоваться этой возможностью, нужно выбрать вкладку *Внешние объекты электронного каталога* и пункт главного меню *Добавить*. В открывшемся диалоговом окне выбрать par-файл исходной базы данных.

Подробнее об особенностях данной возможности см. в подразделе *Включение в полнотекстовую базу текстов из базы данных электронного каталога* статьи *Включение текстов в полнотекстовую базу данных*.

Каждая страница файла как отдельный документ

Соответствующие опции предусмотрены для файлов формата PDF и DJVU и определяют объект полнотекстового поиска: получит ли пользователь в результате поиска ссылку на документ или на отдельную страницу документа.

Технически разбиение файлов на страницы (извлечение страниц) выполняется при добавлении текстов в базу, если выставлена соответствующая опция, а также при создании словаря (для текстов, являющихся страницами многостраничного PDF-файла или DJVU-файлв).

Извлечение страниц (разбиение на страницы) в случае PDF-файлов осуществляется с помощью одной из утилит: `pdftk` или `pdf2pdf`.

Выбор утилиты определяется параметрами конфигурационного файла АРМ Администратор ИРБИС:

- Начиная с версии 2011.1 (а также в последних обновлениях версии 2010.1) подходящая утилита выбирается автоматически путём перебора. Перебор работает следующим образом: если с помощью одной утилиты не удалось извлечь страницу (разбить на страницы), то будет испробована другая. Порядок перебора задаётся с помощью параметра `PDFSplitUtilityOrder`.
- В более ранних версиях утилита выбирается в соответствии со значением параметра `PDFSplitter`.

Описание известных решений проблем извлечения страниц из PDF-файлов см. в подразделе *Разбиение PDF-файлов на страницы при добавлении в базу данных* статьи *Установка и использование ИРБИС 64 для полнотекстовых баз данных*.

Ассоциация текста-подложки с включаемым в полнотекстовую базу документом

Если установлена опция *Искать текст-'подложку'*, то при включении каждого документа в базу осуществляется проверка наличия подложки. Если соответствующая документу подложка найдена, то происходит *ассоциация* включаемого документа с *подложкой*.

Поиск подложки осуществляется в соответствии с правилом: файлы находятся в одной папке; имя файла подложки получается добавлением расширения `.txt` к имени документа. Например: документу `example.pdf` соответствует подложка `example.pdf.txt`.

Примечания:

- *Данная возможность поддерживается с версии 2011.1, а также присутствует в последних обновлениях версии 2010.1.*
- *Если рядом с файлами* текстов *находятся файлы подложек, но при включении* текстов *НЕ установлена опция Искать текст-'подложку', то подложки могут быть включены в базу как самостоятельные текстовые документы. При включённой опции Искать текст-'подложку', файлы, определяемые как подложки, не будут включены в базу данных в качестве самостоятельных документов.*

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.

В зависимости от выбранных опций, при включении документов в базу данных документы могут быть скопированы в специально предназначенный архив – файл с расширением `.izp` в папке базы данных.

Обслуживание словаря полнотекстовых баз данных ИРБИС

Об основных возможностях АРМ Администратор ИРБИС по работе со словарём, общих как для полнотекстовых баз данных, так и для обычных, см. в подразделе *Обслуживание словаря базы данных ИРБИС* статьи *АРМ Администратор*. В данном подразделе описаны особенности работы АРМ Администратор ИРБИС полнотекстовых БД в случае полнотекстовых баз данных.

Общие принципы индексирования полнотекстовых баз данных описаны в подразделе *Индексирование полнотекстовой базы данных* статьи *Полнотекстовые базы данных ИРБИС*.

Извлечение текстовых данных из PDF-файлов

Извлечение текста в процессе создания словаря осуществляется с помощью одной из утилит: `pdftotext.exe` или `docs2text.exe`.

Выбор утилиты определяется параметрами конфигурационного файла АРМ Администратор ИРБИС:

- Начиная с версии 2011.1 (а также в последних обновлениях версии 2010.1) подходящая утилита выбирается автоматически путём перебора. Перебор работает следующим образом: если с помощью одной утилиты не удалось извлечь текст, то будет испробована другая. Порядок перебора задаётся с помощью параметра `PDFTextExtractUtilityOrder`.
- В более ранних версиях утилита выбирается в соответствии со значением параметра `Converter_PDF`.

Примечание: используемые утилиты дополняют друг друга. Из некоторых PDF-файлов текст удаётся извлечь только с помощью `pdftotext.exe`, из других `docs2text.exe`. Существуют PDF-файлы, из которых не удаётся извлечь текст ни одной ни другой утилитой.

Описание известных решений проблем извлечения текста из PDF-файлов см. в подразделе *Извлечение текста из PDF-файлов в процессе создания словаря* статьи *Установка и использование ИРБИС 64 для полнотекстовых баз данных*.

Извлечение текстовых данных из PDF-файлов, защищённых паролем

Извлечение текстовых данных из PDF-файлов, защищённых паролем, позволяет индексировать такие PDF-файлы.

Перед извлечением текста в процессе создания словаря осуществляется снятие защиты с помощью утилиты `pdftk.exe`.

Для указания необходимости снятия защиты и указания пароля используются параметры `isNeedDecryptPDF` и `PDFPassword` в конфигурационном файле АРМ Администратор ИРБИС.

Примечания:

- *работа ИРБИС с защищёнными файлами возможна при их совместимости с "Acrobat 5.0". Если файлы совместимы с "Acrobat 6.0" или "Acrobat 7.0", то они не могут быть проиндексированы.*
- *При указании необходимости работы с защищёнными файлами, незащищённые файлы также будут индексироваться как обычно.*
- *После снятия защиты извлечение текста происходит так, как описано в подразделе Извлечение текстовых данных из PDF-файлов.*
- *Не предусмотрена возможность работы с защищёнными файлами с разбиением на страницы. Для реализации данной возможности пришлось бы отдавать пользователю незащищённые страницы, либо выполнять защиту отдельных страниц.*

Файл журнала

О местонахождении файла журнала см. в соответствующем подразделе статьи *Файлы ИРБИС*.

Файл журнала предназначен для регистрации некоторых событий, происходящих при работе АРМ Администратор:

- начало работы программы;
- окончание работы программы;
- ошибки, происходящие при добавлении файлов в полнотекстовую базу данных;
- сообщения, касающиеся выбора утилиты для разбиения файлов и извлечения текста.

Используемые утилиты

АРМ Администратор использует утилиты:

- при извлечении страниц из многостраничного документа и разбиении документа на страницы;
- при извлечении текста из документов;
- при обработке PDF-файлов, защищённых паролем.

Возможность выполнения пакетных заданий

Подробнее о пакетных заданиях см. в подразделе *Возможность выполнения пакетных заданий* статьи *АРМ Администратор*.

Ниже перечислены специфические команды пакетных заданий, присущие именно *АРМ Администратор полнотекстовых БД*.

Импортировать записи из электронного каталога – `ADDTEXTS_FROM_EK`

Операнды: `DbName, комментарий`

Где:

`DbName` – имя БД электронного каталога.

`комментарий` – будет записан в каждую импортированную запись в соответствующее поле.

Пример использования:

```
ADDTEXTS_FROM_EK IBIS, Тексты из ЭК
```

Ссылки

См. также:

- АРМ Администратор
- Полнотекстовые базы данных ИРБИС
- Конфигурационные параметры ИРБИС для полнотекстовых БД
- Рекомендации по обслуживанию баз данных ИРБИС