Установка и использование ИРБИС 64 для полнотекстовых баз данных

Материал из Wikipedia

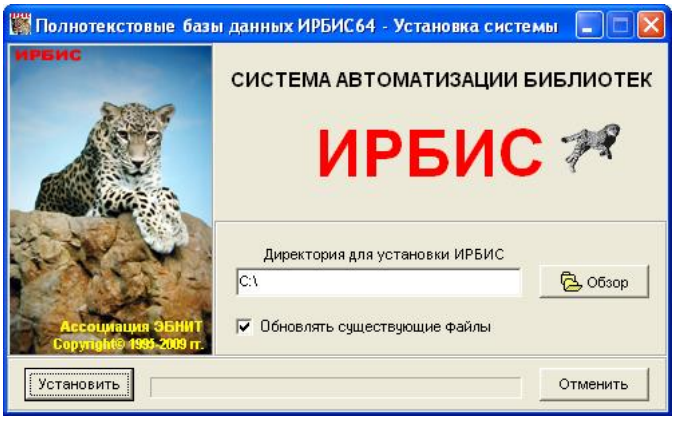
Содержание

Рекомендации по установке

Установочный пакет *Полнотекстовые базы данных ИРБИС 64*

Для установки *АРМ Администратор полнотекстовых БД* (и *АРМ Читатель для полнотекстовых БД*, если он входит в поставку) предназначен установочный пакет *Полнотекстовые базы данных ИРБИС 64* (файл `setup64_FullText.exe`).

Окно установщика *Полнотекстовые базы данных ИРБИС 64*:



Примечание: при появлении в процессе установки сообщения "При регистрации dll библиотеки произошла ошибка. Необходимо в ручном режиме зарегистрировать dll библиотеку ...\doc2text.dll" воспользуйтесь инструкцией Регистрация библиотеки `docs2text.dll` с помощью инструмента Windows `Regsvr32.exe`.

Установка ИРБИС 64 для полнотекстовых баз данных в одну папку с обычным ИРБИС 64

Возможна установка ИРБИС 64 для полнотекстовых баз данных в одну папку с серверной частью ИРБИС 64.

Этот вариант установки полезен тем, что обеспечивает возможность воспользоваться инструментами обычного ИРБИС по отношению к полнотекстовым базам данных. Например, полезной бывает возможность редактировать полнотекстовую базу данных с помощью АРМ Каталогизатор.

Установка Веб-шлюза ИРБИС для полнотекстовых баз данных

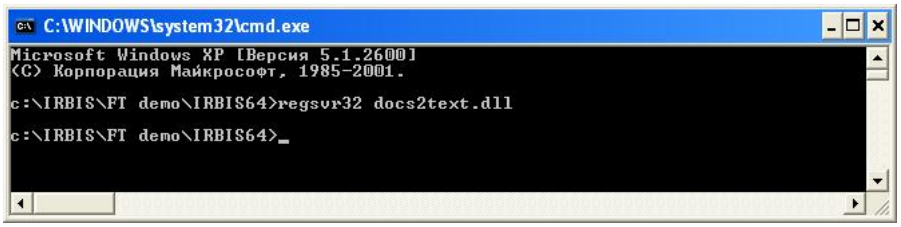
Установка Веб-шлюза ИРБИС для полнотекстовых баз данных осуществляется отдельным установщиком. См. статью Установка Веб-шлюза Ирбис 64.

Регистрация библиотеки `docs2text.dll` с помощью инструмента Windows `Regsvr32.exe`

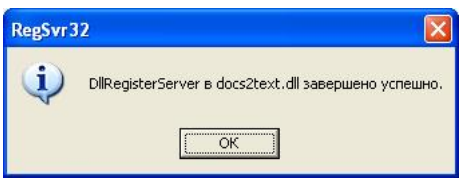
Регистрация библиотеки `docs2text.dll` выполняется автоматически в процессе установки пакета *Полнотекстовые базы данных ИРБИС 64*.

Иногда возникает необходимость зарегистрировать библиотеку `docs2text.dll` в "ручном режиме" с помощью инструмента Windows `Regsvr32.exe`. Для этого достаточно выполнить следующую последовательность шагов:

- В командной строке установить текущей папку, в которой находится файл `docs2text.dll` (это папка, в которую установлена серверная часть ИРБИС).
- Выполнить команду `regsvr32 docs2text.dll`:



- Убедиться, что команда выполнена успешно. Команда выполнена успешно, если получено соответствующее сообщение:



АРМ Читатель для полнотекстовых БД как переносимое приложение

Можно предоставлять доступ пользователей к электронным коллекциям текстовых документов, распространяя полнотекстовые базы данных ИРБИС на CD, DVD (и других съёмных носителях информации) без необходимости установки на жесткий диск компьютера.

При этом для работы с базами используется *АРМ Читатель для полнотекстовых БД*, выступающий в качестве *переносимого приложения*. Не предусмотрены возможность аналогичного использования *АРМ Администратор полнотекстовых БД* и возможность изменения распространяемых таким образом баз данных пользователями. Будем здесь называть распространяемые таким образом базы данных *выпусками*.

Подготовку *выпусков* удобно условно разделить на два шага:

1. Создание копии *заготовки выпуска* (которая включает в себя набор файлов, используемых в любом *выпуске*: *АРМ Читатель для полнотекстовых БД*, необходимые программные библиотеки, а также параметрические и конфигурационные файлы).
2. Подключение базы данных (баз данных) к *заготовке*.

Заготовку выпуска можно сделать, скопировав необходимые файлы из папки, в которой установлены *Полнотекстовые базы данных ИРБИС 64*. *Заготовка* включает в себя следующие файлы:

- borlndmm.dll
- IRBIS64.dll
- MSPRU32.DLL
- qtintf70.dll
- unrar.dll
- unzip.dll
- Unzip32.dll
- zip32.dll
- Irbis64_FullTextReader.exe
- Irbisr_FullText.INI
- MSSP_RU.LEX
- IRBISMSG.TXT

Также в *заготовке* должна присутствовать папка `Datai`, содержащая в себе папку `Deposit`. Папку `Deposit` также можно скопировать из папки, в которой установлены *Полнотекстовые базы данных ИРБИС 64*. Кроме этого в *заготовке* в папке `Datai` не требуется наличия других файлов или папок.

Заготовку можно сделать однократно и пользоваться ей каждый раз при подготовке *выпусков*, копируя её. Также при подготовке нового *выпуска* может быть удобно вместо *заготовки* пользоваться любым другим готовым *выпуском*.

Для подключения базы данных к заготовке необходимо в папку `Datai`:

- скопировать папку базы данных и соответствующий `.par` файл;
- создать файл `Dbnam3_FT.mnu` (справочник, в котором заданы: список доступных баз данных и названия баз данных, которые видит пользователь).

Рекомендации по использованию и настройке

Рекомендации по запуску *АРМ Администратор полнотекстовых БД* и *АРМ Читатель для полнотекстовых БД*

Исходные условия:

- Дистрибутив *Полнотекстовые базы данных ИРБИС 64* установлен на компьютере, который будет выступать в роли *сервера*.
- *Администратору ИРБИС* необходим доступ к *АРМ Администратор полнотекстовых БД*.
- Требуется организовать доступ *читателей* к полнотекстовым базам с компьютеров, выступающих в роли *клиентов*, с помощью *АРМ Читатель полнотекстовых БД*.

АРМ Администратор полнотекстовых БД рекомендуется запускать на *сервере*, где была осуществлена установка дистрибутива *Полнотекстовые базы данных ИРБИС 64*.

Для организации функционирования АРМ Читатель полнотекстовых БД на нескольких компьютерах в сети требуется:

- Открыть сетевой доступ к папке, в которой находится АРМ Читатель для полнотекстовых БД на *сервере*.
- На *клиентах* запускать АРМ Читатель для полнотекстовых БД из этой сетевой папки. Для удобства создать ярлыки на АРМ Читатель для полнотекстовых БД.

Не рекомендуется таким же образом, из сетевой папки, запускать *АРМ Администратор полнотекстовых БД*. В случае редактирования полнотекстовых баз данных с помощью АРМ обычного ИРБИС (например, АРМ Каталогизатор), одновременная с этим работа *АРМ Администратор полнотекстовых БД*, запущенного из сетевой папки с другого компьютера, может привести к порче базы данных. Чтобы при необходимости была возможность восстановления базы данных, рекомендуется регулярно делать резервную копию.

Ссылки на обсуждения по данной теме на форуме ИРБИС:

- Файлы на сетевых дисках (http://irbis.gpntb.ru/read.php?48,52120)

Изменение ограничения допустимого времени работы утилит, используемых для извлечения страниц и текста

Данная настройка выполняется по необходимости – в случаях появления сообщений вида *"Произошло зависание при вызове программы..."*.

Для предотвращения зависания время работы утилит, извлекающих текст и страницы из PDF-файлов, ограничено.

В случае больших PDF-файлов отведённого времени может не хватать. Можно увеличить максимальное время работы утилит с помощью параметра `MAX_TIME_CONVERTING` конфигурационного файла АРМ Администратор ИРБИС.

Рекомендации по созданию полнотекстовой базы данных

Типовые действия администратора при формировании полнотекстовой базы осуществляется очень просто, как описано в подразделе *Типовые действия администратора при формировании полнотекстовой базы* данной статьи.

При этом администратор ИРБИС должен понимать взаимосвязь вопросов размещения полнотекстовых документов, выбора в ИРБИС вида ссылок на полные тексты, а также доступа пользователей к полным текстам через веб-браузер с помощью веб-шлюза ИРБИС. Эта взаимосвязь описывается в нижеследующих подразделах.

Типовые действия администратора при формировании полнотекстовой базы

При формировании полнотекстовой базы данных администратор баз данных ИРБИС выполняет следующие действия:

- Выбор существующей или создание новой полнотекстовой базы данных.
- Добавление текстов в базу данных или удаление текстов из базы.
- Актуализация или создание словаря базы данных (рекомендуется ознакомиться с особенностями обслуживания словаря полнотекстовых баз данных).

Организация доступа пользователей к полнотекстовым базам через веб-браузер с помощью веб-шлюза ИРБИС

Веб-шлюз ИРБИС для полнотекстовых баз данных обеспечивает доступ пользователей к коллекциям полнотекстовых документов через веб-браузер.

Если планируется использовать полнотекстовые базы через Веб, имеет значение вопрос выбора, какого вида ссылки на полные тексты будут использоваться, относительные, абсолютные или URL.

Рекомендуется использовать относительные ссылки по той причине, что в этом случае дополнительных настроек не требуется.

При использовании абсолютных ссылок в случае работы веб-пользователей в рамках локальной сети дополнительных настроек не требуется. Организация работы Интернет-пользователей при использовании абсолютных ссылок возможна при соблюдении следующих условий:

- Полнотекстовые документы должны быть доступны на веб или FTP сервере.
- Возможно преобразование абсолютных ссылок в URL путём замены начала ссылки.
- Чтобы это преобразование происходило, должна быть произведена настройка:
  - В файле `irbis_server.ini`, который был создан при инсталляции в папке веб-сервера для Веб-шлюза ИРБИС, в секции `MAIN` задать параметры (приведен пример значений параметров):

```
FullTextPathDbn=\\Alio1\irbiswrk\lusia\PDF_text\
FullTextPathWeb=ftp:\\ftp.gpntb.ru\pub\irbis\
```

    В первом параметре следует указать сетевой путь на тексты, которые располагались по этому пути при их добавлении, в примере это - `\\Alio1\irbiswrk\lusia\PDF_text\`. Во втором параметре следует указать часть URL текстов (до их названия) их расположения. В примере это - `ftp:\\ftp.gpntb.ru\pub\irbis\`.

  - В формате `BRIEFHTML_ft.pft` имеется вставка (если нет, добавить), которая замещает в адресе текста (в подполе `952^B`) путь на файл при создании на URL местоположения текста. Часть формата:

```
/* возможность замены формата через параметры irbis_server.ini из cgi
      if &uf('IMAIN,FullTextPathDbn,')<>'' and &uf('IMAIN,FullTextPathWeb,')<>''
         then
&uf('+7W10#',&uf('+9I?',,&uf('IMAIN,FullTextPathDbn,'),,'?#',,&uf('IMAIN,FullTextPathWeb,'),,'#',v952^B)),
          '<a  style="border:0px;font-size:12px;" target=_blank href="',G10,
         else
/*
```

При использовании URL-ссылок дополнительных настроек не требуется. Нет возможности группового включения текстов с URL в полнотекстовую базу. Исключение составляют те случаи, когда возможно включение текстов в два этапа, как описано в подразделе *Преобразование ссылок с помощью глобальной корректировки* данной статьи.

Использование относительных ссылок

Особенности относительных ссылок:

- Тексты располагаются в папке базы данных.
- Возможно включение текстов в полнотекстовую базу с помощью АРМ Администратор полнотекстовых БД через диалоговое окно обзора папок и файлов.
- Тексты доступны пользователю опосредовано через Веб-шлюз ИРБИС, их URL представляют собой ссылки на Веб-шлюз ИРБИС с соответствующими командами для получения полного текста.
- Актуально для версий до версии 2012.1 включительно. При включении текстов в полнотекстовую базу с помощью АРМ Администратор полнотекстовых БД путём импорта из электронного каталога администратор ИРБИС должен самостоятельно выполнить действия, описанные в подразделе *Включение текстов из электронного каталога с относительными ссылками* данной статьи.

Использование абсолютных ссылок

Особенности абсолютных ссылок:

- Тексты располагаются в папке, открытой для доступа в локальной сети.
- Возможно включение текстов в полнотекстовую базу с помощью АРМ Администратор полнотекстовых БД через диалоговое окно обзора папок и файлов.
- Возможно включение текстов в полнотекстовую базу с помощью АРМ Администратор полнотекстовых БД путём импорта из электронного каталога.
- Тексты доступны пользователю через веб по URL, который возможно получить путём преобразования абсолютных ссылок, определяемого параметрами `irbis_server.ini`, как описано выше.

Использование URL-ссылок

Особенности URL-ссылок:

- Возможно включение текстов в полнотекстовую базу с помощью АРМ Администратор полнотекстовых БД путём импорта из электронного каталога.
- Нет возможности группового включения текстов с URL в полнотекстовую базу. Исключение составляют те случаи, когда возможно включение текстов в два этапа, как описано в подразделе *Преобразование ссылок с помощью глобальной корректировки* данной статьи.
- Тексты доступны пользователю через Веб по URL, указанному для каждого текста.

Включение текстов из электронного каталога с относительными ссылками

Актуально для версий до версии 2012.1 включительно.

Даже в случае использования в исходном электронном каталоге относительных ссылок, в полнотекстовую базу будут добавлены абсолютные ссылки на тексты – полные пути к текстовым файлам, находящимся в папке базы данных исходного электронного каталога. Это абсолютные пути, начинающиеся с буквы диска, не соответствующие UNC.

Возможно преобразование этих абсолютных ссылок в относительные ссылки с помощью глобальной корректировки, как описано в подразделе *Преобразование ссылок с помощью глобальной корректировки* данной статьи. При этом необходимо скопировать текстовые файлы в папку полнотекстовой базы данных.

Преобразование ссылок с помощью глобальной корректировки

Если возможно преобразование путём замены начала ссылки, оно может быть осуществлено с помощью глобальной корректировки.

Например, это может быть полезно в следующих случаях:

- Включение текстов со ссылками URL в два этапа: 1) включение текстов в полнотекстовую базу с помощью АРМ Администратор полнотекстовых БД через диалоговое окно обзора папок и файлов с относительными или абсолютными ссылками; 2) преобразование ссылок в URL с помощью глобальной корректировки. Тексты при этом должны быть доступны по заданным URL.
- Актуально для версий до версии 2012.1 включительно. Корректировка абсолютных ссылок в относительные после включения текстов из электронного каталога с относительными ссылками. Этот вариант описан в подразделе *Включение текстов из электронного каталога с относительными ссылками* данной статьи.

Пример глобальной корректировки абсолютных ссылок в относительные, которую можно использовать в случае включения текстов из электронного каталога с относительными ссылками:

- оператор

```
REP
```

- поле/подполе

```
952^B
```

- формат

```
'.',&uf('+960*',&uf('+95','c:\irbis64\datai\ibis'),'#',v952^B)
```

где

```
c:\irbis64\datai\ibis
```

это путь к папке базы данных исходного электронного каталога.

Требования и рекомендации по отношению к текстовым документам, включаемым в полнотекстовую базу

Требования и рекомендации по отношению к документам формата PDF

Рекомендуется использовать файлы стандарта PDF/A, специально предназначенного для долгосрочного архивного хранения документов (http://ru.wikipedia.org/wiki/PDF/A) .

Для увеличения скорости загрузки документов пользователем и снижения нагрузки на сервер, где располагаются полные тексты и Веб-шлюз ИРБИС, рекомендуется включать быстрый просмотр в web PDF-документов (http://help.adobe.com/ru_RU/Acrobat/9.0/Standard/WS58a04a822e3e50102bd615109794195ff-7f52.w.html) .

Объём PDF-файлов может достигать десятков и даже сотен мегабайт. В этом случае может потребоваться изменить ограничение допустимого времени работы утилит, используемых для извлечения страниц и текста.

Требования и рекомендации по отношению к документам формата HTML

Если в полнотекстовом документе в формате HTML имеются ссылки на изображения или другие страницы, то они должны быть указаны в форме URL.

Создание полнотекстовой базы на основе ссылок на тексты электронного каталога для распространения на CD/DVD

Для выполнения задачи требуется:

- база данных исходного электронного каталога и связанные с ней текстовые файлы,
- программное обеспечение ИРБИС для работы с полнотекстовыми базами данных,
- АРМ Каталогизатор ИРБИС.

Краткая инструкция:

- Создать новую полнотекстовую базу с помощью АРМ Администратор полнотекстовых баз данных.
- Включить в полнотекстовую базу данных тексты из электронного каталога (см. подраздел *Включение текстов из электронного каталога* статьи *АРМ Администратор полнотекстовых БД*).
- Данный шаг необходим для версий до версии 2012.1 включительно. Преобразовать абсолютные ссылки в относительные с помощью глобальной корректировки и скопировать тексты из исходного электронного каталога в полнотекстовую базу данных (см. подраздел *Включение текстов из электронного каталога с относительными ссылками* данной статьи).

Рекомендации по формированию полнотекстовой БД для ИРБИС 2013.1

Обычные рекомендации по формированию полнотекстовой БД

Данные рекомендации – способ обеспечить работоспособность основных возможностей ИРБИС 64 для ПБД, а именно: 1) сформировать ПБД на основе коллекции текстовых документов, что обеспечивает возможность полнотекстового поиска; 2) обеспечить доступ к ПБД с использованием Веб-шлюза ИРБИС для ПБД; 3) обеспечить доступ к ПБД с использованием АРМ Читатель для полнотекстовых БД.

Локальное размещение

Размещать внешние ресурсы (тексты) следует локально – на том же компьютере, где установлен ИРБИС для полнотекстовых БД.

Относительные пути

Рекомендуется использовать относительные пути.

Правила включения текстов с использованием относительных путей:

- Указать местоположение внешних ресурсов (текстов) В 11 строке .par файла.
- Сделать данную папку доступной через сеть.
- Если тексты находятся в папке БД, то указывать 11 строку .par файла не требуется.
- При включении текстов в полнотекстовую БД использовать относительные пути.

Пример par-файла:

```
1=.\DATAI\Text\
2=.\DATAI\Text\
3=.\DATAI\Text\
4=.\DATAI\Text\
5=.\DATAI\Text\
6=.\DATAI\Text\
7=.\DATAI\Text\
8=.\DATAI\Text\
9=.\DATAI\Text\
10=.\DATAI\Text\
11=\\ComputerName\Texts
```

Абсолютные ссылки

Абсолютные ссылки следует использовать только при невозможности размещения внешних ресурсов (текстов) внутри одной папки.

Известные варианты отступления от обычных рекомендаций по формированию полнотекстовой БД

Отступление от обычных рекомендаций по формированию полнотекстовой БД

Включение текстов с других компьютеров локальной сети

Хотя в ИРБИС ПБД используются *абсолютные* ссылки в формате UNC, есть нюансы их использования.

В рамках обычных рекомендаций их использование ограничено ссылками в пределах локального компьютера.

По умолчанию запрещено включать тексты с других компьютеров в локальной сети.

Известные проблемы и особенности процессов добавления полнотекстовых документов в базу и создания словаря

Разбиение PDF-файлов на страницы при добавлении в базу данных

Для лучшего понимания проблем разбиения на страницы (извлечения страниц из) `PDF`-файлов следует ознакомиться с соответствующим разделом статьи *АРМ Администратор полнотекстовых БД*.

Известные проблемы:

- Утилита `pdftk` не разбивает файлы, содержащие в имени файла русские буквы. Такие файлы следует либо переименовать, либо разбивать с помощью утилиты `pdf2pdf`.
- Утилита `pdf2pdf` не разбивает некоторые файлы. Такие файлы следует разбивать с помощью утилиты `pdftk`.

Извлечение текста из `PDF`-файлов в процессе создания словаря

Для лучшего понимания проблем извлечения текста из `PDF`-файлов следует ознакомиться с соответствующим разделом статьи *АРМ Администратор полнотекстовых БД*.

Известные проблемы:

- Если в словарь не попадают термины из PDF-файла, то в первую очередь необходимо убедиться в наличии в PDF-файле текстовых данных и возможности их извлечения. Наличие текстовых данных и то, что они могут быть корректно извлечены, можно проверить с помощью программы Acrobat Reader. Текст должен выделяться мышью побуквенно; слова из текста должен находить Arcobat Reader своей встроенной системой поиска. Если эти условия выполняются, то текст может быть извлечён, иначе – не может быть извлечён. Если текст может быть извлечён, то следует продолжить диагностировать проблему. Если текст не может быть извлечён, то следует либо подготовить другой PDF-файл вместо проблемного, либо использовать текстовые подложки.
- В случае использования утилиты `docs2text.exe` в системе должна быть зарегистрирована библиотека `docs2text.dll`. Случается так, что библиотека оказывается не зарегистрирована. Если это произошло, то следует её зарегистрировать с помощью инструмента Windows `Regsvr32.exe`.
- Если утилиты `pdftotext.exe` и `docs2text.exe` не могут извлечь текст из PDF-файла, то можно рекомендовать попробовать изменить версию PDF-файла, использовать файлы стандарта PDF/A, специально предназначенного для долгосрочного архивного хранения документов (http://ru.wikipedia.org/wiki/PDF/A) .

Извлечение текста из `DOC`-файлов

Извлечение текста из DOC-файлов выполняется при создании словаря.

Извлечение текста осуществляется с помощью утилиты docs2text.exe или с помощью программы Microsoft Word (с использованием технологии Ole Automation).

Способ извлечения текста выбирается в соответствии со значением параметра Converter_Word в конфигурационном файле АРМ Администратор ИРБИС.

Следует иметь в виду:

- Для извлечения текста вторым способом необходимо наличие установленного приложения Microsoft Word.
- В случае использования утилиты docs2text.exe в системе должна быть зарегистрирована библиотека docs2text.dll. Для регистрации библиотеки используется инструмент Windows Regsvr32.exe.

Известные проблемы:

- Практика показывает, что утилита docs2text.exe не извлекает текст из файлов, содержащих много графических изображений, файлов большого размера (например, десятки мегабайт). При извлечении текста из таких файлов следует выбирать способ с использованием программы Microsoft Word.
- Практика показывает, что при извлечении текста с использованием программы Microsoft Word, не извлекается текст из автофигур. Для извлечении текста из автофигур следует выбирать способ с использованием утилиты docs2text.exe.

Ссылки

См. также:

- Полнотекстовые базы данных ИРБИС
- Установка продуктов ИРБИС
- Установка Веб-шлюза Ирбис 64
- АРМ Администратор полнотекстовых БД
- АРМ Читатель для полнотекстовых БД
- Возможности АРМ Каталогизатор по работе с полнотекстовыми базами данных