Схема полнотекстовой базы данных

Материал из Wikipedia

Структура полнотекстовых баз данных ИРБИС основывается на модели баз данных ИРБИС 64.

Каждой записи базы данных соответствует объект полнотекстового поиска.

Все записи в полнотекстовой базе данных соответствуют определённой структуре, которая включает в себя поля для хранения следующих данных:

- ссылка на объект полнотекстового поиска,
- метаданные, связанные с объектом полнотекстового поиска,
- некоторые технологические данные.

Содержание

Структура записи полнотекстовой базы данных

Ссылка на объект полнотекстового поиска

Соответствие записи базы данных и объекта полнотекстового поиска устанавливается при помощи специальных ссылок.

Ссылка на текст представляет собой структуру, предназначенную для хранения информации, достаточной для обеспечения доступа к объекту полнотекстового поиска.

Поле для хранения ссылки

Метка поля, используемого для хранения ссылки, задаётся в параметре `Full_Text_Name` конфигурационных файлов АРМ Администратор полнотекстовых БД и АРМ Читатель для полнотекстовых БД (описание параметра см. в статье Конфигурационные параметры ИРБИС для полнотекстовых БД).

По умолчанию, для хранения ссылки используется метка поля `952`.

Данное поле не повторяющееся.

Поле доступно на рабочем листе *Технологическая* в АРМ Каталогизатор.

Виды ссылок на полные тексты

Ссылки на тексты из внешних файлов различаются в зависимости от объекта полнотекстового поиска и особенностей доступа.

Виды ссылок по способу доступа к объектам полнотекстового поиска:

- Объект полнотекстового поиска находится на файловой системе.
- Текстовый файл доступен по URL (HTTP или FTP).

Виды ссылок по размещению объектов полнотекстового поиска на файловой системе:

- Текстовый файл на файловой системе.
- Текстовый файл в архиве ZIP или RAR.

Виды ссылок в зависимости от объекта полнотекстового поиска (только в случае размещения текстового файла непосредственно на файловой системе):

- Текстовый файл.
- Страница многостраничного документа PDF или DJVU.
- Файл с текстовой "подложкой".

При нахождении текстового файла непосредственно на файловой системе ссылки различают по способу адресации:

- *относительные* – в ссылке используется относительный путь (начинается с точки, например `.\texts\irbis64_2008.doc`);
- *абсолютные* – в ссылке используется полный путь, включающий имя компьютера, в формате UNC (например, `\\ComputerName\SharedFolder\Resource.pdf`).

Об адресации относительных ссылок см. в подразделе *Относительные ссылки на внешние объекты в базах данных электронного каталога и в полнотекстовых базах* статьи *Относительная адресация в ИРБИС*.

При нахождении текстового файла в архиве ссылки также разделяются на *относительные* и *абсолютные*, в зависимости от того, используется ли в ссылке относительный или абсолютный путь к архиву.

Примечание: до введения в ИРБИС 2011.1 соответствующего запрета в АРМ Администратор было возможно введение абсолютных ссылок, начинающихся с имени диска.

Примечание: 11-я строка `.par`-файла появилась начиная с версии 2012.1, в более ранних версиях относительный путь указывает местоположение файла относительно папки базы данных.

Элементы ссылки

Ссылка на *объект полнотекстового поиска* в общем случае содержит следующие структурные элементы:

- *URL*

- *Путь к текстовому файлу*
- *Номер страницы*
- *Путь к файлу архива*
- *Путь к файлу внутри архива*
- *Имя файла с текстом-"подложкой"*
- *Полный путь для относительной ссылки* (является избыточным и поддерживается по историческим причинам)

В зависимости от вида ссылка содержит те или иные элементы.

Структура, используемая для хранения ссылки в базе данных

Структура, используемая для хранения ссылки в базе данных представляет собой совокупность подполей `^B^C^I^T^U`:

- `B` – в зависимости от вида ссылки это относительный, полный или виртуальный путь к файлу полного текста, или же некоторые данные, дополняющие гиперссылку. Относительный путь используется для полнотекстовых документов, хранящихся в папке базы данных (относительный путь начинается с точки). Полные пути используются для ссылок на полнотекстовые документы, находящиеся вне папки базы данных. Виртуальные пути к текстовым документам используются для ссылок на полнотекстовые документы, хранящиеся в архивах `.zip` и `.rar`, а также в случае ссылок на отдельные страницы многостраничных документов `.pdf` и `.djvu`. Виртуальная ссылка, хранящаяся в этом подполе, позволяет узнать имя файла внутри архива или номер страницы многостраничного документа, но не имя файла архива или многостраничного документа.
- `C` – полный путь к файлу `zip/rar/pdf/djvu`. Данное подполе используется для ссылок на полнотекстовые документы в архиве или отдельные страницы многостраничного документа.
- `I` – URL текста, перенесённого из электронного каталога.
- `T` – ссылка на файл подложки. Представляет собой имя текстового файла, подразумевается, что местонахождение файла подложки соответствует местонахождению полнотекстового документа.
- `U` – введено для технологических целей в версии 2010.1. Подполе `^U` было задумано как универсальная замена подполям `^B^C^I` с возможностью расширения, однако было признано неудобным с точки зрения его разбора средствами языка форматирования. Как следствие, подполе `^U` остаётся вспомогательным, и используется наряду с другими подполями.

Подполе `^U` всегда начинается с префикса

```
uri:irbis:
```

Дальнейшее содержимое зависит от объекта полнотекстового поиска.

Примечание: особенность ссылок на отдельную страницу многостраничного документа (то же касается и текстов в архиве) в том, что относительные или абсолютные ссылки отличаются подполем `^B`, а в подполе `^C` в обоих случаях хранится абсолютный путь (который не используется). Эта избыточность сложилась исторически. Избыточность стала причиной ошибки, которая была исправлена в версии 2011.1 (см. ошибочное использование абсолютного пути из подполя `^C` в случае относительных ссылок с разбиением на страницы*).*

Составление ссылки

Ниже приведены примеры составления ссылки в зависимости от объекта полнотекстового поиска и особенностей доступа.

Ссылка на текстовый файл

Объект полнотекстового поиска: текстовый файл.

Доступ: на файловой системе.

Элементы ссылки: *путь к текстовому файлу*.

Пример относительной ссылки на текстовый файл:

```
^B.\texts\irbis64_2008.doc
```

Пример абсолютной ссылки на текстовый файл:

```
^B\\ComputerName\SharedFolder\FullTexts\doc\ИРБИС_документация\ТЕХНИЧЕСКАЯ ДОКУМЕНТАЦИЯ.doc
```

Ссылка на отдельную страницу многостраничного документа

Объект полнотекстового поиска: отдельная страница многостраничного документа (формата PDF или DJVU).

Доступ: на файловой системе.

Элементы ссылки: *путь к текстовому файлу*, *номер страницы*.

Подполя: `^B^C^U`.

Подполе `^B` конструируется следующим образом:
`<путь к файлу (без имени файла)> + <имя файла (без расширения)> + <суффикс> + <номер страницы> + <расширение файла>`
где:

- `<путь к файлу (без имени файла)>` – путь (относительный или абсолютный) к исходному многостраничному документу, без имени файла;
- `<имя файла (без расширения)>` – имя файла исходного многостраничного документа без расширения;
- `<суффикс>` – последовательность символов, которая отделяет имя файла от номера страницы (по умолчанию два знака подчёркивания __, вообще определяется конфигурационным параметром `FULL_TEXT_FileNamePrefixDiv`);
- `<номер страницы>` – номер страницы, дополненный лидирующими нулями до 4 символов;
- `<расширение файла>` – расширение исходного многостраничного документа `.pdf` или `.djvu`.

Подполе `^C` представляет собой: АБСОЛЮТНЫЙ путь к файлу и имя исходного многостраничного файла. По историческим причинам в данном подполе хранится АБСОЛЮТНЫЙ путь даже в том случае, если ссылка является относительной. *Примечание: хранение пути к файлу в подполях `^B` и `^C` представляет собой факт наличия избыточной информации; более того, АБСОЛЮТНЫЙ путь в относительной ссылке представляет собой не только*

избыточную, но также и потенциально недостоверную информацию, которая не должна использоваться (данная информация игнорируется при интерпретации ссылки).

Подполе `^U` также содержит *путь к текстовому файлу* и *номер страницы*, и игнорируется при интерпретации ссылки.

Пример ссылки на 69-ю страницу `pdf`-документа (относительная ссылка на документ):

```
^B.\FullTexts\pdf\ТЕХНИЧЕСКАЯ ДОКУМЕНТАЦИЯ__0069.pdf^CC:\irbisFT-2010-02-15\IRBIS64\Datai\TEST-PDF\FullTexts\pdf\ТЕХНИЧЕСКАЯ ДОКУМЕНТАЦИЯ.pdf
```

Пример ссылки на 1-ю страницу `pdf`-документа (абсолютная ссылка на документ):

```
^B\\ComputerName\SharedFolder\FullTexts\pdf\ТЕХНИЧЕСКАЯ ДОКУМЕНТАЦИЯ__0001.pdf^C\\ComputerName\SharedFolder\FullTexts\pdf\ТЕХНИЧЕСКАЯ ДОКУМЕНТАЦИЯ.pdf
```

Ссылка на текстовый файл в архиве ZIP или RAR

Объект полнотекстового поиска: текстовый файл.

Доступ: на файловой системе, в архиве ZIP или RAR.

Элементы ссылки: *путь к файлу архива*, *путь к файлу внутри архива*.

Подполе `^B` конструируется следующим образом:
`<путь к файлу архива> + <путь к файлу внутри архива>`
где:

- `<путь к файлу архива>` – относительный или абсолютный путь к файлу архива, без имени архива (записывается с использованием символа `"\"` – обратный слэш, и завершается данным символом);
- `<путь к файлу внутри архива>` – путь к файлу внутри архива, включая имя текстового файла (записывается с использованием символа `"/"` – прямой слэш).

Подполе `^C` представляет собой: АБСОЛЮТНЫЙ полный путь к архиву (путь и имя). По историческим причинам в данном подполе хранится АБСОЛЮТНЫЙ путь даже в том случае, если ссылка является относительной. *Примечание: хранение пути к архиву в подполях `^B` и `^C` представляет собой факт наличия избыточной информации; более того, АБСОЛЮТНЫЙ путь в относительной ссылке представляет собой не только избыточную, но также и потенциально недостоверную информацию, которая не должна использоваться (данная информация игнорируется при интерпретации ссылки).*

Подполе ^U конструируется следующим образом: *<путь к файлу архива>* + ":" + *<путь к файлу внутри архива>*, и игнорируется при интерпретации ссылки.

Примечание: Одновременно с введением подполя `^U` изменилась логика формирования подполя `^B`. С этого момента в подполе `^B` используются только символы `\` и НЕ используются символы `/`. Следствием стало то, что в версиях 2010.1, 2011.1 и 2012.1 присутствует ошибка при интерпретации относительной ссылки, которая может проявляться случае перемещения базы данных. В такой ситуации проблема может быть решена при помощи глобальной корректировки: чтобы ошибка не проявлялась, в подполе `^C` должен присутствовать полный путь к архиву. Именно для устранения данной ошибки в версии 2013.1 при разборе ссылки по возможности используется подполе `^U`.

Пример:

```
^B\\ComputerName\SharedFolder\FullTexts\rar\ТЕХНИЧЕСКАЯ ДОКУМЕНТАЦИЯ.pdf^C\\ComputerName\SharedFolder\FullTexts\rar\ИРБИС_документация.rar
```

Ссылка на файл, ассоциированный с текстом-"*подложкой*"

Объект полнотекстового поиска: файл, ассоциированный с текстом-"*подложкой*".

Доступ: на файловой системе.

Элементы ссылки: *путь к текстовому файлу*, *имя файла с текстом-"подложкой"*. Первый текстовый файл используется для показа пользователю, а индексированию подлежит текст из *"подложки"*.

Подполя: `^B^T^U`. Подполе `^B` содержит *путь к текстовому файлу* (подполе `^U` также содержит данный путь). Подполе `^T` содержит *имя файла с текстом-"подложкой"*.

Пример:

```
^B\\127.0.0.1\FullTexts\!test cases\pdf\external text layer\1.pdf^T1.pdf.txt^Uuri:irbis:\\127.0.0.1\FullTexts\!test cases\pdf\external text layer\1.pdf
```

Ссылка на текстовый файл, доступный по URL

Объект полнотекстового поиска: текстовый файл.

Доступ: по URL.

Элементы ссылки: *URL*.

Подполя: `^B^I`. Подполе `^I` содержит URL. Подполе `^B` содержит дополнительную информацию.

Пример:

```
^Ihttp://www.sweden.se/ru/Start/Education/^Bindex.html : http://www.sweden.se/ru/Start/Education/
```

Устаревшие элементы структуры

Подполе `^A` – имя файла полного текста. Данное подполе используется только для хранения ссылок на полнотекстовые документов в архиве с именем базы данных и расширением `.izp`, находящемся в папке базы данных. Начиная с версии 2010.1 данный вид ссылок не поддерживается.

Пример ссылки на документ в архиве `.izp`:

```
^ATехническая документация для WEB ИРБИС64 и WEB ИРБИС32.doc
```

Интерпретация ссылки

При необходимости доступа к файлу полного текста (при индексации, отображении на экране) элементы ссылки интерпретируются в соответствии с определёнными правилами, которые описаны ниже.

Ссылка на текстовый файл, доступный по URL

Если заполнено подполе `^I`, то интерпретируем данную ссылку как *ссылку на текстовый файл, доступный по URL*. Содержимое подполя `^I` является *URL* (HTTP или FTP).

В этом случае объект полнотекстового поиска – текстовый файл, доступный по указанному URL. На этом интерпретация ссылки заканчивается.

В противном случае продолжаем анализ ссылки.

Примечание: при использовании HTTP-ссылок формат файла определяется по заголовку HTTP, во всех остальных случаях – по расширению файла.

Вопросы интерпретации ссылки

Далее интерпретация ссылки сводится к ответу на следующие вопросы:

- Ссылка является абсолютной в формате UNC или относительной?
- Каков вид объекта полнотекстового поиска (текстовый файл, страница, файл с текстовой подложкой)?
- Находится ли файл в архиве `.zip` или `.rar`?

Объект полнотекстового поиска – текстовый файл (НЕ находящийся в архиве)

Если подполе `^C` не заполнено, то считается, что в подполе `^B` хранится *путь к текстовому файлу* (относительный или абсолютный), и данный файл является объектом полнотекстового поиска.

Файл, ассоциированный с текстом-"*подложкой*"

Если при этом заполнено подполе `^T`, то объект полнотекстового поиска – *файл, ассоциированный с текстом-"подложкой"*.

Первый текстовый файл используется для показа пользователю, а индексированию подлежит текст из *"подложки"*.

Подполе `^T` содержит имя файла с текстом-*"подложкой"*. Местонахождение файла *подложки* соответствует местонахождению основного файла.

Отдельная страница многостраничного документа или текстовый файл в архиве

Если подполе `^C` заполнено, то объект полнотекстового поиска не доступен непосредственно, и является:

- отдельной страницей многостраничного файла (`.pdf` или `.djvu`) или
- текстом в архиве (`.zip` или `.rar`).

В этих случаях для доступа к объекту полнотекстового поиска необходимо соответственно:

- извлечь страницу из исходного многостраничного файла (или из кеша извлечённых страниц), либо
- извлечь текст из архива.

Необходимая для извлечения дополнительная информация содержится в подполях `^B` и `^C`.

Отдельная страница многостраничного документа

Если расширение файла в подполе `^C` – `.pdf` или `.djvu`, то объектом полнотекстового поиска является *отдельная страница многостраничного документа*.

Путь к текстовому файлу (полный, включая имя файла) можно получить следующим образом: из подполя `^B` взять *путь к файлу*, а из подполя `^C` взять *имя файла*.

Номер страницы можно получить из подполя `^B`, которое, как описано в подразделе *Составление ссылки*, составляется следующим образом:
`<путь к файлу (без имени файла)> + <имя файла (без расширения)> + <суффикс> + <номер страницы> + <расширение файла>`

Текстовый файл в архиве

Если в подполе `^C` *расширение файла* `.zip` или `.rar`, то объектом полнотекстового поиска является текстовый файл, находящийся в архиве.

НЕ допускаются ссылки на отдельные страницы многостраничного документа, находящегося в архиве.

Начиная с версии 2013.1 по возможности при интерпретации используется подполе `^U`, иначе подполя `^B^C`.

Получение *пути к файлу архива* и *пути к файлу внутри архива* из подполя `^U` очевидно исходя из его описания в подразделе *Составление ссылки*.

Подполя `^B^C` для получения *пути к файлу архива* и *пути к файлу внутри архива* используются следующим образом:

- Подполе `^B` разбирается в соответствии со следующими правилами: с последним вхождением символа `\` (обратный слэш) заканчивается путь к архиву, дальше начинается относительный путь внутри архива; путь внутри архива записывается с использованием символа `/` (прямой слэш).
- Подполе `^C` содержит имя файла архива. Примечание: данное подполе также содержит также полный путь к архиву, что является, как минимум, избыточной информацией, а в случае использования относительных путей, также и недостоверной информацией.

Метаданные, связанные с объектом полнотекстового поиска

С объектом полнотекстового поиска могут быть связаны метаданные в формате Dublin Core.

Метаданные Dublin Core хранятся в полнотекстовой базе данных ИРБИС 64 в соответствии со следующей схемой:

Метка поля	Элемент метаданных
1	Title — название
2	Creator — создатель
3	Subject — тема
4	Description — описание
5	Publisher — издатель
6	Contributor — внёсший вклад
7	Date — дата
8	Type — тип
9	Format — формат документа
10	Identifier — идентификатор
11	Source — источник
12	Language — язык
13	Relation — отношения
14	Coverage — покрытие
15	Rights — авторские права

Технологические данные

Метка поля	Элемент метаданных
20	Число слов в тексте
21	Индекс естественно-тематического классификатора
22	Первые строки полного текста
23	Комментарий
24	Дата ввода записи в базу данных
25	Размер файла полного текста в байтах
26	Дата создания полного текста в байтах
951	Исходные данные из ЭК
66	Данные о переносе записи из ЭК

Индекс полнотекстовой базы данных

Префикс TXT=

Термин словаря, начинающийся с префикса `TXT=`, предназначен для поиска записи(ей) базы данных, соответствующей(их) известной ссылке на текст.

За префиксом `TXT=` следует:

- Значение подполя `^B` — в том случае, если термин (вместе с префиксом) не превышает определённое количество символов (250 символов).
- Определённым образом укороченное значение подполя `^B` — в том случае, если термин превышает указанное количество символов.

Вторая часть правила позволяет уменьшить необходимость перебора терминов словаря при поиске записей по ссылке на конкретную страницу текста в случае длинных путей и имён файлов (таких, что длина термина превышает 250 символов). Укорочение происходит таким образом, чтобы в поисковом термине фигурировал номер страницы.

Для реализации правила служит UNIFOR `&uf('+3C`.

Примеры результатов вычисления `&uf('+3C',"TXT="v952^b)`:

- В случае короткого значения подполя `^B`

```
c:\TEXTS\example_text__0001.pdf
```

результат будет

```
TXT=c:\TEXTS\example_text__0001.pdf
```

Примеры результатов вычисления `&uf('+3C',"TXT="v952^b)`:

- В случае длинного значения подполя `^B` (числами обозначено количество символов от начала пути файла)

```
c:\TEXTS\example_text=========030=======040=======050=======060=======070=======080=======090=======100=======110=======120=======130=======140=======150=======160=======170=======180
```

результат будет следующим (обратите внимание на укороченное значение подполя `^B`)

```
TXT=c:\TEXTS\example_text=========030=======040=======050=======060=======070=======080=======090=======100=======110=======120=======130=======140=======150=======160=======170======
```

Ссылки

См. также:

- Полнотекстовые базы данных ИРБИС

- Полнотекстовая база данных (вид баз данных ИРБИС)
- Индекс базы данных ИРБИС
- Схема базы данных электронного каталога