CLB Search

Версия 1.1.1.0

Руководство пользователя

http://www.veretennikov.org

Данный документ описывает использование CLB Search.

Последнее обновление документа 2014.01.12.

Оглавление

О программе 5

Лицензия 6

Системные требования 7

Введение 8

Новое в версии 1.1.0.0 9

Установка 10

Запуск 11

Поддерживаемые языки 12

Создание индекса 13

Поиск 18

Просмотр списка проиндексированных документов 26

Открытие исходного файла из результатов поиска 27

Поиск похожих документов 28

Структура каталогов 31

Информация об индексе 34

События 36

История поиска 38

Конфигурационный файл индекса 39

Журнал индекса 40

Настройки программы 43

Дополнительные модули 44

Модуль поддержки форматов 47

Запись диагностической информации 49

Внутреннее устройство 51

Список файлов 53

Программный интерфейс (API) 55

Настройка использования памяти 56

Многопоточность 60

Репозитарий 61

Дополнительные настройки поиска 63

Результаты экспериментов 64

Обработка XML файлов 65

Морфологические словари 68

Новое в версии 1.0.0.4 69

Новое в версии 1.0.0.3 70

Новое в версии 1.0.0.2 71

Лицензионное соглашение 72

Благодарности 74

О программе

CLB Search позволяет создавать индекс по набору текстовых файлов и выполнять с помощью него текстовые запросы.

Возможности

1. Поддержка текстовых файлов, файлов HTML, XML, RTF, CHM, PDF, DJVU, FB2, Microsoft Office (DOC, DOCX, XLS, XLSX, XLSM, PPT, PPTX), Open Office (ODT, ODS, ODP), EPUB.
2. Поиск с учетом морфологии языка.
3. Ранжирование результатов TF-IDF, BM25.
4. Язык запросов с поддержкой логических операторов.
5. Автоматическое распознавание кодировки с учетом морфологии языка, поддерживаемые кодировки: UNICODE, UTF8, CP1251, ASCII, KOI8.
6. Индексирование архивов RAR, CAB, ZIP, 7Z, TAR, GZIP, ARJ.
7. Возможность сохранения полной информации о проиндексированных текстах, что позволяет осуществлять поиск и просмотр документов, даже если исходные документы недоступны.
8. Поиск похожих документов.
9. Доступны x86 и x64 версии программы.

Лицензия

Используя описываемую в данном документе программу CLB Search, Вы автоматически соглашаетесь с лицензией по использованию программы. См. файл license.txt или раздел «Лицензионное соглашение».

Системные требования

1. x86 совместимый процессор.
2. 512 МБ оперативной памяти.

Рекомендуемые системные требования

1. x86 совместимый процессор.
2. 2 ГБ оперативной памяти.

Требования к программному обеспечению

1. Операционная система Windows 2000/XP/2003/Vista/2008/2008 R2/7 и следующих версий.
2. Windows Script Host Версии 5.5 (Microsoft Internet Explorer 5.5 или выше).

   Примечания.

   Windows XP уже содержит Windows Script Host 5.6.

   Windows 2000 по умолчанию содержит Windows Script Host 5.1 ($<$ 5.5), требуется обновление до версии 5.5 (установка Microsoft Internet Explorer 6.0).
3. WSO (WindowSystemObject), включено в дистрибутив, последняя версия доступна по адресу http://www.veretennikov.org.

Введение

CLB Search позволяет создавать индекс по набору текстовых файлов и выполнять с помощью него текстовые запросы.

В CLB Search включены две основные программы. Одна из них создает индекс. На вход программе подается набор файлов или каталогов и текстовый файл — конфигурационный файл индекса, в котором содержится информация о настройках создаваемого индекса. Другая программа позволяет искать в созданном индексе. Конфигурационный файл индекса сопоставлен с программой поиска. Таким образом для поиска в индексе достаточно дважды щелкнуть мышкой по конфигурационному файлу индекса для запуска интерфейса поиска.

Примечание. Не рекомендуется подавать на вход программе создания индекса каталог с произвольными данными, например «Диск C», т. к. в этом случае программа будет обрабатывать все файлы в каталоге, пытаясь извлечь из них текст, в том числе исполняемые файлы, изображения и т. д. Рекомендуется подавать на вход папку, все файлы в которой содержат текст, например папку с текстовой библиотекой. Оптимально подавать на вход папку, в которой большое количество файлов сжато в небольшое количество архивов, т. к. при одинаковом суммарном объеме большое количество небольших файлов считывается с диска существенно медленнее чем несколько больших.

Новое в версии 1.1.0.0

1. TF-IDF, BM25 ранжирование результатов поиска.
2. Новая морфология русского языка.
3. LZMA сжатие репозитария.
4. Автоматическое определение языка документов.
5. Изменение в API, функции морфологического анализатора

Установка

CLB Search устанавливается с помощью программы установки. Для поддержки форматов PDF, DJVU требуется установить отдельный продукт CLB Search Extensions.

Для работы программы требуется установить WSO. Необходимая версия WSO входит в CLB Search и устанавливается автоматически.

Программа установки зарегистрирует COM сервера index.dll, index64.dll (64-я версия).

Программа установки сопоставляет расширение файлов ICF (Index Configuration File) с CLB Search и добавляет в контекстное меню для данного типа файла пункт «Создать индекс». Стандартное действие «Открыть» для данного типа файлов запускает поиск в индексе.

В начале установки программа установки запрашивает язык интерфейса программы.

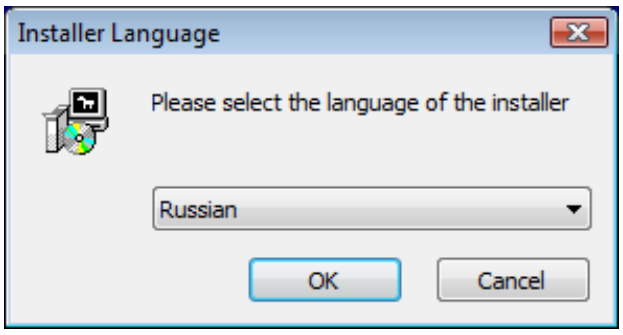


Рис. 1: Установка

Выбранный язык влияет на язык интерфейса программы и названия ярлыков в главном меню.

Запуск

Для запуска программы предназначены скрипты на языке JScript

1. $make.js$ создание индекса
2. $find.js$ поиск в индексе
3. $make\_32.js$ создание индекса, программа запускается в 32-битном режиме
4. $make\_64.js$ создание индекса, программа запускается в 64-битном режиме

Поддерживаемые языки

Сприпты *make.js*, *find.js* поддерживают как русский, так и английский языки. Для смены текущего языка интерфейса следует воспользоваться пунктом меню «Language».

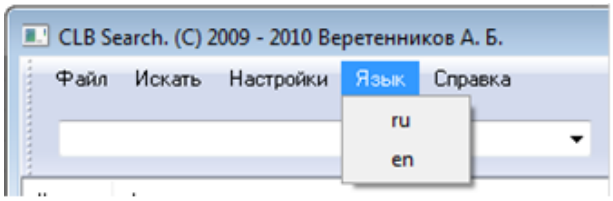


Рис. 2: Выбор языка

Рис. 3: Выбор языка

Создание индекса

Для создания индекса нужно запустить *make.js*.

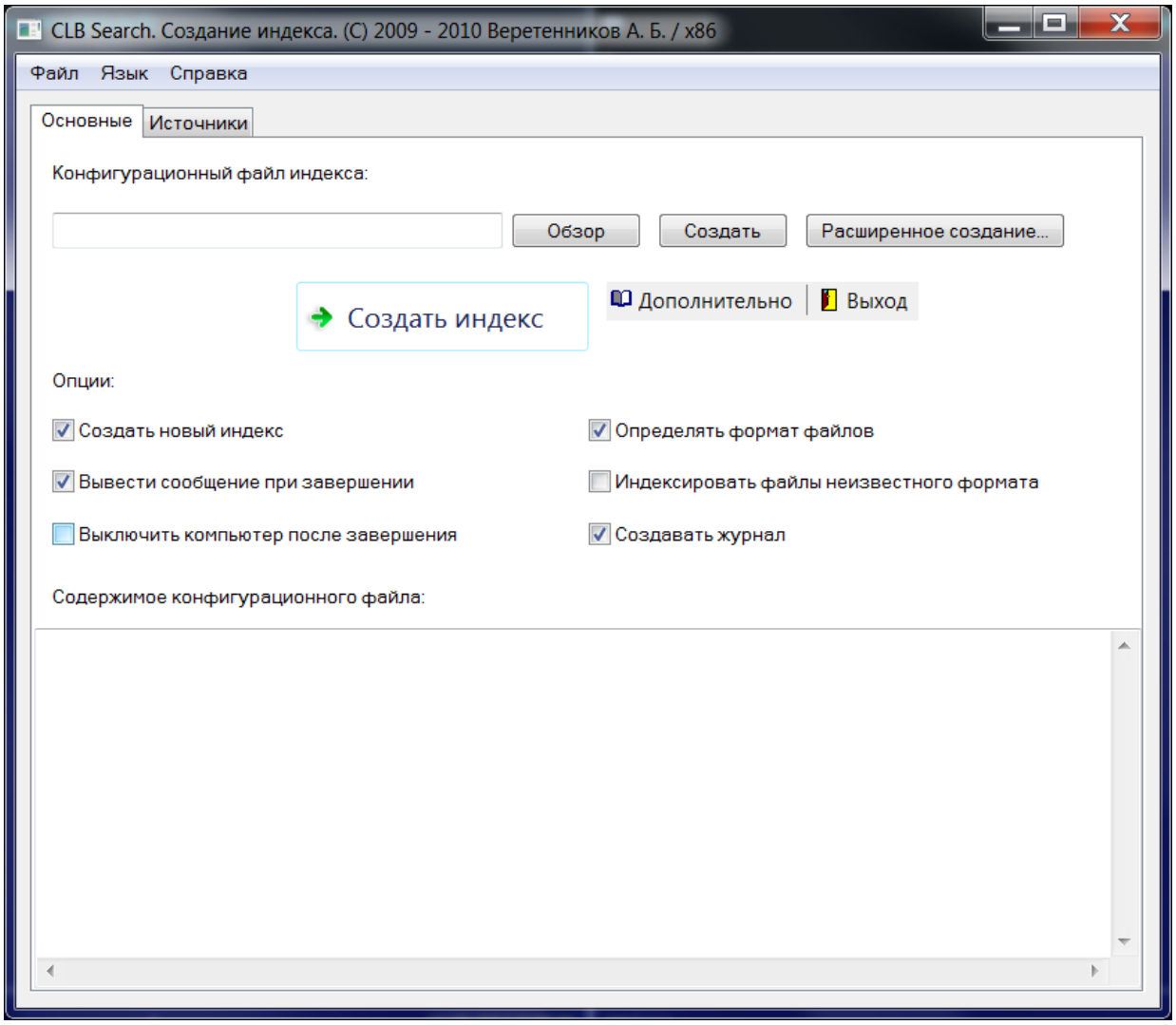


Рис. 4: Создание индекса

Индекс создается на основании конфигурационного файла индекса. Конфигурационный файл содержит в себе различные настройки индекса.

Это текстовый файл со строками вида НАСТРОЙКА=ЗНАЧЕНИЕ. Можно выбрать существующий файл или создать новый.

Нажмите на кнопку «Создать» для создания простого конфигурационного файла индекса.

Простой конфигурационный файл содержит одну запись, которая определяет имя индекса:

NAME = Index\Index

Что означает, что рядом с конфигурационным файлом будет создана папка Index, в которой будет создан индекс с настройками, определяемыми автоматически на основании анализа конфигурации компьютера.

На второй закладке «Источники» можно указать каталоги и файлы, которые следует проиндексировать.

Опции создания индекса:

- Переключатель «Создать новый индекс».

  Создается новый индекс, существующие файлы индекса удаляются. Если данный переключатель отключить и если индекс уже существует, то новые данные будут добавлены в уже существующий индекс.

- Переключатель «Вывести сообщение о завершении».

  После завершения работы будет выведено сообщение об этом. Если переключатель не включен, программа просто завершает свою работу.

- Переключатель «Определять формат файлов».

  Включает режим распознавания кодировки и языка файлов, с использованием анализа морфологии.

- Переключатель «Индексировать файлы неизвестного формата».

  Разрешает индексирование файлов, кодировка и язык которых не распознаны. Учитывается только если включена опция «Определять формат файлов».

- Переключатель «Выключить компьютер после завершения».

  Завершить работу компьютера после завершения индексирования

- Переключатель «Создавать журнал».

  Позволяет отключить создание журнала.

Расширенное создание конфигурационного файла

Нажмите на кнопку «Расширенное создание» для создания конфи-гу-ра-цион-ного файла с выбором дополнительных настроек.

Сверху в окне мастера выбирается имя создаваемого файла. Слева отображается список разделов настройки. В каждом разделе может быть одна или несколько настроек, которые отображаются справа. При наведении мышки на конкретную настройку справка по ней отображается в правом нижнем углу окна. Каждая настройка соответствует одной строке в конфигурационном файле. Описание настроек дано в файле *docs\index.pdf*.

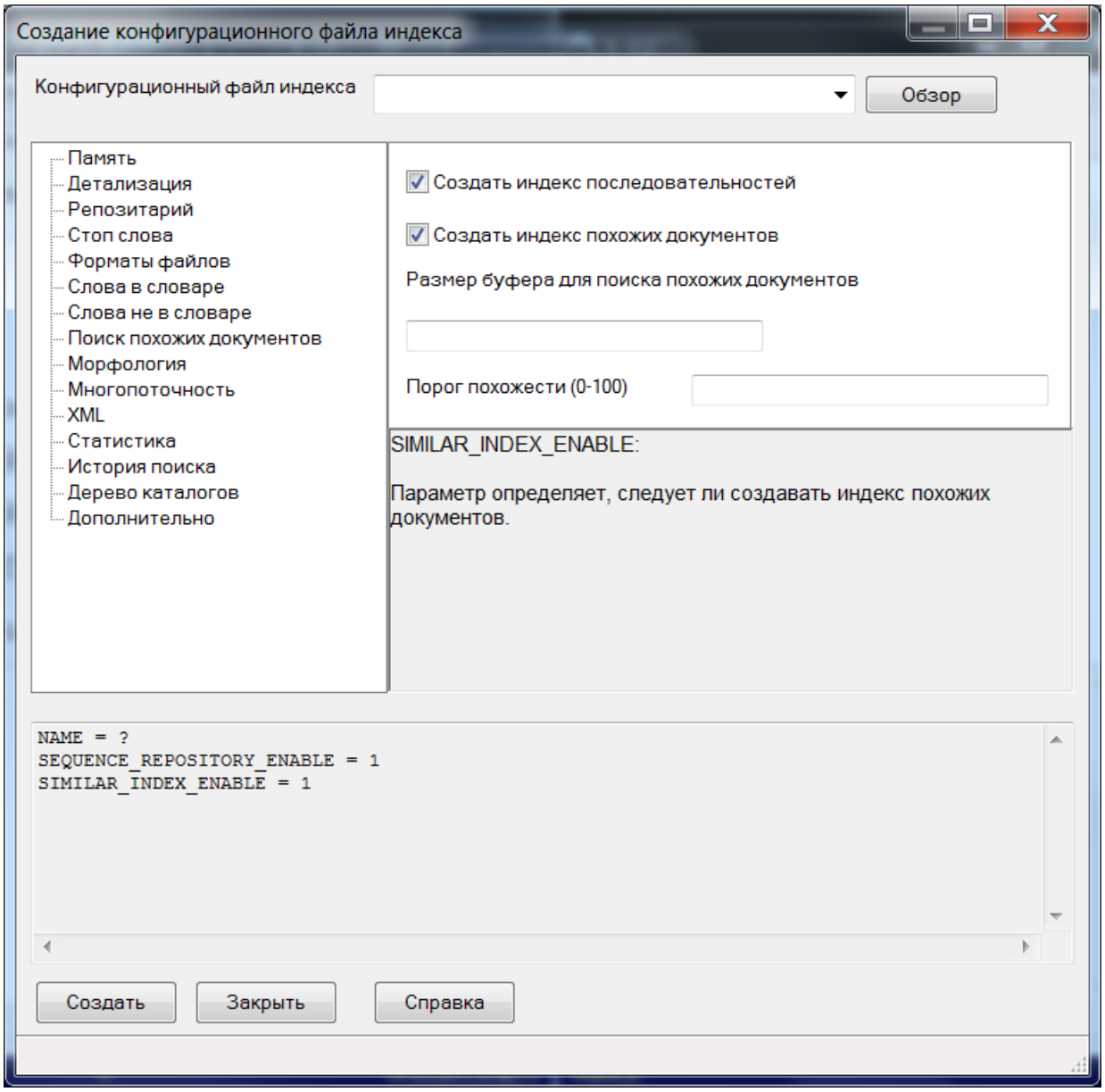


Рис. 5: Расширенное создание индекса

Поле «Конфигурационный файл индекса» является выпадающим списком. В данном списке присутствует несколько вариантов шаблонов, предназначенных для компьютеров разной производительности.

Внизу окна отображается получаемый конфигурационный файл.

При нажатии на кнопку «Создать» будут запрошены имя файла и папка, где будет создан конфигурационный файл.

Дополнительные опции

После нажатия на кнопку «Дополнительно» доступны дополнительные опции.

В поле «Максимальный размер» можно указать максимальный размер файлов, который следует проиндексировать, при достижении данного размера индексирование останавливается.

В поле «Размер сканирования» указывается размер итерации создания индекса. В процессе индексирования указанное число байт считывается во временные файлы, затем из временных файлов информация переносится в индекс, затем процесс повторяется. Чем больше данный параметр – тем больше быстродействие. Рекомендуется не включать данный переключатель, в этом случае будет осуществлена только одна итерация. Размер временных файлов равен 70-90% от исходных файлов.

Поиск

Для поиска можно использовать скрипт $find.js$.

Требуется ввести одно или несколько слов, которые могу встречаться в текстах, соответствующих цели поиска. Будут найдены документы, в которых встречаются все указанные слова.

По умолчанию результаты поиска сортируются по тому, насколько близко в тексте обнаружены найденные слова.

Кроме простого списка слов допустимо указывать набор слов в двойные кавычках, например:

"300 Спартанцев".

В этом случае будет осуществлен точный поиск слов, заключенных в кавычки (т. е. найдены файлы, в которых указанные слова идут подряд). При использовании одинарных кавычек будут искаться файлы, в которых указанные слова идут рядом друг с другом, т. е. между ними могут быть другие слова, но сами искомые слова также должны располагаться в тексте недалеко друг от друга.

Допустимо комбинировать варианты, например:

"300 Спартанцев" Фермопилы Леонид

в этом случае для поиска слов в кавычках осуществляется точный поиск, а для запроса в целом применяются правила по умолчанию, т. е. в данном случае найдутся документы, включающие в себя фразу «300 Спартанцев» и слова Леонид, Фермопилы, при этом результаты поиска

будут отсортированы по максимальному расстоянию между найденными словами в текстах. Пример искомого фрагмента текста:

300 спартанцев во главе с царем Леонидом стойко обороняли Фермопилы.

Допустимо использовать одновременно как одинарные так и двойные кавычки в запросе, например:

’"300 Спартанцев"Фермопилы Леонид’.

При использовании после кавычек восклицательного знака «!» включается поиск с учетом фиксированного порядка слов.

Поддерживаются логические операторы.

Кратко возможности поиска приведены далее:

Операция	Использование
Все слова (Неявно используется логический оператор «И»)	term1 term2
Любое из слов (Логический оператор «ИЛИ»)	term1 | term2
Исключение документов, содержащих заданное слово («-» перед словом)	term1 -term2

Группировка	(term1 term2)
Точный поиск	"term1 term2"
Поиск слов, располагающихся рядом (одинарные кавычки)	'term1 term2'
Поиск слов с фиксированным порядком (добавляется «!»)	"term1 term2"! 'term1 term2'! (term1 term2)!

Основное окно поиска:

В режиме списка для каждого найденного результата приводится фрагмент текста, в котором находятся искомые слова. Также показываются язык, кодировка, размер и имя файла. Для fb2 файлов отображаются заголовок и автор текста, если они указаны в файле.

Нажмите «+» или «Ctrl+Вверх» для того чтобы увеличить фрагмент текста, если текущая длина фрагмента недостаточна. Используйте кнопки со стрелками для перехода к последующим найденным результатам.

Используйте контекстное меню для копирования фрагмента текста. В индексе сохраняется полное имя исходного файла, используйте пункт «Открыть файл» в контекстном меню для открытия файла. Файлы из архивов распаковываются автоматически.

Также можно отображать результаты поиска в виде таблицы.

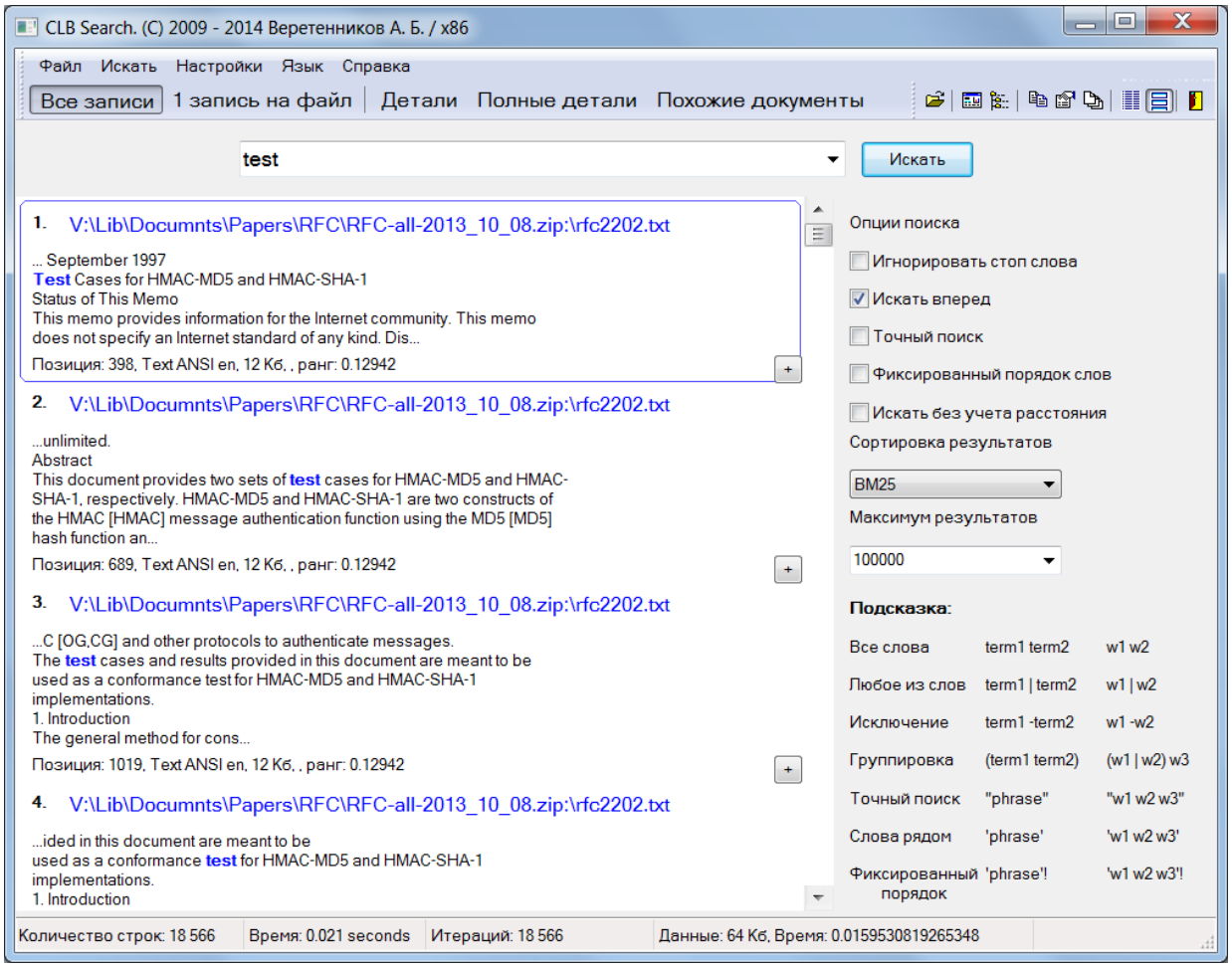


Рис. 6: Результаты поиска в виде списка

Для поиска следует ввести искомые слова или фразу и нажать на кнопку «Искать».

Опции поиска (переопределяются при использовании двойных или одинарных кавычек)

1. Игнорировать стоп слова.

   Игнорирование стоп слов при поиске.

2. Искать вперед.

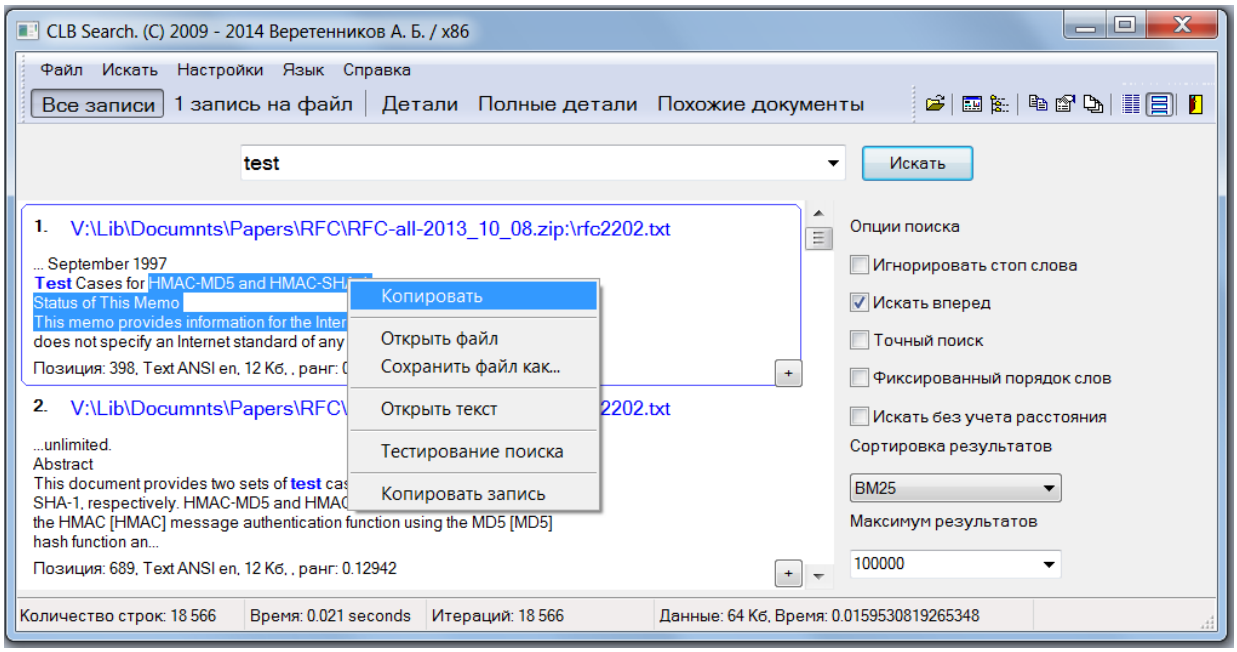


Рис. 7: Работа с результатами поиска

Записи отсортированы в том порядке, в котором были проиндексированы соответствующие файлы, если переключатель не включен – в обратном порядке.

3. Точный поиск.

   Точный поиск фразы. Ищутся только фрагменты, которые не содержат других слов, кроме заданных.

4. Фиксированный порядок слов.

   Порядок слов фиксированный, т. е. находятся только те фрагменты, которые содержат заданные слова в том же порядке, как они указаны.

5. Искать без учета расстояния.

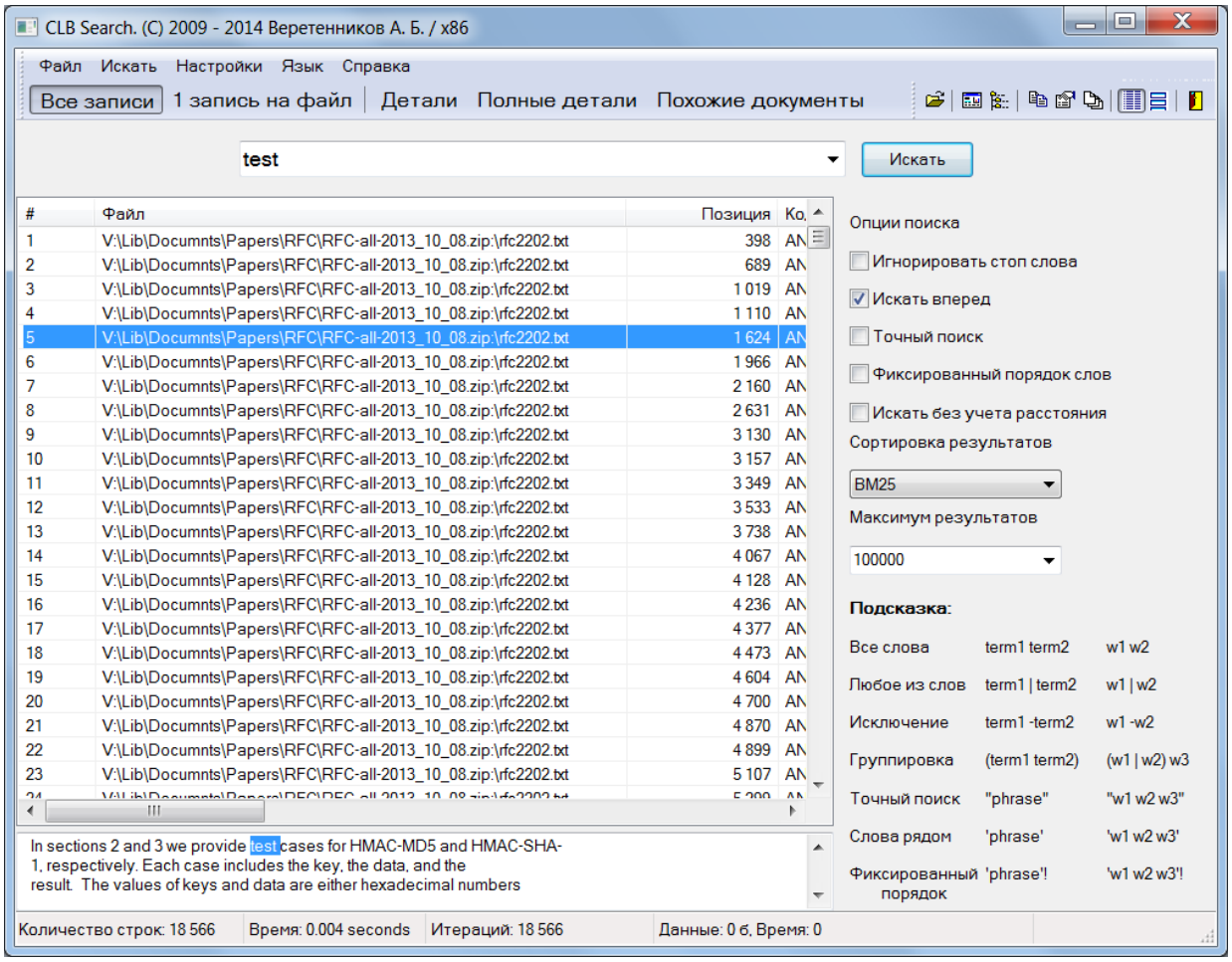


Рис. 8: Результаты поиска в виде таблицы

Расстояние между словами в полученной найденной строке текста значения не имеет, если переключатель не включен – то найденные фрагменты отсортированы по их длине.

6. Сортировка результатов.

   Возможные значения:

   (a) По умолчанию: нет сортировки результатов поиска.

(b) Позиция в документе: найденные записи сортируются по номеру их первого слова.

(c) TF-IDF: сортировка результатов поиска с учетом TF-IDF.

(d) BM25: сортировка результатов поиска с учетом BM25.

(Для индексов, созданных с помощью версий программы до 1.1 требуется пересоздание для поддержки TF-IDF и BM25)

7. Максимум результатов.

   Максимальное количество записей, которое возвращает запрос.

На панели инструментом располагается переключатели «Все записи / 1 запись на файл». Определяет способ вывода записей, если выбран режим «Все записи» выводятся все найденные записи, если «1 запись на файл» – для каждого документа выводится только первая найденная запись. Чтобы посмотреть остальные вхождения фразы в данном файле нужно выбрать файл в таблице и нажать кнопку «Детали» или «Полные детали». При нажатии кнопки «Полные детали» поиск осуществляется повторно, т. к. в первоначальном поиске могли быть включены не все результаты для данного файла (из-за ограничения «Максимум результатов» в результаты поиска могут быть не все возможные фрагменты, которые соответствуют заданным словам).

Сортировка результатов определяется различными параметрами, указанными выше. Если поиск осуществляется с учетом расстояния, то результаты сортируются по длине найденного фрагмента (более короткие фрагменты вначале), если длина одинаковая, то в используется значение

переключателя Сортировка на втором уровне, если он имеет значение «Позиция в документе». Последним применяется значение переключателя «Искать вперед».

Когда кнопка «Искать» нажимается первый раз, происходит загрузка словарей и открытие индекса, если индекс не открыт.

Если результаты поиска отображаются в виде таблицы, то каждая строка таблицы содержит в себе имя файла, который содержит требуемые слова, информацию о местонахождении данных слов в файле и информацию об этом файле, в частности кодировку файла, формат файла, идентификатор файла. Внизу экрана располагается поле для отображения фрагмента файла. При выборе строки таблицы в указанном поле отображается фрагмент файла, содержащий искомые слова. В данном фрагменте найденная фраза выделяется цветом.

Просмотр списка проиндексированных документов

Пункт меню «Файл \Все документы» отображает окно со списком всех проиндексированных документов.

Для каждого документа отображается некоторая информация, например имя документа, кодировка, формат документа, ID документа, размер текста.

Можно просмотреть текст документа, если было включено создание репозитария. Для этого нужно дважды щелкнуть мышкой по строке таблицы, соответствующей документу, или воспользоваться пунктом контекстного меню таблицы «Открыть».

Открытие исходного файла из результатов поиска

При просмотре результатов поиска можно кликнуть правой кнопкой мышки на результате для отображения контекстного меню.

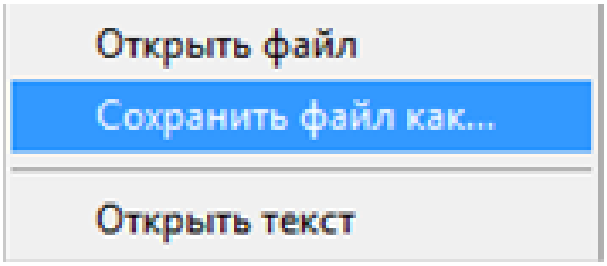


Рис. 9: Просмотр файла из результатов поиска

Пункт меню «Сохранить файл как...» позволяет сохранить текущий файл в заданном месте. Пункт меню «Открыть файл» позволяет открыть файл. Если файл находится в архиве, его извлечение осуществляется автоматически и файл помещается в папку для временных файлов. Удаление временных файлов происходить при завершении работы программы.

Поиск похожих документов

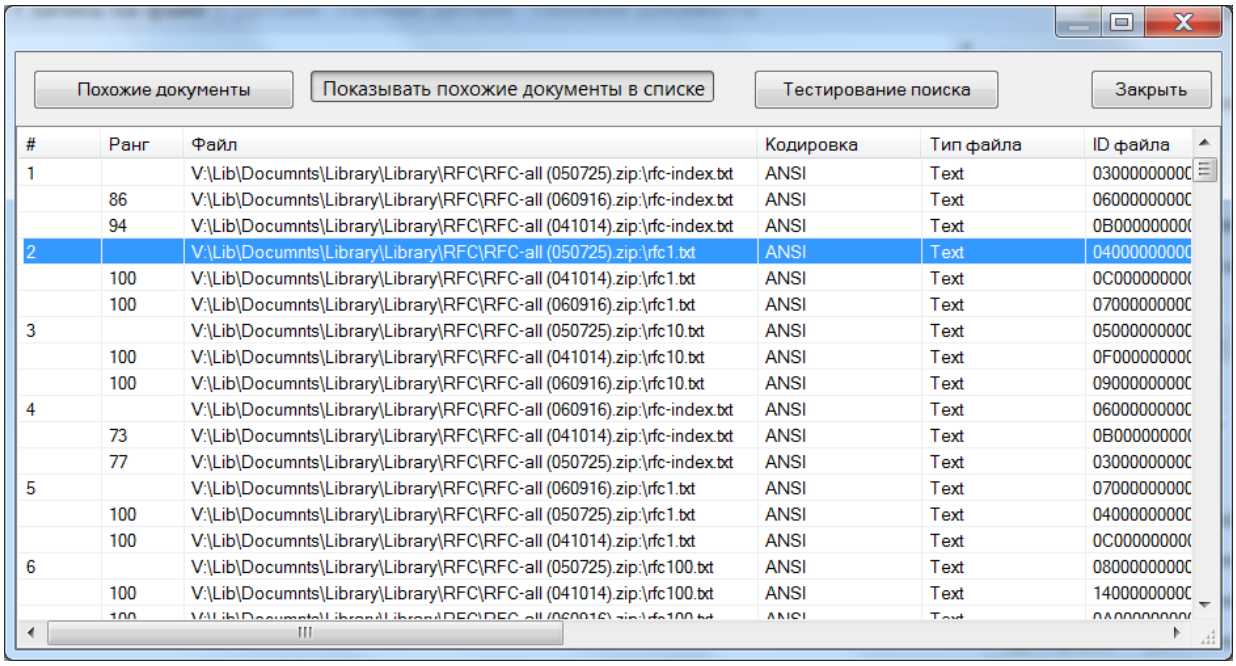
Программа имеет функцию поиска похожих документов. При ее включении после создания индекса ищутся все документы среди проиндексированных, которые могут быть похожими друг другу. Функция требует определенного количества оперативной памяти.

Пункт меню «Файл \Список похожих документов» в $find.js$ позволяет просмотреть похожие документы. В появляющемся окне отображаются все документы, для которых найден хотя бы один похожий документ. Выбрав один из документов можно нажать на кнопку «Похожие документы» чтобы просмотреть список похожих документов, для данного документа. Для каждого похожего документа отображается % похожести.

Если нажата кнопка «Показывать похожие документы в списке», то похожие документы показываются в общем списке, ниже основного документа. Для каждого похожего документа в колонке «Ранг» отображается % похожести.

Индекс похожих документов не включен по умолчанию. Следующие настройки должны быть включены в конфигурационном файле индекса для построения индекса похожих документов (далее указана настройка и пример значения):

1. SEQUENCE_REPOSITORY_ENABLE = 1
2. SIMILAR_INDEX_ENABLE = 1
3. SEQUENCE_BUFFER_SIZE = 1024*1024*1024



#	Ранг	Файл	Кодировка	Тип файла	ID файла
1		V:\Lib\Documnts\Library\Library\RFC\RFC-all (050725).zip:\rfc-index.txt	ANSI	Text	0300000000
	86	V:\Lib\Documnts\Library\Library\RFC\RFC-all (060916).zip:\rfc-index.txt	ANSI	Text	0600000000
	94	V:\Lib\Documnts\Library\Library\RFC\RFC-all (041014).zip:\rfc-index.txt	ANSI	Text	0B000000000
2		V:\Lib\Documnts\Library\Library\RFC\RFC-all (050725).zip:\rfc1.txt	ANSI	Text	0400000000
	100	V:\Lib\Documnts\Library\Library\RFC\RFC-all (041014).zip:\rfc1.txt	ANSI	Text	0C000000000
	100	V:\Lib\Documnts\Library\Library\RFC\RFC-all (060916).zip:\rfc1.txt	ANSI	Text	0700000000
3		V:\Lib\Documnts\Library\Library\RFC\RFC-all (050725).zip:\rfc10.txt	ANSI	Text	0500000000
	100	V:\Lib\Documnts\Library\Library\RFC\RFC-all (041014).zip:\rfc10.txt	ANSI	Text	0F000000000
	100	V:\Lib\Documnts\Library\Library\RFC\RFC-all (060916).zip:\rfc10.txt	ANSI	Text	0900000000
4		V:\Lib\Documnts\Library\Library\RFC\RFC-all (060916).zip:\rfc-index.txt	ANSI	Text	0600000000
	73	V:\Lib\Documnts\Library\Library\RFC\RFC-all (041014).zip:\rfc-index.txt	ANSI	Text	0B000000000
	77	V:\Lib\Documnts\Library\Library\RFC\RFC-all (050725).zip:\rfc-index.txt	ANSI	Text	0300000000
5		V:\Lib\Documnts\Library\Library\RFC\RFC-all (060916).zip:\rfc1.txt	ANSI	Text	0700000000
	100	V:\Lib\Documnts\Library\Library\RFC\RFC-all (050725).zip:\rfc1.txt	ANSI	Text	0400000000
	100	V:\Lib\Documnts\Library\Library\RFC\RFC-all (041014).zip:\rfc1.txt	ANSI	Text	0C000000000
6		V:\Lib\Documnts\Library\Library\RFC\RFC-all (050725).zip:\rfc100.txt	ANSI	Text	0800000000
	100	V:\Lib\Documnts\Library\Library\RFC\RFC-all (041014).zip:\rfc100.txt	ANSI	Text	1400000000

Рис. 10: Список похожих документов

Эта настройка определяет объем памяти, который используется при создании индекса похожих документов.

Пункт меню «Файл \Уникальные документы» отображает список документов, для которых похожие документы отсутствуют.

Дополнительные настройки могут быть указаны в конфигурационном файле индекса:

1. SIMILAR_REQUIRED_PERCENT = 70

   Документы считаются похожими, если их содержимое совпадает на заданное количество процентов, т. е. например, значение 70 для настройки – означает, что документы похожи, если из содержимое одинаково на 70

2. SIMILAR_INDEX_THREADS = 4

   Число потоков для создания индекса похожих документов.

3. SIMILAR_INDEX_MIN_OVERRUN = 1024*1024*1024*5

   Система начинает пропускать часть похожих документов, если их слишком много.

   Это необходимо на случай, если коллекция обрабатываемых документов содержит слишком много похожих документов. Например, если коллекция содержит 100 000 одинаковых документов то в индекс похожих документов требуется поместить 10 000 000 000 записей. За счет пропуска части записей размер индекса похожих документов не будет чрезмерно большим.

   Данный параметр запрещает любой пропуск похожих документов, если размер файла индекса похожих документов меньше данного параметра. То есть пропуск излишних похожих документов начинается при достижения файла индекса данного размера.

Структура каталогов

Пункт меню «Файл \Структура каталогов» показывает структуру каталогов для файлов индекса.

Для каждого каталога показывается количество папок и файлов в каталоге.

Ниже отображается список файлов в каталоге.

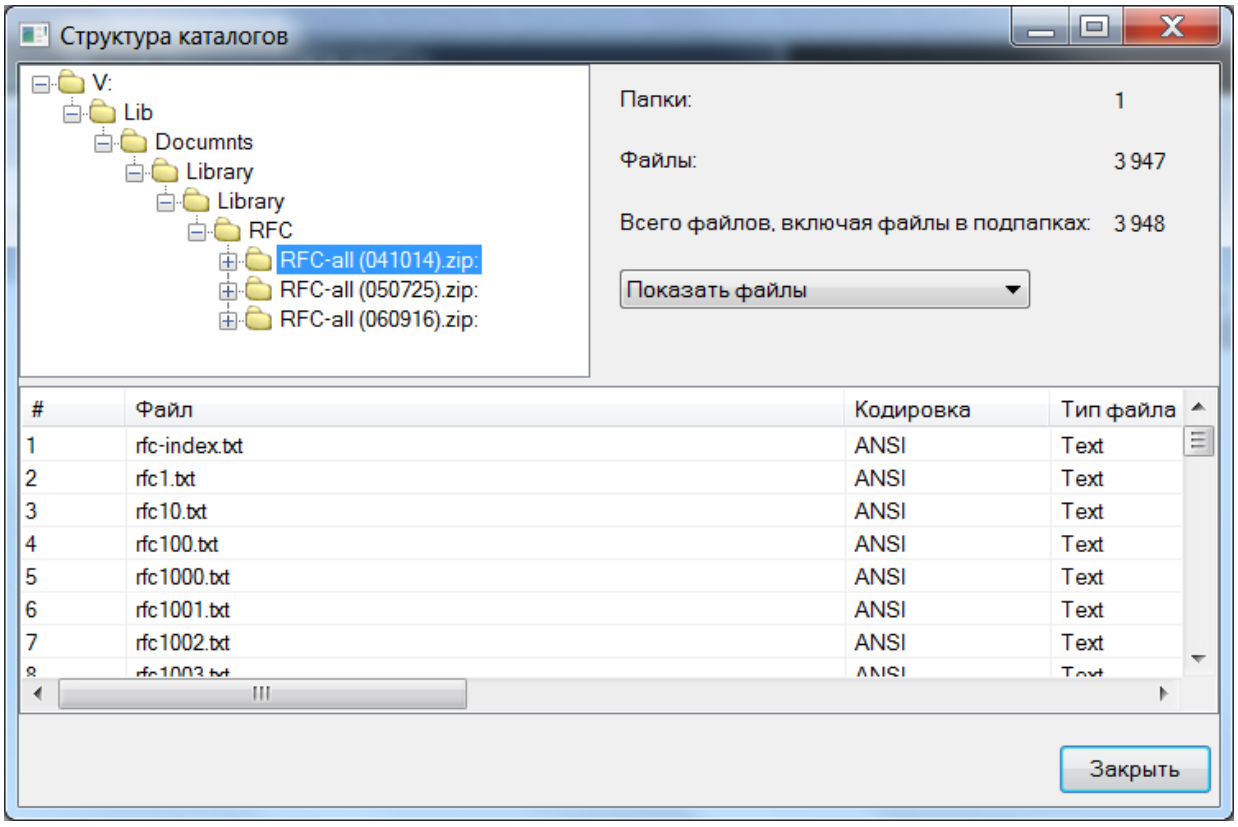


Рис. 11: Структура каталогов

Доступны следующие опции

- Показать файлы

- Показать файлы, включая файлы в подкаталогах
- Уникальные документы

  Этот пункт доступен, только если создан индекс похожих документов. В этом случае показываются документы, для которых не существует похожего документа, находящегося вне выбранной папки.
- Не уникальные документы

  Этот пункт доступен, только если создан индекс похожих документов. В этом случае показываются документы, для которых существует похожий документ, находящийся вне выбранной папки.

В случае отображения в режимах «Уникальные документы» и «Не уникальные документы» доступна кнопка «Показывать похожие документы в списке», которая включает отображение похожих документов после самого документа. Для каждого похожего документа показывается ранг похожести и полный путь.

Примечание. В режиме «Уникальные документы» также могут отображаться документы с похожими документами, если все похожие документы находятся в выбранной папке или ее подпапке.

Настройки в конфигурационном файле индекса:

- DIRECTORY_TREE_ENABLE=1

  Установка этого параметра в 0 запрещает создание индекса дерева каталогов.

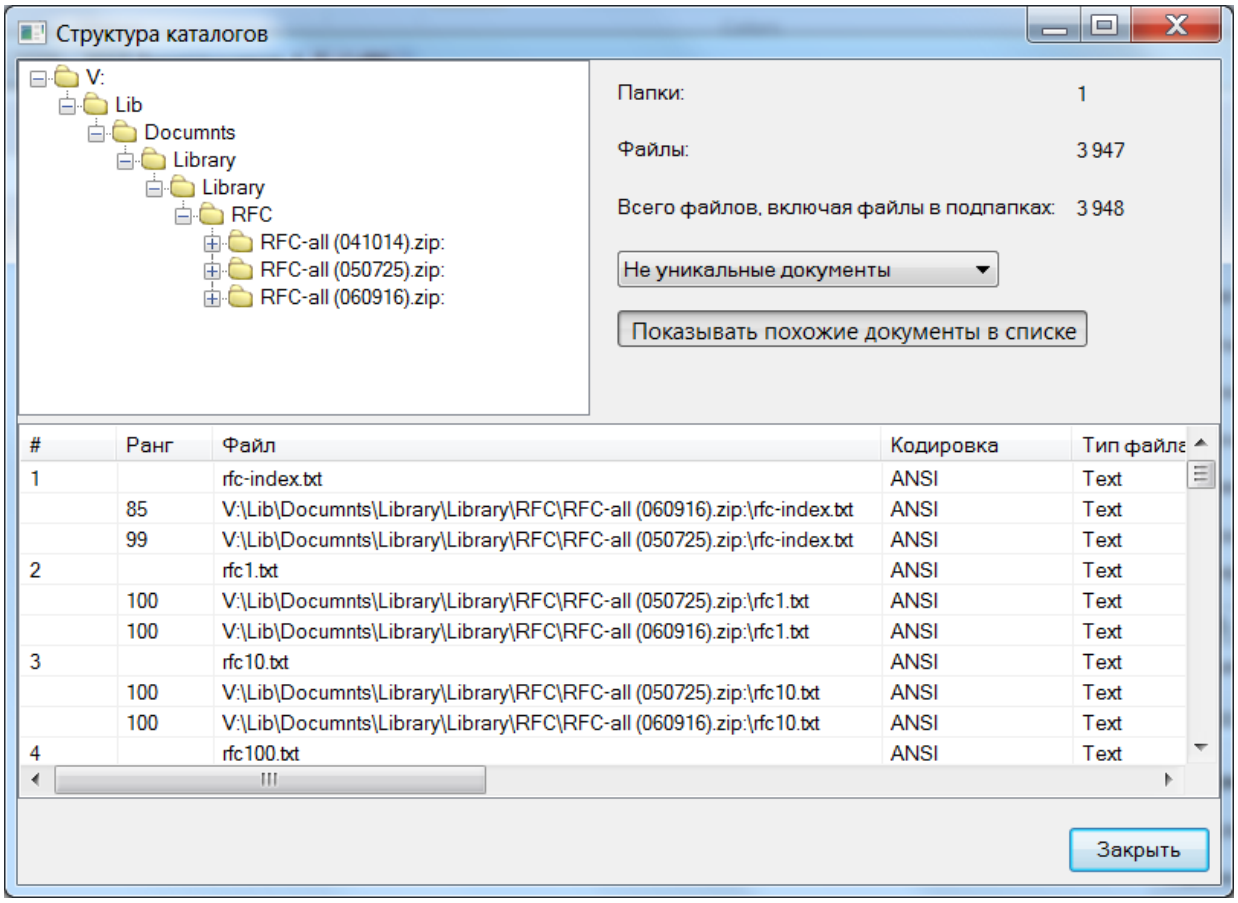


Рис. 12: Структура каталогов, не уникальные документы

Информация об индексе

Пункт меню «Файл \Информация» отображает окно с информацией об индексе.

На первой закладке показаны основные счетчики по обработанным данных, «Всего слов», «Общая длина тексте» и т.д.

Известные слова – слова входящие в словарь морфологического анализатора, неизвестные – все остальные.

Статистика сохраняется для каждого обновления индекса.

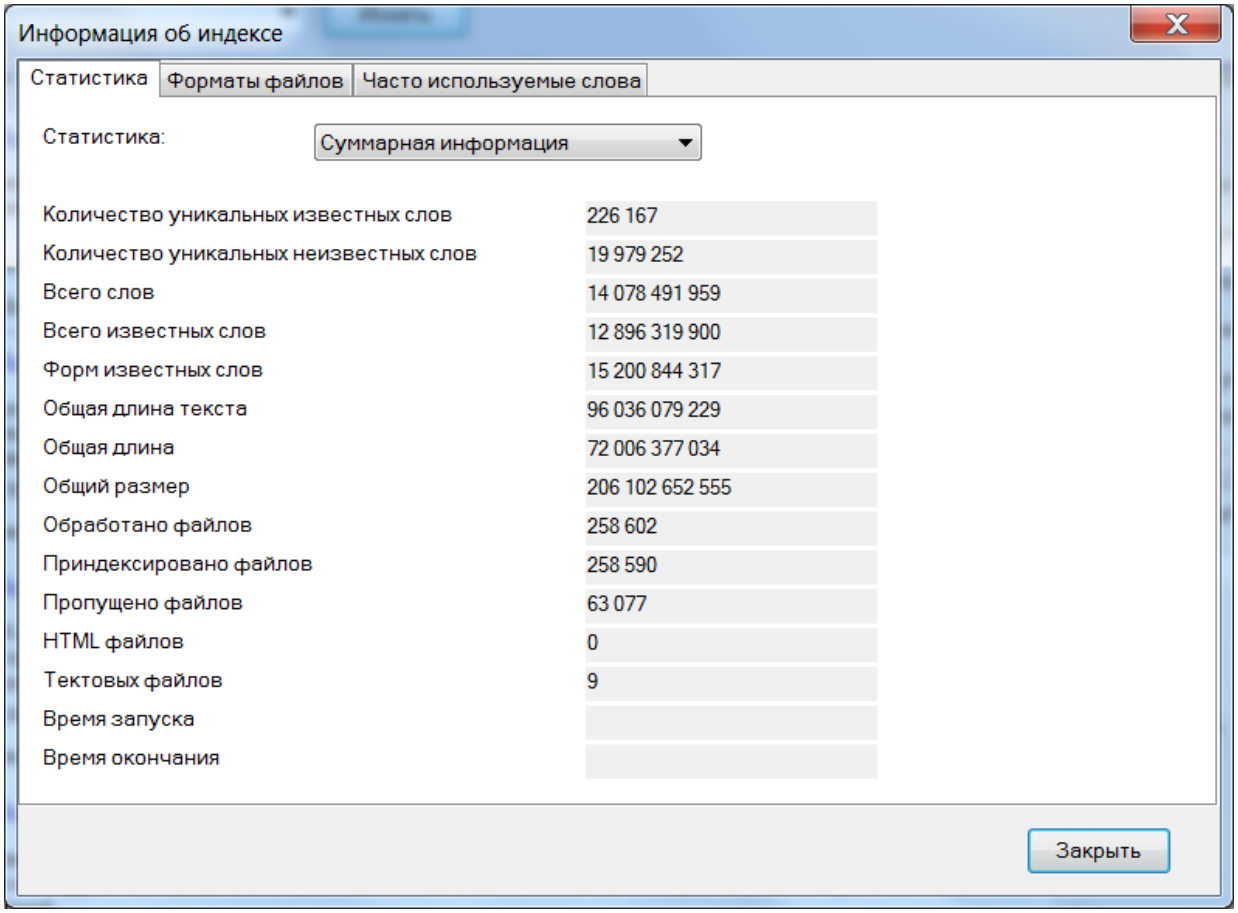


Рис. 13: Информация об индексе

События

Пункт меню «Файл \События» показывает список событий.

Примерами события являются: сообщение сервиса преобразования форматов об ошибке CRC при извлечении документа из архива.

Список возможных событий

- Archive/CRC_Error

  Ошибка CRC файла в архиве.

- Archive/Unsupported_Method

  Неподдерживаемый метод сжатия файла в архиве.

- Archive/Data_Error

  Ошибка данных.

- Archive/Unknown_Error

  Неизвестная ошибка.

- Archive/Unknown_Format

  Неизвестный формат файла.

- Archive/Unknown_Password

  Файл заархивирован с паролем.

- Archive/Volume_Error

  Ошибка многотомного архива.

- Archive/Extract_Partially_Failed

  В процессе распаковки архива возникла ошибка. Распакована только часть файлов.

- Transformation/Memory_Overhead

  Процесс модуля поддержки форматов файлов использовал слишком много памяти и был перезагружен для освобождения ресурсов.

- Indexing/Big_File

  Индексируемый файл слишком большого размера. Превышен предел длины файла или количества слов в файле. Текущие ограничения – максимальное значение позиции слова в файле = 268435456, в зависимости от настроек индекса позиция слова определяется как отступ в байтах от начала файла (DATA_DETAILS = POSITION в конфигурационном файле индекса) или как порядковый номер слова (DATA_DETAILS = WORD в конфигурационном файле индекса). Записи о вхождении слов с позициями, превышающими максимально допустимое значение, пропущены.

- Indexing/Too_Many_Forms

  Словоформа имеет слишком много базовых форм. Это событие не должно возникать с текущими словарями.

История поиска

По умолчанию при поиске история поисковых запросов сохраняется в специальном файле.

Пункт меню «Искать \Очистить историю поиска» в $find.js$ позволяет очистить историю.

Настройки в конфигурационном файле индекса:

- SEARCH_HISTORY_ENABLE=1

  Установка этого параметра в 0 запрещает сохранение истории поиска.

Конфигурационный файл индекса

Конфигурационный файл индекса представляет собой обычный текстовый файл со строками вида

НАСТРОЙКА=ЗНАЧЕНИЕ

Файл может быть создан с помощью мастера, как указано в разделе «Создание индекса» или вручную.

Рекомендуется создавать конфигурационный файл с расширением icf.

Возможные настройки указаны в файле docs\index.pdf.

Правила форматирования

В строковых параметрах можно указывать константы:

1. $\%INDEX\_NAME\%$ Значение параметра NAME
2. $\%INDEX\_PATH\%$ Путь файла, определяемого параметром NAME

Если в параметре – имени файла или пути путь указан не полный, а относительный, то считается что это путь в каталоге, где лежит конфигурационный файл.

При указании значения параметра «Число» можно указывать после числового значения символы K и M – что соответствует килобайтам и мегабайтам.

При указании значения типа BOOL доступные значения: TRUE, FALSE, 0, 1.

В качестве примера можно посмотреть файл index.icf, который располагается в основном каталоге программы.

Журнал индекса

В журнал индекса сохраняется различная статистическая информация, в частности о времени создания индекса, количестве и объеме обработанных файлов, и т. д.

Журнал индекса представляет собой обычный текстовый файл.

Месторасположение журнала индекса определяется настройкой LOG в конфигурационном файле индекса.

Список некоторых позиций в отчете создания индекса:

- Время чтения файлов

  Суммарное время, потраченное на чтение файлов

- Время создания индекса

  Время создания индекса

- Общее время

  Время создания индекса + время чтения файлов

- Полное время операции

  Полное время – общее время + сброс кеша в файлы.

- Время определения формата файлов

  Время, затраченное на определение кодировки файлов

- Суммарный размер

  Суммарный размер обработанных файлов

- Суммарный размер текста

  Суммарный размер текста в обработанных файлах

- Суммарная длина слов

  Суммарная длина словоформ

- Всего слов

  Суммарное количество словоформ

- 'Известных' слов

  Доля словоформ, входящих в словарь морфологического анализатора

- Форм 'известных' слов

  Количество базовых форм слов, для словоформ, входящих в словарь морфологического анализатора

- 'Неизвестных' слов

  Суммарное количество словоформ, не входящих в словарь морфологического анализатора

- Новых 'известных' слов

  Количество новых словоформ, входящих в словарь морфологического анализатора

- Новых 'неизвестных' слов

  Количество новых словоформ, не входящих в словарь морфологического анализатора

- Всего файлов

  Суммарное количество обработанных файлов

- Проиндексировано файлов

  Суммарное количество проиндексированных файлов

- Пропущено файлов

  Суммарное количество пропущенных файлов

- HTML файлов

  Суммарное количество HTML файлов

- Текстовых файлов

  Суммарное количество текстовых файлов

Настройки программы

Настройки программы хранятся в файле index.cfg. Данный файл располагается в основном каталоге программы рядом с файлом index.dll.

Конфигурационный представляет собой обычный текстовый файл со строками вида:

«Настройка» = «Значение»

Возможные настройки описаны в файле library.pdf.

Дополнительные модули

Поддержка текстовых файлов, HTML, XML (FB2), встроена в ядро программы.

С помощью дополнительных модулей осуществляется поддержка различных дополнительных форматов файлов и архивов.

Список модулей:

1. rtf

   Поддержка формата RTF

2. djvu

   Поддержка формата DJVU, входит в CLB Search Extensions.

3. cab\cab.dll

   Поддержка архивов CAB

4. rar\rar.dll

   Поддержка архивов RAR

5. 7Z\7Z.dll

   Поддержка архивов 7Z (и др.)

6. msword

   Поддержка формата Microsoft Word (doc)

7. pdf

   Поддержка формата PDF, входит в CLB Search Extensions.

8. Java:pdf

   Второй вариант поддержки PDF (требуется JAVA)

9. Java:doc

   Поддержка формата Microsoft Word (doc) (требуется JAVA)

10. Java:ppt

    Поддержка формата Microsoft Power Point (ppt) (требуется JAVA)

11. Java:xls

    Поддержка формата Microsoft Excel (xls) (требуется JAVA)

12. Java:docx

    Поддержка формата Microsoft Word 2007 (docx) (требуется JAVA)

13. Java:pptx

    Поддержка формата Microsoft Power Point 2007 (pptx) (требуется JAVA)

14. Java:xlsm

    Поддержка формата Microsoft Excel 2007 с макросами (xlsm) (требуется JAVA)

15. Java:xlsx

    Поддержка формата Microsoft Excel 2007 (xlsx) (требуется JAVA)

16. Java:epub

    Поддержка формата epub (требуется JAVA)

17. Java:odt

    Поддержка формата Open Office Text (odt) (требуется JAVA)

18. Java:ods

    Поддержка формата электронных таблиц Open Office (ods) (требуется JAVA)

19. Java:odp

    Поддержка формата презентаций Open Office (odp) (требуется JAVA)

Для некоторых форматов требуется Java. В CLB Search встроена внутренняя (private) JRE. Внутренняя JRE не регистрируется в системе, не используется в браузерах и не обновляется автоматически. См. подробнее раздел «Модуль поддержки форматов».

Модуль поддержки форматов

Загрузка дополнительных модулей может осуществляться как в основном процессе, так и в отдельном процессе – модуле поддержки форматов. Использование модуля поддержки форматов или сервиса дополнительных модулей контролируется параметром в конфигурационном файле индекса: USE_TRANSFORMATION_SERVICE.

На схеме отображена работа модуля поддержки форматов.

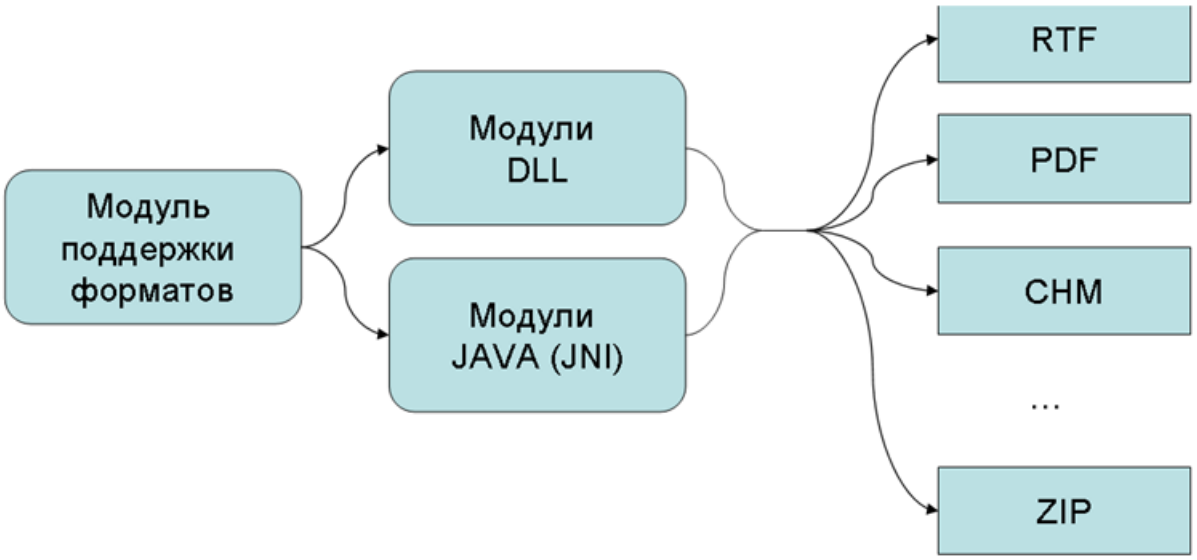


Рис. 14: Модуль поддержки форматов

Модуль поддержки форматов реализует уровень изоляции дополнительных модулей от ядра программы.

При сбое в одном из дополнительных модулей модуль дополнительных

форматов перезапускается автоматически. Таким образом, процесс создания индекса не прерывается.

Модуль дополнительных форматов перезапускается автоматически, если он не отвечает на запросы в течение заданного промежутка времени, определяемого в конфигурационном файле программы index.cfg параметром TRANSFORMATION_SERVICE_TIMEOUT. Значение параметра указывается в миллисекундах, например:
TRANSFORMATION_SERVICE_TIMEOUT = 1000*60*10

В процессе извлечения файлов из архивов небольшие файлы могут обрабатываться в оперативной памяти. Файлы большого размера извлекаются во временную папку. По умолчанию используется папка для временных файлов операционной системы (%TEMP%). При необходимости в конфигурационном файле индекса можно указать другую папку с помощью параметра TRANSFORMATION_SERVICE_TEMP, например:
TRANSFORMATION_SERVICE_TEMP=H:\Temp

Для поддержки модулей форматов, для которых требуется Java каждый процесс модуля поддержки форматов запускает Java машину.

В CLB Search встроена внутренняя (private) JRE. Внутренняя JRE не регистрируется в системе , не используется в браузерах и не обновляется автоматически.

При необходимости можно настроить использование внешней JRE, изменив параметр JAVA_HOME в конфигурационном файле программы.

Запись диагностической информации

Программа осуществляет запись различных сообщений в журнал. Файлы журнала располагаются по умолчанию в каталоге LOG основного каталога программы. При необходимости можно указать другой каталог в конфигурационном файле программы с помощью настройки LOG_PATH.

В зависимости от значения настройки DEBUG в журнал записывается либо только основные сообщения, либо кроме этого, еще дополнительная отладочная информация.

Например, в случае невозможности записи индекса, вследствие ошибок ввода-вывода или других ошибок подробное описание ошибки сохраняется в журнале.

Некоторые сообщения, в частности отчет о создании индекса сохраняется в отдельном журнале индекса. При записи сообщения в журнал индекса, оно дублируется в основной журнал.

Журнал состоит из набора файлов заданного размера. Размеры файлов определяются параметром MAX_LOG_FILE_SIZE. Если параметр не указан, создаются файлы размером по 64 мегабайта.

Как только текущий файл журнала достигает максимального допустимого размера, создается новый файл журнала, и запись отладочной информации с этого момента производится в новый файл. Название файлов журнала формируется на основании текущей даты и времени.

Если задан параметр MAX_LOG_TOTAL_SIZE, то, если суммарный размер всех файлов журнала становится больше значения параметра,

старые файлы журнала удаляются.

Внутреннее устройство

Основная часть программы реализована в виде COM сервера. Оконный интерфейс реализуется с помощью COM сервера WSO.

Программа состоит из нескольких модулей:

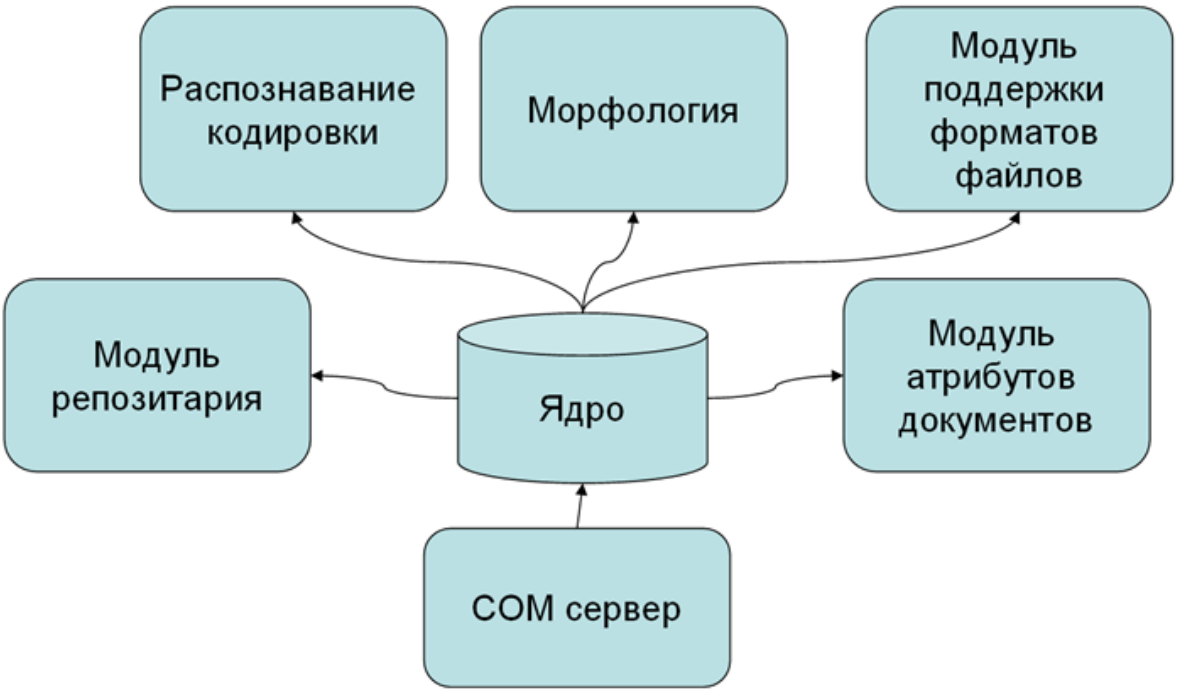


Рис. 15: Модули

1. Ядро – осуществляет создание индекса и поиск.
2. Модуль поддержки морфологии.
3. Модуль распознавания кодировки.

4. Модуль поддержки форматов файлов.

5. Модуль атрибутов документов сохраняет в файле метаданные документов: имя файла, формат и т. д.

6. Модуль репозитария сохраняет текст документов и позволяет извлекать фрагменты текста при поиске, для отображения результатов поиска.

7. Модуль COM позволяет обращаться к программе через COM.

Список файлов

Список файлов, входящих в программу, в данном списке указаны возможно не все файлы.

Имя	Описание
Docs\Help.pdf	Данный файл
Docs\Index.pdf	Описание конфигурационного файла индекса
Docs\Library.pdf	Описание настроек программы
Docs\CLBLibrary.chm	Описание COM интерфейсов программы.
Docs\results.pdf	Результаты некоторых экспериментов по созданию индекса.
license.txt	Лицензия по использованию программы CLB Search
Index.dll	COM сервер CLB
Index64.dll	COM сервер CLB (64-битная версия).
index.icf	Пример конфигурационного файла индекса
find.js	Скрипт для поиска в индексе
Make.js	Скрипт для создания индекса
Dictionaries.js	Скрипт для просмотра морфологических словарей

index.cfg	Конфигурационный файл библиотеки
Transfrm.exe	Сервис дополнительных форматов
RTF.dll	Поддержка формата RTF
cab\cab.dll	Поддержка архивов CAB
rar\rar.dll	Поддержка архивов RAR
7Z\7Z.dll	Поддержка архивов 7Z (и др.)
MSWORD.dll	Поддержка формата MSWORD
Java*	Поддержка форматов (требуется JAVA, не ранее 1.5)
LIB\lib.js	Вспомогательный файл, использующийся в make.js, find.js
NLS*.*	Файлы поддержки различных языков интерфейса в make.js, find.js
Dict*.*	Морфологические словари

Программный интерфейс (API)

Основные функции программы доступны через COM. Описание интерфейсов см. в Docs\ClbLibrary.chm. COM объекты следует использовать только в языках, которые поддерживают автоматизацию (Automation), т. е. обращаться к ним через IDispatch.

Настройка использования памяти

В данном разделе указаны те настройки, которые влияют на использование памяти.

Данные настройки можно не указывать, тогда они определяются автоматически исходя из объема оперативной памяти компьютера.

Параметр	Описание
Размер буфера кеша	Параметр BUFFER в конфигурационном файле индекса. Это размер основного, самого большого буфера, рекомендуется указывать как можно больше, от 300 Мб. до 2 Гб. Большие значения существенного влияния на производительность не оказывают. Рекомендуется 300 мегабайт, если объем оперативной памяти меньше или равен 512 мегабайт, 700 мегабайта, если объем оперативной памяти от 512 до 1024 мегабайта, 1-2 Гб. если оперативной памяти больше.

Дополнительный буфер кеша	Параметр APPEND_BUFFER в конфигурационном файле индекса. Рекомендуется 16-32 мегабайт, если объем оперативной памяти меньше или равен 512 мегабайт, 64 мегабайта, если объем оперативной памяти от 512 до 1024 мегабайта, 128 мегабайт если оперативной памяти больше. Дальнейшее увеличение параметра на производительность существенно не влияет.

Вспомогательный буфер	Параметр UNKNOWN_WORDS_BUFFER_SIZE в конфигурационном файле индекса. Рекомендуется не менее 64M
Размер буфера группы	Параметр GROUP_BUFFER_SIZE в конфигурационном файле индекса. По умолчанию 1M.
Вспомогательный буфер B-дерева	Параметр TREE_BUFFER в конфигурационном файле индекса. Рекомендуется указывать от 8 до 64 Мб.

Создание индекса состоит из следующих этапов

1. Чтение файлов.

   Пиковое использование памяти данном этапе определяется параметром GROUP_BUFFER_SIZE. Все слова текстов на данном этапе разделяются на группы. Для каждой группы слов организуется буфер данного размера. Максимальное количество групп примерно 200-300. К этому добавляется около 50-70 Мб. требуемое для остальных задач, например, для морфологического анализатора.

2. Запись индекса.

   Пиковое использование памяти на этапе создания индекса равно сумме параметров:

   BUFFER,

APPEND_BUFFER,

UNKNOWN_WORDS_BUFFER_SIZE,

TREE_BUFFER.

К этому добавляется около 50-70 Мб. требуемое для остальных задач, например, для морфологического анализатора.

3. Построение индекса похожих документов.

   Пиковое использование памяти на данном этапе определяется параметром SEQUENCE_BUFFER_SIZE, который по умолчанию равен параметру BUFFER.

4. Построение индекса дерева каталогов.

   Пиковое использование памяти на данном этапе определяется параметром DIRECTORY_TREE_BUFFER, который по умолчанию равен параметру BUFFER.

Многопоточность

Начиная с версии 1.0.0.2 поддерживается чтение файлов в многопоточном режиме (включается автоматически если имеется не менее 2 CPU) и создание репозитария в многопоточном режиме (включается автоматически если имеется не менее 3 CPU). Под CPU понимается логический процессор, например, Core 2 Duo E6700 обычно имеет 4 логических процессора, если включен hyperthreading. См. описание параметра MULTITHREADED_SCAN, MULTITHREADED_REPOSITORY и др. в Docs\index.pdf.

При многопоточном чтении файлов, если модули обработки форматов файлов запускаются в отдельном процессе (по умолчанию), т. е. USE_TRANSFORMATION_SERVICE=1 в конфигурационном файле индекса, то для каждого потока создается свой экземпляр сервиса преобразования форматов.

Также по умолчанию в многопоточном режиме создается индекс похожих документов. При необходимости можно указать количество создаваемых потоков в конфигурационном файле индекса, параметром SIMILAR_INDEX_THREADS.

Репозитарий

Репозитарий предназначен для сохранения в сжатом виде проиндексированных текстов и быстрого извлечения из них фрагментов текста при поиске.

Репозитарий необходим, т. к. извлечение фрагмента текста напрямую из исходных файлов требует большего времени, для каждого фрагмента текста потребуется открывать соответствующий файл, что может занять длительное время, если файл располагается в архиве или имеет сложный формат.

Алгоритм сжатия определяется параметром в конфигурационном файле индекса: REPOSITORY_COMPRESSION.

Возможные алгоритмы сжатия:

Значение параметра	Описание
NONE	Нет сжатия
1	LZ1
2	LZ2
3	LZ3
4	LZ4
5	LZ5
6	LZ6
7	LZ7
8	LZ8
9	LZ9
128	Huffman

Алгоритмы сжатия LZMA:

Значение параметра	Описание
100	LZMA0
101	LZMA1
102	LZMA2
103	LZMA3
104	LZMA4
105	LZMA5
106	LZMA6
107	LZMA7
108	LZMA8
109	LZMA9

Самый быстрый алгоритм сжатия Huffman (используется по умолчанию). Алгоритмы LZ (ZIP) обеспечивают хорошее сжатие при хорошей скорости. Алгоритмы LZMA обеспечивают максимальное сжатие, но скорость при этом может быть понижена, особенно на самом сильном варианте LZMA9. Алгоритмы LZ и LZMA не рекомендуется использовать, если отключено создание репозитария в многопоточном режиме или у CPU меньше 4-х ядер.

Дополнительные настройки поиска

Следующие настройки можно задать в конфигурационном файле индекса:

- FILELIST_MAX_MEMORY_SIZE

  Параметр определяет, какой объем файла с описанием документов следует кешировать при поиске. По умолчанию 128 мегабайт. Данный кеш не используется при создании индекса.

  Файл с описанием документов содержит имена и другие атрибуты документов. Кеширование данного файла в памяти ускоряет отображение результатов поиска.

Результаты экспериментов

В файле Docs\results.pdf приведены некоторые результаты экспериментов создания индекса.

Обработка XML файлов

Начиная с версии 1.0.0.4 поддерживается обработка XML файлов с помощью XML парсера (предыдущие версии обрабатывали XML файлы как обычный текст). XML парсер может быть отключен, см. настройки конфигурационного файла индекса, параметр ENABLE_XML_PARSER. Также, по умолчанию включена возможность отображения в результатах поиска Xpath для найденного фрагмента текста за счет сохранения информации о структуре XML документов,. См. параметр XML_INFO_FILE_ENABLE.

Пример результата поиска: 1.xml, /root/section[12]/p[23]. Т. е. искомые слова находятся в документе 1.xml, где внутри тега root, в 12-м по порядку теге section, есть теги p, и в 23-м из них и содержаться искомые слова. Это позволяет пользователю сразу определить смысл найденного фрагмента данных, если известна схема XML документа.

При включенном сохранении информации о структуре XML документа обрабатываются дополнительные конфигурационные файлы, которые располагаются в каталоге XML папки программы. Пример конфигурационного файла

<?xml version="1.0"encoding="windows-1251"?>
<FictionBook xmlns="http://www.gribuser.ru/xml/fictionbook/2.0"
xmlns:clb="http://veretennikov.org»
<binary clb:skip = "true»
</binary>
</FictionBook>

В данном примере для XML элементов с пространством имен «http://www.gribuser.ru/xml/fictionbook/2.0» для тега binary указано, что содержимое тега должно пропускаться. Настройки в данных конфигурационных файлах осуществляются путем указания атрибутов и подтегов, относящихся к пространству имен «http://veretennikov.org». В настоящее время предусмотрена только настройка skip – для пропуска те тегов, которые не требуется индексировать. Планируется расширение набора настроек.

Пункт меню «Файл \Пространства имен XML» отображает все обнаруженные пространства имен XML.

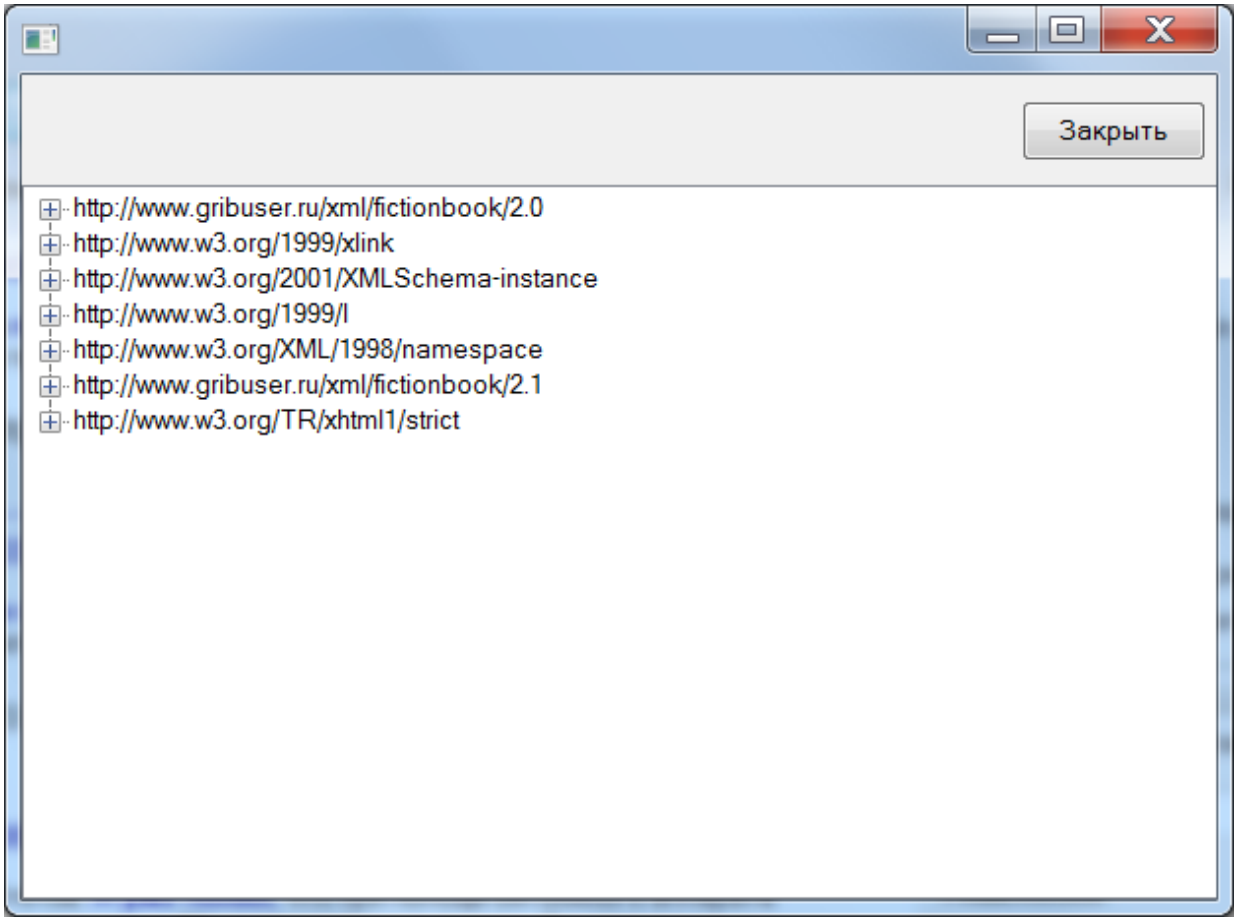


Рис. 16: Пространства имен XML

Морфологические словари

Словарь	Язык	Словоформ	Базовых форм
dic.aot.ru	Русский	3 122 696	171 570
dic.aot.en	Английский	200 562	92 374
dic.ru	Русский	3 496 317	205 362

Состав словарей можно просмотреть с помощью программы *Dictionaries.js*.

Новое в версии 1.0.0.4

- Изменения в интерфейсе программы поиска.
- Улучшенная обработка XML файлов.
- Улучшенная обработка FB2 файлов.
- Отображение XPATH в результатах поиска для XML файлов.
- Возможность просмотра исходного файла, с автоматическим извлечением из архива при необходимости.

Новое в версии 1.0.0.3

Новые возможности поиска. См. раздел «Поиск».

Новое в версии 1.0.0.2

Улучшенная поддержка многопоточности. Чтение документов в многопоточном режиме. Запись репозитария в многопоточном режиме. Чтение документов в многопоточном режиме важно при обработке файлов, которые сохранены не в простых текстовых форматах, например PDF, DJVU. Извлечение текста из таких файлов требует определенного времени и распараллеливание данной задачи значительно ускоряет процесс индексации текстовых коллекций, включающих файлы подобных форматов, в качестве примера см. results.pdf в котором приведены результаты создания индекса 300 ГБ документов, на что потребовалось 8 часов. Создание репозитария в многопоточном режиме может значительно ускорять процесс создания репозитария, если при этом используется сжатие.

Лицензионное соглашение

Благодарности

LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- Copyright (C) 1995-1998 Jean-loup Gailly and Mark Adler

  This software is provided 'as-is', without any express or implied warranty. In no event will the authors be held liable for any damages arising from the use of this software.

  Permission is granted to anyone to use this software for any purpose, including commercial applications, and to alter it and redistribute it freely, subject to the following restrictions:

  1. The origin of this software must not be misrepresented; you must not claim that you wrote the original software. If you use this software in a product, an acknowledgment in the product documentation would be appreciated but is not required.

  2. Altered source versions must be plainly marked as such, and must not be misrepresented as being the original software.

  3. This notice may not be removed or altered from any source distribution.

  Jean-loup Gailly Mark Adler

  jloup@gzip.org madler@alumni.caltech.edu

- Морфологические словари dic.aot.en.dat, dic.aot.ru.dat основаны на словарях AOT engmorph.zip, rusmorph.zip (www.aot.ru) и поставляются под GNU LESSER GENERAL PUBLIC LICENSE т. к. изначальные словари имеют лицензию GNU LESSER GENERAL PUBLIC LICENSE.

  GNU LESSER GENERAL PUBLIC LICENSE

  Version 2.1, February 1999

  Copyright (C) 1991, 1999 Free Software Foundation, Inc. 59 Temple Place, Suite 330, Boston, MA 02111-1307 USA Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

  [This is the first released version of the Lesser GPL. It also counts as the successor of the GNU Library Public License, version 2, hence the version number 2.1.]

  Preamble

  The licenses for most software are designed to take away your freedom to share and change it. By contrast, the GNU General Public Licenses are intended to guarantee your freedom to share and change free software–to make sure the software is free for all its users.

  This license, the Lesser General Public License, applies to some specially designated software packages–typically libraries–of the Free Software Foundation and other authors who decide to use it. You can use it too, but we suggest you first think carefully about whether this license or

the ordinary General Public License is the better strategy to use in any particular case, based on the explanations below.

When we speak of free software, we are referring to freedom of use, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for this service if you wish); that you receive source code or can get it if you want it; that you can change the software and use pieces of it in new free programs; and that you are informed that you can do these things.

To protect your rights, we need to make restrictions that forbid distributors to deny you these rights or to ask you to surrender these rights. These restrictions translate to certain responsibilities for you if you distribute copies of the library or if you modify it.

For example, if you distribute copies of the library, whether gratis or for a fee, you must give the recipients all the rights that we gave you. You must make sure that they, too, receive or can get the source code. If you link other code with the library, you must provide complete object files to the recipients, so that they can relink them with the library after making changes to the library and recompiling it. And you must show them these terms so they know their rights.

We protect your rights with a two-step method: (1) we copyright the library, and (2) we offer you this license, which gives you legal permission to copy, distribute and/or modify the library.

To protect each distributor, we want to make it very clear that there is no warranty for the free library. Also, if the library is modified by someone

else and passed on, the recipients should know that what they have is not the original version, so that the original author's reputation will not be affected by problems that might be introduced by others.

Finally, software patents pose a constant threat to the existence of any free program. We wish to make sure that a company cannot effectively restrict the users of a free program by obtaining a restrictive license from a patent holder. Therefore, we insist that any patent license obtained for a version of the library must be consistent with the full freedom of use specified in this license.

Most GNU software, including some libraries, is covered by the ordinary GNU General Public License. This license, the GNU Lesser General Public License, applies to certain designated libraries, and is quite different from the ordinary General Public License. We use this license for certain libraries in order to permit linking those libraries into non-free programs.

When a program is linked with a library, whether statically or using a shared library, the combination of the two is legally speaking a combined work, a derivative of the original library. The ordinary General Public License therefore permits such linking only if the entire combination fits its criteria of freedom. The Lesser General Public License permits more lax criteria for linking other code with the library.

We call this license the "Lesser"General Public License because it does Less to protect the user's freedom than the ordinary General Public License. It also provides other free software developers Less of an advantage over competing non-free programs. These disadvantages are the reason

we use the ordinary General Public License for many libraries. However, the Lesser license provides advantages in certain special circumstances.

For example, on rare occasions, there may be a special need to encourage the widest possible use of a certain library, so that it becomes a de-facto standard. To achieve this, non-free programs must be allowed to use the library. A more frequent case is that a free library does the same job as widely used non-free libraries. In this case, there is little to gain by limiting the free library to free software only, so we use the Lesser General Public License.

In other cases, permission to use a particular library in non-free programs enables a greater number of people to use a large body of free software. For example, permission to use the GNU C Library in non-free programs enables many more people to use the whole GNU operating system, as well as its variant, the GNU/Linux operating system.

Although the Lesser General Public License is Less protective of the users' freedom, it does ensure that the user of a program that is linked with the Library has the freedom and the wherewithal to run that program using a modified version of the Library.

The precise terms and conditions for copying, distribution and modification follow. Pay close attention to the difference between a "work based on the library"and a "work that uses the library". The former contains code derived from the library, whereas the latter must be combined with the library in order to run.

TERMS AND CONDITIONS FOR COPYING, DISTRIBUTION AND

MODIFICATION

0. This License Agreement applies to any software library or other program which contains a notice placed by the copyright holder or other authorized party saying it may be distributed under the terms of this Lesser General Public License (also called "this License"). Each licensee is addressed as "you".

A "library"means a collection of software functions and/or data prepared so as to be conveniently linked with application programs (which use some of those functions and data) to form executables.

The "Library", below, refers to any such software library or work which has been distributed under these terms. A "work based on the Library"means either the Library or any derivative work under copyright law: that is to say, a work containing the Library or a portion of it, either verbatim or with modifications and/or translated straightforwardly into another language. (Hereinafter, translation is included without limitation in the term "modification".)

"Source code"for a work means the preferred form of the work for making modifications to it. For a library, complete source code means all the source code for all modules it contains, plus any associated interface definition files, plus the scripts used to control compilation and installation of the library.

Activities other than copying, distribution and modification are not covered by this License; they are outside its scope. The act of running a program using the Library is not restricted, and output from such a pro-

gram is covered only if its contents constitute a work based on the Library (independent of the use of the Library in a tool for writing it). Whether that is true depends on what the Library does and what the program that uses the Library does.

1. You may copy and distribute verbatim copies of the Library's complete source code as you receive it, in any medium, provided that you conspicuously and appropriately publish on each copy an appropriate copyright notice and disclaimer of warranty; keep intact all the notices that refer to this License and to the absence of any warranty; and distribute a copy of this License along with the Library.

You may charge a fee for the physical act of transferring a copy, and you may at your option offer warranty protection in exchange for a fee.

2. You may modify your copy or copies of the Library or any portion of it, thus forming a work based on the Library, and copy and distribute such modifications or work under the terms of Section 1 above, provided that you also meet all of these conditions:

a) The modified work must itself be a software library.

b) You must cause the files modified to carry prominent notices stating that you changed the files and the date of any change.

c) You must cause the whole of the work to be licensed at no charge to all third parties under the terms of this License.

d) If a facility in the modified Library refers to a function or a table of data to be supplied by an application program that uses the facility, other

than as an argument passed when the facility is invoked, then you must make a good faith effort to ensure that, in the event an application does not supply such function or table, the facility still operates, and performs whatever part of its purpose remains meaningful. (For example, a function in a library to compute square roots has a purpose that is entirely well-defined independent of the application. Therefore, Subsection 2d requires that any application-supplied function or table used by this function must be optional: if the application does not supply it, the square root function must still compute square roots.)

These requirements apply to the modified work as a whole. If identifiable sections of that work are not derived from the Library, and can be reasonably considered independent and separate works in themselves, then this License, and its terms, do not apply to those sections when you distribute them as separate works. But when you distribute the same sections as part of a whole which is a work based on the Library, the distribution of the whole must be on the terms of this License, whose permissions for other licensees extend to the entire whole, and thus to each and every part regardless of who wrote it.

Thus, it is not the intent of this section to claim rights or contest your rights to work written entirely by you; rather, the intent is to exercise the right to control the distribution of derivative or collective works based on the Library.

In addition, mere aggregation of another work not based on the Library with the Library (or with a work based on the Library) on a volume of a

storage or distribution medium does not bring the other work under the scope of this License.

3. You may opt to apply the terms of the ordinary GNU General Public License instead of this License to a given copy of the Library. To do this, you must alter all the notices that refer to this License, so that they refer to the ordinary GNU General Public License, version 2, instead of to this License. (If a newer version than version 2 of the ordinary GNU General Public License has appeared, then you can specify that version instead if you wish.) Do not make any other change in these notices.

Once this change is made in a given copy, it is irreversible for that copy, so the ordinary GNU General Public License applies to all subsequent copies and derivative works made from that copy.

This option is useful when you wish to copy part of the code of the Library into a program that is not a library.

4. You may copy and distribute the Library (or a portion or derivative of it, under Section 2) in object code or executable form under the terms of Sections 1 and 2 above provided that you accompany it with the complete corresponding machine-readable source code, which must be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange.

If distribution of object code is made by offering access to copy from a designated place, then offering equivalent access to copy the source code from the same place satisfies the requirement to distribute the source code, even though third parties are not compelled to copy the source

along with the object code.

5. A program that contains no derivative of any portion of the Library, but is designed to work with the Library by being compiled or linked with it, is called a "work that uses the Library". Such a work, in isolation, is not a derivative work of the Library, and therefore falls outside the scope of this License.

However, linking a "work that uses the Library"with the Library creates an executable that is a derivative of the Library (because it contains portions of the Library), rather than a "work that uses the library". The executable is therefore covered by this License. Section 6 states terms for distribution of such executables.

When a "work that uses the Library"uses material from a header file that is part of the Library, the object code for the work may be a derivative work of the Library even though the source code is not. Whether this is true is especially significant if the work can be linked without the Library, or if the work is itself a library. The threshold for this to be true is not precisely defined by law.

If such an object file uses only numerical parameters, data structure layouts and accessors, and small macros and small inline functions (ten lines or less in length), then the use of the object file is unrestricted, regardless of whether it is legally a derivative work. (Executables containing this object code plus portions of the Library will still fall under Section 6.)

Otherwise, if the work is a derivative of the Library, you may distribute

the object code for the work under the terms of Section 6. Any executables containing that work also fall under Section 6, whether or not they are linked directly with the Library itself.

6. As an exception to the Sections above, you may also combine or link a "work that uses the Library"with the Library to produce a work containing portions of the Library, and distribute that work under terms of your choice, provided that the terms permit modification of the work for the customer's own use and reverse engineering for debugging such modifications.

You must give prominent notice with each copy of the work that the Library is used in it and that the Library and its use are covered by this License. You must supply a copy of this License. If the work during execution displays copyright notices, you must include the copyright notice for the Library among them, as well as a reference directing the user to the copy of this License. Also, you must do one of these things:

a) Accompany the work with the complete corresponding machine-readable source code for the Library including whatever changes were used in the work (which must be distributed under Sections 1 and 2 above); and, if the work is an executable linked with the Library, with the complete machine-readable "work that uses the Library", as object code and/or source code, so that the user can modify the Library and then relink to produce a modified executable containing the modified Library. (It is understood that the user who changes the contents of definitions files in the Library will not necessarily be able to recompile the

application to use the modified definitions.)

b) Use a suitable shared library mechanism for linking with the Library. A suitable mechanism is one that (1) uses at run time a copy of the library already present on the user's computer system, rather than copying library functions into the executable, and (2) will operate properly with a modified version of the library, if the user installs one, as long as the modified version is interface-compatible with the version that the work was made with.

c) Accompany the work with a written offer, valid for at least three years, to give the same user the materials specified in Subsection 6a, above, for a charge no more than the cost of performing this distribution.

d) If distribution of the work is made by offering access to copy from a designated place, offer equivalent access to copy the above specified materials from the same place.

e) Verify that the user has already received a copy of these materials or that you have already sent this user a copy.

For an executable, the required form of the "work that uses the Library"must include any data and utility programs needed for reproducing the executable from it. However, as a special exception, the materials to be distributed need not include anything that is normally distributed (in either source or binary form) with the major components (compiler, kernel, and so on) of the operating system on which the executable runs, unless that component itself accompanies the executable.

It may happen that this requirement contradicts the license restrictions of

other proprietary libraries that do not normally accompany the operating system. Such a contradiction means you cannot use both them and the Library together in an executable that you distribute.

7. You may place library facilities that are a work based on the Library side-by-side in a single library together with other library facilities not covered by this License, and distribute such a combined library, provided that the separate distribution of the work based on the Library and of the other library facilities is otherwise permitted, and provided that you do these two things:

a) Accompany the combined library with a copy of the same work based on the Library, uncombined with any other library facilities. This must be distributed under the terms of the Sections above.

b) Give prominent notice with the combined library of the fact that part of it is a work based on the Library, and explaining where to find the accompanying uncombined form of the same work.

8. You may not copy, modify, sublicense, link with, or distribute the Library except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, link with, or distribute the Library is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

9. You are not required to accept this License, since you have not signed it. However, nothing else grants you permission to modify or distribute the

Library or its derivative works. These actions are prohibited by law if you do not accept this License. Therefore, by modifying or distributing the Library (or any work based on the Library), you indicate your acceptance of this License to do so, and all its terms and conditions for copying, distributing or modifying the Library or works based on it.

10. Each time you redistribute the Library (or any work based on the Library), the recipient automatically receives a license from the original licensor to copy, distribute, link with or modify the Library subject to these terms and conditions. You may not impose any further restrictions on the recipients' exercise of the rights granted herein. You are not responsible for enforcing compliance by third parties with this License.

11. If, as a consequence of a court judgment or allegation of patent infringement or for any other reason (not limited to patent issues), conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot distribute so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence you may not distribute the Library at all. For example, if a patent license would not permit royalty-free redistribution of the Library by all those who receive copies directly or indirectly through you, then the only way you could satisfy both it and this License would be to refrain entirely from distribution of the Library.

If any portion of this section is held invalid or unenforceable under any particular circumstance, the balance of the section is intended to apply,

and the section as a whole is intended to apply in other circumstances.

It is not the purpose of this section to induce you to infringe any patents or other property right claims or to contest validity of any such claims; this section has the sole purpose of protecting the integrity of the free software distribution system which is implemented by public license practices. Many people have made generous contributions to the wide range of software distributed through that system in reliance on consistent application of that system; it is up to the author/donor to decide if he or she is willing to distribute software through any other system and a licensee cannot impose that choice.

This section is intended to make thoroughly clear what is believed to be a consequence of the rest of this License.

12. If the distribution and/or use of the Library is restricted in certain countries either by patents or by copyrighted interfaces, the original copyright holder who places the Library under this License may add an explicit geographical distribution limitation excluding those countries, so that distribution is permitted only in or among countries not thus excluded. In such case, this License incorporates the limitation as if written in the body of this License.

13. The Free Software Foundation may publish revised and/or new versions of the Lesser General Public License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

Each version is given a distinguishing version number. If the Library spec-

ifies a version number of this License which applies to it and "any later version", you have the option of following the terms and conditions either of that version or of any later version published by the Free Software Foundation. If the Library does not specify a license version number, you may choose any version ever published by the Free Software Foundation.

14. If you wish to incorporate parts of the Library into other free programs whose distribution conditions are incompatible with these, write to the author to ask for permission. For software which is copyrighted by the Free Software Foundation, write to the Free Software Foundation; we sometimes make exceptions for this. Our decision will be guided by the two goals of preserving the free status of all derivatives of our free software and of promoting the sharing and reuse of software generally.

NO WARRANTY

15. BECAUSE THE LIBRARY IS LICENSED FREE OF CHARGE, THERE IS NO WARRANTY FOR THE LIBRARY, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE LIBRARY "AS IS"WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE LIBRARY IS WITH YOU. SHOULD THE LIBRARY PROVE DEFECTIVE, YOU ASSUME THE

COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

16. IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MAY MODIFY AND/OR REDISTRIBUTE THE LIBRARY AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE LIBRARY (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE LIBRARY TO OPERATE WITH ANY OTHER SOFTWARE), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

END OF TERMS AND CONDITIONS

- Данный продукт создан с использованием библиотеки 7-ZIP. Некоторые модули 7-ZIP включены в дистрибутив на основании пункта 6 GNU LESSER GENERAL PUBLIC LICENSE (т. е. в данном случае модули CLB Search, созданные автором CLB Search) являются «work that uses the Library»).

- Данный продукт создан с использованием библиотеки Xerces.

==========

== NOTICE file corresponding to section 4(d) of the Apache License, == == Version 2.0, in this case for the Apache Xerces distribution. ==

==========

This product includes software developed by The Apache Software Foundation (http://www.apache.org/). Portions of this software were originally based on the following: - software copyright (c) 1999, IBM Corporation., http://www.ibm.com.

Apache License Version 2.0, January 2004 http://www.apache.org/licenses/

TERMS AND CONDITIONS FOR USE, REPRODUCTION, AND DISTRIBUTION

1. Definitions.

"License"shall mean the terms and conditions for use, reproduction, and distribution as defined by Sections 1 through 9 of this document.

"Licensor"shall mean the copyright owner or entity authorized by the copyright owner that is granting the License.

"Legal Entity"shall mean the union of the acting entity and all other entities that control, are controlled by, or are under common control with that entity. For the purposes of this definition, "control"means (i) the power, direct or indirect, to cause the direction or management of such entity, whether by contract or otherwise, or (ii) ownership of fifty percent (50outstanding shares, or (iii) beneficial ownership of such entity.

"You"(or "Your") shall mean an individual or Legal Entity exercising permissions granted by this License.

"Source"form shall mean the preferred form for making modifications, including but not limited to software source code, documentation source, and configuration files.

"Object"form shall mean any form resulting from mechanical transformation or translation of a Source form, including but not limited to compiled object code, generated documentation, and conversions to other media types.

"Work"shall mean the work of authorship, whether in Source or Object form, made available under the License, as indicated by a copyright notice that is included in or attached to the work (an example is provided in the Appendix below).

"Derivative Works"shall mean any work, whether in Source or Object form, that is based on (or derived from) the Work and for which the editorial revisions, annotations, elaborations, or other modifications represent, as a whole, an original work of authorship. For the purposes of this License, Derivative Works shall not include works that remain separable from, or merely link (or bind by name) to the interfaces of, the Work and Derivative Works thereof.

"Contribution"shall mean any work of authorship, including the original version of the Work and any modifications or additions to that Work or Derivative Works thereof, that is intentionally submitted to Licensor for inclusion in the Work by the copyright owner or by an individual or Legal Entity authorized to submit on behalf of the copyright owner. For the purposes of this definition, "submitted"means any form of electronic,

verbal, or written communication sent to the Licensor or its representatives, including but not limited to communication on electronic mailing lists, source code control systems, and issue tracking systems that are managed by, or on behalf of, the Licensor for the purpose of discussing and improving the Work, but excluding communication that is conspicuously marked or otherwise designated in writing by the copyright owner as "Not a Contribution."

"Contributor"shall mean Licensor and any individual or Legal Entity on behalf of whom a Contribution has been received by Licensor and subsequently incorporated within the Work.

2. Grant of Copyright License. Subject to the terms and conditions of this License, each Contributor hereby grants to You a perpetual, worldwide, non-exclusive, no-charge, royalty-free, irrevocable copyright license to reproduce, prepare Derivative Works of, publicly display, publicly perform, sublicense, and distribute the Work and such Derivative Works in Source or Object form.

3. Grant of Patent License. Subject to the terms and conditions of this License, each Contributor hereby grants to You a perpetual, worldwide, non-exclusive, no-charge, royalty-free, irrevocable (except as stated in this section) patent license to make, have made, use, offer to sell, sell, import, and otherwise transfer the Work, where such license applies only to those patent claims licensable by such Contributor that are necessarily infringed by their Contribution(s) alone or by combination of their Contribution(s) with the Work to which such Contribution(s) was sub-

mitted. If You institute patent litigation against any entity (including a cross-claim or counterclaim in a lawsuit) alleging that the Work or a Contribution incorporated within the Work constitutes direct or contributory patent infringement, then any patent licenses granted to You under this License for that Work shall terminate as of the date such litigation is filed.

4. Redistribution. You may reproduce and distribute copies of the Work or Derivative Works thereof in any medium, with or without modifications, and in Source or Object form, provided that You meet the following conditions:

(a) You must give any other recipients of the Work or Derivative Works a copy of this License; and

(b) You must cause any modified files to carry prominent notices stating that You changed the files; and

(c) You must retain, in the Source form of any Derivative Works that You distribute, all copyright, patent, trademark, and attribution notices from the Source form of the Work, excluding those notices that do not pertain to any part of the Derivative Works; and

(d) If the Work includes a "NOTICE"text file as part of its distribution, then any Derivative Works that You distribute must include a readable copy of the attribution notices contained within such NOTICE file, excluding those notices that do not pertain to any part of the Derivative Works, in at least one of the following places: within a NOTICE text file distributed as part of the Derivative Works; within the Source form or

documentation, if provided along with the Derivative Works; or, within a display generated by the Derivative Works, if and wherever such third-party notices normally appear. The contents of the NOTICE file are for informational purposes only and do not modify the License. You may add Your own attribution notices within Derivative Works that You distribute, alongside or as an addendum to the NOTICE text from the Work, provided that such additional attribution notices cannot be construed as modifying the License.

You may add Your own copyright statement to Your modifications and may provide additional or different license terms and conditions for use, reproduction, or distribution of Your modifications, or for any such Derivative Works as a whole, provided Your use, reproduction, and distribution of the Work otherwise complies with the conditions stated in this License.

5. Submission of Contributions. Unless You explicitly state otherwise, any Contribution intentionally submitted for inclusion in the Work by You to the Licensor shall be under the terms and conditions of this License, without any additional terms or conditions. Notwithstanding the above, nothing herein shall supersede or modify the terms of any separate license agreement you may have executed with Licensor regarding such Contributions.

6. Trademarks. This License does not grant permission to use the trade names, trademarks, service marks, or product names of the Licensor, except as required for reasonable and customary use in describing the

origin of the Work and reproducing the content of the NOTICE file.

7. Disclaimer of Warranty. Unless required by applicable law or agreed to in writing, Licensor provides the Work (and each Contributor provides its Contributions) on an "AS IS"BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied, including, without limitation, any warranties or conditions of TITLE, NON-INFRINGEMENT, MERCHANTABILITY, or FITNESS FOR A PARTICULAR PURPOSE. You are solely responsible for determining the appropriateness of using or redistributing the Work and assume any risks associated with Your exercise of permissions under this License.

8. Limitation of Liability. In no event and under no legal theory, whether in tort (including negligence), contract, or otherwise, unless required by applicable law (such as deliberate and grossly negligent acts) or agreed to in writing, shall any Contributor be liable to You for damages, including any direct, indirect, special, incidental, or consequential damages of any character arising as a result of this License or out of the use or inability to use the Work (including but not limited to damages for loss of goodwill, work stoppage, computer failure or malfunction, or any and all other commercial damages or losses), even if such Contributor has been advised of the possibility of such damages.

9. Accepting Warranty or Additional Liability. While redistributing the Work or Derivative Works thereof, You may choose to offer, and charge a fee for, acceptance of support, warranty, indemnity, or other liability obligations and/or rights consistent with this License. However, in ac-

cepting such obligations, You may act only on Your own behalf and on Your sole responsibility, not on behalf of any other Contributor, and only if You agree to indemnify, defend, and hold each Contributor harmless for any liability incurred by, or claims asserted against, such Contributor by reason of your accepting any such warranty or additional liability.

END OF TERMS AND CONDITIONS

APPENDIX: How to apply the Apache License to your work.

To apply the Apache License to your work, attach the following boilerplate notice, with the fields enclosed by brackets "[]"replaced with your own identifying information. (Don't include the brackets!) The text should be enclosed in the appropriate comment syntax for the file format. We also recommend that a file or class name and description of purpose be included on the same "printed page"as the copyright notice for easier identification within third-party archives.

Copyright [yyyy] [name of copyright owner]

Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at

http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS"BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions

and limitations under the License.

- Морфологический словарь dic.ru включен с разрешения автора словаря Лукача Юрия Сауловича. Екатеринбург.

- http://www.oracle.com/technetwork/java/javase/terms/license/index.html

  Oracle Binary Code License Agreement for the Java SE Platform Products and JavaFX

  ORACLE AMERICA, INC. ("ORACLE"), FOR AND ON BEHALF OF ITSELF AND ITS SUBSIDIARIES AND AFFILIATES UNDER COMMON CONTROL, IS WILLING TO LICENSE THE SOFTWARE TO YOU ONLY UPON THE CONDITION THAT YOU ACCEPT ALL OF THE TERMS CONTAINED IN THIS BINARY CODE LICENSE AGREEMENT AND SUPPLEMENTAL LICENSE TERMS (COLLECTIVELY "AGREEMENT"). PLEASE READ THE AGREEMENT CAREFULLY. BY SELECTING THE "ACCEPT LICENSE AGREEMENT"(OR THE EQUIVALENT) BUTTON AND/OR BY USING THE SOFTWARE YOU ACKNOWLEDGE THAT YOU HAVE READ THE TERMS AND AGREE TO THEM. IF YOU ARE AGREEING TO THESE TERMS ON BEHALF OF A COMPANY OR OTHER LEGAL ENTITY, YOU REPRESENT THAT YOU HAVE THE LEGAL AUTHORITY TO BIND THE LEGAL ENTITY TO THESE TERMS. IF YOU DO NOT HAVE SUCH AUTHORITY, OR IF YOU DO NOT WISH TO BE BOUND BY THE TERMS, THEN SELECT THE "DECLINE LICENSE AGREEMENT"(OR THE EQUIVALENT) BUTTON AND YOU MUST NOT USE THE SOFTWARE ON THIS

SITE OR ANY OTHER MEDIA ON WHICH THE SOFTWARE IS CONTAINED.

1. DEFINITIONS. "Software"means the software identified above in binary form that you selected for download, install or use (in the version You selected for download, install or use) from Oracle or its authorized licensees, any other machine readable materials (including, but not limited to, libraries, source files, header files, and data files), any updates or error corrections provided by Oracle, and any user manuals, programming guides and other documentation provided to you by Oracle under this Agreement. "General Purpose Desktop Computers and Servers"means computers, including desktop and laptop computers, or servers, used for general computing functions under end user control (such as but not specifically limited to email, general purpose Internet browsing, and office suite productivity tools). The use of Software in systems and solutions that provide dedicated functionality (other than as mentioned above) or designed for use in embedded or function-specific software applications, for example but not limited to: Software embedded in or bundled with industrial control systems, wireless mobile telephones, wireless handheld devices, kiosks, TV/STB, Blu-ray Disc devices, telematics and network control switching equipment, printers and storage management systems, and other related systems are excluded from this definition and not licensed under this Agreement. "Programs"means (a) Java technology applets and applications intended to run on the Java Platform, Standard Edition platform on Java-enabled General Purpose

Desktop Computers and Servers; and (b) JavaFX technology applications intended to run on the JavaFX Runtime on JavaFX-enabled General Purpose Desktop Computers and Servers. "Commercial Features" means those features identified in Table 1-1 (Commercial Features In Java SE Product Editions) of the Java SE documentation accessible at http://www.oracle.com/technetwork/java/javase/documentation/index.html. "README File" means the README file for the Software accessible at http://www.oracle.com/technetwork/java/javase/documentation/index.html.

2. LICENSE TO USE. Subject to the terms and conditions of this Agreement including, but not limited to, the Java Technology Restrictions of the Supplemental License Terms, Oracle grants you a non-exclusive, non-transferable, limited license without license fees to reproduce and use internally the Software complete and unmodified for the sole purpose of running Programs. THE LICENSE SET FORTH IN THIS SECTION 2 DOES NOT EXTEND TO THE COMMERCIAL FEATURES. YOUR RIGHTS AND OBLIGATIONS RELATED TO THE COMMERCIAL FEATURES ARE AS SET FORTH IN THE SUPPLEMENTAL TERMS ALONG WITH ADDITIONAL LICENSES FOR DEVELOPERS AND PUBLISHERS.

3. RESTRICTIONS. Software is copyrighted. Title to Software and all associated intellectual property rights is retained by Oracle and/or its licensors. Unless enforcement is prohibited by applicable law, you may not modify, decompile, or reverse engineer Software. You acknowledge that the Software is developed for general use in a variety of information

management applications; it is not developed or intended for use in any inherently dangerous applications, including applications that may create a risk of personal injury. If you use the Software in dangerous applications, then you shall be responsible to take all appropriate fail-safe, backup, redundancy, and other measures to ensure its safe use. Oracle disclaims any express or implied warranty of fitness for such uses. No right, title or interest in or to any trademark, service mark, logo or trade name of Oracle or its licensors is granted under this Agreement. Additional restrictions for developers and/or publishers licenses are set forth in the Supplemental License Terms.

4. DISCLAIMER OF WARRANTY. THE SOFTWARE IS PROVIDED "AS IS"WITHOUT WARRANTY OF ANY KIND. ORACLE FURTHER DISCLAIMS ALL WARRANTIES, EXPRESS AND IMPLIED, INCLUDING WITHOUT LIMITATION, ANY IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NONINFRINGEMENT.

5. LIMITATION OF LIABILITY. IN NO EVENT SHALL ORACLE BE LIABLE FOR ANY INDIRECT, INCIDENTAL, SPECIAL, PUNITIVE OR CONSEQUENTIAL DAMAGES, OR DAMAGES FOR LOSS OF PROFITS, REVENUE, DATA OR DATA USE, INCURRED BY YOU OR ANY THIRD PARTY, WHETHER IN AN ACTION IN CONTRACT OR TORT, EVEN IF ORACLE HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. ORACLE'S ENTIRE LIABILITY FOR DAMAGES HEREUNDER SHALL IN NO EVENT EX-

CEED ONE THOUSAND DOLLARS (U.S. $1,000).

6. TERMINATION. This Agreement is effective until terminated. You may terminate this Agreement at any time by destroying all copies of Software. This Agreement will terminate immediately without notice from Oracle if you fail to comply with any provision of this Agreement. Either party may terminate this Agreement immediately should any Software become, or in either party's opinion be likely to become, the subject of a claim of infringement of any intellectual property right. Upon termination, you must destroy all copies of Software.

7. EXPORT REGULATIONS. You agree that U.S. export control laws and other applicable export and import laws govern your use of the Software, including technical data; additional information can be found on Oracle's Global Trade Compliance web site (http://www.oracle.com/products/export). You agree that neither the Software nor any direct product thereof will be exported, directly, or indirectly, in violation of these laws, or will be used for any purpose prohibited by these laws including, without limitation, nuclear, chemical, or biological weapons proliferation.

8. TRADEMARKS AND LOGOS. You acknowledge and agree as between you and Oracle that Oracle owns the ORACLE and JAVA trademarks and all ORACLE- and JAVA-related trademarks, service marks, logos and other brand designations ("Oracle Marks"), and you agree to comply with the Third Party Usage Guidelines for Oracle Trademarks currently located at http://www.oracle.com/us/legal/third-party-

trademarks/index.html . Any use you make of the Oracle Marks inures to Oracle's benefit.

9. U.S. GOVERNMENT LICENSE RIGHTS. If Software is being acquired by or on behalf of the U.S. Government or by a U.S. Government prime contractor or subcontractor (at any tier), then the Government's rights in Software and accompanying documentation shall be only those set forth in this Agreement.

10. GOVERNING LAW. This agreement is governed by the substantive and procedural laws of California. You and Oracle agree to submit to the exclusive jurisdiction of, and venue in, the courts of San Francisco, or Santa Clara counties in California in any dispute arising out of or relating to this agreement.

11. SEVERABILITY. If any provision of this Agreement is held to be unenforceable, this Agreement will remain in effect with the provision omitted, unless omission would frustrate the intent of the parties, in which case this Agreement will immediately terminate.

12. INTEGRATION. This Agreement is the entire agreement between you and Oracle relating to its subject matter. It supersedes all prior or contemporaneous oral or written communications, proposals, representations and warranties and prevails over any conflicting or additional terms of any quote, order, acknowledgment, or other communication between the parties relating to its subject matter during the term of this Agreement. No modification of this Agreement will be binding, unless in writing and signed by an authorized representative of each party.

SUPPLEMENTAL LICENSE TERMS

These Supplemental License Terms add to or modify the terms of the Binary Code License Agreement. Capitalized terms not defined in these Supplemental Terms shall have the same meanings ascribed to them in the Binary Code License Agreement. These Supplemental Terms shall supersede any inconsistent or conflicting terms in the Binary Code License Agreement, or in any license contained within the Software.

A. COMMERCIAL FEATURES. You may not use the Commercial Features for running Programs, Java applets or applications in your internal business operations or for any commercial or production purpose, or for any purpose other than as set forth in Sections B, C, D and E of these Supplemental Terms. If You want to use the Commercial Features for any purpose other than as permitted in this Agreement, You must obtain a separate license from Oracle.

B. SOFTWARE INTERNAL USE FOR DEVELOPMENT LICENSE GRANT. Subject to the terms and conditions of this Agreement and restrictions and exceptions set forth in the README File incorporated herein by reference, including, but not limited to the Java Technology Restrictions of these Supplemental Terms, Oracle grants you a non-exclusive, non-transferable, limited license without fees to reproduce internally and use internally the Software complete and unmodified for the purpose of designing, developing, and testing your Programs.

C. LICENSE TO DISTRIBUTE SOFTWARE. Subject to the terms and conditions of this Agreement and restrictions and exceptions set forth

in the README File, including, but not limited to the Java Technology Restrictions and Limitations on Redistribution of these Supplemental Terms, Oracle grants you a non-exclusive, non-transferable, limited license without fees to reproduce and distribute the Software, provided that (i) you distribute the Software complete and unmodified and only bundled as part of, and for the sole purpose of running, your Programs, (ii) the Programs add significant and primary functionality to the Software, (iii) you do not distribute additional software intended to replace any component(s) of the Software, (iv) you do not remove or alter any proprietary legends or notices contained in the Software, (v) you only distribute the Software subject to a license agreement that: (a) is a complete, unmodified reproduction of this Agreement; or (b) protects Oracle's interests consistent with the terms contained in this Agreement and that includes the notice set forth in Section H, and (vi) you agree to defend and indemnify Oracle and its licensors from and against any damages, costs, liabilities, settlement amounts and/or expenses (including attorneys' fees) incurred in connection with any claim, lawsuit or action by any third party that arises or results from the use or distribution of any and all Programs and/or Software. The license set forth in this Section C does not extend to the Software identified in Section G.

D. LICENSE TO DISTRIBUTE REDISTRIBUTABLES. Subject to the terms and conditions of this Agreement and restrictions and exceptions set forth in the README File, including but not limited to the Java Technology Restrictions and Limitations on Redistribution of these Sup-

plemental Terms, Oracle grants you a non-exclusive, non-transferable, limited license without fees to reproduce and distribute those files specifically identified as redistributable in the README File ("Redistributables") provided that: (i) you distribute the Redistributables complete and unmodified, and only bundled as part of Programs, (ii) the Programs add significant and primary functionality to the Redistributables, (iii) you do not distribute additional software intended to supersede any component(s) of the Redistributables (unless otherwise specified in the applicable README File), (iv) you do not remove or alter any proprietary legends or notices contained in or on the Redistributables, (v) you only distribute the Redistributables pursuant to a license agreement that: (a) is a complete, unmodified reproduction of this Agreement; or (b) protects Oracle's interests consistent with the terms contained in the Agreement and includes the notice set forth in Section H, (vi) you agree to defend and indemnify Oracle and its licensors from and against any damages, costs, liabilities, settlement amounts and/or expenses (including attorneys' fees) incurred in connection with any claim, lawsuit or action by any third party that arises or results from the use or distribution of any and all Programs and/or Software. The license set forth in this Section D does not extend to the Software identified in Section G.

E. DISTRIBUTION BY PUBLISHERS. This section pertains to your distribution of the JavaTM SE Development Kit Software ("JDK") with your printed book or magazine (as those terms are commonly used in the industry) relating to Java technology ("Publication"). Subject to

and conditioned upon your compliance with the restrictions and obligations contained in the Agreement, Oracle hereby grants to you a nonexclusive, nontransferable limited right to reproduce complete and unmodified copies of the JDK on electronic media (the "Media") for the sole purpose of inclusion and distribution with your Publication(s), subject to the following terms: (i) You may not distribute the JDK on a stand-alone basis; it must be distributed with your Publication(s); (ii) You are responsible for downloading the JDK from the applicable Oracle web site; (iii) You must refer to the JDK as JavaTM SE Development Kit; (iv) The JDK must be reproduced in its entirety and without any modification whatsoever (including with respect to all proprietary notices) and distributed with your Publication subject to a license agreement that is a complete, unmodified reproduction of this Agreement; (v) The Media label shall include the following information: “Copyright [YEAR], Oracle America, Inc. All rights reserved. Use is subject to license terms. ORACLE and JAVA trademarks and all ORACLE- and JAVA-related trademarks, service marks, logos and other brand designations are trademarks or registered trademarks of Oracle in the U.S. and other countries.” [YEAR] is the year of Oracle’s release of the Software; the year information can typically be found in the Software’s “About” box or screen. This information must be placed on the Media label in such a manner as to only apply to the JDK; (vi) You must clearly identify the JDK as Oracle’s product on the Media holder or Media label, and you may not state or imply that Oracle is responsible for any third-party software contained on the Media; (vii) You may not include any third party software on the

Media which is intended to be a replacement or substitute for the JDK; (viii) You agree to defend and indemnify Oracle and its licensors from and against any damages, costs, liabilities, settlement amounts and/or expenses (including attorneys' fees) incurred in connection with any claim, lawsuit or action by any third party that arises or results from the use or distribution of the JDK and/or the Publication; ; and (ix) You shall provide Oracle with a written notice for each Publication; such notice shall include the following information: (1) title of Publication, (2) author(s), (3) date of Publication, and (4) ISBN or ISSN numbers. Such notice shall be sent to Oracle America, Inc., 500 Oracle Parkway, Redwood Shores, California 94065 U.S.A , Attention: General Counsel.

F. JAVA TECHNOLOGY RESTRICTIONS. You may not create, modify, or change the behavior of, or authorize your licensees to create, modify, or change the behavior of, classes, interfaces, or subpackages that are in any way identified as "java", "javax", "sun", "oracle" or similar convention as specified by Oracle in any naming convention designation.

G. LIMITATIONS ON REDISTRIBUTION. You may not redistribute or otherwise transfer patches, bug fixes or updates made available by Oracle through Oracle Premier Support, including those made available under Oracle's Java SE Support program.

H. COMMERCIAL FEATURES NOTICE. For purpose of complying with Supplemental Term Section C.(v)(b) and D.(v)(b), your license agreement shall include the following notice, where the notice is displayed in a manner that anyone using the Software will see the notice:

Use of the Commercial Features for any commercial or production purpose requires a separate license from Oracle. "Commercial Features" means those features identified Table 1-1 (Commercial Features In Java SE Product Editions) of the Java SE documentation accessible at http://www.oracle.com/technetwork/java/javase/documentation/index.html

I. SOURCE CODE. Software may contain source code that, unless expressly licensed for other purposes, is provided solely for reference purposes pursuant to the terms of this Agreement. Source code may not be redistributed unless expressly provided for in this Agreement.

J. THIRD PARTY CODE. Additional copyright notices and license terms applicable to portions of the Software are set forth in the THIRDPARTYLICENSEREADME file accessible at http://www.oracle.com/technetwork/java/javase/documentation/index.html. In addition to any terms and conditions of any third party opensource/freeware license identified in the THIRDPARTYLICENSEREADME file, the disclaimer of warranty and limitation of liability provisions in paragraphs 4 and 5 of the Binary Code License Agreement shall apply to all Software in this distribution.

K. TERMINATION FOR INFRINGEMENT. Either party may terminate this Agreement immediately should any Software become, or in either party's opinion be likely to become, the subject of a claim of infringement of any intellectual property right.

L. INSTALLATION AND AUTO-UPDATE. The Software's installation and auto-update processes transmit a limited amount of data to Ora-

cle (or its service provider) about those specific processes to help Oracle understand and optimize them. Oracle does not associate the data with personally identifiable information. You can find more information about the data Oracle collects as a result of your Software download at http://www.oracle.com/technetwork/java/javase/documentation/index.html.

For inquiries please contact: Oracle America, Inc., 500 Oracle Parkway, Redwood Shores, California 94065, USA.

Last updated 02 April 2013