

Поисковая система  
*Avalanche 2.5*



# Руководство пользователя

# Содержание

<b>1</b>	<b>Введение</b>	<b>6</b>
<b>2</b>	<b>Общая информация</b>	<b>9</b>
2.1	Системные требования . . . . .	10
2.2	Установка и удаление . . . . .	10
2.3	Запуск программ Avalanche и Spider . . . . .	10
<b>3</b>	<b>Словарь терминов</b>	<b>11</b>
<b>4</b>	<b>Работа с Avalanche</b>	<b>12</b>
4.1	Интерфейс программы . . . . .	12
4.2	Главное меню . . . . .	12
4.3	Горячие клавиши . . . . .	16
4.4	Панели инструментов . . . . .	17
4.4.1	Панель управления . . . . .	17
4.4.2	Панель отображения . . . . .	17
4.5	Строка состояния . . . . .	18
4.6	Боковая панель . . . . .	19
4.6.1	Источники . . . . .	19
4.6.2	Рубрики . . . . .	19
4.6.3	Закладки . . . . .	19

4.7	Просмотр документов . . . . .	20
4.7.1	Паспорт информационного материала . . . . .	21
4.7.2	Характеристика источника . . . . .	22
4.8	Настройка рубрик . . . . .	23
4.8.1	Добавление и удаление рубрик . . . . .	23
4.8.2	Досье персоны и компании . . . . .	23
4.8.3	Карточка объекта . . . . .	24
4.8.4	Календарь Событий . . . . .	24
4.8.5	Импорт и Экспорт рубрик . . . . .	25
4.9	Язык рубрикации . . . . .	26
4.10	Настройки Avalanche . . . . .	28
4.10.1	Основные . . . . .	28
4.10.2	База данных . . . . .	28
4.10.3	Хранилище . . . . .	28
<b>5</b>	<b>Работа со Spider</b>	<b>29</b>
5.1	Интерфейс программы . . . . .	29
5.2	Главное меню . . . . .	29
5.3	Горячие клавиши . . . . .	33
5.4	Панели инструментов . . . . .	34
5.4.1	Панель управления . . . . .	34

5.4.2	Панель отображения . . . . .	34
5.5	Боковая панель . . . . .	35
5.6	Рабочий цикл программы . . . . .	35
5.7	Настройка источников . . . . .	36
5.7.1	Добавление и удаление источников . . . . .	36
5.7.2	Импорт и Экспорт источников . . . . .	36
5.7.3	Использование поисковых сервисов . . . . .	37
5.8	Регламент обхода . . . . .	40
5.9	Тонкая настройка . . . . .	42
5.9.1	Ручная настройка . . . . .	42
5.9.2	Помощник по тонкой настройке . . . . .	43
5.10	Язык робота . . . . .	44
5.10.1	Операторы робота . . . . .	44
5.10.2	Специальные символы . . . . .	45
5.10.3	Регулярные выражения . . . . .	45
5.10.4	Пример — Шаблон ссылки на новость . . . . .	46
5.10.5	Пример — Шаблон текста новости . . . . .	47
5.10.6	Пример — Регулярное выражение в ссылке . . . . .	48
5.11	Настройки Spider . . . . .	49
5.11.1	Основные . . . . .	49
5.11.2	Сеть . . . . .	49

5.11.3 База данных . . . . .	50
5.11.4 Хранилище . . . . .	50
5.11.5 Фильтр стоп-ссылок . . . . .	51
<b>6 Авторизация</b>	<b>52</b>
6.1 Добавить пользователя . . . . .	52
6.2 Сменить пользователя . . . . .	52
6.3 Изменить пароль . . . . .	52
6.4 Блокировка . . . . .	52
<b>7 Работа с несколькими базами</b>	<b>53</b>
7.1 Создание, удаление и настройка базы . . . . .	53
7.2 Сохранение и восстановление базы . . . . .	53
<b>8 Хранилище документов</b>	<b>54</b>
8.1 Перенос и установка базы . . . . .	55
<b>9 Структура индекса</b>	<b>56</b>
<b>10 База данных</b>	<b>57</b>
10.1 Структура базы данных . . . . .	57
10.2 Интеграция с Oracle . . . . .	58
10.3 Отчеты в Microsoft Office . . . . .	59
10.4 Отчеты в OpenOffice . . . . .	60

<b>11</b>	<b>Форматы файлов</b>	<b>61</b>
11.1	Источники . . . . .	61
11.2	Рубрики . . . . .	61
11.3	База . . . . .	61
<b>12</b>	<b>Дополнения</b>	<b>62</b>
12.1	Адресная книга . . . . .	62
12.2	Экспресс-аудит . . . . .	62
12.3	Cache Viewer . . . . .	62
12.4	DOM Inspector . . . . .	63
12.5	JavaScript Debugger . . . . .	63
12.6	SQLite Manager . . . . .	63
<b>13</b>	<b>Примеры использования</b>	<b>64</b>
13.1	Оружие службы безопасности . . . . .	64
13.2	Оружие маркетолога . . . . .	64
13.3	Оружие аналитика . . . . .	64
13.4	Оружие трейдера . . . . .	65
13.5	Оружие журналиста . . . . .	65
13.6	Оружие тендер-менеджера . . . . .	65
13.7	Оружие менеджера по персоналу . . . . .	65

# 1 Введение

## Как и зачем мы создали *Avalanche*

В настоящее время в Интернете насчитывается около полутора тысяч популярных поисковых систем (термином «популярный» мы определяем системы, которые хотя бы единожды похвалил кто-то, кроме их создателей. Вообще-то поисковиков значительно больше).

Поисковые системы отлично справляются с простыми однократными запросами. Однако если информационный поиск надо повторять постоянно, если предметная область сложна по структуре и если от результатов поиска зависит ваш доход или заработок — вы довольно быстро обнаружите, что:

- Поисковики перегружают вас тысячами бесполезных ссылок.
- Интернет не помнит, что вы уже видели, а что нет, и завтра принесет вам тот же миллион уже просмотренных ссылок.
- Поисковики не отличают действительно важную для вас информацию от шелухи.
- Поисковики не умеют правильно сортировать полученную информацию и раскладывать ее по нужным рубрикам.
- Поисковики не видят свежих тематических новостей. Задержка в индексировании конкретного сообщения может достигать до двух недель.
- Поисковики принципиально не видят некоторых сайтов (например, большинства сборников компромата). А пользователи Интернета, наоборот — видят. И увидят компромат на вас раньше вас, если вы решите положиться на любимую поисковую систему.
- Результаты работы любого поисковика можно купить или подделать (помните, как накануне 8 марта один популярный российский поисковик на запросы о цветах давал адрес только одного поставщика, пусть самого крупного, но все равно обидно).

- Поисковая система в Интернет выполняет поиск по вашему запросу, а значит, нагружает вас повторяющейся рутинной работой.
- А если на фирме работают несколько аналитиков, то однотипные запросы нескольких человек многократно увеличивают ваш трафик.

И список можно продолжить.

Люди долго мирились с такими неудобствами, пока финансовые аналитики (для которых и время — деньги, и результаты поиска — деньги) не сформировали спрос на более умную поисковую систему, которая бы решала хотя бы часть перечисленных выше проблем.

И вот в 98-м году появился пакет Enfish Tracker. Он чуть лучше формировал запросы, чуть удобнее хранил результаты и сам лазил в Интернет за обновлениями. За это «чуть» авторитетнейшая Investor's Business Daily объявила Enfish «Программой года».

Однако проблемы с поиском оставались. И в конце 98-го группа аналитиков Гарвардского университета сформулировала российским разработчикам постановку задачи на создание более совершенной поисковой системы. Вот так и появился пакет Avalanche (что означает — «лава»).

Что умеет пакет Avalanche в отличие от других систем поиска в Интернет?

Во-первых, вы формируете модель предметной области в виде набора «умных папок» (в американском патенте они называются Smart Folders). Каждая папка «знает», что именно должно в нее попадать и, естественно, способна проследить, чтобы не было дублирования.

Во-вторых, наполнением этих умных папок занимается специализированный поисковый робот, который запускается с вашего компьютера с вашими настройками. Его нельзя обмануть или подкупить - он принесет ровно то, что просили.

В-третьих, робот может запускаться и автоматически, принося и раскладывая по папкам свежие новости для вас аккуратно к вашему приходу на



работу. Есть и еще несколько маленьких приятных особенностей, благодаря которым Аваланч сегодня используют не только в аналитических или консалтинговых компаниях, но и в торговых фирмах.

Зачем? Например, один из самых крупных поставщиков супов в пакетах, чье имя вы слышите в каждой ТВ-рекламе, с помощью пакета Avalanche решает три основные задачи:

- Ведет мониторинг своей популярности, автоматически собирая все свежие упоминания о фирме в Интернет.
- Автоматически пополняет досье на основных конкурентов, мгновенно фиксируя появление любых новых материалов.
- Фильтрует результаты поиска других поисковых систем, устраняя ненужные ссылки (например, упоминания о своей фирме в прайс-листах многочисленных дилеров) — такая настройка тоже есть в Аваланче.

Конечно, не стоит ждать чуда — стопроцентная полнота и релевантность результатов поиска в Интернете в принципе недостижима. Avalanche — не более чем инструмент, легкий и гибкий, который избавляет аналитика от рутины, а результаты Интернет-поиска делает более точными и удобными для работы.

Попробуйте сами оценить его полезность - и вы поймете, почему пакет Avalanche установлен во Внешторгбанке и Суперкомпьютерном центре РАН, «Российской газете» и агентстве МИЭЛЬ, а также во многих других организациях. (В 2007 году было поставлено более 100 новых копий Avalanche PE).

## 2 Общая информация

Программа построена на движке **Mozilla XULRunner**. Ядро написано на **C++** и скомпилировано в **Microsoft Visual Studio 2008**. Клиентская часть программы написана на **JavaScript**, а диалоги пользователя спроектированы на **XUL**. Следующие технологии используются в программе:

- **XML** – Extensible Markup Language.
- **XUL** – XML User Interface Language.
- **XBL** – Extensible Binding Language.
- **RDF** – Resource Description Framework.
- **DOM** – Document Object Model.
- **AJAX** – Asynchronous JavaScript and XML.
- **XPCOM** – Cross-Platform Component Object Model.

По умолчанию движок XULRunner ставится вместе с программой.

В качестве сервера базы данных используется **MySQL Server**, системой управления базой данных является **MySQL Connector/ODBC**. Оба компонента входят в дистрибутив и устанавливаются по умолчанию.

Компанента **Java SE Runtime Environment** может понадобиться для работы некоторых дополнений.

## 2.1 Системные требования

Программа работает под управлением операционной системы **Microsoft Windows Vista**. Также доступны дистрибутивы программы для **MacOS (Intel)** и **Linux**.

Для установки программы на жестком диске требуется около 80 Мб. Кроме того необходимо выделить достаточно места для хранилища загружаемых документов.

Программа использует много поточную архитектуру и оптимально распределяет загрузки с нескольких серверов. Рекомендованная скорость канала связи от 3 Мбит/с.

## 2.2 Установка и удаление

Чтобы установить на компьютере поисковую систему **Avalanche**, необходимо запустить файл с названием **avalanche-mozilla-win32.exe**. Далее следуйте за мастером установки.

Для удаления программы выберите в меню **Пуск → Все программы → Avalanche 2.5 → Удаление**. Следуйте за мастером.

## 2.3 Запуск программ **Avalanche** и **Spider**

Для запуска поискового агента **Spider** выберите **Пуск → Все программы → Avalanche 2.5 → Spider**.

Для запуска менеджера документов **Avalanche** **Пуск → Все программы → Avalanche 2.5 → Avalanche**.

Кроме того на рабочем столе доступны иконки **Avalanche** и **Spider**.

### 3 Словарь терминов

- **Avalanche** – менеджер для работы с загруженными документами.
- **Spider** – поисковый агент (робот, паук) для загрузки документов из Internet.
- **База** – независимая база данных, имеющая собственные источники, рубрики и хранилище документов.
- **Документ** – новость или страница.
- **Закладка** – закладка на документ (новость или страницу).
- **Источник** – ссылка на Internet сайт.
- **Паспорт информационного материала** – параметры документа.
- **Регламент обхода** – настройка способа обхода источника.
- **Рубрика** – умная папка (Smart Folder).
- **Тонкая настройка** – список настроек по извлечению текста новости с сайта.
- **Фильтр документов** – отображение документов в менеджере Avalanche по заданному временному периоду.
- **Фильтр ссылок** – список стоп-ссылок настраиваемый в поисковом агенте Spider.
- **Характеристика источника** – атрибуты источника.
- **Язык робота** – шаблон извлечения ссылки или текста новости. Содержит фрагменты привязки к HTML коду с использованием зарезервированных операторов для извлечения ссылок, заголовков и текста.
- **Язык рубрикации** – шаблон поиска, по которому определяется попадает ли полученный документ в рубрику. Содержит слова, фразы и операторы поиска.

## 4 Работа с Avalanche

### 4.1 Интерфейс программы

В главном окне программы отображаются:

- Главное меню
- Панель управления
- Боковая панель (левое окно)
- Список документов (правое верхнее окно)
- Текст документа (правое нижнее окно)
- Панель отображения
- Строка состояния

### 4.2 Главное меню

Меню – Файл:

- **Создать базу** – создание новой базы.
- **Открыть базу** – открытие существующей базы.
- **Заккрыть** – закрыть программу.
- **Импорт...** – импортировать рубрики из файла.
- **Экспорт...** – экспортировать рубрики в файл.
- **Выход** – выход из программы.

### Меню – Правка:

- **Вырезать** – вырезать документы или текст.
- **Копировать** – копировать документы или текст.
- **Вставить** – вставить документы или текст.
- **Удалить** – удалить документы или текст.
- **Выделить всё** – выделить весь текст.
- **Найти** – поиск по странице документа.

### Меню – Вид:

- **Панели инструментов** – включение/отключение панелей управления и отображения.
- **Строка состояния** – включение/отключение строки состояния.
- **Боковая панель** – переключение на Источники, Рубрики или Закладки в боковой панели.
- **Представление документа** – переключение представления документа в виде HTML страницы, исходника HTML или обычного текста.
- **Кодировка** – изменение кодировки документа.
- **Блокировка** – блокировка пользователя.
- **Обновить** – обновить базу (после работы поискового агента Spider).

### Меню – Рубрикация:

- **Рубрицировать** – рубрицировать новые документы.
- **Перерубрицировать все** – перерубрикация всей базы (в случае изменения настроек рубрик).

- **Добавить рубрику** – добавление рубрики.
- **Удалить рубрику** – удаление рубрики.
- **Настроить рубрику** – настройка рубрики.

#### Меню – Закладки:

- **Добавить документ в закладки** – добавление текущего документа в закладки.

#### Меню – База:

- **Создать базу** – создание новой базы.
- **Удалить базу** – удаление базы.
- **Настроить базу** – настройка базы.
- **Сохранить базу** – сохранение базы в файл.
- **Восстановить базу** – восстановление базы из файла.  
*Внимание! Существующая база будет удалена.*

#### Меню – Инструменты:

- **Очистка** – очистка базы от нерубрицированных (или всех) документов с возможностью задания периода очистки. Внизу указывается количество новостей, страниц, а также суммарный размер предстоящей очистки.
- **Отчеты** – создание отчетов рубрицированных документов.
- **Дополнения** – просмотр и настройка дополнений.  
*Подробнее смотрите в разделе **Дополнения**.*
- **Консоль ошибок** – консоль ошибок JavaScript.  
*Используется для служебных нужд.*

- **Редактор настроек** – редактор настроек Avalanche.  
*Используется для служебных нужд.*
- **Авторизация** – управление пользователями. Добавление и смена пользователей, а также изменение пароля.
- **Настройки...** – настройки программы.  
*Подробнее смотрите в разделе **Настройки Avalanche**.*

#### Меню - Справка:

- **Совет дня** – советы по работе с программой.
- **О Avalanche** – информация о программе.



### 4.3 Горячие клавиши

- **Ctrl+N** – создание новой базы.
- **Ctrl+W** – закрыть программу.
- **Ctrl+X** – вырезать документы или текст.
- **Ctrl+C** – копировать документы или текст.
- **Ctrl+V** – вставить документы или текст.
- **Ctrl+A** – выделить весь текст.
- **Ctrl+F** – поиск по странице.
- **Ctrl+U** – переключиться на Источники в боковой панели.
- **Ctrl+R** – переключиться на Рубрики в боковой панели.
- **Ctrl+B** – переключиться на Закладки в боковой панели.
- **Ctrl+H** – переключиться на представление документа в виде HTML страницы.
- **Ctrl+K** – переключиться на представление документа в виде исходника HTML.
- **Ctrl+T** – переключиться на представление документа в виде обычного текста.
- **Ctrl+L** – блокировка.
- **Ctrl+D** – добавить документ в закладки.
- **Del** – удалить документ или текст.

## 4.4 Панели инструментов

### 4.4.1 Панель управления

- **Spider** – запустить поисковый агент Spider.
- **Обновить** – обновить базу (после работы поискового агента Spider).
- **Рубрицировать** – рубрицировать новые документы.
- **Перерубрицировать все** – перерубрикация всей базы (в случае изменения настроек рубрик).
- **Добавить рубрику** – добавление рубрики.
- **Удалить рубрику** – удаление рубрики.
- **Настроить рубрику** – настройка рубрики.
- **Фильтр документов** – показать **Все документы**, **За сегодня**, **За последние семь дней** или **За период**. В последнем случае необходимо установить дату начала и окончания периода.

### 4.4.2 Панель отображения

- **Источники** – переключиться на Источники в боковой панели.
- **Рубрики** – переключиться на Рубрики в боковой панели.
- **Закладки** – переключиться на Закладки в боковой панели.
- **HTML** – переключиться на представление документа в виде HTML страницы.
- **Коды** – переключиться на представление документа в виде исходника HTML.
- **Текст** – переключиться на представление документа в виде обычного текста.

## 4.5 Строка состояния

Используется для вывода статистики - количества новостей, страниц, а также суммарного размера документов. Для каждого отображения боковой панели (**Источники**, **Рубрики** или **Закладки**) вычисляется своя статистика. Кроме того при подсчете статистики учитывается стоящий временной фильтр (**Все документы**, **За сегодня**, **За последние семь дней** или **За период**).

## 4.6 Боковая панель

### 4.6.1 Источники

Список источников состоит из трех уровней вложенности. На первом уровне вложенности строятся названия источников с количеством страниц и новостей (в скобках). Далее для каждого источника строится список загруженных страниц в следующем формате: дата загрузки, количество загруженных новостей, заголовок страницы. И наконец для каждой страницы строится список, загруженных с этой страницы новостей в формате: дата загрузки, заголовок новости. Менеджер Avalanche отслеживает историю загруженных документов и повторяющиеся документы будут проигнорированы при построении списка. Кроме того список источников строится с учетом временного фильтра, определенного в панели управления (по умолчанию отображаются все документы).

### 4.6.2 Рубрики

Корневая рубрика предопределена системой и называется **Каталог**. Вы можете создавать дерево рубрик любой степени вложенности.

### 4.6.3 Закладки

Закладки образуют список из двух стоящих друг за другом блоков — новостей и страниц.

Для удаления закладки кликните правой кнопкой мыши на ней и во всплывающем контекстном меню выберите **Удалить закладку**.

## 4.7 Просмотр документов

Выбрать документ для просмотра можно в боковой панели, при выборе папки в боковой панели построится список документов этой папки в правом верхнем окне, где кликнув на соответствующий документ можно также просмотреть его содержимое. Каждый документ в списке документов содержит следующие поля:

- **Тип** – тип документа — страница или новость.
- **Дата** – дата загрузки документа.
- **Заголовок** – заголовок документа.
- **Источник** – источник загрузки документа.
- **Просмотрено** – зеленый маркер указывает на то, что документ ни разу не просматривался. Маркер можно изменить на противоположный, кликнув на нем левой кнопкой мыши.
- **Рубрицировано** – синяя иконка указывает на то, что документ находится в какой-либо рубрике.
- **Закладка** – звездочка указывает на то, что документ находится в закладках.

Список документов может быть отсортирован по каждой колонке в возрастающем и убывающем порядках. Для этого кликните левой кнопкой мыши на заголовке соответствующей колонки. Первый клик мыши отсортирует колонку в возрастающем порядке, второй - в убывающем, а третий отменит сортировку по этой колонке.

С помощью **Drag and Drop** можно изменить порядок отображения колонок. Для этого кликните на заголовке соответствующей колонки левой кнопкой мыши и удерживая кнопку в нажатом положении переведите курсор мыши в нужное место (серая линия между колонками укажет новое положение).

Крайняя правая иконка в заголовке списка документов служит для настройки отображения колонок (колонок можно отключить). Кликните эту иконку и в появившемся меню настройте отображение. Последний пункт этого меню — **Восстановить обычный порядок** используется для восстановления обычного порядка колонок (после перестановки Drag and Drop).

Текст документа отображается в правом нижнем окне. С помощью панели отображения документ представляется в виде HTML страницы или в виде обычного текста. Каждый документ имеет **Заголовок**, **Ссылку** и **Дату** загрузки. Оригинал документа можно открыть в браузере кликнув по ссылке левой кнопкой мыши.

#### 4.7.1 Паспорт информационного материала

- **Материал №** – уникальный номер документа.
- **Источник** – источник загрузки документа.
- **Заголовок** – заголовок документа.
- **Ссылка** – ссылка на документ в Internet.
- **Файл** – путь на документ в Хранилище.
  
- **Сессия загрузки** – каждой сессии загрузки назначается уникальный номер.
- **Идентификатор сайта** – уникальный номер источника в сессии загрузки.
- **Родительская страница** – номер родительской страницы.
- **Глубина скачивания** – число последовательных переходов по вложенным ссылкам до этой страницы.

- **Дата загрузки** – дата загрузки документа.
- **Дата публикации** – дата публикации документа, извлекается из оператора языка робота (**date**).
- **Язык страницы** – определяется по кодировке документа.
- **Статус загрузки** – статус загрузки документа.

#### 4.7.2 Характеристика источника

- **№** – уникальный номер источника в сессии загрузки (идентификатор сайта).
- **Источник** – источник загрузки документа.
- **Ссылка** – ссылка на источник в Internet.
- **Дата** – дата загрузки источника.
- **Страна** – определяется по доменному имени ссылки.
- **IP адрес** – определяется с помощью DNS-сервиса.
- **Информация о владельце...** – определяется с помощью WHOIS-сервиса.

## 4.8 Настройка рубрик

### 4.8.1 Добавление и удаление рубрик

Для добавления рубрики выберите родительскую рубрику и нажмите кнопку **Добавить рубрику** на панели управления (можно также воспользоваться меню **Рубрикация** → **Добавить рубрику**) и введите название рубрики.

*Внимание! Название рубрики в данной ветке рубрик должно быть уникальным.*

Введенная рубрика появится в списке на боковой панели. Теперь нажмите на кнопку **Настроить рубрику** и задайте шаблон поиска (см. следующий раздел **Язык рубрикации**).

*Внимание! Если шаблон поиска не корректен, вы получите сообщение о синтаксической ошибке с указанием места ошибки (на месте ошибки стоит точка).*

Для удаления рубрики выберите ее из списка на боковой панели и нажмите на кнопку **Удалить рубрику**. Все вложенные подрубрики также будут удалены.

### 4.8.2 Досье персоны и компании

Для прикрепления досье выберите рубрику и нажмите на кнопку **Настроить рубрику** и далее нажмите на соответствующую кнопку:

- **Прикрепить персональное досье** - для ведения досье персоны в этой рубрике.
- **Прикрепить досье компании** - для ведения досье компании в этой рубрике.

В результате этой операции иконка соответствующей рубрики изменится, а на месте кнопок прикрепления появятся кнопки **Карточка объекта** и **Календарь Событий**.



### 4.8.3 Карточка объекта

Выберите рубрику и нажмите на кнопку **Настроить рубрику** и далее нажмите на кнопку **Карточка объекта**.

*Внимание! Если данная кнопка отсутствует, то необходимо сначала прикрепить досье персоны или компании.*

1. **Персональная карточка** — Содержит основные данные физического лица (персоны), такие как — фамилия, имя, отчество, год и место рождения, паспортные данные, адрес регистрации и т. д. Кроме того можно прикрепить и фотографию персоны, кликнув на картинку в правом верхнем углу.
2. **Карточка компании** — Содержит основные данные юридического лица (компании), такие как — регистрация, юридический адрес, банковский счет и т. д. Кроме того можно прикрепить логотип компании, кликнув на картинку в правом верхнем углу.

### 4.8.4 Календарь Событий

Выберите рубрику и нажмите на кнопку **Настроить рубрику** и далее нажмите на кнопку **Календарь Событий**.

*Внимание! Если данная кнопка отсутствует, то необходимо сначала прикрепить досье персоны или компании.*

- **Добавить событие** — для добавления нового события.
- **Удалить событие** — для удаления, выделенного мышкой, события.
- **Редактировать событие** — для редактирования, выделенного мышкой, события.

#### 4.8.5 Импорт и Экспорт рубрик

Для импорта дерева рубрик выберите в меню **Файл** → **Импорт...** и в появившемся диалоге откройте ваш файл (расширение **.avr**). В случае если название рубрики импортируемое из файла совпадает с уже существующей рубрикой, появится диалог, где необходимо будет изменить название импортируемой рубрики на уникальное.

Для экспорта дерева рубрик выберите в меню **Файл** → **Экспорт...** и в появившемся диалоге введите имя экспортируемого файла.

## 4.9 Язык рубрикации

Шаблон поиска содержит слова, цитаты и зарезервированные операторы.

1. Слова могут обрываться символом **\***. В этом случае окончание слова может быть любым. Для слова **Путин\*** найдутся все документы, содержащие: **Путин**, **Путина**, **Путину** и т. д.
2. Цитаты указываются в двойных кавычках. В этом случае поиск слов в тексте идет в порядке слов цитаты. Слова в цитате можно также задавать с помощью символа **\***. Для цитаты **"Един\* Рос-си\*"** найдутся действительно все упоминания этой партии.
3. Если все буквы слова маленькие, поиск слова осуществляется без учета регистра (в любом написании). Если в слове используется хотя бы одна большая буква, поиск слова происходит с учетом регистра (как есть). Для **кпрф** будут найдены документы с любым написанием, а для **КПРФ** только в заглавном написании.
4. Операторы шаблона поиска:
  - **OR** – логическое ИЛИ.
  - **AND** – логическое И.
  - **NOT** – логическое НЕ.

С помощью скобок операторы можно группировать в логические группы.

(Медведев\* OR Путин\*) AND NOT кризис\*

Здесь будут найдены все документы с упоминанием Медведева или Путина, не посвященные кризису.

5. Слова через пробел интерпретируются так, как если бы между ними, стоял оператор **AND**.

Обам\* Клинтон

Обам\* AND Клинтон

Эти два написания аналогичны.

6. Поиск слов и цитат, стоящих на ограниченном расстоянии друг от друга, осуществляется с помощью конструкции:

`[(словоформа OR словоформа) AND (словоформа OR словоформа)] N`

В этом случае шаблону поиска будет удовлетворять любое сочетание словоформ из левой и правой части, стоящих друг от друга не далее чем на **N** слов. Например, поиск всевозможных упоминаний Центрального банка Российской Федерации осуществляется с помощью следующего шаблона:

`[(ЦБ OR "центральный* банк*" OR Банк*) AND  
(РФ OR "Российск* Федераци*" OR России)] 1`

## 4.10 Настройки Avalanche

### 4.10.1 Основные

- **Показывать неизмененные документы** – включите эту опцию и менеджер Avalanche покажет все неизмененные с прошлой загрузки документы.

### 4.10.2 База данных

Источник данных (ODBC):

- **Источник данных** – название источника данных.
- **Имя пользователя** – имя пользователя.
- **Пароль** – пароль.

Вы можете настроить собственный источник данных ODBC и даже подключить собственный сервер базы данных (поддерживающий ODBC).

**Панель управления → Администрирование → Источники данных (ODBC) → Системный DSN**

*Внимание! В директории установки **Avalanche 2.5** в папке **database** вы найдете SQL скрипт **avalanche.sql** создающий базу данных программы. Кроме того в этой директории находятся два командных файла **create.bat** и **drop.bat** для создания и удаления базы данных. Теперь вы сможете настроить базу данных программы самостоятельно.*

### 4.10.3 Хранилище

Путь в Хранилище документов. Подробнее смотрите в разделе **Хранилище документов**.

## 5 Работа со Spider

### 5.1 Интерфейс программы

В главном окне программы отображаются:

- Главное меню
- Панель управления
- Боковая панель (левое окно)
- Данные загрузки (правое окно)
- Панель отображения
- Строка состояния

### 5.2 Главное меню

Меню – Файл:

- **Создать базу** – создание новой базы.
- **Открыть базу** – открытие существующей базы.
- **Заккрыть** – закрыть программу.
- **Импорт...** – импортировать источники из файла.
- **Экспорт...** – экспортировать источники в файл.
- **Выход** – выход из программы.

### Меню – Правка:

- **Вырезать** – вырезать текст.
- **Копировать** – копировать текст.
- **Вставить** – вставить текст.
- **Удалить** – удалить текст.
- **Выделить всё** – выделить весь текст.

### Меню – Вид:

- **Панели инструментов** – включение/отключение панелей управления и отображения.
- **Строка состояния** – включение/отключение строки состояния.
- **Боковая панель** – включение/отключение боковой панели (списка Источников).
- **Данные** – отображение общей или текущей статистики, также сообщений об ошибках.
- **Блокировка** – блокировка пользователя.

### Меню – Управление:

- **Старт** – запустить загрузку источников.
- **Стоп** – остановить загрузку источников.
- **Активировать таймер** – запустить периодическую загрузку источников. Период обхода задается в регламенте обхода каждого источника.
- **Завершить работу таймера** – остановить периодическую загрузку источников.

- **Добавить источник** – добавление источника.
- **Удалить источник** – удаление источника.
- **Настроить источник** – настройка источника.

#### Меню – База:

- **Создать базу** – создание новой базы.
- **Удалить базу** – удаление базы.
- **Настроить базу** – настройка базы.
- **Сохранить базу** – сохранение базы в файл.
- **Восстановить базу** – восстановление базы из файла.  
*Внимание! Существующая база будет удалена.*

#### Меню – Инструменты:

- **Дополнения** – просмотр и настройка дополнений.  
*Подробнее смотрите в разделе **Дополнения**.*
- **Консоль ошибок** – консоль ошибок JavaScript.  
*Используется для служебных нужд.*
- **Редактор настроек** – редактор настроек Avalanche.  
*Используется для служебных нужд.*
- **Авторизация** – управление пользователями. Добавление и смена пользователей, а также изменение пароля.
- **Настройки...** – настройки программы.  
*Подробнее смотрите в разделе **Настройки Spider**.*



**Меню – Справка:**

- **Совет дня** – советы по работе с программой.
- **О Spider** – информация о программе.

### 5.3 Горячие клавиши

- **Ctrl+N** – создание новой базы.
- **Ctrl+W** – закрыть программу.
- **Ctrl+X** – вырезать текст.
- **Ctrl+C** – копировать текст.
- **Ctrl+V** – вставить текст.
- **Ctrl+A** – выделить весь текст.
- **Ctrl+L** – блокировка.
- **Del** – удалить текст.

## 5.4 Панели инструментов

### 5.4.1 Панель управления

- **Старт** – запустить загрузку источников.
- **Стоп** – остановить загрузку источников.
- **Активировать таймер** – запустить периодическую загрузку источников. Период обхода задается в регламенте обхода каждого источника.
- **Завершить работу таймера** – остановить периодическую загрузку источников.
- **Добавить источник** – добавление источника.
- **Удалить источник** – удаление источника.
- **Настроить источник** – настройка источника.

### 5.4.2 Панель отображения

- **Общая статистика** – отображение общей статистики, а также списка загружаемых источников.
- **Текущая статистика** – отображение текущей статистики по каждому источнику. Для просмотра текущей статистики выберите необходимый источник в боковой панели.
- **Сообщения об ошибках** – сообщения об ошибках в режиме периодической загрузки источников выводятся в этом окне.

## 5.5 Боковая панель

Боковая панель служит для формирования списка источников загрузки. Слева от каждого источника находится иконка активации (зеленая стрелка или красная точка). Если источник включен в список обхода, то используется зеленая стрелка. Красная точка указывает на то, что источник временно исключен из списка. Для изменения состояния источника кликните на нем дважды левой кнопкой мыши.

## 5.6 Рабочий цикл программы

Поисковый агент работает в двух режимах — ручном и автоматическом. Для ручного запуска нажмите кнопку **Старт** на панели управления (можно также воспользоваться меню **Управление** → **Старт**).

*Внимание! Если статистика не начнет меняться в течении десяти секунд, проверьте соединение с Internet на вашем компьютере (например, запустите программу Internet Explorer). И если Internet работает, а статистика по-прежнему не меняется, то проверьте настройки прокси сервера на вашем компьютере. Подробнее смотрите пункт **Сеть** в разделе **Настройки Spider**.*

Для запуска поискового агента в автоматическом режиме нажмите кнопку **Активировать таймер** на панели управления (либо в меню **Управление** → **Активировать таймер**). В этом режиме поисковый агент будет запускаться автоматически, период обхода задается для каждого источника в регламенте обхода. По умолчанию период обхода устанавливается в одни сутки, это значит что каждый день вы будете автоматически получать последнюю информацию из Internet.

В конце загрузки поисковый агент автоматически проведет рубрикацию только, что загруженных документов.

## 5.7 Настройка источников

### 5.7.1 Добавление и удаление источников

Для добавления источника нажмите кнопку **Добавить источник** на панели управления (можно также воспользоваться меню **Управление** → **Добавить источник**) и введите название источника.

*Внимание! Название источника должно быть уникальным.*

Введенный источник появится в списке на боковой панели. Теперь нажмите на кнопку **Настроить источник** и введите Internet ссылку.

*Внимание! Ссылка должна содержать префикс протокола **http://** (в противном случае вы получите сообщение об ошибке **Неправильная ссылка**).*

Для добавления источника в группу источников нажмите правой кнопкой мыши на корневом источнике группы и в появившемся контекстном меню выберите **Добавить источник в группу**. Корневой источник задает регламент обхода всей группе, также как и определяет временное исключение группы из списка обхода.

Для удаления источника выберите его из списка на боковой панели и нажмите на кнопку **Удалить источник**.

### 5.7.2 Импорт и Экспорт источников

Для импорта списка источников выберите в меню **Файл** → **Импорт...** и в появившемся диалоге откройте ваш файл (расширение **.avs**). В случае если название источника импортируемое из файла совпадает с уже существующим источником, появится диалог, где необходимо будет изменить название импортируемого источника на уникальное.

Для экспорта списка источников выберите в меню **Файл** → **Экспорт...** и в появившемся диалоге введите имя экспортируемого файла.

### 5.7.3 Использование поисковых сервисов

В настройке источников можно использовать, как встроенный поисковой сервис **Google** и **Yandex**, так и Ваш собственный. Для этого установите флажок "Использовать поисковый сервис" в настройках источника.

При использовании встроенного поискового сервиса поле «Ссылка» заполнится автоматически, и Вам останется ввести только Ваш запрос в поле «Поиск». Для поисковых сервисов **Google** и **Yandex** тонкая настройка уже сделана, она позволяет извлекать страницы из выдачи по вашему запросу. По необходимости вы можете изменить, как ссылку заданных поисковых сервисов, так и их тонкую настройку.

В ссылке необходимо указать оператор (**query**), на место этого оператора будет подставлен поисковый запрос из поля «Поиск». Изменение ссылки может понадобиться если потребуется изменить место поиска (например на поиск в Новостях Google) или задать дополнительные параметры поиска (например – фильтр поиска по времени).

Получить необходимую ссылку для настройки в Avalanche можно, осуществив необходимый поиск со всеми требуемыми опциями и после выдачи результата скопировать полученную ссылку в настройку источника Avalanche. Данная ссылка будет содержать ваш поисковый запрос и вам потребуется заменить его на оператор (**query**). Сам же поисковый запрос введите в поле «Поиск». Таким образом ссылка уже не будет требовать изменения и вы сможете удобно экспериментировать с вашим поисковым запросом.

Для передачи параметров через метод POST, используйте соответствующую настройку. Раскройте дополнительные параметры (нажатием на кнопку с изображением галочки) и укажите значение метода POST в появившейся группе дополнительных настроек. Строку POST можно получить с помощью плагина httpFox для Mozilla Firefox.

По аналогии с методом POST можно установить строку Cookie. Ее значение также можно взять из плагина httpFox.

Для правильного обращения к поисковому сервису установите кодиров-

ку, на которой он работает. Например **Google** и **Yandex** работают на кодировке **UTF-8**. Кодировку страницы вашего поискового сервиса можно узнать из меню браузера **Mozilla Firefox** или **Internet Explorer**.

Кроме стандартных поисковых сервисов Google и Yandex вы можете настроить и собственный (по умолчанию это первая иконка с увеличительным стеклом). Настройте «Ссылку» и «Поиск», а также в тонкой настройке определите шаблон по извлечению заголовков и ссылок из выдачи вашего поискового сервиса. Подробнее смотрите в разделе **Тонкая настройка**. Это динамичекий способ настройки вашего собственного поискового сервиса.

Часто используемый поисковый сервис, такой как Google или Yandex вы можете статически прописать в файле с настройками поисковых сервисов **chrome/rdf/spider/search.rdf** (в директории установки **Avalanche 2.5**). Это файл в формате RDF (Resource Description Framework). Например, для добавления нового поискового сервиса **Новотека** понадобится добавить следующий блок в этот файл:

```
...
<RDF:li>
  <RDF:Description NC:name="Novoteka"
    NC:logo="resource://spider/novoteka.png"
    NC:url="http://www.novoteka.ru/search?query=(query)"
    NC:post=""
    NC:cookie=""
    NC:charset=""
    NC:start_string=""
    NC:end_string=""
    NC:page_pattern="<h2 class=p01><a class=news_name_out t
arget=_blank href=(url)>(title)</a></h2>"
    NC:news_pattern=""
    NC:auto_detect="1"/>
</RDF:li>
...
```

где,

- **NC:name** – название поискового сервиса.
- **NC:logo** – создайте и укажите логотип вашего поискового сервиса (высота иконки должна быть 32 пикселя). Файл с иконкой логотипа должен быть расположен в папке **chrome/rdf/spider/novoteka.png** и иметь соответствующее имя **"resource://spider/novoteka.png"**.
- **NC:url** – ссылка поискового запроса с использованием оператора (**query**) на месте подстановки поискового запроса.
- **NC:post** – параметры метода POST. Здесь, также допустимо использование оператора (**query**) для подстановки поискового запроса.
- **NC:cookie** – строка Cookie.
- **NC:charset** – кодировка страницы поискового сервиса.
- **NC:start\_string** – начинать обрабатывать документ с этой строки.
- **NC:end\_string** – не обрабатывать документ после этой строки.
- **NC:page\_pattern** – шаблон по извлечению заголовка и ссылки.
- **NC:news\_pattern** – шаблон по извлечению текста документа (как правило список выдачи поискового сервиса имеет ссылки из разных источников и такого шаблона не существует).
- **NC:auto\_detect** – автоматическое извлечение текста новости, в этом случае текст документа извлекается автоматически и необходимости задания параметра **NC:news\_pattern** нет.

Такой статический способ задания поискового сервиса имеет преимущество перед динамическим, тем что, настроив единожды, он появится в списке предопределенных поисковых сервисов и вам не потребуется каждый раз задавать его «Ссылку» и тонкую настройку. Все это, будет автоматически заполняться из файла **search.rdf** в момент клика по соответствующему логотипу.



## 5.8 Регламент обхода

Выберите источник на боковой панели и нажмите на кнопку **Настроить источник**, в появившемся диалоге нажмите на кнопку **Регламент обхода...**

- **Глубина сбора** – определяет максимальное число последовательных переходов по вложенным ссылкам.
- **Приоритет** – загрузки источников образуют очередь потоков. Приоритет загрузки источника влияет, как на очередность исполнения потоков, так и на их приоритет при параллельном исполнении. Потоки с приоритетом **9** выполняются быстрее всего, далее по уменьшению приоритета выполняются все остальные потоки.
- **Период обхода** – интервал времени между автоматическими загрузками источника и используется при обходе источника с использованием таймера.
- **Временно исключить из списка** – временно исключить источник из списка обхода.
- **Не переходить на другие сайты** – обходить источник в пределах доменного имени.
- **Загрузить сайт полностью** – загрузить все найденные страницы в пределах доменного имени.
- **Не загружать изображения** – изображения не будут загружаться, тем самым обеспечится наибольшая скорость загрузки.
- **Собирать новости со страницы** – включение/отключение загрузки новостей.
- **Извлекать текст со страницы** – текст новости будет извлекаться непосредственно с главной страницы, не загружая страницы новостей.

- **Тонкая настройка...** – тонкая настройка на новости (см. следующий раздел **Тонкая настройка**).

## 5.9 Тонкая настройка

### 5.9.1 Ручная настройка

Выберите источник на боковой панели и нажмите на кнопку **Настроить источник**, в появившемся диалоге нажмите на кнопку **Регламент обхода...** и далее, поставив галочку напротив опции **Ручная настройка**, на кнопку **Тонкая настройка...**

- **Обрабатывать документ начиная со строки** – строка на HTML странице с которой необходимо начать искать новости. Если это поле не задано, то робот ищет новости с начала страницы.
- **Не обрабатывать документ после строки** – строка на HTML странице на которой следует остановиться искать новости. Если это поле не задано, то робот ищет новости до конца страницы.
- **Шаблон ссылки на новость** – содержит фрагменты привязки к HTML коду с использованием зарезервированных операторов для извлечения ссылок, заголовка и даты (см. следующий раздел **Язык робота**).
- **Автоматическое определение текста новости** – в случае если робот не смог автоматически извлечь текст новости, необходимо отключить эту опцию и в ручную задать шаблон текста новости в появившемся окне.
- **Шаблон текста новости** – содержит фрагменты привязки к HTML коду с использованием зарезервированных операторов для извлечения текста (см. следующий раздел **Язык робота**).

При нажатии на кнопку **Просмотр** откроется окно, в котором будет подсвечен текст в исходном коде страницы, удовлетворяющий указанным шаблонам. Для удаления настройки выберите ее из списка и нажмите на кнопку **Удалить настройку**.

### 5.9.2 Помощник по тонкой настройке

Вызывается аналогично, но при отключенной опции **Ручная настройка**. Настройка производится в 4 последовательных этапа:

- **1** — Показывается главная страница с новостями. Пользователю необходимо кликнуть мышкой по двум, желательно рядом стоящим, ссылкам.
- **2** — Автоматически ищутся остальные ссылки, и подсвечиваются их предполагаемые заголовки. Если заголовки определены неправильно, нажав на кнопку **Указать самостоятельно** можно выделить один заголовок, и после нажатия кнопки **Вперед** программа привяжет его к ссылке, после чего снова подсветятся все заголовки. Если результат устраивает пользователя, нажав на кнопку **Вперед** можно перейти к следующему шагу.
- **3** — Открывается страница с текстом новости. Автоматически подсвечивается предполагаемый полезный текст. Аналогично с предыдущим случаем, пользователь может сам указать текст, выделив его. При этом старайтесь выделять так, чтоб конструктивные элементы начала и конца текста (абзацы, блоки, отдельные надписи и т. п.) присутствовали на всех подобных страницах с новостями. На этом шаге можно закончить настройку, нажав кнопку **ОК**, или перейти к определению даты, нажав на кнопку **Вперед**.
- **4** — Определение даты, все аналогично определению текста. Кнопка **ОК** — закончить настройку.

На всех шагах кнопкой **Отмена** отменяется настройка и закрывается окно, при этом никаких изменений в способе обработки документа не произойдет. По кнопке **Назад** можно вернуться к предыдущему шагу, отменив настройки на текущем.

## 5.10 Язык робота

### 5.10.1 Операторы робота

- **(url)** – извлечение ссылки из-под этого оператора.
- **(title)** – извлечение заголовка из-под этого оператора.
- **(date)** – извлечение даты из-под этого оператора.
- **(text)** – извлечение текста из-под этого оператора.
- **(...)** – из-под этого оператора ничего не извлекается (служит для пропуска фрагмента).

**Точка привязки** новостного блока — одинаковый для каждой новости фрагмент HTML кода. Для разных от новости к новости фрагментов HTML кода будем использовать операторы языка робота. Встретив такой оператор робот будет пропускать фрагмент HTML кода до следующей точки привязки. Очевидным становятся два правила написания шаблона поиска:

1. Началом и концом шаблона поиска являются точки привязки к HTML коду.
2. Операторы робота не могут идти вместе. Между ними обязательно должна стоять точка привязки к HTML коду.

Итак необходимо найти первую точку привязки новости в новостном блоке. Ссылка на новость для каждой новости будет уникальна и эта ссылка, как раз и нужна роботу. Ставим на месте ссылки оператор **(url)** и робот будет знать, что HTML код до следующей точки привязки будет ссылкой на новость.

Кроме ссылки необходимо задать роботу заголовок новости, для этого используем оператор **(title)**. И весь текст до следующей точки привязки будет извлекаться роботом, как заголовок новости.

### 5.10.2 Специальные символы

- `\n` – перевод строки.
- `\t` – символ табуляции.

Для задания точки привязки, состоящей из нескольких строк используется символ перевода строки. Аналогично задается символ табуляции.

### 5.10.3 Регулярные выражения

Регулярные выражения могут использоваться во всех операторах робота:

- `(url|match|replace)` – для преобразования ссылки.
- `(title|match|replace)` – для преобразования заголовка.
- `(date|match|replace)` – для преобразования даты.
- `(text|match|replace)` – для преобразования текста.
- `(...|match|)` – для проверки фрагмента на соответствие **match**.

Регулярные выражения служат для преобразования кода, извлеченного из-под оператора. С помощью директивы **match** задается шаблон регулярного выражения, которому должен соответствовать фрагмент. А директива **replace** формирует результат, который и определяет данный оператор.

#### 5.10.4 Пример — Шаблон ссылки на новость

```
...  
<a href="/press/news/2008/11/1/2.html"><noindex>Альфа-Банк и InOut  
объединяют свои дисконтные программы</noindex></a>  
...  
<a href="/press/news/2008/11/1/1.html"><noindex>Йоханн Йонах занял  
пост Председателя Совета директоров <nobr>Альфа-Банка</nobr>  
</noindex></a>  
...  
<a href="/press/news/2008/10/23/1.html"><noindex><nobr>Альфа-Банк  
</nobr> и Цезарь-Сателлит объявляют о начале совместной акции «Ваш  
автомобиль под охраной»</noindex></a>  
...
```

Выделим точки привязки — в начале, в середине и в конце каждой из этих трех новостей, а на месте ссылки и заголовка поставим соответствующие операторы. Таким образом шаблон ссылки на новость будет:

```
<a href="(url)"><noindex>(title)</noindex></a>
```

### 5.10.5 Пример — Шаблон текста новости

```
...
<table cellpadding="0" cellspacing="0"><tbody><tr><td>
<p>
<!--medialand_ru_context_start-->
<p align=justify><font color="blue">18.04.2009, Порт-оф-Спейн</font>
Президент США Барак Обама и президент Венесуэлы Уго Чавес пожали друг другу
руки на открытии саммита стран Америки в республике Тринидад и Тобаго.
"Я приветствовал президента Джорджа Буша этой самой рукой восемь лет назад.
Я хочу быть вашим другом", - сказал У.Чавес президенту Б.Обаме, передает
Reuters со ссылкой на заявление секретариата венесуэльского лидера.</p>
<!--medialand_ru_context_end-->
</p>
</td></tr></tbody></table>
...
```

Можно легко заметить HTML фрагменты, которые будут встречаться от новости к новости. Искомый шаблон текста новости будет:

```
<!--medialand_ru_context_start-->(text)<!--medialand_ru_context_end-->
```



### 5.10.6 Пример — Регулярное выражение в ссылке

В некоторых сложных случаях на странице отсутствуют непосредственные ссылки на новостные страницы. Вместо прямых ссылок используется директива **javascript**:

```
<a href="javascript:click('D9307A1C-9237-264A-A7FB-D09B39ECBA00');">
```

Страница, которая загружается по этому клику имеет адрес:

```
/cgi-bin/news.pl?D9307A1C-9237-264A-A7FB-D09B39ECBA00
```

С помощью специального оператора (**url|match|replace**) можно осуществить данное преобразование, где:

**match** — является регулярным выражением ссылки в HTML коде,

**replace** — полученная ссылка новостной страницы.

В нашем примере оператором, выполняющим необходимое преобразование будет:

```
<a href="(url|javascript:click\('(.*)'\);|/cgi-bin/news.pl?$1);">
```

где \$1 соответствует **(.\*)** - первому аргументу регулярного выражения.

## 5.11 Настройки Spider

### 5.11.1 Основные

- **Таймаут соединения** – таймаут ожидания загрузки страницы. Если страница не загружается по истечению этого времени считается, что страница не доступна.
- **Авторубрикация** – включение/отключения режима автоматической рубрикации в конце загрузки.

### 5.11.2 Сеть

- **Использовать прокси сервер** – включите эту опцию если вы выходите в Internet через прокси сервер.
- **Адрес сервера и порт** – введите адрес прокси сервера и порт. Эти параметры можно узнать у вашего системного администратора или посмотреть в настройках вашего Internet браузера.

### 5.11.3 База данных

Источник данных (ODBC):

- **Источник данных** – название источника данных.
- **Имя пользователя** – имя пользователя.
- **Пароль** – пароль.

Вы можете настроить собственный источник данных ODBC и даже подключить собственный сервер базы данных (поддерживающий ODBC).

**Панель управления → Администрирование → Источники данных (ODBC) → Системный DSN**

*Внимание! В директории установки **Avalanche 2.5** в папке **database** вы найдете SQL скрипт **avalanche.sql** создающий базу данных программы. Кроме того в этой директории находятся два командных файла **create.bat** и **drop.bat** для создания и удаления базы данных. Теперь вы можете настроить базу данных программы самостоятельно.*

### 5.11.4 Хранилище

Путь в Хранилище документов. Подробнее смотрите в разделе **Хранилище документов**.

### 5.11.5 Фильтр стоп-ссылок

Фильтр ссылок содержит список ссылок, которые исключаются поисковым роботом из списка загрузки. Список стоп-ссылок задается в текстовом файле RDF (Resource Description Framework). Файл находится в директории установки **Avalanche 2.5** и далее по пути **chrome/rdf/spider/filter.rdf**

В нижеследующем примере заданы три стоп ссылки:

```
<?xml version="1.0"?>
<RDF:RDF xmlns:RDF="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
          xmlns:NC="http://home.netscape.com/NC-rdf#">
  <RDF:Seq RDF:about="urn:root">
    <RDF:li><RDF:Description NC:filter="spylog.ru"/></RDF:li>
    <RDF:li><RDF:Description NC:filter="top100.rambler.ru"/></RDF:li>
    <RDF:li><RDF:Description NC:filter="top.list.ru"/></RDF:li>
  </RDF:Seq>
</RDF:RDF>
```

## 6 Авторизация

### 6.1 Добавить пользователя

В программе Avalanche можно ввести учетную запись, ограничив круг лиц, которые будут пользоваться Avalanche на данном компьютере. В процессе регистрации, Вы должны будете ввести ваш логин и пароль.

По умолчанию создается единственный пользователь с именем **Аналитик** и без пароля.

### 6.2 Сменить пользователя

Вы можете сменить пользователя не выходя из программы, для этого выберите соответствующий пункт в меню Avalanche, введите логин и пароль.

### 6.3 Изменить пароль

Если по каким-либо причинам, вас не устраивает пароль, то вы можете изменить его. Для этого Вам нужно будет ввести свой логин, старый и новый пароли.

### 6.4 Блокировка

Для блокировки доступа посторонних лиц к Вашим данным в программе существует блокировка.

При нажатии клавиши **Ctrl+L** появится окошко, в котором для продолжения дальнейшей работы необходимо будет ввести ваш пароль, который Вы указали при регистрации. В случае ввода неверных данных программа закроется.

## 7 Работа с несколькими базами

### 7.1 Создание, удаление и настройка базы

Для создания новой базы выберите в меню **База → Создать базу** и введите название базы.

*Внимание! Название базы должно быть уникальным.*

По названию базы создается папка для сохранения загруженных документов в Хранилище. Подробнее смотрите в разделе **Хранилище документов**.

Для удаления базы выберите в меню **База → Удалить базу**. Открытая в данный момент база будет удалена.

Изменить название базы можно выбрав в меню **База → Настроить базу**.

### 7.2 Сохранение и восстановление базы

Для сохранения базы выберите в меню **База → Сохранить базу** и в появившемся диалоге введите имя сохраняемой базы (расширение **.avb**).

Для восстановления базы выберите в меню **База → Восстановить базу** и в появившемся диалоге откройте файл с вашей базой.

*Внимание! Существующая база будет удалена.*

## 8 Хранилище документов

Документы загружаемые из Internet сохраняются в Хранилище в двух форматах — в виде HTML файлов, а также в виде RSS ленты новостей. По умолчанию Хранилище создается в папке текущего пользователя Windows с именем **.avalanche**. Корневая папка может настраиваться в меню **Инструменты** → **Настройки...** → **Хранилище**.

- Для каждого пользователя в Хранилище создается своя папка, куда помещаются папки с базами этого пользователя.
- В базе для каждого Источника загрузки создается папка по имени этого источника.
- В папке источника для каждой сессии загрузки создается папка, название которой формируется из даты и времени загрузки.
- В папку сессии помещаются все загруженные документы. Стартовая страница источника записывается с именем **index.html**. Новости и страницы последовательно нумеруются и записываются с именами **news<номер>.html** и **page<номер>.html** соответственно. Также в папке **files** сохраняются все используемые в HTML ресурсы (таким образом менеджер Avalanche уже не требует подключения к Internet для корректного отображения HTML документов).

Каждый загруженный документ снабжается файлом индекса (файл с расширением **.xml**). Индекс используется при рубрикации документов. Индекс строится в алфавитном порядке, предоставляя возможность двойного поиска по нему ((см. следующий раздел **Структура индекса**)).

Помимо HTML все документы сохраняются в виде обычного текста. Файлы именуются аналогичным способом, расширение этих файлов **.txt**.

Новости каждого источника сохраняются в виде RSS ленты новостей в файле **news.rss**.

Информация о владельце (WHOIS-сервис) сохраняется в файле **whois.txt**.

## 8.1 Перенос и установка базы

Для переноса базы необходимо:

1. Сохранить базу, используя меню **База** → **Сохранить базу** на внешний носитель.
2. Скопировать папку базы в Хранилище документов на внешний носитель.

Для установки базы необходимо:

1. Восстановить базу, используя меню **База** → **Восстановить базу** с внешнего носителя.
2. Скопировать папку базы с внешнего носителя по указанному в настройках пути:  
**Инструменты** → **Настройки...** → **Хранилище**



## 9 Структура индекса

Каждая загруженная из Internet страница или новость снабжается файлом индекса. Файл индекса это XML файл, содержащий все найденные в документе слова, отсортированные в алфавитном порядке.

Если в слове используется хотя бы одна большая буква, то кроме оригинала в индекс попадет это-же слово в нижнем регистре.

Слова в исходном документе последовательно нумеруются от начала к концу (начиная с единицы). В результате для каждого слова строится список позиций, где это слово встретилось.

Такая структура позволяет быстро осуществлять поиск слов в документе.

```
<?xml version="1.0" encoding="utf-8"?>
<dictionary>
  ...
  <word value="реальная" pos="1192"/>
  <word value="регулярно" pos="743"/>
  <word value="резервы" pos="273"/>
  <word value="результат" pos="417"/>
  <word value="результатами" pos="108"/>
  <word value="результате" pos="318,771,1063,1143"/>
  <word value="результативность" pos="1024"/>
  <word value="рекомендаций" pos="934"/>
  <word value="реорганизация" pos="999"/>
  <word value="ресурсам" pos="289"/>
  <word value="ресурсы" pos="262,272,800"/>
  <word value="решение" pos="877,1042"/>
  <word value="решения" pos="1012"/>
  <word value="решить" pos="424"/>
  ...
</dictionary>
```

## 10 База данных

### 10.1 Структура базы данных

Таблицы базы данных Avalanche:

- **av\_user** – список пользователей.
- **av\_base** – список баз.
- **av\_source** – список источников.
- **av\_rubric** – дерево рубрик.
- **av\_session** – сессии загрузки.
- **av\_site** – загруженные сайты.
- **av\_page** – загруженные страницы.
- **av\_rubricated** – список рубрицированных документов.
- **av\_bookmark** – список закладок на документы.
- **av\_person** – персональное досье.
- **av\_company** – досье компании.
- **av\_calendary** – календарь событий.
- **av\_image** – изображение в досье.

## 10.2 Интеграция с Oracle

1. Создайте базу данных Oracle с помощью SQL скрипта **oracle.sql**, находящегося в папке с установленным **Avalanche 2.5**, и далее в папке **database**.
2. Пропишите в источнике данных ODBC созданную базу данных Oracle.
3. Установите флажок **Force SQL\_WCHAR Support** в **Oracle ODBC Driver Configuration** на закладке **Workarounds**.
4. В настройках базы данных Avalanche (**Инструменты** → **Настройки...** → **База данных**) укажите настроенные источник данных Oracle ODBC, а также логин и пароль пользователя Oracle в котором вы создали базу данных Avalanche.
5. В редакторе настроек Avalanche (**Инструменты** → **Редактор настроек**) измените настройку **avalanche.database.server** на строку-значение **oracle**.
6. Перезапустите Avalanche.

## 10.3 Отчеты в Microsoft Office

Стыковка базы данных **Avalanche** и **Microsoft Office Excel**:

1. Запускаем **Excel**, выбираем пункт меню **Данные**, раздел **Получить внешние данные** и кнопку **Из других источников**.
2. Выбираем **Из мастера подключения данных** и указываем тип источника данных **ODBC DSN**.
3. Выбираем источник данных **avalanche**.
4. Выбираем нужную нам таблицу (пусть это будет таблица **av\_page**).
5. Щелкаем на следующем окне по кнопке **Готово**, а на форме **Импорт данных** нажимаем на **ОК**.
6. Теперь у нас на листе отображаются данные из этой таблицы. Причем они связаны постоянной связью (т.е. поддерживается подключение к этой таблице). И нажав на кнопку **Обновить все** можно получить самую свежую версию данных из таблицы.

*Внимание! Источник данных **avalanche** использует логин и пароль. Смотрите в директории **Avalanche 2.5** папку **defaults** и далее в папке **preferences** файл **prefs.js***

## 10.4 Отчеты в OpenOffice

Стыковка базы данных **Avalanche** и **OpenOffice Calc**:

1. Запускаем **OpenOffice Base** и в **Мастере баз данных** выбираем пункт **Подключиться к существующей базе данных**, далее пункт **ODBC** и нажимаем на кнопку **Далее**.
2. На этом шаге **Установка подключения к ODBC в Имени ODBC источника данных в вашей системе** выбираем **avalanche** и нажимаем **Далее**.
3. После ввода **Имени пользователя** и установки флажка **Требуется пароль** нажимаем **Далее**.
4. На шаге **Сохранить и выполнить** вводим имя файла **avalanche.odt** и нажимаем **Сохранить**.
5. Вводим пароль и нажимаем на **ОК**.
6. Теперь запускаем **OpenOffice Calc** и в меню **Вид** выбираем **Источники данных**.
7. В проводнике выбираем **avalanche**, далее **Таблицы** и после ввода пароля выбираем таблицу (пусть это будет таблица **av\_page**).
8. Теперь у нас на листе отображаются данные из этой таблицы. Причем они связаны постоянной связью (т.е. поддерживается подключение к этой таблице). И нажав на кнопку **Обновить** можно получить самую свежую версию данных из таблицы.

*Внимание! Источник данных **avalanche** использует логин и пароль. Смотрите в директории **Avalanche 2.5** папку **defaults** и далее в папке **preferences** файл **prefs.js***

## 11 Форматы файлов

### 11.1 Источники

Файл с расширением **.avs** является zip архивом списка источников — **av\_source.xml**.

### 11.2 Рубрики

Файл с расширением **.avr** является zip архивом дерева рубрик — **av\_rubric.xml**.

### 11.3 База

Файл с расширением **.avb** является zip архивом всей базы.

## 12 Дополнения

Дополнения устанавливаются в программу «на лету» (без необходимости что-либо перекомпилировать или переустанавливать). Для управления дополнениями выберите в меню **Инструменты** → **Дополнения**.

Кнопка **Установить...** используется для установки новых дополнений.

Закладка **Расширения** служит для настройки установленных дополнений, которые можно временно отключить или удалить. Отключение дополнения временно удалит дополнение из меню **Инструменты**.

### 12.1 Адресная книга

Адресная книга является дополнением менеджера Avalanche и запускается из меню **Инструменты** → **Адресная книга**. Адресная книга позволяет находить адреса и телефоны людей, живущих в Москве, Санкт-Петербурге и других городах России.

### 12.2 Экспресс-аудит

Экспресс-аудит является дополнением менеджера Avalanche и запускается из меню **Инструменты** → **Экспресс-аудит**. Экспресс-аудит предназначен для обнаружения уязвимостей источника.

### 12.3 Cache Viewer

Показывает содержимое кэша в памяти и на диске. Позволяет вытащить из кэша любую информацию.

## **12.4 DOM Inspector**

Предназначен для проверки и редактирования DOM-дерева какого-либо веб-документа или XUL-приложения.

## **12.5 JavaScript Debugger**

Средство отладки Java Script.

## **12.6 SQLite Manager**

Управление базой данных SQLite.



## **13 Примеры использования**

Поисковая система Avalanche является мощным оружием в руках специалистов из самых разных областей. Далее перечисляются некоторые области и характерные портреты успешного использования системы.

### **13.1 Оружие службы безопасности**

1. Решает задачу кадровой безопасности, путем сбора досье на служащих компании.
2. Осуществляет мониторинг безопасности в отношении конкурентов, партнеров и смежных организаций.
3. Ведет конкурентную разведку в Internet.

### **13.2 Оружие маркетолога**

1. Позволяет проводить маркетинговые исследования.
2. Собирает информацию о деятельности конкурентов и поведении потребителей.
3. Осуществляет мониторинг имиджа компании.

### **13.3 Оружие аналитика**

1. Настраивается на необходимые источники информации.
2. Автоматически собирает информацию для анализа.
3. Автоматически извлекает и структурирует полезную информацию.

### **13.4 Оружие трейдера**

1. Собирает новости о ситуации на рынке и в мире.
2. Систематизирует данные для фундаментального и технического анализа.
3. Оценивает настроение рынка, тенденции и прогнозы.

### **13.5 Оружие журналиста**

1. Настраивается на необходимые источники информации.
2. Автоматически собирает информацию для статей.
3. Фильтрует и рубрицирует искомую информацию.

### **13.6 Оружие тендер-менеджера**

1. Настраивается на информационные порталы государственных, муниципальных и коммерческих организаций.
2. Позволяет автоматически загружать списки тендеров по проводимым аукционам и конкурсам.
3. Отбирает и распределяет тендеры по заданным рубрикам.

### **13.7 Оружие менеджера по персоналу**

1. Настраивается на необходимые кадровые агентства, специализированные порталы и форумы.
2. Автоматически собирает картотеку возможных кандидатур.
3. Извлекает и рубрицирует необходимые компании кандидатуры.