

ООО «ЛАН-ПРОЕКТ»

Руководство по эксплуатации «ЛАН.Студия управления источниками»

2021

1. Общие сведения

Программный компонент «ЛАН.Студия управления источниками» предназначена для настройки информационных источников в сети Интернет. Каждый источник представляет собой группы настроек, которые регламентируют выполнение различных процессов.

Набор модулей формируется индивидуально в зависимости от назначения программного компонента и может отличаться от примеров, указанных в данном руководстве. Далее приведено описание отдельных модулей, которые могут быть включены в программный компонент «ЛАН.Студия управления источниками».

Графический интерфейс программного компонента «ЛАН.Студия управления источниками» представлен на рисунке 1 и состоит из вкладок:

- «Выделение текста» – вкладка, предназначенная для просмотра и добавления главных страниц, выбора режима и перехода к правилам сбора источников;

- «Общие параметры» – вкладка, предназначенная для просмотра и изменения настроек процессов;

- «Описание» – в данной вкладке расположены прописываемые вручную параметры, описывающие выбранный источник;

- «XML-представление» – отображение источника в виде XML конфигурации;

- «Результаты тестирования» – результаты работы правил сбора после тестирования источника;

- «Рубрики» – вкладка, предназначенная для классификации источников по тематическим классификаторам;

- «Привязки» – связывание правил и главных страниц, для оптимизации процессов сбора с источника.

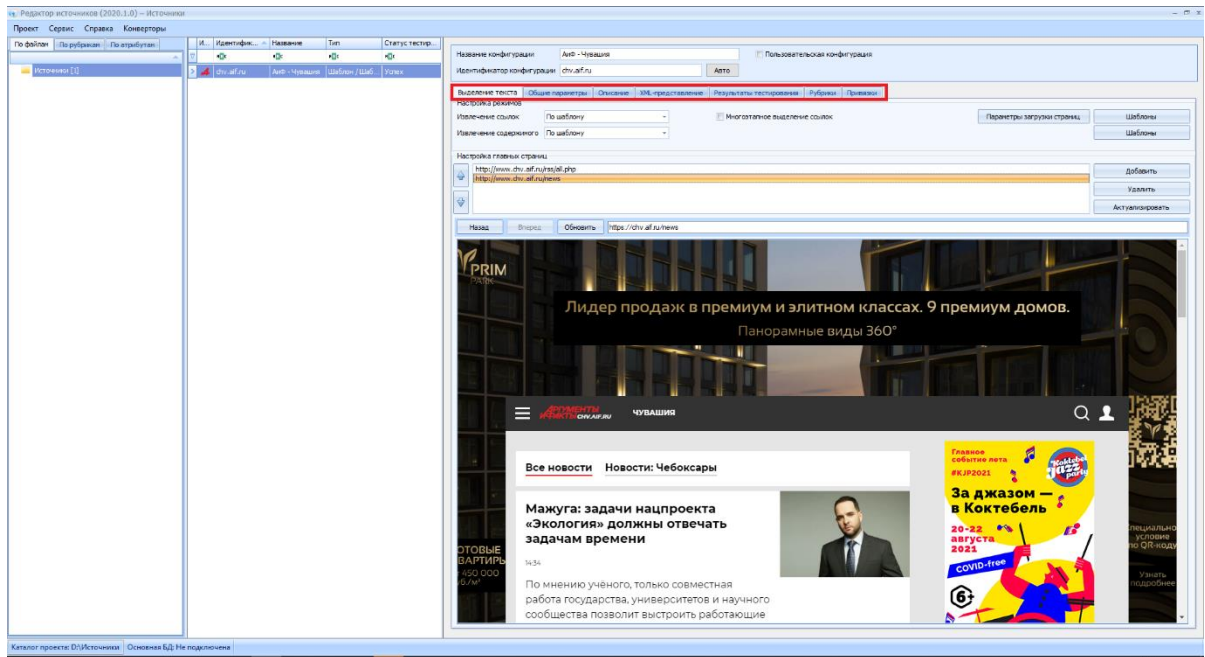


Рисунок 1 – Графический интерфейс программного компонента
«ЛАН.Студия управления источниками»

2. Вкладка «Выделение текста»

«Выделение текста» представлена на рисунке 2.

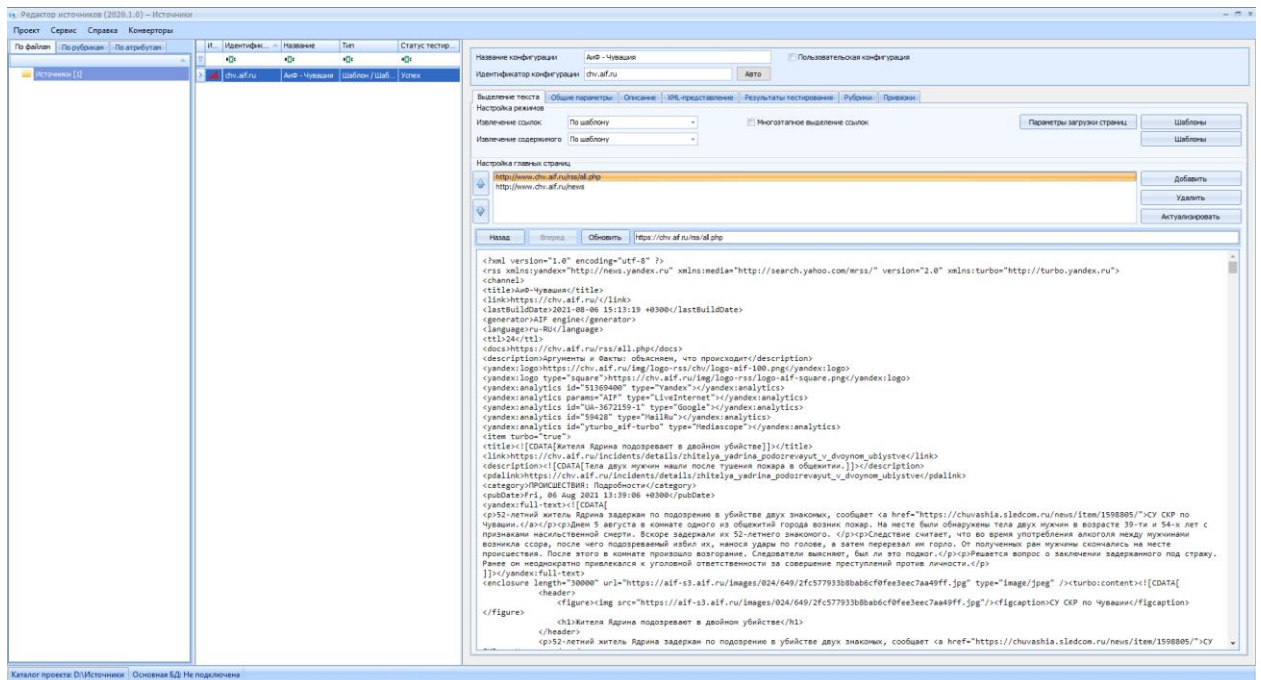


Рисунок 2 – Вкладка «Выделение текста»

По умолчанию, для источников задается режим обхода «Авто». В этом случае программный компонент загрузки документов будет пытаться

самостоятельно выделить ссылки на новости, расположенные на странице, указанной при создании, затем переходить по ним и пытаться выделить там текст новости, дату публикации и заголовок. Данный способ не всегда дает удовлетворительные результаты. Это обусловлено фактическим отсутствием каких-либо единых стандартов верстки, применяемых при создании страниц различных сайтов. Для более точной обработки источников рекомендуется использовать режим «По шаблону», где необходимо самостоятельно создать правила обхода для выделения ссылок, заголовков, дат, текстов и авторов сообщений. Переключение режимов производится отдельно для страницы со ссылками на новости и для страниц с самими новостями.

Для первоначального создания нового правила обхода источника необходимо переключить режим обхода с «Авто» на «По шаблону» и нажать на кнопку «Шаблоны», после чего откроется окно «Редактора шаблонов», а главное окно программы заблокируется до закрытия данного окна.

Если требуется получить информацию с сайта, структура страниц которого представляет «дерево», необходимо установить метку «Многоэтапное выделение ссылок». Данный параметр изменяет режим обработки шаблонов извлечения ссылок таким образом, чтобы каждый шаблон выделял ссылки для своего уровня вложенности: первый шаблон – для корня сайта, второй шаблон – для второго уровня сайта, третий – для третьего и так далее до конца списка шаблонов.

Как правило, многоэтапный режим выделения ссылок необходим для обработки сайтов-форумов.

На вкладке «Выделение текста» также имеется возможность редактировать список главных страниц источника, подлежащих обходу. Для добавления новой страницы необходимо зайти на нее во встроенном браузере и нажать на кнопку «Добавить» напротив списка главных страниц. Следует отметить, что если просто вставить желаемую ссылку в адресную строку браузера и сразу нажать на данную кнопку, то добавится не желаемая ссылка, а ссылка на ту страницу, которая в данный момент открыта в

браузере. Поэтому после копирования ссылки необходимо сначала нажать на клавишу «Enter» и только затем добавлять желаемую ссылку в список.

Кнопка «Удалить» удаляет ту страницу, которая выделена курсором в списке уже добавленных главных страниц.

Кнопка «Актуализировать» выполняет поочередный переход на каждую из главных страниц и, в случае, если страница недоступна или перемещена, отображает соответствующее информационное окно с возможностью удалить или изменить главную страницу.

Кнопка «Параметры загрузки страниц» выполняет функцию перехода к окну настроек загрузки страницы таких как: правило ожидания элемента, настройки скриптов, настройки скриптов авторизации, настройки обхода капчи.

3. Вкладка «Общие параметры»

На данной вкладке, представленной на рисунке Рисунок 3 – Вкладка «Общие параметры», расположены настраиваемые параметры, используемые при обходе сайта программой «ЛАН.Интернет-мониторинг».

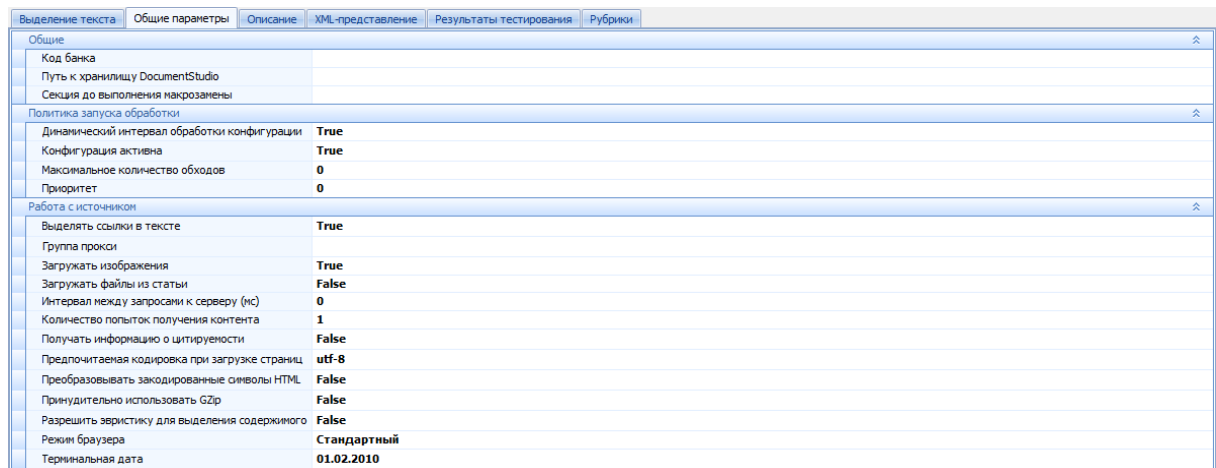


Рисунок 3 – Вкладка «Общие параметры»

4. Вкладка «Описание»

На данной вкладке, представленной на рисунке 4, расположены прописываемые вручную параметры, описывающие выбранный источник.

При нажатии на кнопку «Получить изображение» будет выполнена попытка получения текущего логотипа сайта.

При нажатии на кнопку «Получить описание источника» будет произведена «выкачка» информации о нем со специализированного сайта «Whois».

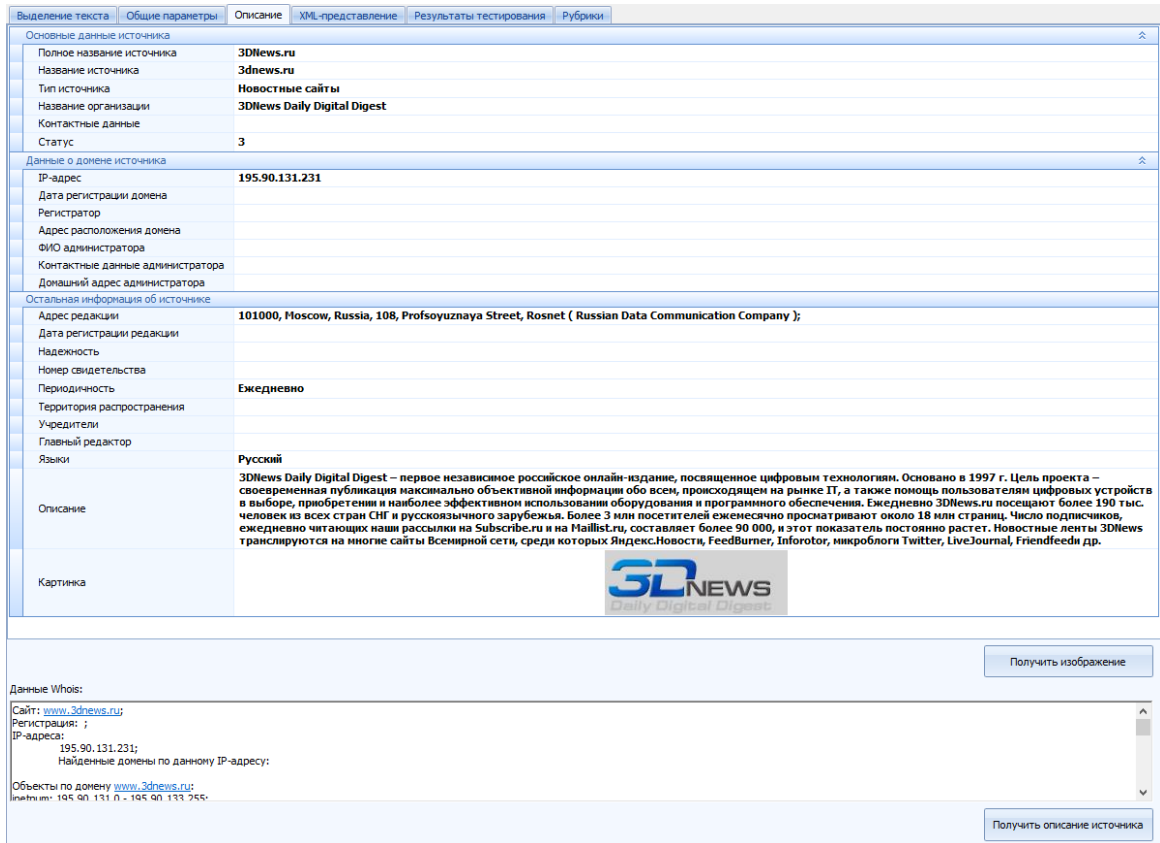


Рисунок 4 – Вкладка «Описание»

5. Вкладка «Рубрики»

На вкладке «Рубрики», представленной на рисунке 5, расположены все классификаторы, загруженные в Систему.

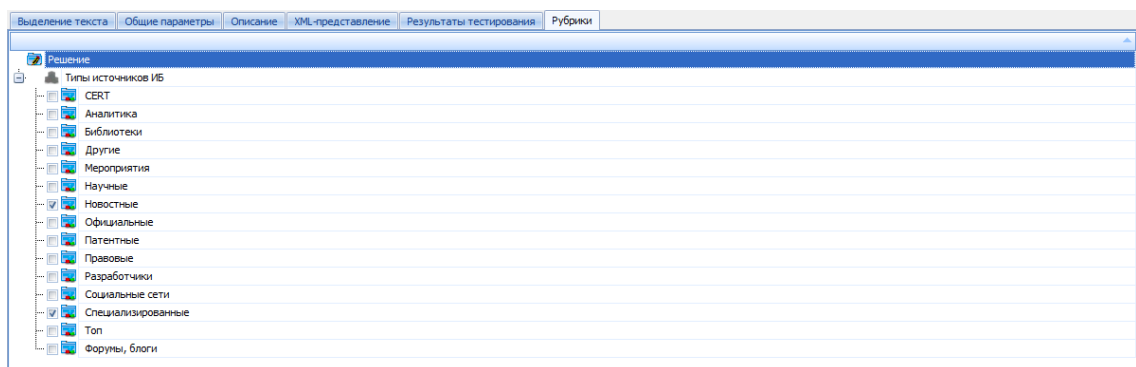


Рисунок 5 – Вкладка «Рубрики»

В данной вкладке можно выбрать желаемые рубрики классификаторов, которые будут автоматически соотнесены со всеми документами, скачанными из выбранного источника, независимо от фактического содержания текста.

6. Вкладка «XML-представление»

На данной вкладке, представленной на рисунке 6, расположен исходный код конфигурационного файла (источника). В программе показывается версия, находящаяся в оперативной памяти, которая будет сохранена при нажатии на кнопку «Сохранить проект», и ее вид может отличаться от версии, которая будет видна при одновременном открытии этого конфигурационного файла в текстовом редакторе до выполнения сохранения проекта.

```

1 <Site SiteEncoding="utf-8" Name="3dnews.ru" ExternalName="3DNews.ru" CreateHttp="false" Downloading="true" Section="" MaxCycles="90" Priority="0" T
2 <Description Description="3DNews Daily Digital Digest - первое независимое российское онлайн-издание, посвященное цифровым технологиям. Основан
3 </Description>
4 <WhoisText "CaZr: www.3dnews.ru;#xD;#xA;Регистрация: ;#xD;#xA;IP-адреса:#xD;#xA;#x9;195.90.131.231;#xD;#xA;#x9;Найденные домены п
5 </WhoisText>
6 <Attributes>
7 <Attribute Code="Address" StringValue="101000, Москва, Russia, 108, Profsoyuznaya Street, Rosnet ( Russian Data Communication Company ); " La
8 </Attribute>
9 <Attribute Code="Founders" StringValue="" Label="Учредители" />
10 <Attribute Code="Languages" StringValue="Русский" Label="Языки" />
11 <Attribute Code="Number" StringValue="" Label="Номер свидетельства" />
12 <Attribute Code="Period" StringValue="Ежедневно" Label="Периодичность" />
13 <Attribute Code="RegistrationDate" StringValue="" Label="Дата регистрации редакции" />
14 <Attribute Code="Reliability" StringValue="" Label="Надежность" />
15 <Attribute Code="Territory" StringValue="" Label="Территория распространения" />
16 <Attribute Code="OrganizationName" StringValue="3DNews Daily Digital Digest" Label="Название организации" />
17 <Attribute Code="IPAddress" StringValue="195.90.131.231" Label="IP-адрес" />
18 <Attribute Code="DomainRegistrationDate" StringValue="" Label="Дата регистрации домена" />
19 <Attribute Code="DomainAddress" StringValue="" Label="Адрес расположения домена" />
20 <Attribute Code="AdminIO" StringValue="" Label="ИЮ администратора" />
21 <Attribute Code="AdminContacts" StringValue="" Label="Контактные данные администратора" />
22 <Attribute Code="AdminAddress" StringValue="" Label="Домашний адрес администратора" />
23 <Attribute Code="Registrar" StringValue="" Label="Регистратор" />
24 <Attribute Code="WhoisText" StringValue="Сайт: www.3dnews.ru;#xD;#xA;Регистрация: ;#xD;#xA;IP-адреса:#xD;#xA;#x9;195.90.131.231;#xD;#xA;#x9;Найденные домены п
25 </Attribute>
26 <Attribute Code="Location" StringValue="" Label="Местоположение" />
27 <Attribute Code="Latitude" StringValue="" Label="Широта" />
28 <Attribute Code="Longitude" StringValue="" Label="Долгота" />
29 </Attributes>
30 <Classificators>
31 <Classifier UniqueID="a4fe9338-8887-4047-8aff-138a4db20859" Name="Типы источников ИБ">
32 <Rubric GlobalID="c25be3e5-249f-4281-9e03-1f7ea93e2347" Name="Специализированные" />
33 <Rubric GlobalID="93551b9b-b94a-40b8-94fe-8704df990832" Name="Новостные" />
34 </Classificators>
35 </Classificators>
36 <RegExSection AutoSearch="false" AutoParse="false" SequentialProcessing="false">
37 <ArticlesSearchTemplates>
38 <RegTemplate Regexp="<title>.*&lt;title&gt;[^\&lt;]*&lt;/title&gt;.*&lt;link&gt;{[^\&lt;]}*&lt;/link&gt;.*?&lt;pubDate&gt;{[^\&lt;]}*&lt;/pubDate&gt;{[^\&lt;]}*&lt;/pubDate&gt;{[^\&lt;]}*&lt;/pubDate&gt;{[^\&lt;]}*&lt;/pubDate&gt;{[^\&lt;]}*&lt;/pubDate&gt;{[^\&lt;]}*&lt;/pubDate&gt;{[^\&lt;]}*&lt;/pubDate&gt;{[^\&lt;]}*&lt;/pubDate&gt;{[^\&lt;]}*&lt;/pubDate&gt;{[^\&lt;]}*" />
39 </ArticlesSearchTemplates>
40 <ArticlesParseTemplates>
41 <RegTemplate Regexp="itemprop="&quot;articleBody&quot;&gt;(.*)&lt;/itemprop="&quot;rate-box&quot;&gt;" DateTemplate="" URLTemplate="" TitleTemp
42 </ArticlesParseTemplates>
43 <ArticlePagesSearchTemplates>
44 <RegTemplate NextPage="True" Regexp="&lt;a href="&quot;{[^\&lt;]*&gt;}&quot;&gt;&lt;img alt="&quot;MaximumLinkCo
45 <RegTemplate NextPage="True" Regexp="&lt;a href="&quot;{[^\&lt;]*&gt;}&quot;&gt;&lt;img alt="&quot;MaximumLinkCo
46 </ArticlePagesSearchTemplates>
47 </RegExSection>
48 <MainLinks>
49 <MainPage URL="http://www.3dnews.ru/news/zss/" />
50 <MainPage URL="http://www.3dnews.ru/news/" />
51 </MainLinks>
52 <RequestTemplates />
53 <SocialCrawlerRequests RequestFlags="" ExportDir="" OutputType="" Login="" Password="" GrabFriends="false" EngineType="" SelectedFileFormats=""
54 </Site>

```

Рисунок 6 – Вкладка «XML-представление»

7. Вкладка «Результаты тестирования»

После завершения создания шаблона необходимо протестировать его на работоспособность. Для этого следует закрыть окно редактора шаблонов, выбрать требуемый конфигурационный файл в панели «Список

конфигурационных файлов», нажать на него правой кнопкой «мыши» и выбрать пункт «Протестировать», как показано на рисунке 7.

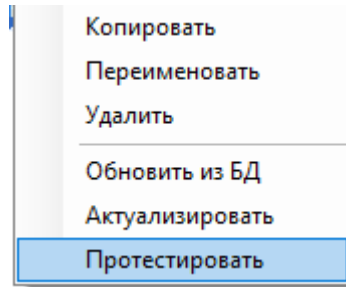


Рисунок 7 – Выбор тестирования

Выбранное поле таблицы окрасится в голубой цвет, что будет означать, что происходит тестирование, затем цвет изменится либо на ярко-зеленый, что будет означать, что тестирование было успешным, либо на красный, что означает, что с помощью созданного шаблона не удалось загрузить ни одной новости.

Предусмотрена возможность выделить несколько файлов и запустить тестирование сразу всего списка. Для этого необходимо воспользоваться стандартными командами среды Windows («Ctrl+A» - чтобы выделить все, «Ctrl+клик» - чтобы выделить желаемые позиции и «Shift+клик» - чтобы выделить промежуток).

После успешного завершения тестирования на вкладке «Результаты тестирования», представленной на рисунке 8, станет доступна подробная информация по проведенному тестированию. Стоит заметить, что после закрытия программы вместе с проектом сохраняется только статус тестирования, но не результаты тестирования.

Выделение текста | Общие параметры | Описание | XML-представление | Результаты тестирования | Рубрики

Выполнено 16.07.2018 в 14:25. Успех. Найдено: 50. Успешно обработано: 50.

Ссылка	Дата публикации	Автор	Заголовок	Текст
https://3dnews.ru/972660	16.07.2018 12:51:04		Sony выпустила улучшенную в...	Компания Sony, по сообщениям сетевых источников, выпустила улучшенную версию компактной фото...
https://3dnews.ru/972666	16.07.2018 13:32:59		«ВыпелКом» стал эксклюзивн...	«ВыпелКом», оказывающий услуги связи под маркой «Билайн», и финская компания HMD Global, выпу...
https://3dnews.ru/972639	16.07.2018 14:04:24		Складной планшет Microsoft An...	Пару лет назад впервые прозвучало кодовое имя Andromeda, за которым якобы скрывается проект M...
https://3dnews.ru/972654	16.07.2018 11:41:24		Материнские платы на Intel Z37...	Большинство крупных производителей материнских плат начали выпускать новые версии BIOS для св...
https://3dnews.ru/972656	16.07.2018 11:10:57		AMD Ryzen 5 2500X: новые под...	Уже довольно давно стало известно, что компания AMD готовит пополнение семейства процессоров R...

Содержимое

Заголовок	Sony выпустила улучшенную версию фотоаппарата RX100 V
Ссылка	https://3dnews.ru/972660
Дата публикации	16.07.2018 12:51
Статус	OK
Автор	
Текст	<p>Компания Sony, по сообщениям сетевых источников, выпустила улучшенную версию компактной фотокамеры RX100 V, которая дебютировала осенью 2016 года.</p> <p>Напомним, что фотоаппарат RX100 V (DSC-RX100M5) оснащён 20,1-мм КМОП-матрицей Exmor RS формата 1" с интегрированным в неё чипом памяти. За обработку данных отвечает процессор BIONZ X. Применён гибридный автофокус с 315 точками, покрывающими 65 % кадра. Устройство оснащено 0,39" OLED-видеоискателем с разрешением XGA (≈2,35 млн точек), модулями беспроводной связи Wi-Fi и NFC, а также 3-дюймовым ЖК-дисплеем с разрешением 1,2 млн пикселей и изменяемым положением.</p> <p>Улучшенная модификация камеры получила обозначение DSC-RX100M5A. Она несёт на борту усовершенствованный процессор, позаимствованный у модели RX100 VI. Чип обеспечивает повышенную производительность на ключевых операциях.</p> <p>Ещё одно изменение коснулось буфера памяти при последовательной фотосъёмке: если раньше он был рассчитан на 150 кадров, то теперь — на 233. Реализовано новое меню, которое предлагает настраиваемый раздел My Menu. Режим Proxru Movie Mode позволяет записывать видео 720p вместе с 4K-роликом. Прочие характеристики изменений не претерпели. Цена обновлённого фотоаппарата составляет около 1000 долларов США.</p> <p>Источники: DPRreview Sony</p>
Комментарий	
Объекты	<pre> <TasObjects> <TasObject Type="PUBLICATION" Id="12807406537165620878"> <TasProperty Type="Title" Value="Sony выпустила улучшенную версию фотоаппарата RX100 V" /> <TasProperty Type="PUBLICATION_Header" Value="Sony выпустила улучшенную версию фотоаппарата RX100 V" /> <TasProperty Type="PUBLICATION_PublicationDate" Value="2018-07-16T12:51:04" /> <TasProperty Type="LoadDate" Value="1900-01-01T00:00:00" /> <TasProperty Type="Source" Value="3dnews.ru" /> <TasProperty Type="URI" Value="https://3dnews.ru/972660" /> </TasObject> </pre>




Рисунок 8 – Вкладка «Результаты тестирования»

Данная вкладка используется при поиске ошибок выделения даты новости, сравнения количества найденных новостей с их фактическим количеством на главной странице, для выяснения правильности нахождения конца текста новости и выявления возможных проблем с кодировкой текста новости.

На вкладке «Результаты тестирования» только после успешного выполнения тестирования выводятся результаты выделения, как если бы программный компонент «ЛАН.Интернет-мониторинг» скачивал документы с сайта с использованием текущего конфигурационного файла. Причем тестирование показывает актуальные результаты как в режиме обхода «Авто», так и в режиме «По шаблону».

8. Вкладка «Привязки»

Вкладка «Привязки» (рисунок Рисунок 9 – Вкладка «Привязки») предназначена для связывания правил обработки источников с главными страницами.

У каждой главной страницы может быть своя верстка и исходя из этого нужен свой набор правил сбора информации.

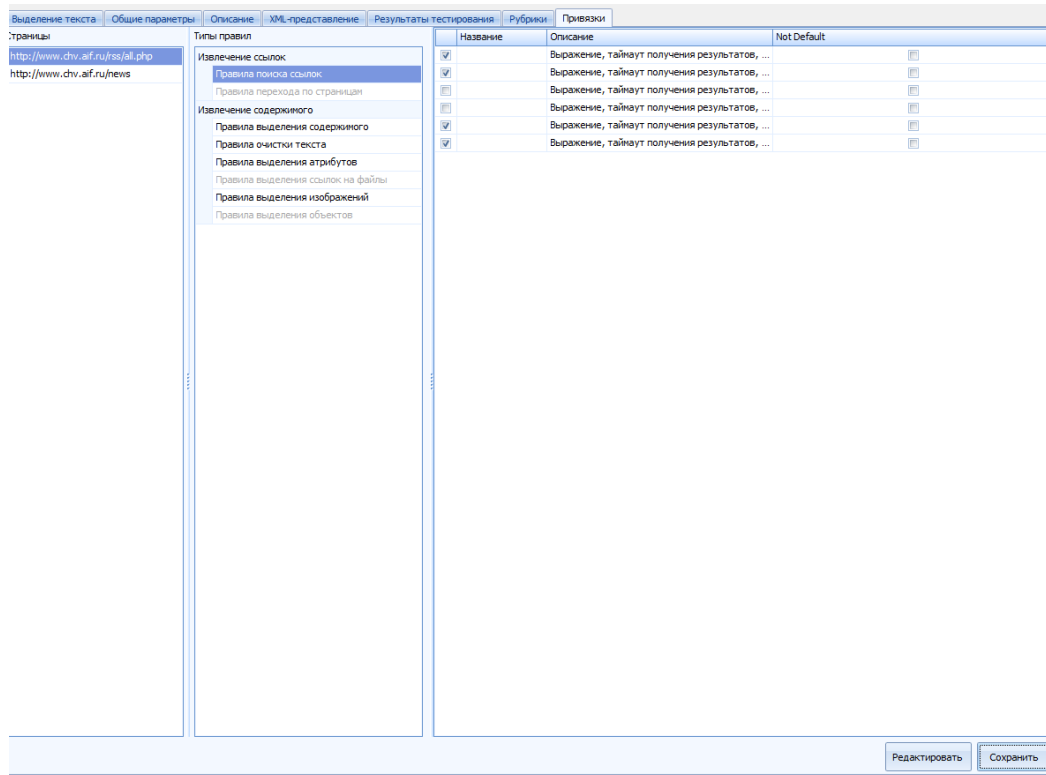


Рисунок 9 – Вкладка «Привязки»