Министерство образования Республики Беларусь

Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники»

Факультет компьютерных систем и сетей Кафедра информатики Дисциплина: Объектно-ориентированное программирование (ООП) ПОЯСНИТЕЛЬНАЯ ЗАПИСКА к курсовому проекту на тему «Архивирование интернет-новостей» Выполнил: студент гр. 953501 Гришаев Н.Г. ассистент кафедры информатики Руководитель: Рогов М.Г.

Минск

2021

Содержание

Введение		3
	Постановка задачи	
	1.1.Описание предметной области	4
	1.2. Цели и задачи проектирования	4
	1.3. Требования к разрабатываемому программному обеспечению	5
2.	Проектирование задачи	7
	2.1.Проектирование архитектуры приложения	7
	2.2.Обоснование выбора технологии	8
	2.3.Описание сред реализации	8
3.	Программная реализация	12
	3.1. Физическая структура приложения	12
	3.2. Архитектура приложения	13
4.	Описание применения	17
Заключение		26
Список использованных источников		27

Введение

В настоящее время актуальной является проблема ограничения государством Республики Беларусь доступа к независимым новостным интернет-порталам, таким как «Хартия'97», «Радыё Свабода», «Белсат», «Еврорадио», «Медиазона. Беларусь», «Тhe Village Беларусь». В любой момент времени материал любого издания, сервера которого располагаются в Беларуси, может быть ограничен в свободном доступе, как это произошло с порталом «ТUT.ВУ».

Возникает необходимость удобного хранения материалов различных независимых изданий, позволяющего пополнять архив новыми новостными статьями и производить поиск по ранее сохранённым материалам.

Решением поставленного вопроса может стать приложение, предоставляющее пользователю указанный функционал.

1. Постановка задачи

1.1. Описание предметной области

В общем случае новостная статья представляет собой отдельную вебстраницу с уникальный URL-ссылкой, где в одинаковом для каждого новостного портала формате располагаются заголовок, текст статьи и время публикации.

Доступ к отдельным новостным страницам осуществляется посредством перехода с другой страницы, в том числе главной страницы портала.

Для поиска по ключевым словам и упорядочивания результатов по релевантности используется полнотекстовый поиск.

1.2. Цели и задачи проектирования

Так как каждый новостной сайт имеет уникальную разметку, то требуется создать возможность расширения приложения путем добавления новых программных единиц для каждого сайта. Так как каждый такой модуль имеет общий интерфейс (получает полный html-код страницы, и возвращает заполненный объект NewsArticle), то имеет смысл объединить каждый такой модуль — SpecificInstruction — путём создания абстрактного класса, содержащего методы, необходимые или опциональные для переопределения.

Микросервис, позволяющий перерабатывать текущее состояние новостного сайта в набор NewsArticle, должен быть способен принять в запросе ссылку на новостной сайт для скачивания статей и параметры загрузки. Далее он должен определить, какая из имеющихся SpecificInstruction, применима к переданному сайту, если таковая найдётся.

Микросервис для работы с базой данных предоставляет возможности добавлять новые статьи и производить поиск по старым.

Клиентское приложение должно иметь удобный интерфейс и позволять отправить запрос на скачивание последних новостей с доступного сайта, а также производить параметризированный поиск по архивным статьям.

Микросервис для соединения воедино вышеперечисленным частей в должен служить для перенаправления запросов в нужный микросервис, а также может включать в себя дополнительную логику, например, ограничение API или требование аутентификации.

1.3. Требования к разрабатываемому программному обеспечению

Готовый программный продукт должен удовлетворять следующим требованиям:

- 1. Серверная часть должна быть представлена микросервисами на технологии ServiceStack
- 2. Для скачивания веб-страниц использовать компонент «abot»
- 3. Доступ к базе данных должен осуществляться посредством технологии объектно-реляционного отражения (ORM), в качестве которой выступает ServiceStack.OrmLite;
- 4. СУБД PostgreSQL
- 5. Клиентская часть не должна знать информацию об используемой базе данных
- 6. Клиентское приложение должно иметь простой, интуитивно понятный интерфейс с использованием технологии WinForms
- 7. Приложение должно уметь выделять из интернет-порталов новостные статьи и сохранять их
- 8. Приложение должно быть расширяемым в плане добавления новых новостных сайтов для архивирования

- 9. Приложение должно позволять производить поиск статей по ключевым словам
- 10. Приложение должно позволять производить фильтрацию статей по дате публикации и сайту-источнику
- 11. Приложение должно позволять использовать постраничный просмотр краткой информации о хранящихся статьях

Таким образом, задача данного курсового проекта сводится к разработке приложения на основе микросервисов для архивирования новостных статей с различных интернет-порталов. Готовое приложение должно иметь понятный интерфейс и быть удобным в использовании.

2. Проектирование задачи

2.1. Проектирование архитектуры приложения

На рисунке 1 изображена диаграмма основных классов, используемых в приложении. В центре располагаются классы-сущности, хранимые в базе данных и используемые во всём приложении.

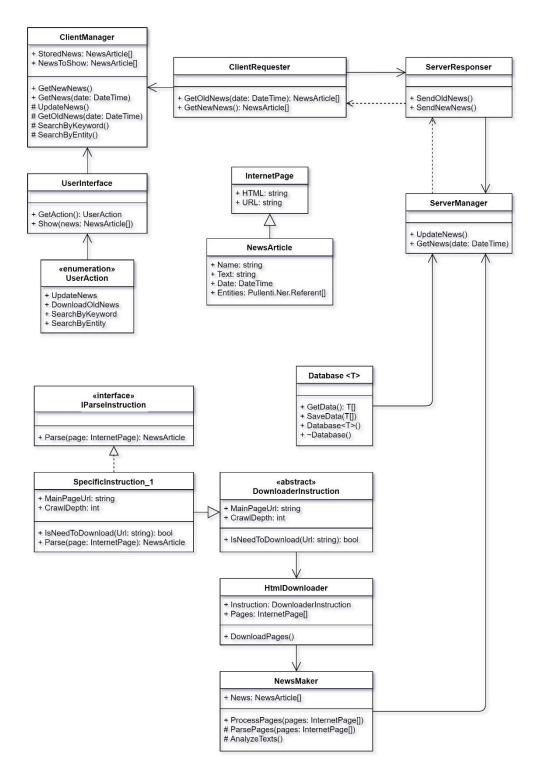


Рисунок 1 — UML-диаграмма классов

2.2. Обоснование выбора технологии

Веб-сервис (англ. web service) — идентифицируемая веб-адресом программная система со стандартизированными интерфейсами. Веб-службы могут взаимодействовать друг с другом и со сторонними приложениями посредством сообщений, основанных на определённых протоколах.

Язык С# представляет собой мощный инструмент для создания вебсервисов, а с использованием фреймворка ServiceStack их разработка значительно упрощается.

PostgreSQL является одной из наиболее популярных систем управления базами данных, дающей возможность использовать основные и дополнительные функции язык программирования структурированных запросов SQL, а расширение ServiceStack.OrmLite предоставляет ORM-функционал для взаимодействия с базами данных.

2.3. Описание сред реализации

Місгоsoft Visual Studio. Microsoft Visual Studio — это новая разработка компании Microsoft, позволяющая создавать приложения, работающие на платформе .NET. Особенность этой платформы заключается в широком наборе сервисов, которые доступны в различных языках программирования. При этом сервисы реализуются в виде промежуточного кода, который не зависит от базовой архитектуры. Едва ли не главной целью создания такой платформы было оснащение разработчиков специальными сервисноориентированными приложениями, которые могли бы работать на любой платформе, начиная от персонального компьютера и заканчивая мобильным устройством.

Microsoft Visual Studio объединяет в себе огромное количество функций, позволяющих осуществлять разработки для Windows всех версий, в том числе и Windows 8, интернета, SharePoint, различных мобильных устройств и

облачных технологий. В Visual Studio реализуется новая среда разработчика, благодаря которой создавать приложения стало проще. Microsoft Visual Studio – это обновленная и упрощенная программная среда, для которой характерна высокая производительность, причем она не зависит от особенностей оборудования.

С#. С# разрабатывался как язык программирования прикладного уровня для CLR и, как таковой, зависит, прежде всего, от возможностей самой CLR. Это касается, прежде всего, системы типов C#, которая отражает BCL. Присутствие или отсутствие тех или иных выразительных особенностей языка особенность быть диктуется тем, может ЛИ конкретная языковая транслирована в соответствующие конструкции CLR. Так, с развитием CLR от версии 1.1 к 2.0 значительно обогатился и сам С#; подобного взаимодействия следует ожидать и в дальнейшем. (Однако эта закономерность была нарушена выходом С# 3.0, представляющего собой расширения языка, опирающиеся на расширения платформы .NET.) CLR предоставляет C#, как и всем другим .NET-ориентированным языкам, многие возможности, которых лишены «классические» языки программирования. Например, сборка мусора не реализована в самом C#, а производится CLR для программ, написанных на С# точно так же, как это делается для программ на VB.NET, J# и др.

WinForms. Windows Forms — это платформа пользовательского интерфейса для создания классических приложений Windows. Она обеспечивает один из самых эффективных способов создания классических приложений с помощью визуального конструктора в Visual Studio. Такие функции, как размещение визуальных элементов управления путем перетаскивания, упрощают создание классических приложений.

B Windows Forms можно разрабатывать графически сложные приложения, которые просто развертывать, обновлять, и с которыми удобно работать как в автономном режиме, так и в сети. Приложения Windows Forms

могут получать доступ к локальному оборудованию и файловой системе компьютера, на котором работает приложение.

Аbot. Abot — это фреймворк С#, представляющий собой поискового робота с открытым исходным кодом, созданный для обеспечения скорости и гибкости. Он заботится о низкоуровневых вещах (многопоточность, НТТР-запросы, планирование, синтаксический анализ ссылок и т.д.). Вы просто получаете события для обработки данных страницы. Вы также можете добавлять свои собственные реализации основных интерфейсов, чтобы полностью контролировать процесс сканирования. Версия пакета Abot Nuget ≥ 2.0 нацелена на .NET Standard 2.0, а версия пакета Abot Nuget < 2.0 нацелена на .NET версии 4.0, что обеспечивает высокую совместимость со многими реализациями .NET Framework / Core.

ServiceStack. ServiceStack — это простая, быстрая, универсальная и высокопроизводительная полнофункциональная инфраструктура вебсервисов и веб-служб, продуманная с целью снижения искусственной сложности и продвижения передовых методов удаленных служб с помощью дизайна на основе сообщений, который позволяет максимально многократно использовать, что может использовать интегрированный сервисный шлюз для создания слабосвязанных модульных сервисных архитектур. Сервисы ServiceStack можно использовать с помощью массива встроенных форматов быстрых данных (включая JSON, XML, CSV, JSV, ProtoBuf, Wire и MsgPack), а также XSD / WSDL для конечных точек SOAP и хостов Rabbit MQ, Redis MQ и Amazon SQS MQ.

Его дизайн и простота ориентированы на беспрецедентный набор функций производительности, которые можно декларативно включить без кода, от создания полностью запрашиваемых веб-API с помощью всего одного типизированного запроса DTO с автоматическим запросом, поддерживающим все основные СУБД, до встроенной поддержки автоматических пакетных запросов или без особых усилий включить богатое кэширование HTTP и

шифрованный обмен сообщениями для всех ваших существующих служб через плагины.

ServiceStack Services также максимизирует продуктивность для потребителей, предоставляя мгновенный сквозной типизированный API без генерации кода, что обеспечивает наиболее продуктивный опыт разработки для разработки веб-служб в .NET.

PostgreSQL. PostgreSQL — это объектно-реляционная система управления базами данных (OPCУБД, ORDBMS), основанная на POSTGRES, Version 4.2 — программе, разработанной на факультете компьютерных наук Калифорнийского университета в Беркли. В POSTGRES появилось множество новшеств, которые были реализованы в некоторых коммерческих СУБД гораздо позднее.

PostgreSQL — СУБД с открытым исходным кодом, основой которого был код, написанный в Беркли. Она поддерживает большую часть стандарта SQL и предлагает множество современных функций: сложные запросы, внешние ключи, триггеры, изменяемые представления, транзакционная целостность, многоверсионность.

Кроме того, пользователи могут всячески расширять возможности PostgreSQL, например, создавая свои типы данных, функции, операторы, агрегатные функции, методы индексирования, процедурные языки.

А благодаря свободной лицензии, PostgreSQL разрешается бесплатно использовать, изменять и распространять всем и для любых целей — личных, коммерческих или учебных.

3. Программная реализация

3.1. Физическая структура приложения

Прямое взаимодействие с базой данных осуществляет статический класс Database. При первом обращении к любому из его методов вызывается статический конструктор, в котором происходит первоначальная настройка подключения к базе данных при помощи строки с параметрами доступа. Результат этой настройки сохраняется в readonly переменную dbFactory типа OrmLiteConnectionFactory. Далее любое взаимодействие с базой данных осуществляется следующим образом: 1) происходит открытие соединения с базой данных с помощью метода dbFactory.open(), 2) выполняются необходимые операции, 3) соединение закрывается, чтобы не нагружать базу данных. Язык С# предоставляет удобный способ использования данной схемы через конструкцию using: using (var db = dbFactory.Open()) { /* работа с БД через переменную db */ }. По окончании блока кода происходит автоматическое закрытие соединения с базой данных.

После первоначальной настройки проверяется, существует ли в базе данных таблица NewsArticle. В случае, если такой таблицы нет, сбрасывается таблица text_analysis, и далее происходит загрузка в базу данных последней резервной копии материалов. Резервная копия представляет собой отсортированные по дате публикации сериализованные новостные статьи, и хранится в текстовом файле в формате JSON, обновляясь каждый раз при запуске микросервиса.

Интерфейс взаимодействия с классом Database состоит из следующих методов:

void Add(NewsArticle article) — метод для внесения новой статьи в базу данных. После открытия соединения с базой данных происходит проверка на наличие в базе данных уже существующей статьи с такой же URL-ссылкой. В случае, если такой статьи в базе нет, она добавляется, после чего заголовок и текст статьи анализируются в

специальный формат tsvector для осуществления дальнейшего полнотекстового поиска.

- NewsArticle GetOne(string url) метод для загрузки из базы данных новостной статьи с URL-ссылкой, соответствующей переданной строке. Для ускорения поиска по строке-ссылке она индексируется средствами базы данных.
- List<NewsArticle> GetFilteredList(ListRequest request) метод для получения краткого (заголовок, дата публикации, ссылка) списка статей, соответствующих заданному фильтру. Сначала проверяется, что дата публикации включена в заданный отрезок времени, а ссылка начинается с указанного префикса (например, ссылка на основную страницу сайта). Далее, если запрос содержит поиск по ключевым словам, осуществляется полнотекстовый поиск по заготовленной характеристике tsvector, а результаты поиска сортируются по убыванию степени соответствия поисковому запросу. Если запрос не содержит такого поиска, то отфильтрованные статьи сортируются по дате публикации. Получившийся список ограничивается некоторым количеством статей с некоторым смещением относительно начала, переданными в запросе. Это позволяет осуществлять просмотр отфильтрованного списка статей постранично. Статьи, отобранные для показа на странице, выгружаются из базы частично, без полного текста и HTML, что позволяет снизить нагрузку на БД и сеть.

3.2. Архитектура приложения

Приложение представляет собой три заменяемых микросервиса, предоставляющих следующие интерфейсы: АРІ для работы с серверной частью, загрузка новостных статей, работа с базой данных. Клиентское приложение представляет собой десктопное приложение. Все взаимодействия между клиентским приложением и сервером, а также между микросервисами

осуществляются посредством веб-запросов. Схема взаимодействия частей приложения представлена на рисунке 2.

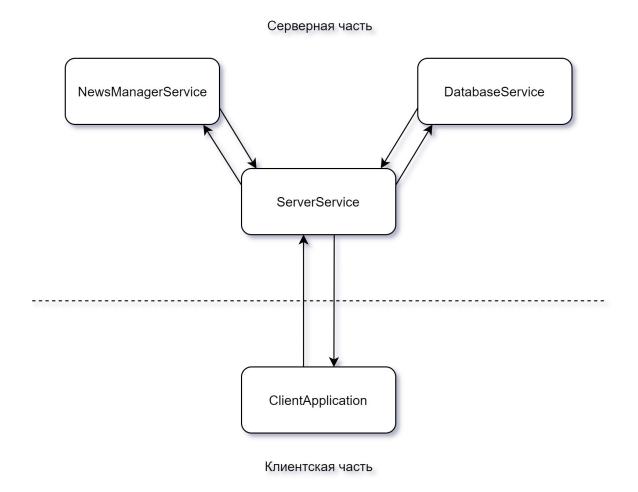


Рисунок 2 — Схема взаимодействия микросервисов и клиентской части

Веб-запросы представляют собой следующие классы:

• Для взаимодействия с базой данных:

```
// Сохранение статьи в базу
public class SaveRequest : IReturn<SaveResponse>
{
    public NewsArticle Article { get; set; }
}
public class SaveResponse
{
}
```

```
// Получение списка новостей по запросу
   public class ListRequest : IReturn<ListResponse>
    {
        public string Url { get; set; } = null;
        public DateTime LeftBoundDate { get; set; } =
DateTime.MinValue;
        public DateTime RightBoundDate { get; set; } =
DateTime.MaxValue;
        public string Keywords { get; set; } = null;
        public string[] Entitities { get; set; } = null;
        public int Skip { get; set; } = 0;
        public int Count { get; set; } = 10;
        public bool OldestFirst { get; set; } = false;
    }
    public class ListResponse
        public NewsArticle[] Result { get; set; }
    }
    // Получение полной статьи по ссылке
    public class ArticleRequest : IReturn<ArticleResponse>
    {
        public string Url { get; set; }
   public class ArticleResponse
        public NewsArticle Result { get; set; }
    }
    • Для взаимодействия со скачивателем новых новостей:
    // Получение списка доступных сайтов
    public class AvailableSitesRequest :
IReturn<AvailableSitesResponse>
   public class AvailableSitesResponse
        public string[] Result { get; set; }
    }
```

```
// Загрузка новых статей
    public class NewsAsyncRequest : IReturnNewsAsyncResponse>
    {
        public string Url { get; set; } = null;
        public int MaxNewsCount { get; set; } = 1;
        public int CrawlMinimumDelayMilliSeconds { get; set; } =
1000;
    public class NewsAsyncResponse
    {
        public NewsArticle[] Result { get; set; }
    }
    • Для взаимодействия с серверной частью:
    // Перенаправление запросов к загрузчику и базе данных
    public class AvailableSitesRequest :
ManagerService.AvailableSitesRequest { }
    public class AvailableSitesResponse :
ManagerService.AvailableSitesResponse { }
    public class ListRequest : DatabaseService.ListRequest { }
    public class ListResponse : DatabaseService.ListResponse { }
    // Загрузка новых статей
    public class NewsRequest : IReturnNewsResponse>
    {
        public string Url { get; set; } = null;
        public int Count { get; set; } = 1;
    }
    public class NewsResponse
    {
        public NewsArticle[] ResultNews { get; set; }
        public bool[] ResultIsSaved { get; set; }
    }
```

4. Описание применения

Клиентское приложение имеет графический интерфейс и является в целом интуитивно понятным в использовании.

Запуск

После запуска приложения необходимо дождаться установления подключения к серверной части для дальнейшей работы. В процессе подключения отображается значок ожидания (рисунок 1), в случае успеха он заменяется на надпись «ОК» (рисунок 2). Если в дальнейшем в процессе работы соединение будет утеряно, оно будет пытаться восстановиться автоматически.

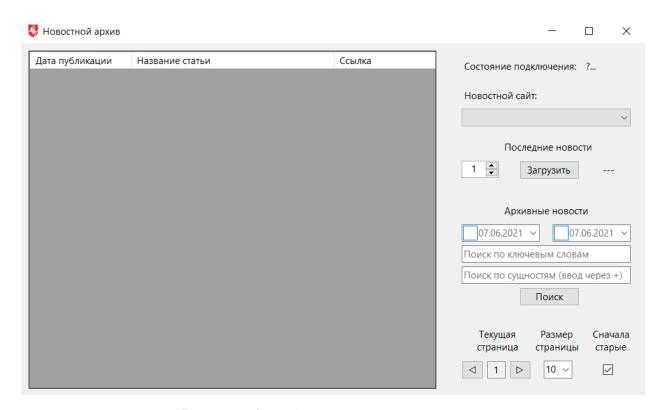


Рисунок 1 — Ожидание подключения

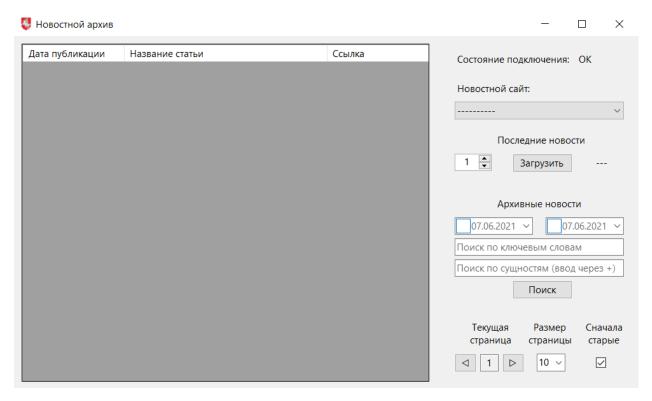


Рисунок 2 — Подключение установлено

Загрузка последних новостей

Для загрузки последних новостей необходимо выбрать нужный новостной сайт из доступных с помощью выпадающего списка, затем установить необходимое количество новостей для скачивания, и нажать кнопку «Загрузить» (рисунок 3), после чего дождаться (рисунок 4) окончания работы команды: в случае успешного скачивания статус ожидания сменится на «ОК», а в поле просмотра новостей появятся скачанные статьи (рисунок 5).

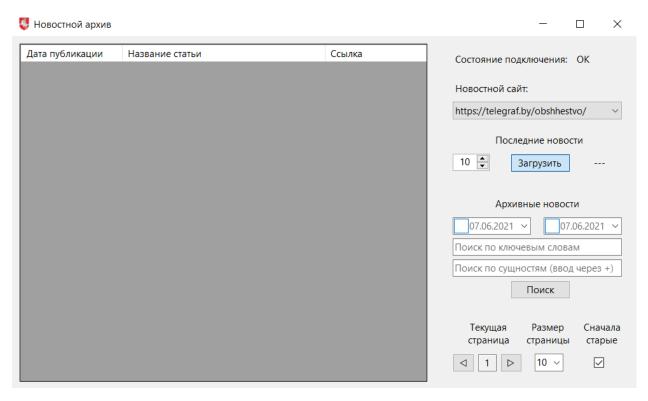


Рисунок 3 — Настойка загрузки последних новостей

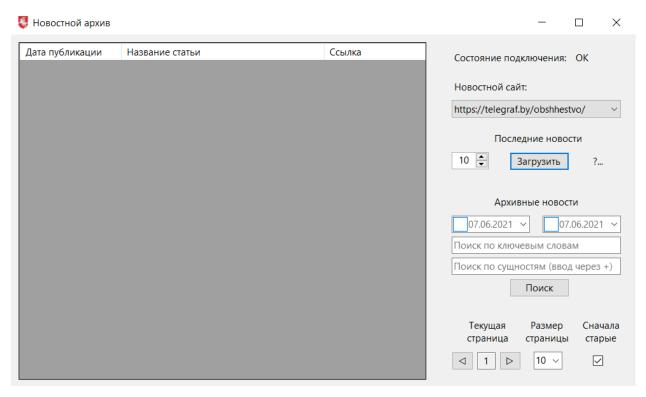


Рисунок 4 — Ожидание скачивания последних новостей

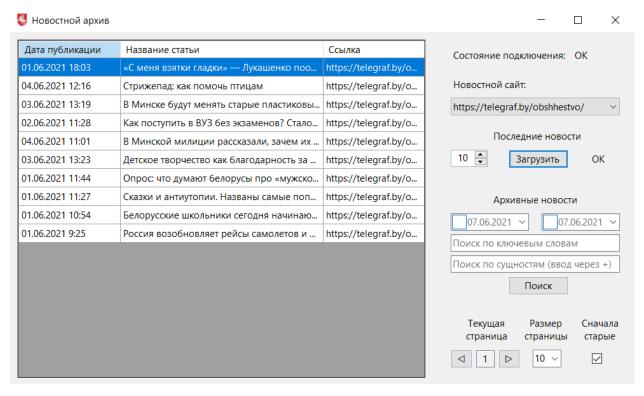


Рисунок 5 — Успешная загрузка последних новостей с указанного сайта

Поиск по архивным статьям

Для поиска по архивным новостям можно установить параметры поиска (фильтр по дате публикации, поиск по ключевым словам, фильтр по сущностям). После ввода необходимых параметров следует нажать на кнопку «Поиск» (рисунок 6), после чего в поле вывода статей будут показаны результаты поиска (рисунок 7). В случае поиска по ключевым словам результаты будут отсортированы по релевантности, иначе — по дате публикации.

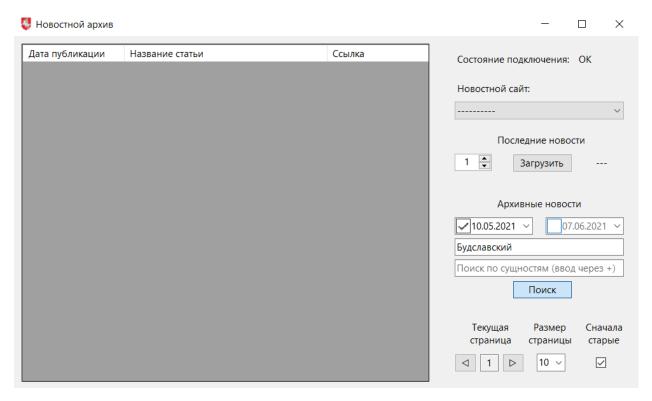


Рисунок 6 — Параметризация поиска

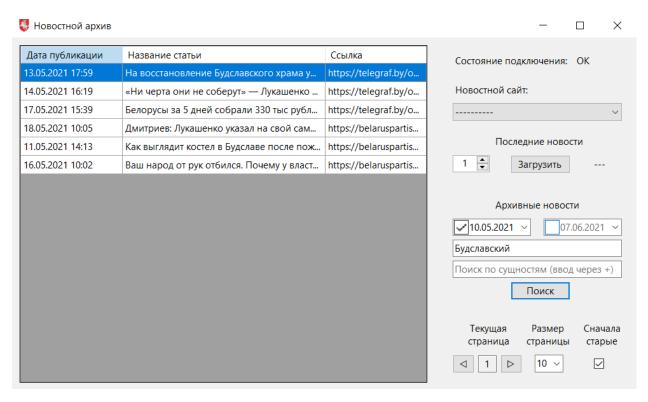


Рисунок 7 — Результаты поиска

Масштабирование и параметры вывода

Для удобства просмотра приложение возможно расширить на весь экран. Тогда появится больше места для вывода информации о новостных статьях (рисунок 8). В параметрах вывода можно изменить количество статей, выводимых на одной странице (рисунок 9). В случае, если выбранное количество статей не помещается на экран полностью, появляется полоса прокрутки (рисунок 10). По страницам можно перемещаться с использованием стрелочек (рисунок 11). Можно изменить порядок вывода новостных статей — либо сначала выводить последние новости, либо самые старые хранящиеся в архиве (рисунок 12).

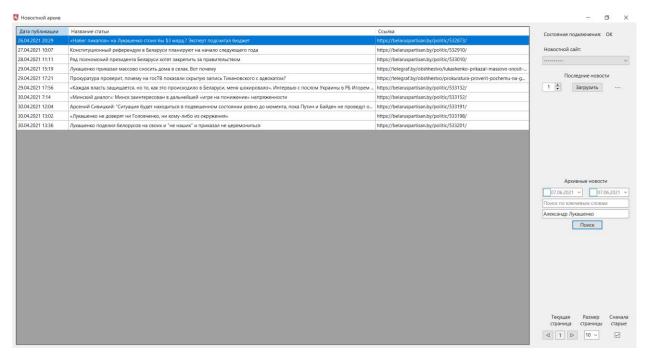


Рисунок 8 — Масштабирование во весь экран

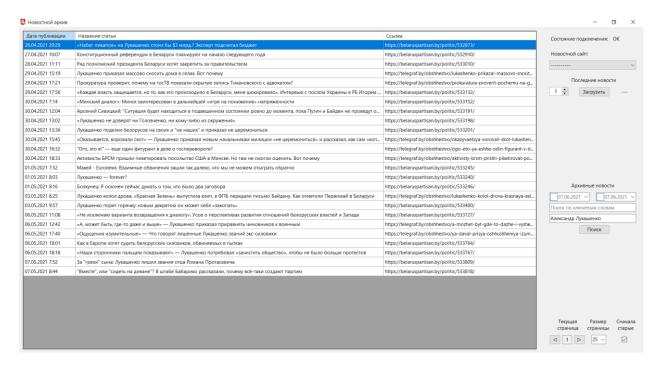


Рисунок 9 — Изменение размера страницы



Рисунок 10 — Полоса прокрутки страницы

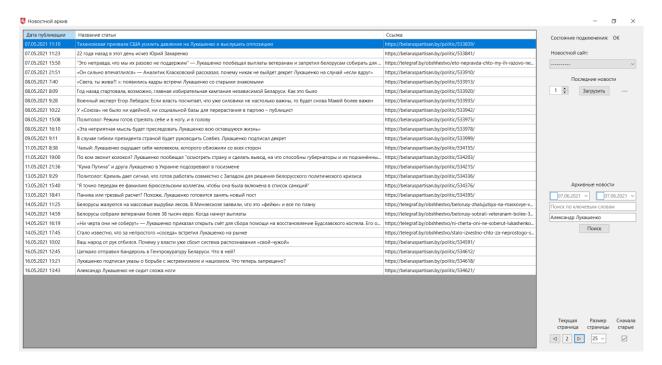


Рисунок 11 — Просмотр различных страниц поиска

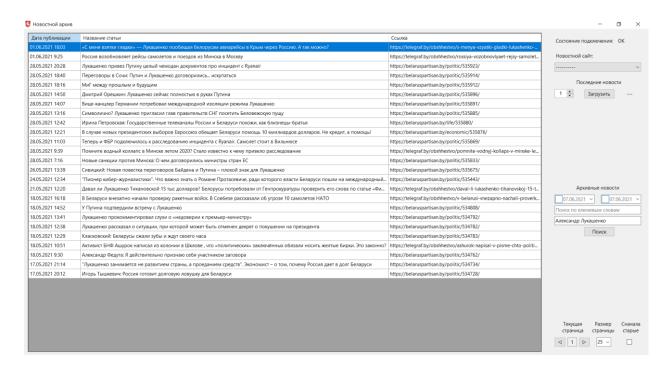


Рисунок 12 — Изменение порядка вывода статей

Получение полного текста статьи

Для прочтения полного текста архивной статьи необходимо совершить двойное нажатие левой кнопкой мыши на нужной статье из отображаемого списка. После этого откроется окошко, в котором будут отображены ссылка на статью, дата публикации, её название, и полный текст (рисунок 13).

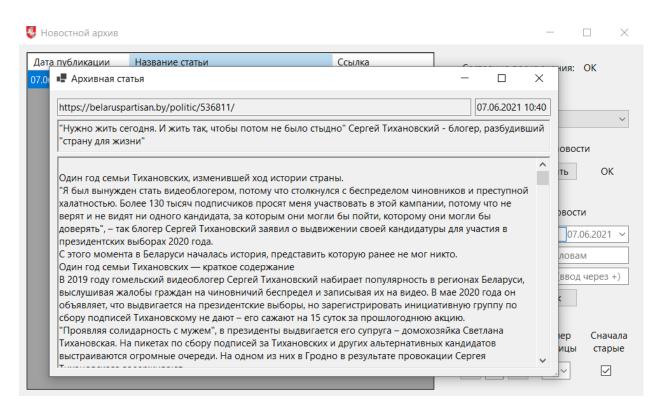


Рисунок 13 — Просмотр полного текста статьи

Заключение

Таким образом, в результате выполнения курсовой работы было реализовано приложение, способное скачивать, хранить и производить поиск по новостным статьям интернет-изданий.

За время написания проекта были получены навыки проектирования и написания приложения на основе веб-сервисов на языке C# с использованием фреймворка ServiceStack, эксплуатации базы данных PostgreSQL с применением ORM-технологий и фреймворка ServiceStack.OrmLite.

Разработанное программное средство представляет собой законченный программный продукт, готовый к введению в эксплуатацию. Также при желании в приложение могут быть добавлены новые программные модули, расширена функциональность, изменён дизайн клиентской части и пр.

Исходный код программы доступен по ссылке https://github.com/Rigorich/NewsSaver_Project

Список использованных источников

- ServiceStack Documentation [Электронный ресурс]: https://docs.servicestack.net/
- 2. ServiceStack / GitHub [Электронный ресурс]: https://github.com/ServiceStack/ServiceStack
- 3. ServiceStack.OrmLite / GitHub [Электронный ресурс]: https://github.com/ServiceStack/ServiceStack.OrmLite
- 4. Документация PostgreSQL и Postgres Pro [Электронный ресурс]: https://postgrespro.ru/docs/
- 5. Руководство по PostgreSQL / Metanit [Электронный ресурс]: https://metanit.com/sql/postgresql/
- 6. Полнотекстовый поиск в PostgreSQL / Александр Алексеев (Postgres Professional) / YouTube [Электронный ресурс]: https://www.youtube.com/watch?v=5ZATiyvPQ-A
- 7. Готовим полнотекстовый поиск в Postgres. Часть 1 / Хабр [Электронный ресурс]: https://habr.com/ru/post/442170/