Development of a web-based system of automatic content retrieval database

Olha V. Korotun[0000-0003-2240-7891], Tetiana A. Vakaliuk[0000-0001-6825-4697] and Viacheslav A. Oleshko[0000-0001-6434-250X]

Zhytomyr Polytechnic State University, 103, Chudnivska Str., Zhytomyr, 10005, Ukraine
{olgavl.korotun, tetianavakaliuk, vladolleshko19}@gmail.com

Abstract. In this work, the database was designed and implemented in accordance with the requirements of the relational model, which ensures the storage and collective access to the information of the auto-filling system and CMS WordPress data. Algorithms of system functioning were developed, the order of interaction of classes during program code execution was determined, as a result of which the application was implemented. Template Method architectural pattern was chosen to implement the web-based automatic content filling system. The following tools and technologies were selected to create the software package HTML markup language for HTML documents; programming language PHP; MySQL database management environment; Apache web server; the OpenServer package. The algorithms of the basic processes of content filling automation were considered and the interaction of the system classes during the processes of parsing, filtering and storing of information were analyzed. The developed system does not require specialized hardware, additional settings and deployment tools other than the standard ones for such plugins. This application is mostly for the site administrator and does not have user interface. That is why the features of the plugin automation system configuration interface; RSS feeds view and management interface, as well as the RSS feed configuration interface are described in detail. In the future, this system can be improved by introducing new functionality and improving the algorithm for reading data.

Keywords: system, content, automatic content, development.

1 Introduction

1.1 Formulation of the problem

Professional SEO and website promotion are long processes. In such circumstances, it is difficult to predict the timeframe within which a project will start to return investment and generate profit. To increase the load, you can work for three or hire a copywriter, programmer, and marketer.

An automatic content filling system is a cost-effective alternative that will save you unnecessary costs and reduce the time spent filling the site with content. The secret of

auto-filling is extremely simple – the staff is replaced by a special program or plugin, customized for the project's requests. Its functions are to collect, adapt and publish content from competing for RSS feeds on a web resource.

The urgency of the chosen topic is that automation of the automatic filling system will provide information to the web resource without the help of a moderator, which will greatly simplify the maintenance of the web resource with minimal interventions in the process. The functionality of the system for automatic filling of information will allow using it according to the needs of the user.

1.2 Analysis of recent research and publications

The problem of development of the system of automatic filling of the context was investigated in various aspects: application of the information system of content management of a web resource for conducting e-commerce [1]; unified methods of processing information resources in systems of electronic content commerce [8]; peculiarities of formation and analysis of content of Internet newspaper of music news [4]; intellectual content management system for e-business sites [2]; application of content analysis of textual information in e-commerce systems [3], etc.

In particular, in paper [8] is described the formal model of information resource processing in e-commerce systems that simplifies the technology of content formation, management and implementation, and proposes methods for solving e-commerce problems and functional content management services.

That is why the purpose of this article is the design of architecture, the development of algorithms and the implementation of software complex information retrieval by parameters and automation systems for information processing.

2 Methods

Methods of research: theoretical analysis of scientific literature to clarify the state of the problem under study, systematization, generalization. The design method was also used to develop the architecture of the application, the methods of algorithm design and object-oriented programming – to develop algorithms for the operation of individual blocks and the application as a whole.

3 Results

The main purpose of the implementation is to simplify the work of filling the site with information. First, implementation of the system will help the site administrator to automate their functional tasks: fully automate the process of finding the necessary information, automate and organize the storage of data, reduce the time of work with the site, the time of their processing, as well as save money in the promotion of the site.

The result of this task is a comprehensive web-based content automation system,

which contains a server structure of data storage, a multi-user client application for the implementation of functionality and means of control and access control [6].

Content Filling Automation automates content collection and publishing on a web-based resource. The modern software market features a wide variety of tools and technologies that help you solve problems related to the automatic search and content parsing processes.

WP RSS Aggregator is the most popular, easy to use and effective plugin for news aggregation. Its main functions are the ability to specify multiple sources, update interval, hide or no source, control the display of material. The plugin is free, but for some add-ons that extend its functionality, you will have to pay. The disadvantages include a small number of content post-processing features.

The FeedWordPress plugin is one of the news aggregators. The news collected by the plugin is copied to the database in the form of notes of a separate type, with the assignment of appropriate tags. If the required tag is not already in the database, the plugin will create one automatically. However, the plugin is very cumbersome and has many settings that will not be clear and useful to the potential user.

WPeMatico is an easy-to-use news aggregator that automatically publishes content from various sources, combining them into so-called “campaigns” according to your chosen topic. It can use keywords, phrases, and regular expressions to filter material, but most of the functionality is paid.

The Push Syndication plug-in has been specifically designed to manage to autocomplete across multiple sites. With one click, you can post to multiple platforms (up to more than 100 sites). The solution can be used to generate API tags used to promote blog content on WordPress, but the plugin does not have content settings.

The Syndicate Out plugin allows blog owners to auto-aggregate or creates content blogs from any number of different sources without relying on RSS feeds. However, there is no media-parsing, configuration, and post processing of the content.

CyberSyn is a powerful, easy and easy-to-use Atom / RSS posting plugin. It allows you to automatically receive and embed videos from YouTube channels. It does not have any problems with the syndication of various types of embedded media content. The disadvantages include storing all links from the source, inability to add multiple RSS feeds at a time.

Therefore, the main features of the new system should be the presence of a web interface, the module for parsing and storage of content and media data, the module for generating articles by parameters, the module for filtering information for parsing by parameters.

The Template Method architectural pattern (Fig. 1, 2) was chosen [7]. The Template Method pattern is widely used in application frameworks. Each framework implements immutable pieces of architecture in the domain and identifies those parts that can or should be customized by the client.

The component designer decides which algorithm steps are unchanged (or standard) and which are variable (or custom). The abstract base class implements standard algorithm steps and can provide (or not) the default implementation for custom steps. Variable steps can (or should) be provided by a component client in specific derived classes.

The component designer defines the required steps of the algorithm, the order in which they are performed, but allows the component clients to extend or replace some of these steps.

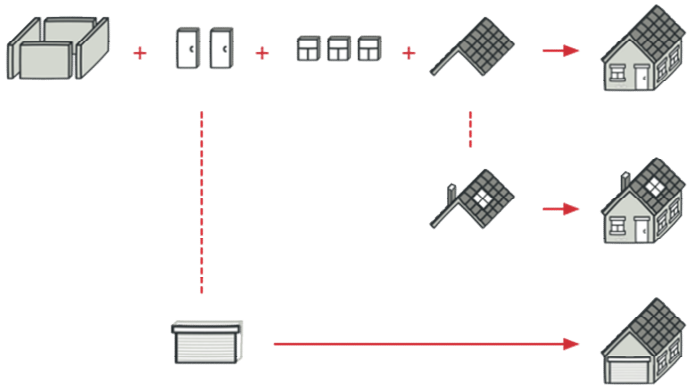

Fig. 1. An analogy to the life pattern of the Template Method

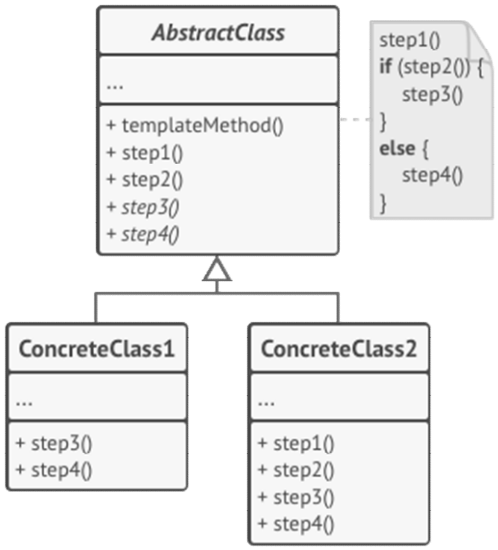

Fig. 2. Template Method Pattern Structure

The abstract class defines the steps of the algorithm and contains a template method consisting of the calls of these steps. The steps can be both abstract and include a default implementation.

The specific class overrides some (or all) steps of the algorithm. Specific classes do not outweigh the template method itself. Concerning the platform on which the system is built, the most popular CMS in the world has been selected today.

In general, Content Management is a web application that allows site owners, editors, authors to manage their sites and publish content without any programming knowledge.

Word Press uses PHP and MySQL, which is supported by virtually all hosting providers [5].

Typically, this CMS is used to create a blog, but a WordPress site can easily be turned into an online store, a portfolio, a periodic site that is indisputably suited to the subject matter of a web-based content filling system.

One of the important features of WordPress is its intuitive and friendly interface.

The important thing is that WordPress is an open-source system and is free for everyone. In addition, it allows millions of people around the world to create modern, high-quality sites that can easily connect to the automatic content filling system and fill your site with content within minutes.

The following tools and technologies were selected to create the software package:

- markup language for hypertext HTML documents;
- PHP programming language;
- MySQL database management environment;
- Apache web server;
- OpenServer package.

The system of automatic filling of content has the main purposes: the project is created to automate the collection and publication of unique content online resource.

User requirements:

External users – User:

1. Two-way communication with the administrator via the email contact form.
2. Getting information about the actual content (on the site).
3. Getting information about current content changes (on the site).
4. The user has the opportunity to post comments about the published content on the site.
5. Provide useful links to related sites.
6. Provide background information on related topics in the form of articles.
7. View the latest news of the site: information about the new features of the information system available to the user.

Internal admin users:

1. Add, remove, and edit content published by the system.
2. Change the content status (delay posting).
3. Database editing.
4. View content that is being processed or published with the participation of this information system.
5. Information exchange with external users via email correspondence.

Characterization of the object of computerization:

The user on the site will be able to view the content that was published by the system of automatic filling of content about current articles, to leave relevant comments on the received content, as well as to receive answers to questions via the contact form.

In turn, the administrator has the opportunity to customize the system to specific content topics, select sources of information, keywords to search, organize a template for the appearance of content design.

Functional requirements:

1. Authorization of users in the system: The system must have the function of authorizing the user and assigning him the appropriate role.
2. Maintenance of the working directory: A set of articles on specific topics, designed by the system. Content management tools should be provided in IP.
3. Ability to store information: The system must store the information and allow the administrator to manage it.
4. Creating conditions for online communication of users: The system should allow users to communicate in the mode of email correspondence.

Non-functional requirements:

1. Perception

- It takes 1 hour for ordinary users to learn application tools and 20 minutes for experienced users.
- The system response time for normal requests should not exceed 1 second and for more requests that are complex 20 seconds.
- The application presentation interface must be intuitive to the user and require no further training.

2. Reliability

- Availability – the time required for system maintenance should not exceed 1% of the total operating time.
- Average continuous working time is 20 working days.
- The maximum rate of errors and defects in the system operation is 1 error per 1000 user requests.

3. Productivity

The system must support a minimum of 100 concurrent users associated with a shared database.

4. Ability to operate

- Scaling – the system should be able to increase capacity (productivity), with the increase of users in such a way that it does not negatively affect its performance.

- Version Updates – Updates should be updated automatically depending on the preferences of the users and the expansion of the list of scheduled content.

The CMS WordPress platform and the PHP programming language [9] were chosen to implement the project.

Analysis of functional requirements allowed us to distinguish the following entities that will provide the implementation of the software system. In Fig. 3 presented a diagram of the classes of the system controller level.

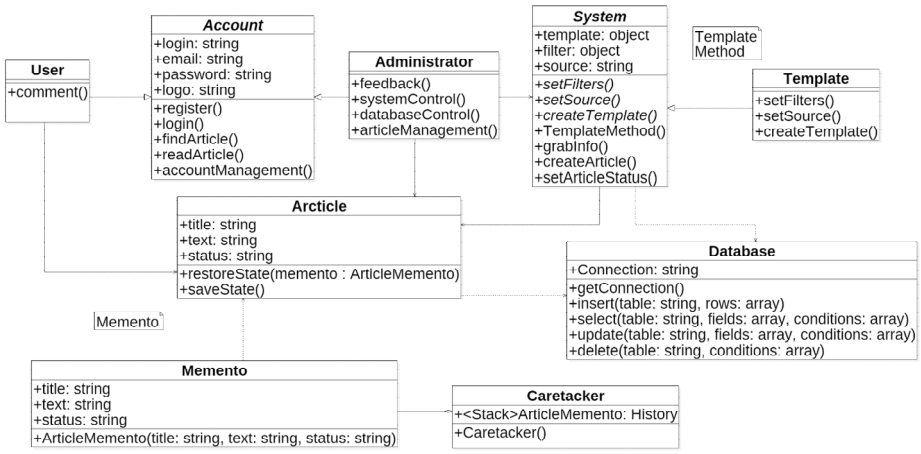

Fig. 3. Class diagram

The following classes can be distinguished in this figure:

- User – user class. The class has the following methods: comment - for commenting on a newly created article.
- Account – a class for saving user data. The class has the following methods: register – for user registration, login – authorization on the news site, findArticle, readArticle, accountManagement – editing user account data.
- Administrator – the site admin class. The class has the following methods: feedback – response to user messages, systemControl – entering and editing the automatic content filling system, databaseControl – receiving and editing database information, articleManagement – viewing and editing articles received by the automatic content filling system.
- System – a class of automatic content filling system. The class has the following methods: setFilters – installing and editing content filters, setSource – installing and editing an RSS feed of a donor site, createTemplate – creating an article template from an RSS feed to create articles, grabInfo – parsing content from a donor site, createArticle, setArticleStatus.
- Article – a class of articles on the automatic content filling system.

Thus, this system implements the functionality of content parsing from the donor site, filtering information, saving data in the form of articles organized on the WordPress platform.

The implemented MySQL database consists of 12 tables that contain all the data for the program. The bulk of these tables were created and maintained automatically by CMS WordPress, so only those that use SANC will be considered. The database is named wp-auto.

The structure of the database is shown in Fig. 4.

Consequently, the database, following the requirements of the relational model, provides the storage and collective access to the information of the autofill system and CMS WordPress data. The database consists of 12 tables. The main ones are wp-posts, wp-postmeta, wp-terms, wp-options.

Design and implementation of algorithms for system operation

The main modules of the system are the Parsing and Storage module in the CMS WordPress database.

User activity:

- when logging in to the site, the user can log in under the rights of the user or administrator, if he has such access;
- the user is logged in as a user, he or she can search for articles in the list available;
- the user opens the article and reads the information;
- the user has the opportunity to leave a comment under the article;
- the user also can manage their account data;
- the user is logged on as an administrator, besides all features of the user under the rights of the user, he/she is additionally able to control the system of automatic filling of content;
- the administrator can edit the information received for the article from the automatic content filling system;
- admin can set the status for the article (delay posting);
- the administrator has the opportunity to customize an article template that copies the auto-fill system;
- the administrator can add, delete and edit sources of information from which the auto-fill system copies the content.

The implementation of the activity of the system provides for the interaction of models Account, Administrator, User, Article, and System (Fig. 5). The main methods used in this process are System – for parsing content from RSS feeds into the system, Template – method of information storage template, Article – a method that saves filtered information in the form of CMS WordPress article.

Thus, the algorithms of the basic processes of content filling automation were considered and the interaction of system classes during the processes of parsing, filtering and information storage was analyzed.

Content Filling Automation is a plugin based on CMS WordPress and created in PHP – it has configuration files that spell out the domain name or path to the files. Before transferring the system to another hosting, you must save the location of all additional libraries to the plugin. The system uses CMS WordPress, so you need to

move the plugin to a special plugin folder for the system to work properly wp-content/plugins/.

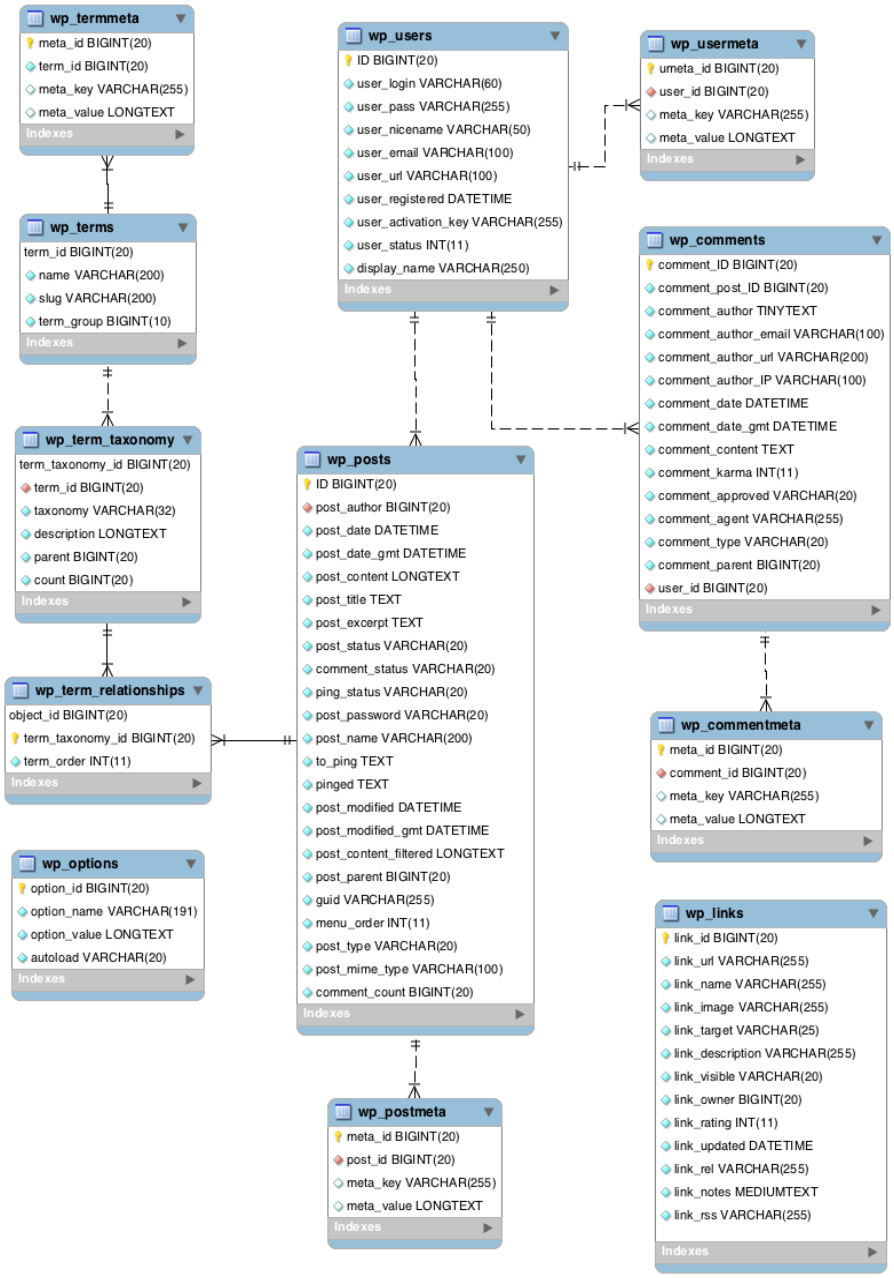

Fig. 4. Structural diagram of the database

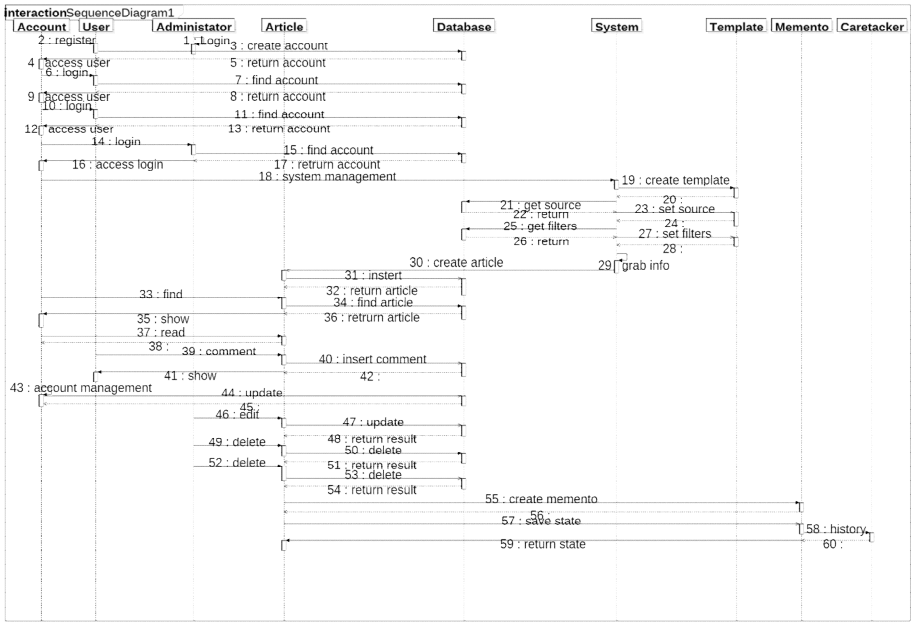

Fig. 5. Sequence diagram for autofill activity

The system does not need to create a copy of the database; instead, it will add the necessary fields and database data to MySQL on WordPress on the new hosting.

Thus, the system developed does not require specialized hardware, additional configuration, and deployment tools beyond the standard for such plug-ins.

In Fig. 6 shown a diagram of system deployment. The diagram shows that this system has three nodes: the program, the interface and the user.

When loading the start page the user has the opportunity to select any of the suggested menu items (Fig. 7).

The main menu of the site displays the main categories of articles that were generated by the system. From the main menu, the user can navigate to a specific category of articles of interest.

In the top menu, the user can change the language of the site and go to the social networks where the site is registered (see Fig. 8).

To get started with the automatic content filling system, you must log in to the site and log in to the WordPress admin panel, and the site administrator must fill in the login information at wp-auto/admin/.

After authorization from the side menu, you need to find the item “Plugins” and go to the page of installed WordPress plugins (Fig. 9), then find the “PRJ-Parser” on the installed plugins page and click the Activate activation link (Fig. 10).

After the plugin is successfully activated, a new item of the automatic content filling system “PRJ-Parser” will appear in the left menu (Fig. 11).

When selecting the PRJ-Parser menu item, the site administrator will be taken to the main page of the automatic content filling system, where he will be able to view

the existing list of feeds registered to the systems when they were the last read and when the next content reading is scheduled. The administrator can also add his RSS feed to read it, delete the selected RSS feeds and related articles, delete only the selected RSS feeds, delete the articles related to the selected RSS feeds (Fig. 12).

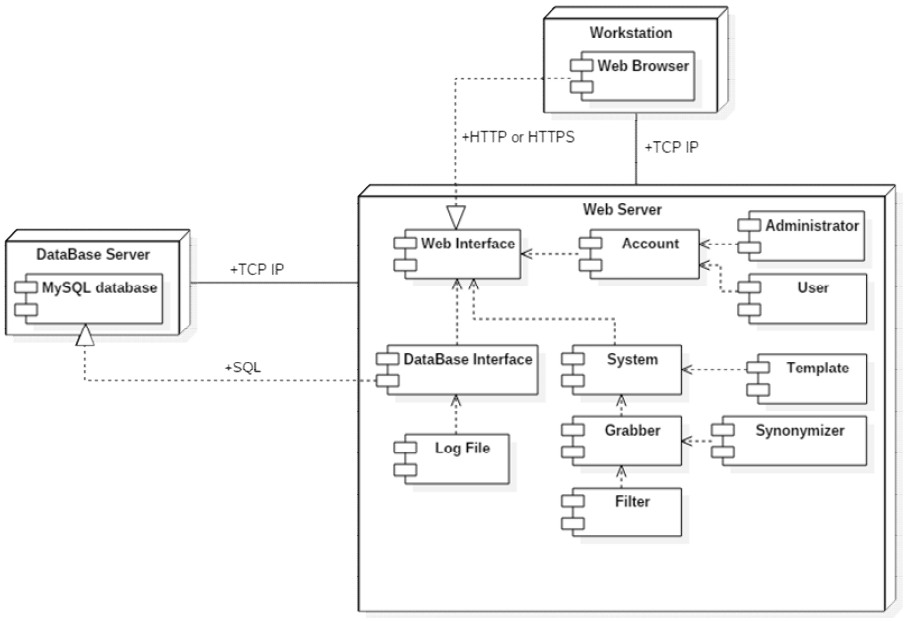

Fig. 6. System Deployment Diagram

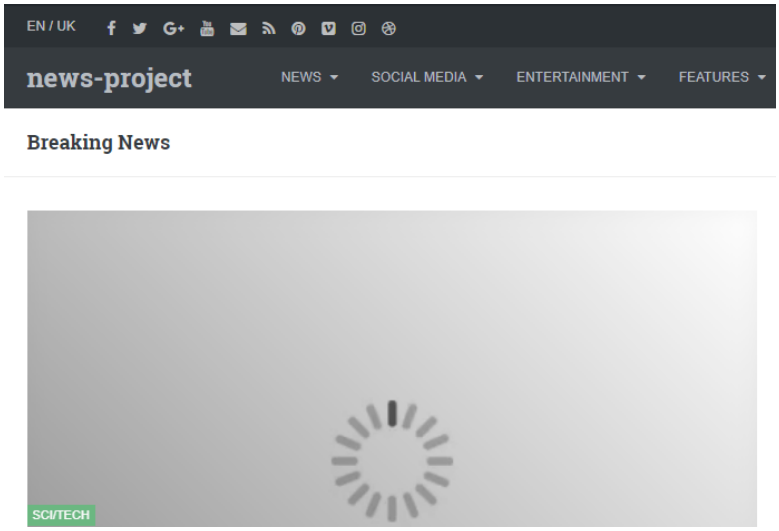

Fig. 7. The main menu of the site

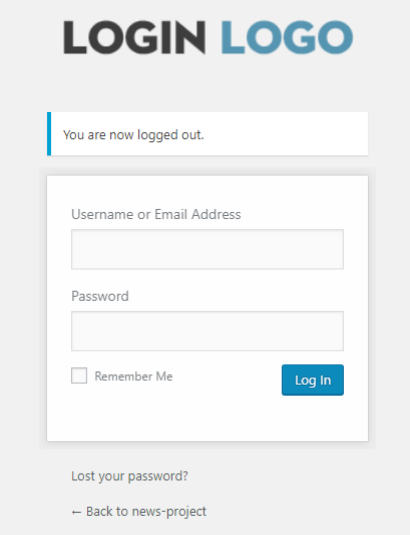

Fig. 8. Authorization to the admin panel.

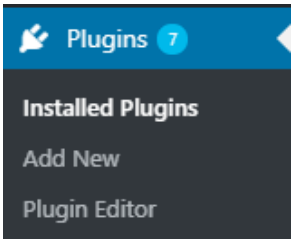

Fig. 9. Plugins menu item

Fig. 10. Activate the plugin in the WordPress administration panel

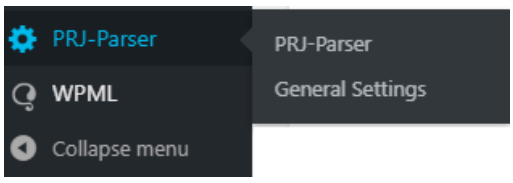

Fig. 11. The plugin is in the administration menu list

After entering the new RSS feed address and clicking the “Add a new feed” button, the administrator will go to the fine-tuning page of the new RSS feed (Fig. 12).

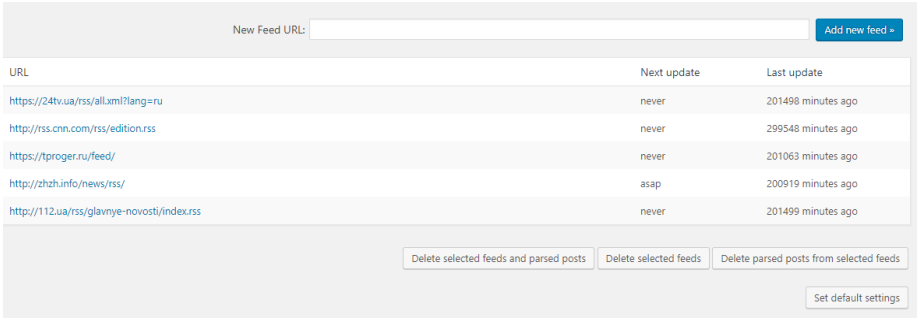

Fig. 12. The homepage of the automatic content filling system

On this page, an administrator can to:

- View and edit the name of the new RSS feed that will be displayed on the system homepage in the feed table;
- View URL of future RSS feed;
- Add categories to which articles from this RSS feed will be published;
- Enable universal reading mode by going to the full article page with additional filters;
- Specify on whose behalf the articles of this RSS feed will be published.
- Enable article tagging;
- Add your tags in the article while reading the current RSS feed;
- Activate the duplicate article check function;
- Activate the automatic reading function for a specified period;
- Specify the number of articles that will be published in a single RSS feed;
- Specify the status of articles to be read. Articles can be published immediately, left for review, saved to draft or hidden;
- Allow or deny users comments;
- Specify what date to use when publishing a read article; The date of publication may be the original date from the donor site or the date of reading by the automatic content filling system;
- Specify where to insert the article attachment. Attachments may be placed at the top of the article, bottom or not at all;
- Activate the thumbnail generation feature of the article. The thumbnail may be generated from the first image in the article, from the last image, read from the donor site or not generated at all;
- Activate the mandatory feature of the article thumbnail. If there is no article thumbnail, then the article will be removed from the site;
- Activate the article coding conversion feature. If RSS feed encoding is different from admin site encoding, then RSS feed will be converted to UTF-8 encoding;

- Enable local image saving from articles. If this feature is not activated, images will be sent to the donor site;
- Determine the range of minutes through which read articles will be randomly published in a single reading;
- Specify the source at the end of the article where the article was read;
- Insert the source at the end of the lifts;
- Activate the function of removing links from words and make links simple;
- Insert video resources into reading articles.

After saving the RSS feed settings, it will appear in the RSS feeds table on the main system page (Fig. 13).

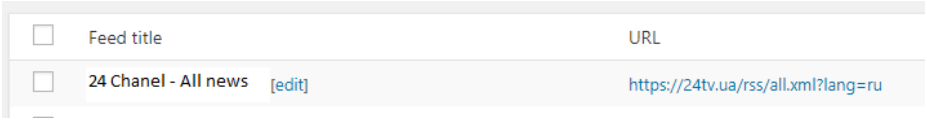

Fig. 13. New RSS feed

The site administrator can change the selected settings at any time by clicking on the "edit" link (see Figure 13), and then a fine-tuning page with the last saved settings for a particular RSS feed will be opened (see Figure 12).

To read RSS feeds on existing articles, select the required RSS feeds and click the Start Parser button (Fig. 14). Reading will be done according to the settings of each of the RSS feeds marked.

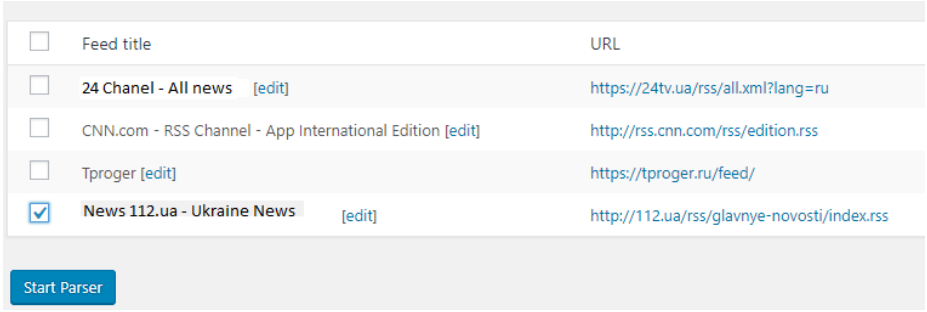

Fig. 14. Reading RRS feeds

After the action is completed, the system will notify the successful reading and articles will be generated with the corresponding status, which was specified in the RSS feed settings.

Automatic filling systems also have a settings page (Fig. 15), where the site administrator can configure the automatic start of the plugin after a certain period of time, specify mandatory to leave a link to the source at the end of the article, the path

to the full function reading libraries and the ability to disable check for duplicate articles.

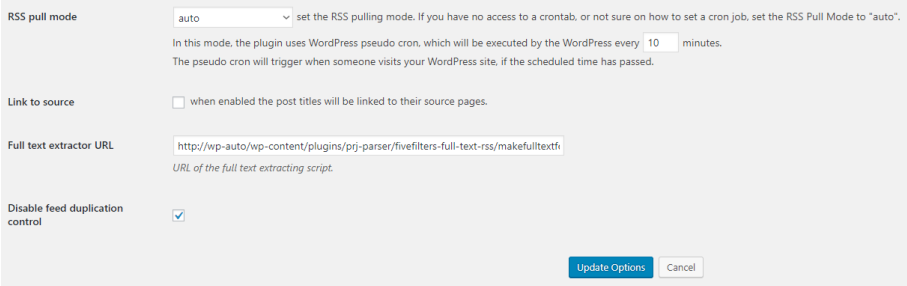

Fig. 15. AutoFill Setup Page

4 Conclusions

In this paper, a database was designed and implemented following the requirements of a relational model, which ensures the storage and collective access to information of the autofill system and CMS WordPress data. Algorithms of system functioning have been developed, the order of interaction of classes during program code execution has been determined, and the application has been implemented. Most of the application is intended for the site administrator and has no user interface. The administrator has a plug-in configuration interface for the plug-in automation system, an interface for viewing and managing RSS feeds, as well as an interface for configuring RSS feeds.

In the future, this system can be improved by introducing new functionality and improving the algorithm for reading data.

References

1. Averianov, A.Ye.: Zastosuvannia informatsiinoi systemy upravlinnia kontentom veb-resursu dlia vedennia elektronnoi komertsii (Implementation of web content management information system for e-commerce). Formuvannia rynkovykh vidnosyn v Ukraini 9(160), 171–174 (2014)
2. Berko, A.Yu., Dorosh, V.M., Chirun, L.V.: Intelektualna systema upravlinnia kontentom saitiv elektronnoho biznesu (Intelligent Content Management System for E-Business Websites). Visnyk Natsionalnoho universytetu "Lvivska politekhnika" **715**, 13–23 (2011)
3. Chyrun, L.V.; Vysotska, V.A.: Zastosuvannia kontent-analizu tekstovoi informatsii v systemakh elektronnoi komertsii (Application of content analysis of text information in e-commerce systems). Visnyk Natsionalnoho universytetu "Lvivska politekhnika" **689**, 332–347 (2010)
4. Korobchinsky, M.V., Chirun, L.B., Vysotska, V.A., Kondratiev, E.A.: Osoblyvosti formuvannia ta analizu kontentu internet-hazety muzychnykh novyn (Of content formation and analysis features in online newspaper of music news). Radio Electronics, Computer Science, Control 4, 139–150 (2017)

5. MacDonald, M.: WordPress: The Missing Manual. O'Reilly Media, Sebastopol (2014)
6. Morozov, A.V., Kuzmenko, O.V., Danilchenko, A.A.: Osnovy veb-rozrobky dlia Wordpress ta Yii (Web Development Basics for Wordpress and Yii). ZhDU, Zhytomyr (2018)
7. Sanders, W.: Learning PHP Design Patterns. O'Reilly Media, Sebastopol (2013)
8. Vysotska, V.A., Chirun, L.B., Chirun, L.V.: Unifikovani metody opratsiuvannia informatsiinykh resursiv u systemakh elektronnoi kontent-komertsii (Unified methods of processing information resources in systems of electronic content commerce). Naukovi pratsi [Chornomorskoho derzhavnoho universytetu imeni Petra Mohyly] **213**(201), 13–23 (2013)
9. Williams, B.: WordPress Plugin Development. Wiley Publishing Inc., Indianapolis (2011)