Web Scraper Project Brief (v1)

Prepared by Dr Mohamed Mahyoub Version 1: 29/04/2023

INDUSTRY 4.0 CONSULTANTS

Introduction

Before reading please review the draft mock-up and notes guide here. The project at hand entails the construction of a responsive web application that will function as a web scraper. This application is exclusively accessible through a web browser on a mobile or desktop device, thereby negating the need for publication on iOS or Android app stores. The essence of this project is to establish a web scraping mechanism that can extract content from third-party websites. This mechanism enables the admin user to oversee and modify the content obtained from the scraped pages before it is published and submitted to a third-party WordPress RESTful API.

Workflow

The proposed workflow commences with the user inserting a domain name into the system. Subsequently, our system maps the entered domain using a third-party tool or our inbuilt code to generate a sitemap. Users have the flexibility to modify the sitemap for each domain, either by deleting or adding a new URL. The web scraper then pulls content from every URL listed in the sitemap, enabling the user to moderate and edit the obtained article. This process involves transforming the article title into an article slug, utilising a content editor to modify the content, as well as editing SEO keywords into custom SEO titles, custom SEO descriptions, and meta tags. Users can also select a predefined category that aligns with the target WordPress categories, and then post excerpts.

Once the content editing process is complete, users can save their work as a draft, allowing them to make further edits at a later time before publishing. However, once the 'publish' button is clicked, the content becomes uneditable through the web scraper system and is transmitted to the WordPress API. Post-publication, the 'edit' button will only open a new tab redirecting users to the editing page in the WordPress system.

1) **Layout and Interface -** **Dashboard Area**

The layout of the user interface consists of a menu section on the left-hand side, featuring various menu items. The centre of the interface presents content that changes based on the link selected in the menu. The first item in the menu is the dashboard. On selecting it, users are provided with shortcuts and buttons leading them to different areas within the system. The dashboard also provides statistical insights in four boxes, along with two columns. One column displays the most recent publications sent to WordPress, while the other showcases the most recent content scraped.

2) **Edit Content & SEO -** **Pending Approval Section**

The second menu item is the 'Pending Approval' section, which contains all the content scraped from the sitemap URLs. This section contains a table displaying all the scraped content, which can be filtered using a search box, allowing users to find specific content within the table. The table provides details such as the article name, the source URL where it was scraped from, and the time and date of the scraping. Users can view the source article, edit the scraped content via the content editor, or choose to ignore the content from that URL. Ignored content is added to an 'ignored list,' ensuring the scraper does not revisit it. Above the table, there's a button to add new articles, leading to the content editor, which initially appears empty. When adding a new article, the user can specify the content details and publish.

The content editor form is a crucial part of the system that allows users to moderate and modify scraped content before it gets published. This form is accessible when the user chooses to 'edit' a scraped content from the 'Pending Approval' page, or if they decide to 'Add New Articles'.

Here are the main fields that would be present in the content editor form:

- **Article Slug:** The user can create a URL-friendly version of the article's title, also known as the slug. This usually consists of lowercase letters, numbers, and hyphens.
- **Article Content:** This is a text editor area that provides the users with the full range of editing options. Here, the users can adjust the formatting of the text, add bullet points, hyperlinks, images, etc. If the user is editing an existing article, this field is pre-populated with the scraped content.
- **SEO Title:** This field allows users to specify a title that will be used specifically for SEO purposes. The SEO title is the title that appears in the search engine results pages (SERPs) and is crucial for attracting clicks from potential visitors.

- **SEO Description:** This is a text input field that allows the user to write a brief summary of the article. This description appears beneath the title in SERPs, providing potential visitors with a snapshot of what the article entails.
- **Meta Tags:** This is where the user can enter relevant keywords and phrases that best describe the content of the article. These tags help search engines understand and categorise the page content, thereby improving its visibility.
- **Category:** This dropdown field allows the user to choose a category that matches the article's content. The category should be in sync with the WordPress categories on the destination website.
- **Post Excerpt:** This is a short summary or snippet of the article that can be used as a preview on the website.

Once all fields are completed, the user can save the article as a draft and come back to it later or publish it right away. When the user presses 'Publish', the article content is sent to the WordPress RESTful API and published on the corresponding WordPress site.
The content editor form ensures the articles are correctly formatted, SEO-optimised, and correctly categorised before they are published, which is integral for site organisation and search engine visibility.

3) **Published Content -** **Published Articles Section**

In the third link of the left-hand side menu table, there's a section showcasing all the published articles. This section contains a table that displays the article name, the original source URL, the destination URL where it was published, the scraping date and time, and an action column. The action column directs the user to the URL where the content was published. The 'edit' button in this column doesn't open the content editor within the system but instead redirects the user to the WordPress editor where the content was published. This table can be filtered, searched, and supports pagination. It also provides the user with the option to select multiple rows and perform bulk actions.

4) **Domain Sitemaps -** **Sitemaps Section**

The fourth menu item pertains to the sitemaps section. This section displays all the sitemaps created for the domains imported by the user in a table format. This table can be filtered and searched, and also allows for bulk actions on selected rows. The table columns include the site name, the number of URLs in the sitemap, the number of active URLs for the sitemap, the number of ignored URLs for the sitemap, the last time the sitemap was regenerated, and the last time it

was scanned. The final column includes action buttons allowing the user to view the entire URL list in the sitemap, delete the sitemap, or regenerate it.

Clicking on the 'view sitemap' action opens a sub-page with a breadcrumb navigation aid situated above the table. The sub-page contains a table of URLs in the sitemap. This table shows the URL pertaining to the sitemap, the date and time it was added, its current status (whether it's been ignored or remains active), the last time it was scanned, and the target HTML tag ID used to scrape content from the webpage. This ID, for instance, might correspond to a blog post's container ID if we only wish to scrape that specific part of the page. The final column provides action buttons allowing the user to view the URL in the targeted domain, ignore the URL, or edit it.

In the view sitemap subpage the user should get ‘Add new’ button above the URL table and breadcrumb. Or be able to edit if they click edit in the action column. If the user clicks the 'edit' button for a particular URL or decides to add a new URL to the sitemap, an add/edit form will popup from the right hand side as shown in the mockup design. Here are the form's essential fields:

- **Domain Name:** This dropdown field is pre-populated with the domain names that are already registered in the system. If the user is editing a URL, the domain name can be pre-selected based on the sitemap that the URL belongs to.
- **URL**: This text input field is where the user enters the URL to be scrapped. If the user is editing an existing URL, this field is pre-populated with that URL.
- **Target Area in DOM:** This is an advanced setting where the user specifies the ID of the HTML container in the web pages of the URL from which content is to be scrapped.
- **Scraping Frequency**: This dropdown field allows the user to set how frequently the URL should be scrapped. Options could range from every hour, every day, every week, or manual selection.

After filling out these fields, the user can save the form. This will add the new URL to the sitemap or update the existing URL. The web scraping system will then use this information for its scraping operation, ensuring that the right content is scraped at the right intervals.

5) **Domains List -** **My Domains Section**

The fifth menu item, found on the left-hand side navigation panel, is the 'My Domains' list. This section features a table listing the domains for which the user wishes to generate a sitemap and scrape each URL contained within the sitemap. The table, filterable with a search function, displays the domain name, the domain type, the target area ID (which identifies the section in the domain's DOM file to be targeted), the frequency at which the sitemap for this domain should be generated, the last time the sitemap was scanned, and options to edit or delete the domain.

The domain editing form is accessible via the 'My Domains' list in the left-hand side navigation panel. Here, the user can modify existing domains or add new ones to the system.

Upon clicking the 'edit' action next to a domain in the 'My Domains' table, or upon clicking the 'Add new domain' button, the user is taken to the domain editing form (pop-up from right hand side as shown in the mockup designs). The form includes the following fields:

- **Domain Name:** This is the full URL of the domain the user wants to add to the system. This field is a text input field.
- **Domain Type:** This dropdown field allows the user to choose from a list of predefined types. These types could be defined based on the platform used for the domain (like WordPress, Joomla, Drupal, etc.), or they could be defined based on the structure of the website (like Blog, News, E-commerce, etc.).
- **Target Area ID:** This is an advanced setting. Here, the user specifies the ID of the HTML container in the web pages of the domain, from which content is to be scrapped. This field ensures that only the relevant content is scraped and not the entire webpage.
- **Sitemap Generation Frequency:** This is a dropdown field where the user can choose how often the sitemap for the domain should be automatically generated by the system. Options might include Daily, Weekly, Monthly, or Manual.

On filling out all these fields, the user can save the form by clicking on the 'Save' button. Upon saving, the domain is either updated (if it already existed in the system) or added to the 'My Domains' list. This action triggers the system to generate a sitemap for the domain based on the frequency set by the user, and the web scraping begins as per the established workflow.

The domain editing form is crucial as it gives the user control over the domains to be scraped and the frequency of scraping, ensuring the process is tailored to their specific requirements.

6) **Blacklisted Pages -** **Ignored URLs Section**

Finally, the last menu item is the 'Ignored URLs' list. This table displays the URL, the date and time it was added to the system and when it was marked to be ignored, the last time it was scanned or scraped, and the target area it was intended to target. The action column provides the option to view this URL or to unignore it, thereby reinstating it back into the sitemaps as an active URL to be scrapped. This table, like others in the system, allows users to select a row or multiple rows and perform bulk actions.

7) **User Settings -** **Account & Settings Section**

Alongside these functionalities, the bottom left corner of the menu provides a section for users to modify their account settings. This section allows users to update personal details like their first name, last name, email, and password. It also presents settings for the targeted WordPress API and lets them establish different frequency levels for the domains. This section can also incorporate other settings as needed.