# Otomatika.

# News Collection

Process Definition Document - **Otomatika News Collection**

**DOCUMENT VERSION HISTORY**

**Version : 1.0.0**
**Author : Gundo Sifhufhi**
**Date: 01 March 2023**

**SIGNED OFF BY**

**Name: Gundo**
**Function: All**
**Responsibility: Developer**

**CONTRIBUTORS**

**Name: Gundo**
**Function: All**
**Responsibility: Developer**

**Process Overview:**

This process involves collecting the latest news from The New York Times website and storing them in an excel file for further analysis. The following steps are involved in the process:

1. Open The New York Times website by following the link (Ideally we head straight to www.nytimes.com/search)
2. Enter a phrase in the search field
3. On the result page, apply the following filters:
4. a. Select a news category or section
5. b. Choose the latest news
6. c. Get the values: title, date, and description.
7. Store the following information in an excel file:
8. a. Title
9. b. Date
10. c. Description (if available)
11. d. Picture filename
12. e. Count of search phrases in the title and description
13. f. True or False, depending on whether the title or description contains any amount of money
14. Download the news picture and specify the file name in the excel file
15. Repeat steps 4-5 for all news that fall within the required time period.

Process Steps:
Step 1: Open The New York Times website

- Navigate to the website using the following link: //www.nytimes.com/search

Step 2: Enter a phrase in the search field

- Enter the required search phrase in the search field provided on the website.

Step 3: Apply filters and get required information

- Select the required news category or section from the available options
- Choose the latest news
- Get the following information: title, date, and description.

Step 4: Store information in an excel file

- Create an excel file with the following columns: Title, Date, Description, Picture filename, Search phrase count, Contains money
- Store the values obtained in Step 3 in the respective columns
- Count the number of times the search phrase appears in the title and description and store it in the "Search phrase count" column.
- Check if the title or description contains any amount of money and store True or False in the "Contains money" column.

Step 5: Download news picture

- Download the news picture using the link provided on the website
- Specify the file name in the "Picture filename" column in the excel file

Step 6: Repeat for all news within the required time period

- Repeat Steps 3-5 for all news that fall within the required time period.

Process Maintenance:

- This process needs to be kept up to date with any changes in the website structure or procedure.
- Any changes made to the involved systems or procedures must be updated in the Process Definition Document.
- All changes need to be signed off by all parties involved in the process.

Current process analysis:

At this time, someone within the company is manually scraping articles from the New York Times website and saving them to an Excel spreadsheet. The process involves logging into the website and manually searching for articles that match the desired criteria. The employee then copies and pastes the relevant information from each article into the spreadsheet.

High-level description:
The process is used to scrape articles from the New York Times website and save them to an Excel spreadsheet. It is run once a week on Tuesdays at 10 am by an employee within the company. The process involves logging into the website, searching for articles that match the desired criteria, and copying and pasting the relevant information into the spreadsheet.

Systems involved:

- New York Times website: Used to access and search for articles. The operator needs to have a valid login and password to access the website.
- Microsoft Excel: Used to save the information from the articles. The operator needs to have access to the spreadsheet and the ability to add new rows and columns.

Process flow:
Here is a description of the process in the form of a flow chart:

1. Log in to the New York Times website using valid credentials.
2. Search for articles that match the desired criteria.
3. Open each relevant article.
4. Copy the desired information from the article.
5. Paste the information into the Excel spreadsheet.
6. Repeat steps 2-5 for all relevant articles.
7. Save the Excel spreadsheet.

Detailed steps:

1. Log in to the New York Times website using valid credentials.
2. Navigate to the New York Times website.
3. Enter valid login credentials and click "Sign In".
4. Search for articles that match the desired criteria.
5. Enter search terms into the search bar.
6. Select any desired filters (e.g. date range, section, etc.).
7. Click "Search".
8. Open each relevant article.
9. Click on the title of each relevant article to open it.
10. Copy the desired information from the article.
11. Highlight the desired information in the article.
12. Right-click and select "Copy".
13. Paste the information into the Excel spreadsheet.
14. Navigate to the relevant row and column in the spreadsheet.
15. Right-click and select "Paste".
16. Repeat steps 2-5 for all relevant articles.
17. Continue searching for and opening relevant articles until all desired information has been copied and pasted.
18. Save the Excel spreadsheet.
19. Click "File" > "Save" to save the spreadsheet.

Possible exceptions:

Logic exceptions:

- No articles match the desired criteria: The operator will need to adjust the search terms and filters and try again.
- Relevant information is missing from an article: The operator will need to skip that article and move on to the next one.
- Business rules dictate a maximum number of articles to scrape: The operator will need to stop scraping once the maximum number of articles has been reached.

System exceptions:

- Website is down or inaccessible: The operator will need to wait for the website to come back online or contact technical support for assistance.
- Excel crashes or becomes unresponsive: The operator will need to save any unsaved work and restart Excel.

Conclusion:

The Process Definition Document provides a clear, explicit, and shared understanding of the Otomatika News Collection process. It outlines the steps involved in collecting the latest news from The New York Times website, storing them in an excel file, and analyzing them. This document needs to be updated and maintained throughout the entire lifecycle of the automation process.