Google Scholar Citation Scraper & Validator (REFGoogleScholar)

An automated solution built on the **Robotic Enterprise Framework** to streamline academic data collection. The robot autonomously navigates to a researcher's Google Scholar profile, extracts high-level metrics (h-index, i10-index), and scrapes the complete list of publications. It then individually processes each paper to validate citation counts, ensuring accurate, granular reporting without manual effort.

Contents

1. Introduction

1.1 Purpose

The Process Definition Document outlines the business process chosen for automation. The document describes the sequence of actions performed as part of the business process, the conditions and rules of the process prior to automation (AS IS) as well as the new sequence of actions that the process will follow as a result of preparation for automation (TO BE).

The PDD is a communication document between:

- The RPA Business Analyst and the SME/Process Owner. The goal is to ensure that the RPA Business Analyst has the correct understanding of the process and has represented it accurately.
- The RPA Business Analyst and the Development team (represented by the Solution Architect and RPA Development Lead). The goal is to ensure that the process is documented appropriately and to a sufficient level of detail so that the Solution Architect can then create the solution based on the PDD content.

1.2 Objectives

The business objectives and benefits expected by the Business Process Owner after automation of the selected business process are:

- **Efficiency:** Reduce the time required to compile citation metrics by approximately **80%** compared to manual navigation and data entry.
- **Accuracy:** Ensure **100% data accuracy** by eliminating human error associated with manual copy-pasting of paper titles and citation counts.
- **Scalability:** Enable the processing of researchers with extensive publication lists (500+ papers) without increasing the workload on human staff.
- **Monitoring:** Provide detailed logs and transaction statuses via UiPath Orchestrator to track successful data extractions and identify failed URLs.

1.3 Key Contacts

Add here any stakeholders that need to be informed or to approve changes to the process:

Role	**Name**	**Contact Details** (email, phone number)	**Notes**

1.4 Minimum Pre-requisites for the Automation

a) Filled in Process Definition Document
b) Test Data to support development
c) User access and user accounts creations (licenses, permissions, restrictions to create accounts for robots)
d) Credentials (user ID and password) required to logon to machines and applications

2. AS IS Process Description

In this section the Business Analyst will document the process. This section will serve as the starting point for the re-engineering and automation effort.

2.1 Process Overview

Section contains general information about the process before automation.

Item	Description/Answer
Process Full Name	**Google Scholar Citation Scraper & Validator**
Process Area	**Academic Research / Data Analysis**
Department	**Research Administration**
Short Description (operation, activity, outcome)	The Process Owner manually navigates to Google Scholar, searches for a specific researcher, copies their high-level metrics (h-index) and the list of publications into an Excel sheet. They then manually click each individual paper link to verify the specific "Cited By" count, ensuring the data is not cached or estimated.
Role(s) required in applications to perform the process	**Standard User** (Requires a valid Google Account to access full profile details).
Process schedule and frequency	**Ad-hoc / Monthly** (Typically run when grant reports or academic performance reviews are due).
Number of times the process is ran by selected frequency	**1 per Researcher** (Repeated for every faculty member required).
Process execution time	**~45 - 60 Minutes per Researcher** (Based on a profile with ~50-100 papers. Manual navigation and data entry is slow).
Process Restrictions	**1. Rate Limiting:** The user cannot click too fast or Google Scholar presents a CAPTCHA. **2. Access:** Must be performed on a machine with open internet access (no firewall blocking Google). **3. Accuracy:** Manual copy-pasting is prone to transcription errors.
Peak Period (s)	End of Academic Year / Grant Deadlines
Peak Volume Approximate increase	**~200% increase** in requests during reporting seasons.
Number of persons performing the process	**1** (Research Assistant or the Researcher themselves).
Expected Volume increase during next periods	**10-15%** (As researchers publish more papers, the list grows longer to process).
Percentage Un-handled exceptions	**~10%** (Human error in data entry or skipping papers by mistake).
Input data description	**1. Researcher Name/URL** (Source to scrape). **2. Target Excel Template** (Where data is manually pasted).
Output Data description	**Completed Excel Report** containing the verified citation count for every published paper.

2.2 Applications Used

The table includes a comprehensive list of all the applications that are used as part of the process to be automated to perform the given actions in the flow.

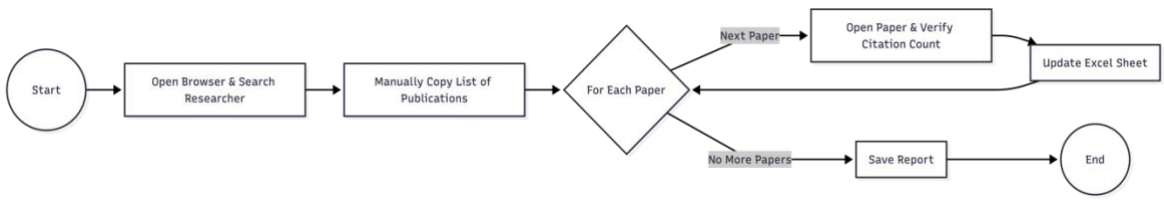
Application Name	Version	Application Language	Thin/Think Client	Environment/ Access method	Comments
Google Chrome	Latest (Auto-updates)	English (US)	Thin Client	Web / Public Internet	Used to access scholar.google.com. The UI structure of Google Scholar may change slightly over time, requiring selector updates.
Microsoft Excel	Office 365 / 2016+	English (US)	Think Client	Desktop / Local Install	Used to store the extracted citation data. *Note:* Regional settings (comma vs. dot decimals) must be consistent with the robot's environment.

2.3 AS IS Process Map

This section contains various process maps contributing to a better understanding of how the process is performed pre-automation.

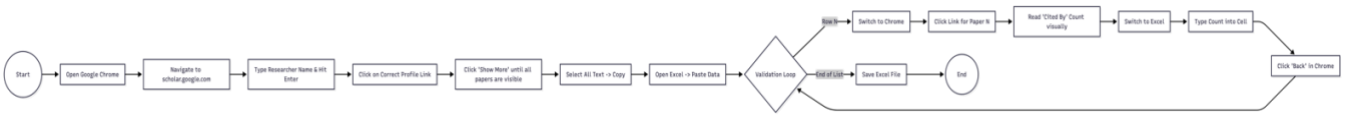
2.3.1 High Level Process Map

This section is useful for the Business Analyst in presentations and discussions with management to underline areas of weakness, inefficiency or to demonstrate which actions could be in scope for automation.



2.3.2 Detailed Level Process Map

This section describes the process at key-stroke level and is an essential part for the communication with the developers.



2.4 Process Statistics

High Level statistics

Processes	Windows	Actions	Mouse clicks	Keys pressed	Text entries	Hotkeys used	Time
2	2	~250	~210	~150	52	3	45 mins

Detailed statistics

Window name	Mouse clicks	Text entries	Key pressed
Google Chrome	110 (Navigation & Clicking Links)	1 (Researcher Name)	20 (Typing URL/Name)
Microsoft Excel	100 (Cell Selection)	51 (Paste + 50 Citation Counts)	130 (Typing numbers)

2.5 Detailed AS IS Process Actions

#Action	Input	Description	Details (Screen/Video Recording Index	Exception Handling	Possible Actions
1	Researcher Name	Open Google Chrome and navigate to scholar.google.com. Type the researcher's name.	Screen: Google Scholar Home	If internet is down, retry.	Type, Click
2	Profile Link	Identify the correct researcher from the search results (check affiliation) and click the profile link.	Screen: Search Results	If profile not found, verify spelling.	Click
3	Profile Page	Click "Show More" repeatedly until the full list of publications is loaded.	Screen: User Profile	If CAPTCHA appears, solve it manually.	Click, Scroll
4	Publication Data	Select the table of publications, Copy, and Paste it into the Excel Template.	Screen: Profile / Excel	If Paste format is broken, retry as "Text Only".	Copy, Paste
5	Excel List	**Loop for every paper:** Click the paper link in Chrome, read the "Cited by" count, switch to Excel, and update the cell.	Screen: Paper Detail / Excel	If link is broken, mark cell as "Error".	Clikc

{#sequenceLayout}

1. {sequenceTitle}	
{sequenceDescription}	Est. time: {sequence_execution_time}

{#actionLayout}

1.1 {actionTitle}	
{actionDescription}	Est. time: {action_execution_time}
{%actionImage}	{#action_metadata} Action: {action_type} {/action_metadata}

{/actionLayout}

{/sequenceLayout}

2.6 Input Data Description

The following table should contain details regarding the inputs that every action of the process takes.

#Action	Sample	Input Type	Location	Are inputs Natively Digital*?	Are the inputs Structured*?
1	“John Doe”	Text	Email Request / Ticket	Yes	Yes
2	https://scholar.google.com	URL	Web Browser	Yes	No (Search result vary)
3	Title, Year, Cited By	Text/Table	Web Browser (Profile)	Yes	Semi-Structured (HTML Table)
4	"Cited by 45"	Text/Number	Web Browser (Article Page)	Yes	Semi-Structured (HTML Element)
5	Report.xlsx	File	Local Desktop	Yes	Yes

** Native Digital: This is data that was originally created digitally e.g. excel, database or application reports etc. The non-native digital inputs are usually scanned images.*

** Structured Data: has a predictable format and exists in fixed fields (e.g. an excel cell or a field in a form) and is easily detectable via search algorithms.*

3 TO BE Process Description

In this section the proposed improvements to the process, actions to the process will be outlined as well as the actions proposed for automation and the type of robot required. **This will be cross-checked by the Solution Architect.**

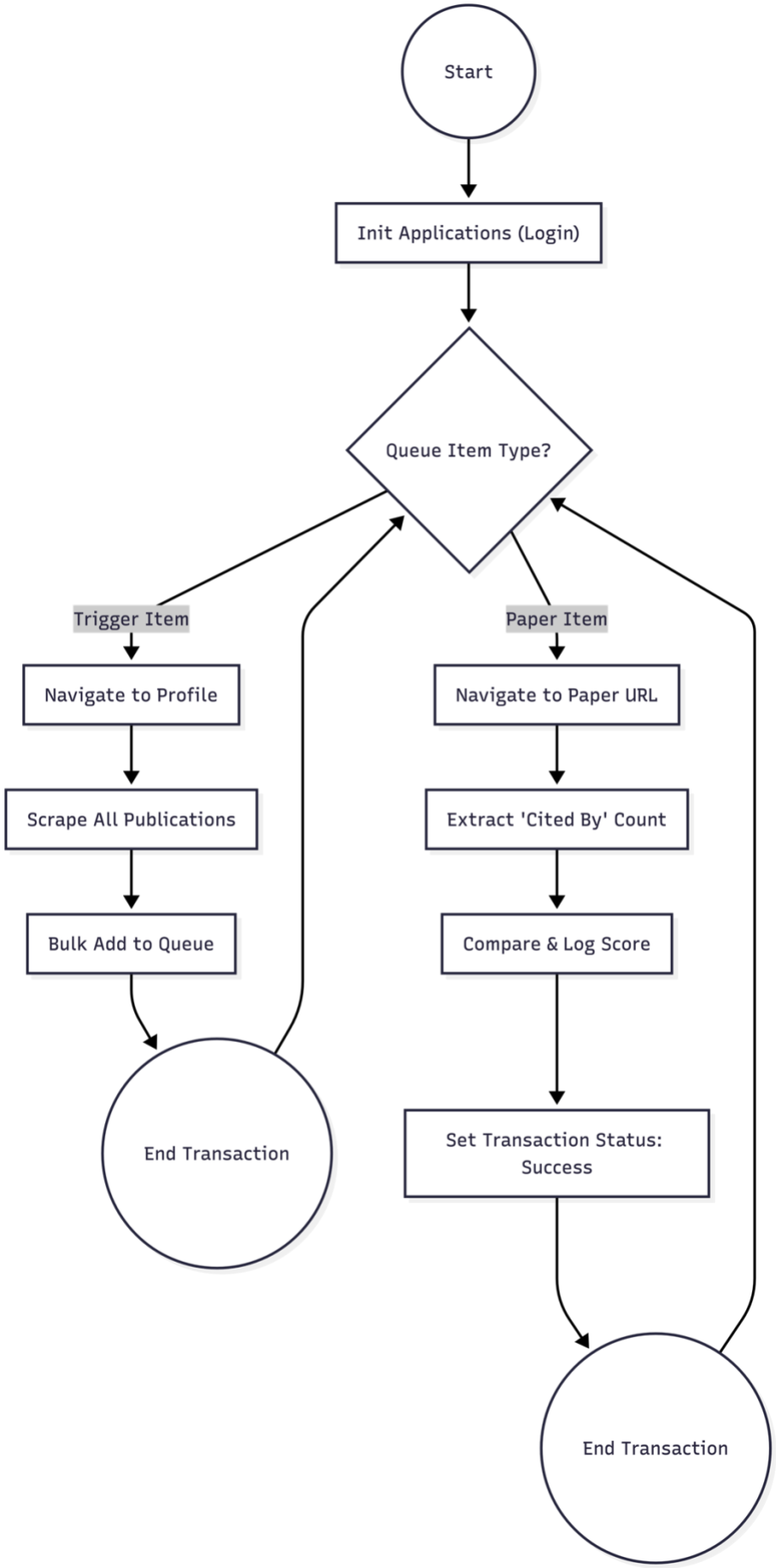
3.1 Detailed TO BE Process Map

A detailed process map of the process as it will look like post-automation will be outlined here.

Highlight Bot interventions/ To-Be automated actions with different legend/ icon (purple).

Mention below if process improvements were performed on the To-Be design and provide details.

Legend	Description
Direct Navigation	Instead of visually searching for a paper in a list, the bot navigates directly to the specific URL for each validation step.
Bulk Processing	The entire list of publications is scraped in one action (Producer), reducing the time spent loading the profile page repeatedly.
Orchestrator Queues	Decouples the "Search" phase from the "Validate" phase, allowing for better error handling and retry mechanisms.


Start
Init Applications (Login)
Queue Item Type?
Trigger Item
Navigate to Profile
Scrape All Publications
Bulk Add to Queue
End Transaction
Paper Item
Navigate to Paper URL
Extract 'Cited By' Count
Compare & Log Score
Set Transaction Status: Success
End Transaction

3.2 Parallel Initiatives

The table below will capture the proposed Business, Process or Application changes to be made in the near future that would impact the process at hand (if any).

Initiative Name	Process Action(s) where it is identified	Impact on current Automation Request	Expected Completion Date	Contact Person
Google Scholar UI Update	**Scraping / Navigation**	**High:** If Google changes the CSS selectors for the "Cited By" field or the Profile Table, the robot will fail and require maintenance.	Unknown	Google (External Vendor)
Authentication Policy Change	**Login (Init)**	**Medium:** If 2FA becomes mandatory for all Google Accounts, the login workflow will require an update to handle OTPs.	Unknown	IT Security

3.3 In Scope for RPA

The actions in scope for RPA should be listed below:

- **Login:** Securely logging into Google Scholar using credentials from Orchestrator.
- **Navigation:** Searching for a specific researcher profile or navigating to a specific paper URL.
- **Data Extraction:** Scraping the full table of publications (Title, Year, URL) and extracting the specific "Cited By" count from article pages.
- **Data Validation:** Parsing the extracted string (e.g., "Cited by 45" -> 45) to ensure it is a valid integer.
- **Queue Management:** Adding items to the Queue and updating their status (Success/Failed) based on processing results.

3.4 Out of Scope for RPA

The actions **out of scope** for RPA should be listed in the table below together with the reasoning.

Activity/Action	Reason for out of scope	Impact on the TO BE	Possible measures to be taken into consideration for future automation
Creating Google Accounts	Security / Complexity	The robot requires an existing account. It cannot register new users.	Automating account creation (highly restricted by Google).
Solving CAPTCHAs	Technical Constraint	If Google Scholar presents a "I am not a robot" challenge, the bot cannot solve it natively.	Integrate 3rd party API (e.g., 2Captcha) in a future phase.
Subjective Data Analysis	Cognitive	The robot cannot determine if a paper is "relevant" based on its abstract, only purely quantitative citation counts.	Use AI/ML models to analyze abstract text for relevance.

3.5 Exception Handling

The Business Process Owner and Business Analysts are expected to document below all the business exceptions identified in the automation process. Exceptions are of 2 types and both need to be addressed:

Known exceptions = previously encountered. A scenario is defined with clear actions and workarounds for each case.

Unknown = New situation that was not encountered before. It cannot be predicted and in case it happens it needs to be flagged and communicated to an authorized person for evaluation.

3.5.1 Known Business Exceptions

Details regarding how the robot should handle the exceptions.

Exception Name	Action	Parameters	Actions to be taken
Profile Not Found	Search Researcher	Researcher Name	Log a **Business Rule Exception**: "Researcher profile not found for [Name]". Mark transaction as Failed. Send email to admin.
Zero Citations	Extract Score	Citation Count	If the "Cited By" field is missing (new paper), default the score to **0**. Do not fail the transaction; mark as **Success** with a note.
Invalid URL	Navigate to Paper	Paper URL	If the specific paper link is broken (404 Error), Log a **Business Rule Exception**: "Invalid URL". Mark transaction as Failed. Proceed to next paper.

3.5.2 Unknown Business Exceptions

An umbrella rule that includes a notification needs to be designed for all other exceptions that could happen and cannot be anticipated.

Strategy: For any unhandled error (e.g., Browser Crash, Network Timeout, Selector Not Found):

1. **Capture Screenshot:** Save a screenshot of the desktop at the moment of failure.
2. **Retry Mechanism:** The robot will retry the transaction **2 times** (as defined in Config.xlsx).
3. **Final Failure:** If it fails after retries, mark the transaction as **System Exception**.
4. **Notification:** (Optional) Send an email to the support team with the subject "System Exception in REFGoogleScholar" and attach the screenshot.
5. **Recovery:** Close and Re-open Google Chrome (KillAllProcesses -> InitAllApplications) and attempt the next transaction.

3.6 Applications Errors & Exceptions Handling

A comprehensive list of all errors, warnings or notifications should be consolidated here together with the action to be taken for each by the Robot. There are 2 types of exceptions/errors:

Known = Previously encountered and action plan or workaround available for it (e.g. SAP unresponsive during peak times)

Unknown = these are exceptions and errors that cannot be anticipated but for which the robot needs to have a rule so that the RPA solution is sustainable.

3.6.1 Known Applications Errors and Exceptions

Details regarding how the robot should handle the exceptions.

Error/Exception Name	Action	Parameters	Actions to be taken
Selector Not Found	Navigating to Google Scholar	TimeoutMS (30s)	If the robot cannot find the "Search" bar or the "Profile Link", it likely means the page did not load correctly. **Action:** Retry the current transaction up to 2 times (REFramework default). If it fails again, throw a System Exception.
Google CAPTCHA	Accessing Profile / Paper	Element Exists	If the "I am not a robot" screen appears, the robot cannot proceed. **Action:** Throw a System Exception immediately to stop the process and alert the administrator. Wait 5 minutes before retrying (via Queue Retry).
Chrome Not Responding	Init / Navigation	Browser Object	If Chrome freezes or crashes: **Action:** The Global Exception Handler will catch this. Use KillAllProcesses to force-close chrome.exe, wait 5 seconds, and restart the application from scratch.
Login Failed	initAllApplications	Invalid Credentials	If the robot cannot log in (e.g., "Wrong password" text appears): **Action:** Throw a System Exception. Do NOT retry (retrying might lock the account). Send a high-priority alert email.

3.6.2 Unknown Applications Errors and Exceptions

An umbrella rule that includes a notification needs to be designed for all other exceptions that could happen and cannot be anticipated.

Rule: For any unforeseen application error (e.g., internet connection drop, Windows update restart, unknown Google Scholar popup):

1. **Terminate:** The robot will catch the exception in the **Process State**.
2. **Screenshot:** It will take a screenshot of the entire desktop and save it to the Exceptions_Screenshots folder (defined in Config).
3. **Reset:** It will invoke CloseAllApplications and KillAllProcesses to clean the environment.
4. **Retry:** It will attempt to process the transaction again (up to the MaxRetryNumber defined in Config).
5. **Alert:** If the final retry fails, the job status is set to **Faulted** in Orchestrator.

3.7 Reporting

In this section all the reporting requirements of the business should be detailed so that when the RPA solution is moved to production the administrators can track the performance of the solution.

Report Type	Update frequency	Details	Monitoring Tool to visualize the data
Transaction Logs	**Real-Time**	Tracks every individual paper processed. Includes the **Paper Title**, **URL**, and the extracted **Citation Score**. Used to verify specific data points.	**UiPath Orchestrator** (Transactions View)
Process Summary	**Daily/Weekly** (Per Job)	Shows the Success Rate vs. Exception Rate. (e.g., "Processed 50 items, 48 Successful, 2 Business Exceptions").	**UiPath Orchestrator** (Dashboard / Jobs)
Exception Report	**Daily**	List of all Business Rule Exceptions (e.g., "Profile not found") and System Exceptions. Helps identify if Google is blocking the bot.	**UiPath Orchestrator** (Logs) or Export to Excel

** For complex reporting requirements, include them into a separate document and attach it to the present documentation*

4 Other

In this section the proposed improvements to the process, actions to the process will be outlined as well as the actions proposed for automation and the type of robot required. **This will be cross-checked by the Solution Architect.**

4.1 Additional sources of process documentation

If there is additional material created to support the process automation please mention it here, along with the supported documentation provided.

Additional Process Documentation		
Other documentation (Optional)	README.md	Located in the project root folder. Contains technical setup instructions for developers.
Standard Operating Procedure(s) (Optional)		Insert any relevant comments
Input Files (Optional)	Config.xlsx	Configuration file containing Queue names and URL settings.
Output Files (Optional)	Report_[Date].xlsx	(If local reporting is enabled) The final Excel sheet containing the validated citation data.