A blue and red logo

Description automatically generated

image001

EVIDENCE LIBRARY 2.0 Data Scraping Automation

SOLUTION DEFINITION DOCUMENT

Table of Contents

[**1.** **Introduction** 3](#_Toc114685402)

[**1.1** **Purpose of the Document** 3](#_Toc114685403)

[**1.2** **Objectives/Goals** 3](#_Toc114685404)

[**1.3** **Primary Process Key Contact** 3](#_Toc114685405)

[**1.4** **Prerequisite for Automation** 3](#_Toc114685406)

[**2.** **As-Is Process** 4](#_Toc114685407)

[**2.1** **Overview** 4](#_Toc114685408)

[**2.1.1** **Details of the Application Used in the Process** 4](#_Toc114685409)

[**2.2** **As-Is Process Mapping** 4](#_Toc114685410)

[**2.3** **Explanation of As-Is Process** 5](#_Toc114685411)

[**3.** **To-Be Process** 5](#_Toc114685412)

[**3.1** **To-Be Process Map** 5](#_Toc114685413)

[**3.2** **In-Scope of Automation** 6](#_Toc114685414)

[**3.2.1** **Search PubMed(Web Reader) Bot** 6](#_Toc114685415)

[**3.2.2** **Web Extractor Bot** 6](#_Toc114685416)

[**3.2.3** **Excel with Additional publications Bot** 7](#_Toc114685417)

[**3.2.4** **Merge Result list Bot** 7](#_Toc114685418)

[**3.2.5** **ML & Rule Engine Bot** 7](#_Toc114685419)

[**3.2.6** **Excel Writer Bot** 7](#_Toc114685420)

[**3.2.7** **AI Feedback Excel** 8](#_Toc114685421)

[**3.2.8** **API Bot** 8](#_Toc114685422)

[**3.2.9** **API Trigger Bot** 8](#_Toc114685423)

[**3.2.10** **API Config Bot** 9](#_Toc114685424)

[**3.3** **To Be Process in detail** 10](#_Toc114685425)

[**3.3.1** **Extraction Template Structure** 10](#_Toc114685426)

[**3.3.2** **Extraction (Output) Template screenshot:** 11](#_Toc114685427)

[**3.3.3** **Business rules for deriving the fields:** 11](#_Toc114685428)

[**3.3.4** **Source:** 11](#_Toc114685429)

[**3.3.5** **Output Document storage location:** Share Point location 11](#_Toc114685430)

[**3.3.6** **Searching information from the website-** 12](#_Toc114685431)

[**3.3.7** **Scraping rules (for all the fields in extraction template)** 14](#_Toc114685432)

[**5.** **Out of Scope Automation** 27](#_Toc114685433)

[**6.** **Errors and Exception Handling** 27](#_Toc114685434)

[**7.** **Reports** 27](#_Toc114685435)

[**8.** **Appendix** 28](#_Toc114685436)

# **Introduction**

## **Purpose of the Document**

The Purpose of this document is to provide details of a Robotic Process Automation based solution to automate the current manual process of collating the data for various literae from PUBMED & the international conference, validating it & providing the data into output excel in the required format. The document shall briefly cover the AS-IS process. The proposed automated process, TO-BE process, shall be covered in depth.

## **Objectives/Goals**

To automate the manual process of collating the data for various literature and accelerate the process of web scraping of the literature from PUBMED & the international conference and providing the scraped data into the output excel.

## **Primary Process Key Contact**

Jesvin

Email id:[jesvin@accessinfinity.com](mailto:jesvin@accessinfinity.com)

## **Prerequisite for Automation**

1. PubMed Link
2. Conference link
3. Drug name to be searched & search criteria
4. Business rules for deriving the fields.

# **As-Is Process**

## **Overview**

Access Infinity is a Pharma Tech company and a specialist consultancy in the ‘Market Access’ domain.

The current Project focus of AIQoD is scraping the data from PubMed & enter the extracted data in a specific format into Evidence Library Publication database.

Evidence Library is a product offered by Access Infinity to its customers (Pfizer) to track the publications of the drugs launched in the market. Evidence Library is a repository of various publications related to a specific drug / product search.

## **Details of the Application Used in the Process**

1. PubMed Link
2. Evidence Library Application (Drop2,3)
3. Conference Link

## **As-Is Process Mapping**

**Diagram

Description automatically generated with medium confidence**

## **Explanation of As-Is Process**

* Access Infinity Data Analyst clicks on the PubMed URL.
* The drug information is searched based on the source URL.
* The Data Analyst goes through each publication **abstract and/or full text to analyze and extract the data** & its Full text to analyze the data.
* The resultant output analyzed is entered into Evidence Library ERP by the Data Analyst.

# **To-Be Process**

## **To-Be Process Map**

**A diagram of a software flowchart

Description automatically generated**

## **In-Scope of Automation**

* Search string given by AI shall be configured in the system
  + Provision for updating the input shall be provided to Access infinity team.
* Brand (INN) + specific indication
  + Search strings pre-defined, replicating past searches.
  + The following filters must be applied to PubMed.
    - Species: Human
    - Article language: English
  + Applying filters in international conference literature.
  + **English language + In Humans + Last 5 years**
  + **Optional filter:  Study type - Clinical studies, observational studies, claims databases**
  + **Could also add countries: AND [US or EU or China or Japan or UAE or Saudi Arabia**]
* The process shall be scheduled to run as required (configured)
* The process can also be run manually; it should not run when any other process is running (already in progress).
* Extraction rules shall be written as per business rules provided by Access Infinity.
* The most occurring PubMed website shall be visited for extraction of data of Full text.
* 29 websites will be studied by AIQoD to extract the full text. If any website does not fall under these websites and don’t have similar structure, then it will be considered as a change request.

### **Search PubMed (Web Reader) Bot**

* Once the process is triggered the Bot will visit the PubMed URL to search the string
* given by AI [shall be configured in the system This string can be updated as required e.g. (tofacitinib OR xeljanz OR "JAK inhibitor") AND (rheumatoid arthritis OR psoriatic arthritis OR ulcerative colitis OR ankylosing spondylitis OR juvenile idiopathic arthritis) to be scrapped.
* An email shall be sent to the concerned business user when the process is triggered. The AI email list is configurable.
* A loop is run to search for information from all the publications displayed on the webpage for that drug and read by the Bot.
* If No information is found on the drug (no publications available on a webpage) then a notification email shall be sent to Access Infinity for the same.

### **Web Reader Bot**

* The HTML link shall be directly downloaded and processed by the GenAI.
* Web reader Bot extracts data from webpage for each publication based on the business rules provided by Access Infinity.
* All the basic information required is extracted from the PubMed webpage (Abstract & Title).
* **Incase Additional information not available in Title & abstract then it shall be extracted from the full text** link.

### **Excel with Additional publications Bot**

* The Excel with Additional Publications Bot shall extract the links from the “Additional Publications” excel and passes to the Merge Result list Bot.
* This excel shall be on AI share point.
* Output will be provided in different excel.



* The Excel with poster and abstract publication types from different conferences will be provided in the Excel (with the title of the abstract/poster and the conference. Each conference has a different website/s to be looked into for the abstract and poster processing. Below are the details of the conference and the source links
* This excel shall be on AI SharePoint
* The output is required in a different Excel

### **Merge Result list Bot**

* The Merge Result list bot will compare the websites link (refer 3.2.2) and the additional publications excel links and passes the search for the unique websites to the run a loop for that drug/search string and read by the Bot.

### **GenAI Bot**

* GenAI Model & Rule Engine Bot is applied on the extracted data.
* Data extracted from the sections (refer to section **3.2.2**), additional rules are applied on this data , to extract the specific details.
* Business rules are applied as provided by Access Infinity.

### **Excel Writer Bot**

* The validated output data is shared in Excel in prescribed format (section 3.3.2) on SharePoint (Drop1)
* The extracted output is sent to business users by email along with the reports.

## **To Be Process in Detail**

### **Extraction Template Structure**

**Field names and Purpose:**

* Describes the respective field name and its purpose.

|  |  |
| --- | --- |
| **Column Names** | **Description** |
| Title | Title of the article |
| Indication | Broad therapy area (medical condition or disease) addressed in the article |
| Patient population | Specific patient population that the article reports on. Patient descriptors can be split as shown to the right |
|
|
|
| Geography | Countries or regions from where participants in the study were included. Can be global, regional, selection of countries, or specific country |
| Publication date | Date on which the publication is published in the journal |
| Authors | List of all authors listed for the publication |
| Journal name | Full title of article or shortform |
| Data Type | Observational study, RWE, Review |
| Study link | PubMed or article URL |
| Study identifier | PubMed ID or any specific identifier mentioned |
| Data Extracted Date |  |
| Study Methodology |  |
| Study Size |  |

### **Extraction (Output) Template screenshot:**

A screenshot of a computer

Description automatically generated

### **Business rules for deriving the fields:**

|  |  |  |
| --- | --- | --- |
| **Column Names** | **Description** | Rules – GenAI Based |
| Title | Title of the article | Include the full article title, including any identifiers (e.g. poster number) |
| Indication | Broad therapy area (medical condition or disease) addressed in the article | Overlap with below is possible; redundancy is acceptable and needed |
| Patient population | Specific patient population that the article reports on. Patient descriptors can be split as shown to the right | 1.Age – Key word like adult (≥18y) or all patients ≥6y |
| 2. Specific disease by stage – keywords- severe, refractory etc |
| 3. Treatment status or eligibility criteria – keywords - previously treated, previously treated with etc |
| Need to extract **comparator drug name**.  Keyword (compared with, versus) |
| Geography | Countries or regions from where participants in the study were included. Can be global, regional, selection of countries, or specific country | This must be on where the specific data is reported from; not authors’ location. Limited to the Countries specified. |
| Publication date | Date on which the publication is published in the journal | - |
| Authors | List of all authors listed for the publication |  |
| Journal name | Full title of article or shortform |  |
| Data Type | Observational study, RWE, Review |  |
| Study link | PubMed or article URL |  |
| Study identifier | PubMed ID or any specific identifier mentioned |  |
| Data Extracted Date |  | GenAI baed |
| Study Methodology |  |  |
| Study Size |  | GenAI based |

### **Source:**

|  |  |
| --- | --- |
| **Particulars** | **Details** |
| Source weblink | [**https://pubmed.ncbi.nlm.nih.gov/**](https://pubmed.ncbi.nlm.nih.gov/) |
| Log in | Free Source |
| Document type | NA (Online database) |

### **Output Document storage location:** SharePoint location

### **Scraping rules – GenAI-based Extraction**

* The data from the PubMed or any conference link shall be downloaded as HTML and GenAI-based extraction shall be done using the prompt engineering.

# **Out of Scope Automation**

* Evidence Library integration shall be out of scope for Drop 1.
* Change in structure of the webpage or Business rules :This shall be considered as change request.

# **Errors and Exception Handling**

* Website or webpage is down: e.g. Server down, Internet issue, Session terminates for known or unknown exception. This is handled by the recorder Bot through screenshots. A screenshot shall be taken when the error occurs & the file path of the screenshot is stored in the output Excel file.
* The Exceptions description shall be added in the output Excel file.
* Webpage scraping or data upload failure exception count : The limit of the count of exception encountered while uploading data or scrapped count from webpages
* Notification email shall be sent to Access Infinity Business user incase of above error and exceptions with details.

# **Reports**

* Number of records scrapped for the drug.
* Number of records pushed successfully into Evidence Library.
* Publications detailed report
* Publications scrapping status



# **Appendix**

**Phase 1A Contains**

1. Data scraping from PubMed URL
2. Output data in excel
3. Feedback loop from csv
4. Configure product search criteria from AIQoD application
5. Configure scheduler frequency from AIQoD application
6. On-demand process trigger from AIQoD application
7. TRIO NLP model support

**Phase 1B Contains**

1. Phase 1A
2. Excel input for additional publication URL
3. Full text format support for PubMed URL (at least 1 format with maximum occurrence)



**Phase 2**

1. Integration with Evidence Library via API and output pushed in Evidence Library..

**Phase 3**

1. Full text for top 29 websites PubMed links.

**Note:**

* If no information is available on the webpage of the publication for the drug/search string ,the fields shall be marked as ‘NA’ in the output excel.
* All the content should be extracted from Title and Abstract, if the content is not available in Title and Abstract as per business rules, look in full text.