Product Design

Team: 5 *Kowshik, Chaitanya, Anandhini, Vidhatri*

Design Overview

Architectural design

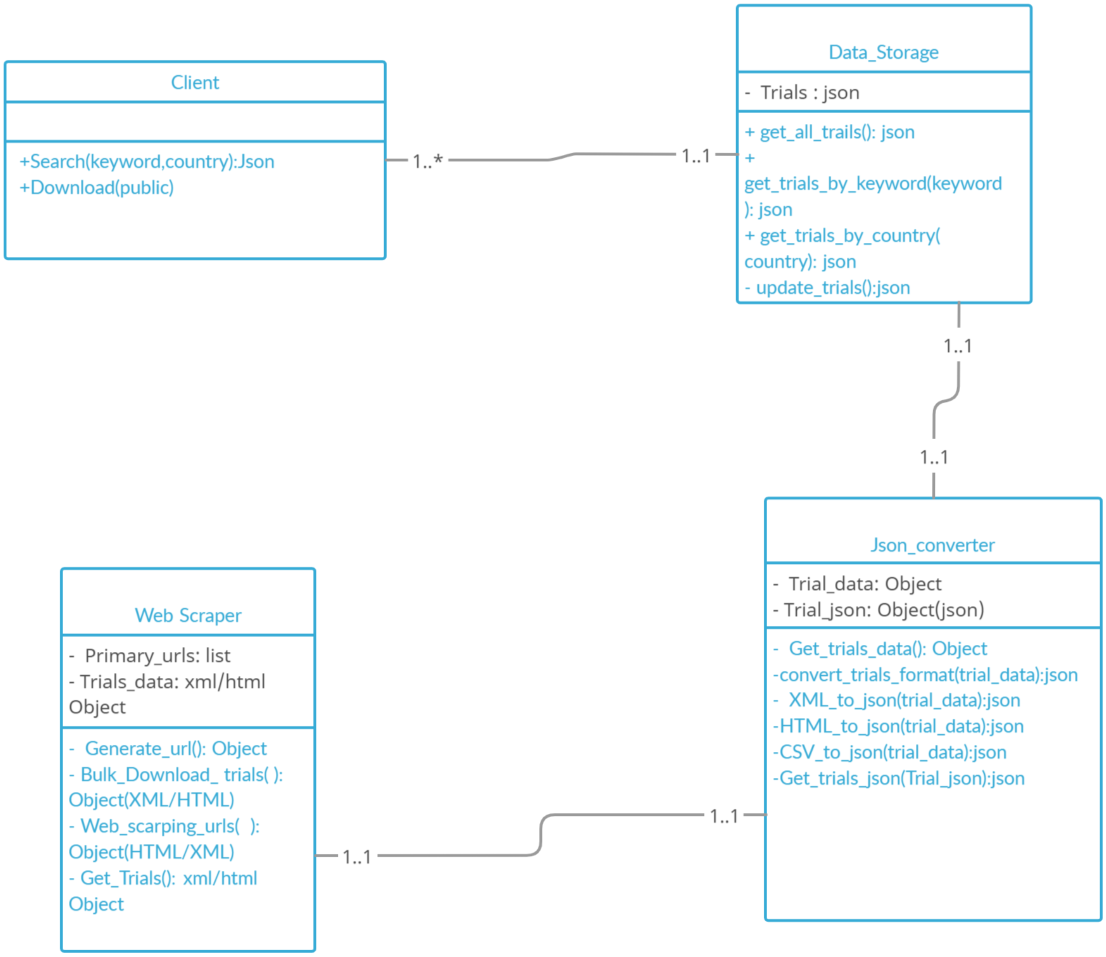
There are 18 different modules Each work individually

1. Australian New Zealand Clinical Trials Registry (ANZCTR)
2. Brazilian Clinical Trials Registry (ReBec)
3. Chinese Clinical Trial Registry (ChiCTR)
4. Clinical Research Information Service (CRiS), Republic of Korea
5. Clinical Trials Registry - India (CTRI)
6. Cuban Public Registry of Clinical Trials(RPCEC)
7. EU Clinical Trials Register (EU-CTR)
8. German Clinical Trials Register (DRKS)
9. Iranian Registry of Clinical Trials (IRCT)
10. ISRCTN
11. Japan Primary Registries Network (JPRN)
12. Lebanese Clinical Trials Registry (LBCTR)
13. Thai Clinical Trials Registry (TCTR)
14. The Netherlands National Trial Register (NTR)
15. Pan African Clinical Trial Registry (PACTR)
16. South African Clinical Trial Registry(SANCTR)
17. Peruvian Clinical Trial Registry (RePEc)
18. Sri Lanka Clinical Trials Registry (SLCTR)

Each module was named after the web scraped Registry which has codes to download/web scrape the clinical trials data from their official website and has methods to convert the extracted data into JSON format. The analysis for each website is present in a separate document used for checking the data patterns, which are helpful for web scraping registries. The downloaded trials are present in separate docker containers. All the coding part was done in a separate AWS server provided by the client.

System interfaces

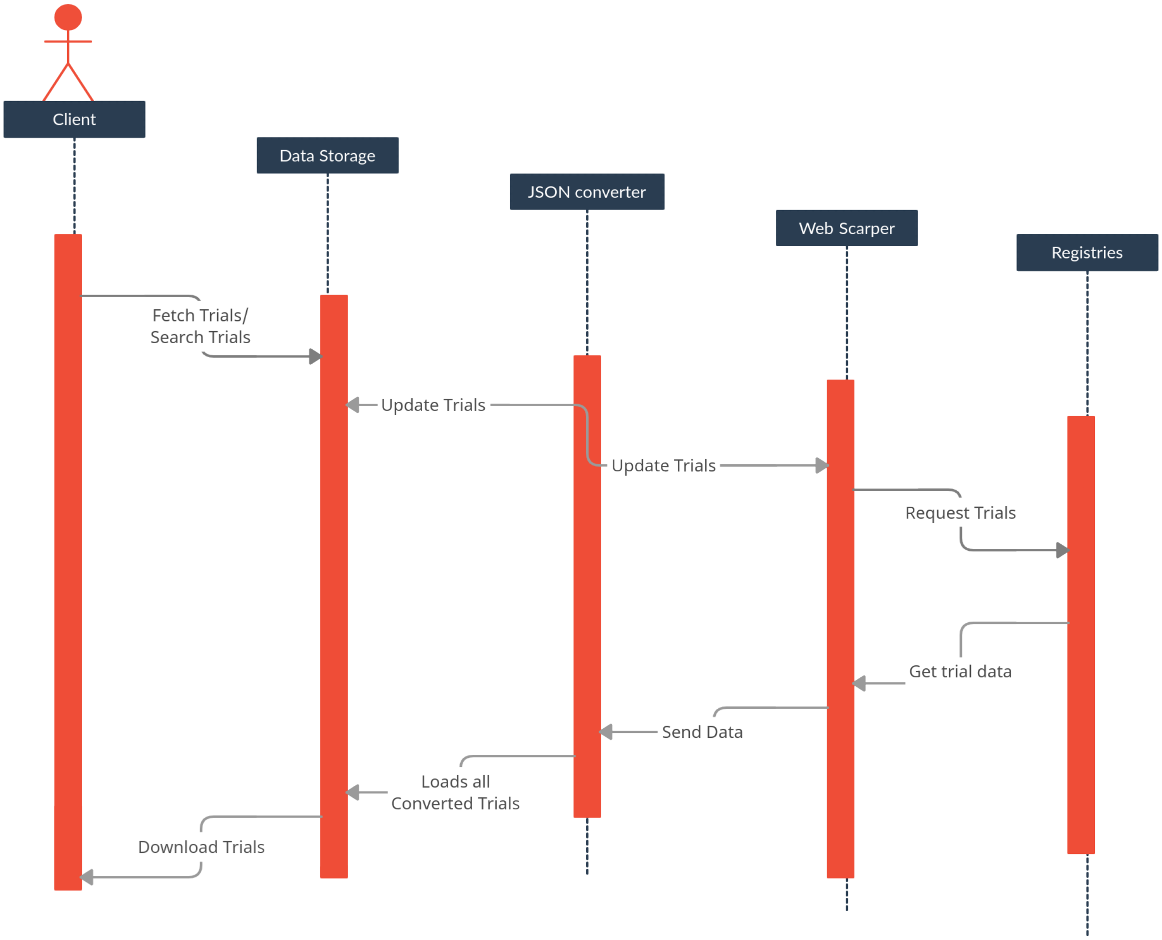
Model


Client
+Search(keyword,country):Json
+Download(public)
1..*
1..1
Data_Storage
- Trials : json
+ get_all_trails(): json
+ get_trials_by_keyword(keyword): json
+ get_trials_by_country(country): json
- update_trials():json
1..1
1..1
Json_converter
- Trial_data: Object
- Trial_json: Object(json)
- Get_trials_data(): Object
-convert_trials_format(trial_data):json
- XML_to_json(trial_data):json
-HTML_to_json(trial_data):json
-CSV_to_json(trial_data):json
-Get_trials_json(Trial_json):json
1..1
1..1
Web Scraper
- Primary_urls: list
- Trials_data: xml/html Object
- Generate_url(): Object
- Bulk_Download_ trials( ): Object(XML/HTML)
- Web_scarping_urls( ): Object(HTML/XML)
- Get_Trials(): xml/html Object

Client	Class state ● contains information about specific trials like name, id, age group, start date... Class behavior ● used to search for trails ● Can download the required trials
Data_Storage	Class state ● contains all trails in JSON format

	Class behavior ● This is used to get all trials information ● Get trials information by country ● Update trial information ● Get trials using keywords.
Web Scraper	Class state ● used to get url for web scraping and the registries contain the trials in XML/HTML/CSV/TEXT/PDF formats. Class behavior ● This can be used to download the trial data as a bulk download ● The class can be used to obtain the HTML page URL to scrape for all necessary information ● The trials can be obtained by iterative download using web scraping tools
Json_converter	Class state ● Contains trail data in JSON format Class behavior ● It is used to convert the web scraped data from XML/HTML/TEXT/CSV format to JSON format ● get the trials in JSON format.

Sequence Diagram(s)



Design Rationale

REGISTRIES	ISSUES	SOLUTION/APPROACH USED
Australian New Zealand Clinical Trials Registry (ANZCTR)	Couldn't find the download location of the files(downloaded using selenium) on the server.	Used a Docker container to create the environment required for the selenium
Brazilian Clinical Trials Registry (ReBec)	Download from trial search failed for an individual trial.	Found an alternate api to download all trial data.
Chinese Clinical Trial Registry (ChiCTR)	Chinese clinical trials website has a limit to view number of trials per day which halts the downloading processes	Yet to be resolved

Clinical Research Information Service (CRiS), Republic of Korea	Korean clinical trials website contains a direct download of all the clinical trials but it leads to an empty file	Downloaded trials using individual download
Clinical Trials Registry - India (CTRI)	required data is present in different tables with different keys.	scrapy
EU Clinical Trials Register (EU-CTR)	It has a direct download of all the trials in text format but the conversion of text to JSON has issues	done using beautiful soup.
German Clinical Trials Register (DRKS)	The first approach used using beautiful soup failed since we couldn't convert the extracted HTML to JSON format. The second approach using selenium but due to conversion problems from HTML to JSON, this approach also failed.	done using beautiful soup.
Iranian Registry of Clinical Trials (IRCT)	No issues faced	done
Thai Clinical Trials Registry (TCTR)	Individual downloads for each trial in XML format had Unicode characters from HTML which can not be parsed by the python xmltodict library hence creating an error in conversion from XML to JSON.	One solution is to replace all the Unicode special characters in all files which seems infeasible due to the abundance of these characters in the files and the large scale of files to be converted. Still finding ways to solve this problem. Currently suggested is to either omit these characters or to hard code to replace them.
The Netherlands National Trial Register (NTR)	Source API is not present in its HTML source code	Got information using requests for each individual trial.
Pan African Clinical Trial Registry (PACTR)	Direct Download of the trials is available but it was blocked	It has an option of downloading selected trials. So selected trials in each page and downloaded them page wise
Peruvian Clinical Trial Registry (RePEc)	Direct Download of the trials is available but the website has server issues.	Can’t be resolved by us as it is the server issue from the website
South africa	no issues	
South Africa Clinical Trials Registry (SACTR)	Direct Download of the trials is available but it was blocked	It has an option of downloading selected trials. So selected trials on each page and downloaded them page wise
US	Bulk download of all trials available but it is folder in folder format	bulk download of trials in xml format using selenium.

Cuban	no issues	
European Clinical Trials Registry(EUCTR)	results page slow and classes not defined properly.	had to hard code beautiful soup extraction due to poor definition of classes. Separate code for results page.
ISRCTN	Normal API requests are not allowed by the website.Limit of only 100 registry scraping	Scraped only 100 registry server issues couldn't be solved.
JPRN	no issues	
Lebanese Clinical trial registry(LBCTR)	Classes not defined properly.	Hard coded the beautiful soup conversion to json
Sri Lankan	All trials are not in the same format.	Used both beautiful soup and selenium to account for different formats.