Fetching Articles

Database of scientific articles

We give the list of database of article where we fetch the article.

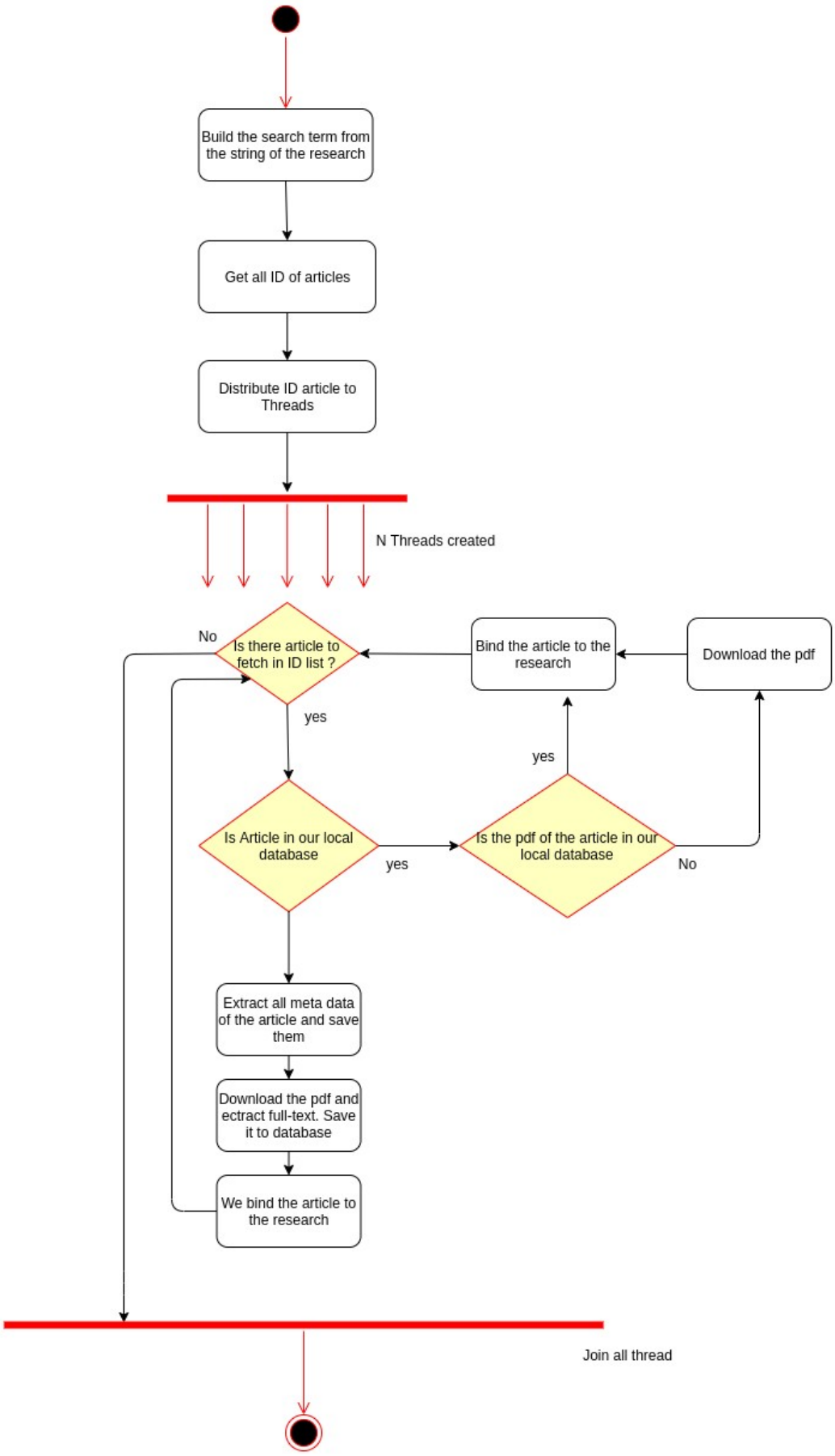
- ArXiv
- BiorXiv
- MedrXiv
- Paperity
- PMC
- PubMed

Each database is different and so the method to fetch too. For ArXiv, PMC and PubMed, they have an api that return an xml file. For Paperity, this is an api that returns a JSON file and they want that between each http request there is at most one second of waiting. Otherwise, they block your ip adress. For BiorXiv and MedrXiv, the funniest, we directly fetch from the html page of their site because there isn’t aby api. So, we have to be careful than the tags, class name or id change in their page.

The data that are fetching are :

- Title
- DOI
- Publication date
- Abstract
- Authors
- URL to pdf
- the article itself

Even if handling each one is different, we create a systematic way to fetching the articles :


Build the search term from the string of the research
Get all ID of articles
Distribute ID article to Threads
N Threads created
No
Is there article to fetch in ID list ?
Bind the article to the research
Download the pdf
yes
yes
Is Article in our local database
yes
Is the pdf of the article in our local database
No
Extract all meta data of the article and save them
Download the pdf and ectract full-text. Save it to database
We bind the article to the research
Join all thread

We try to follow this way for each article database. For paperity, we distribute a list of page rather than a list of id beacause to get the total number of page use only one http request. As we have to wait at most one second between each sending, we have to save up some time as much as it is possible.

To fetch articles, this is done one by one article database. We wait to terminate fetching from one before to fetch to another one. So we can easily control the number of thread used and see in which step we are. But, to fetching to all database in same time could be an improvement for futur and have a progress bar of each database fetching too.

There is an additional utility we must have for each article’s database. This is the one that gives us the total number of article that we can fetch from this database for our research. This permit us to calcul an estimation of total of article for a research.