RDP - Round 2: Internal Review

What are the assumptions you made?

The user will provide a domain name to search from and not the name of a researcher. I have assumed that only the critical information about a researcher is to be provided and for further contact, the user can use a ORCID page or the IRINS page for the researchers.

What are the constraints you assumed?

I have assumed that the user will provide a complete search query and not a partial word. These will not return any relevant results and the searching will return any researchers who can fulfill all the terms mentioned in the search query and not a subpart of them.

What stopped you from incorporating more features?

Time. The time to implement new scrapers is short but integrating them into the product with support to return a finite and fixed amount of data per page was difficult. The other constraint was time the scrapers took to return a result. As such only the IRINS dataset was integrated completely as the other scrapers returned data only after a few seconds.

What is the size of the database created using IRINS and where it is stored?

The database is stored in the cloud using the MongoDB Atlas system. It currently contains around 37000 records and has a size of 12.37 MB. The MongoDB can store up to 50 MB of data.

How did you crawled IRINS DB, did you used any API?

I created a custom crawler that can scrape the needed data and write to the DB when needed. The start and the end page can be set and the crawler will function automatically without any user intervention. This script can be run as per administrator convenience to refresh the dataset.

The scraper first takes all the records in the results page specified. Then it goes to every result in the page and scrapes the data required. This data is then written to the database with a unique ID that can help in rendering results in the React JS front-end.

What extra value add your tool provides over running my search query directly on IRINS website?

The IRINS search results are not based of the expertise but rather the qualifications of the researcher. Hence not all results are returned. The application accounts for this and hence more relevant results are returned. In addition to this the IRINS website does not provide a sort by H-Index or by Citations which is provided in my tool for ease of understanding crucial attributes of a researcher.

Are you using any other sources to further enrich your data, if yes please specify along with an approach to use this source?

The ACM scraper is another source that has been implemented, however it is disabled in the code and does not contribute to the results. This can be improved upon by changing the structure of the scraper and then by enabling it in the front-end.

How will scale this data without putting substantial resources (Money/Time)?

One of the most important attributes of a good dataset is reliability. Most of the other indexing services do not return enough relevant data to make a proper decision. However, more scrapers can be incorporated to return more results as well.

The basic structure of a scraper is mostly the same. The only real changes will be to the elements of the page that must be grabbed. Once the required number of records are taken the scraper can be killed. For more information any unique identifying ID can be used to create an ORCID or a SCOPUS link to know further details.

These sources can all be disabled in the sources tab in the application by default. If the user wants more results, then, these can be enabled and these scrapers would return the results in real time.
