

ASSIGNMENT-2

# DS Tools and Techs

USAMA KHALID

I19-1236

Saood Khan

19I-2128

MS(DS)-B

# **Background of the problem**

Having all the dblp data organized in a relational database the current task is to identify the

FoR (Focus of Research) of each author. The data about FoR is not present in dblp but can be obtained from <http://portal.core.edu.au>. Each author can publish in a journal or conference and each of them has an FoR. The most number of publications of an author in an FoR is their main Field of Research and that is the current focus of this assignment.

# **Issues faced and their solutions**

1. Data of Journals was not exportable: The data of conferences was easily exportable as a csv file from the core portal but exporting of data of journals was not possible.

Solution: We wrote a python script to scrape data of both Journals and conferences.

1. Title of FoRs were not present in search table: The main page containing table of journals or conferences only contained ids of FoR not their titles. They were available on the journal and conferences separate pages. Moreover Those titles were not easy to scrape because their tag wasn't uniquely identifiable.

Solution: We modified our scrapping script to go on each journal and conference page (The link to them was available in the table) And extract the title. Because the FoR title was mentioned in the same tag as the FoR id and we had the FoR id so we could search for the tag.

1. Data about conferences was not stored properly in dblp database: The dblp.xml stored data about each conference in an ‘Inproceeding’ super tag instead of a sub tag inside an article or any other publication. When we stored data about inproceedings in Assignment 1 we hadn’t linked it with any publication like we did for journals. This resulted in not being able to find the publications or the authors who published in that conference.

Solution: The ‘key’ attribute of each publication contained the acronym of the journal or conference it belonged to. So we extracted those acronyms and luckily we also had acronyms from core’s data of conferences. So we were able to link up publications with their conferences and find their FoRs.

1. Titles of Journals were abbreviated in dblp data: Due to this abbreviation the titles could not be matched exactly with the titles of Journals present in core’s data. Moreover all of the Journals in dblp were not present in core’s data.

Solution: We used SQL’s ‘LIKE’ operator that can match words of variable length to match up the titles. For eg. A title of a journal in dblp was “Inf. Sci.” and real title was “Information Sciences”.

So we collected all the matching titles and selected the title with the least difference in length between those two.

1. Android App faced severe connectivity issues with local web server: First we hosted our Python Web server locally and used a local IP to access it from an Android app. But that caused a lot of problems as the PC had to be continuously running and local IP’s are dynamically assigned by the router.

Solution: We had to host the server as well as the 4 GB dblp data somewhere online to ensure 24 hours access for the Android App. So we used Google Cloud SQL to host our database and Heroku to host our python web server.

**Pseudocode of the programming logic designed to identify FoR of each author**

1. **Identify the journals and the no. of publications in them for an author.**

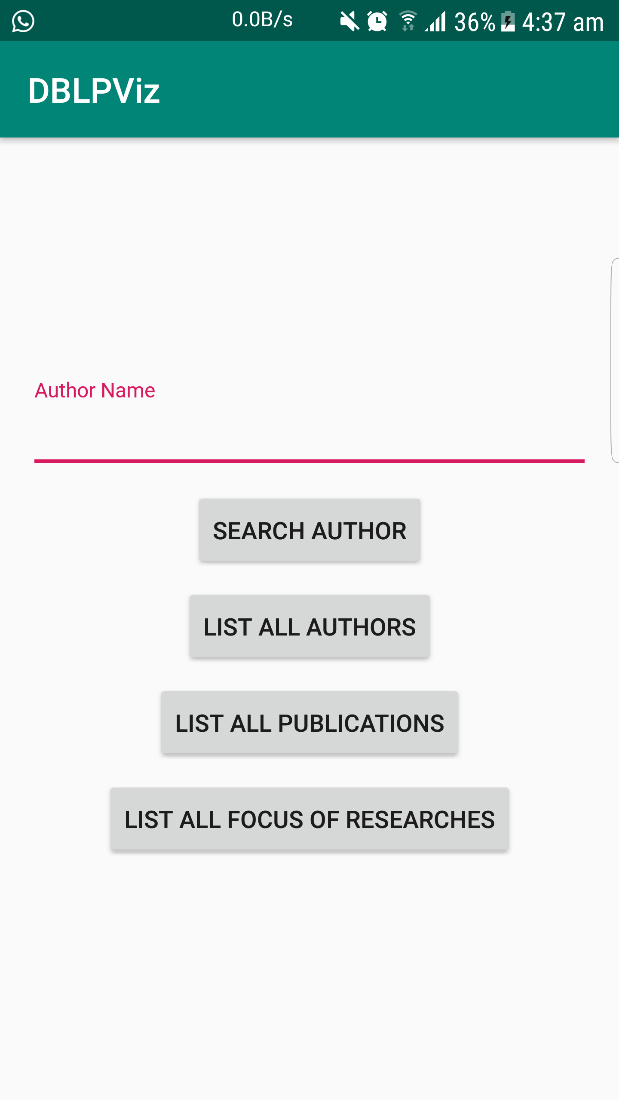
**This is done by joining the author, publication, authors\_publications and journal tables and matching the foreign keys.**

1. **Identify the publications and the count of them in each conference for the same author.**

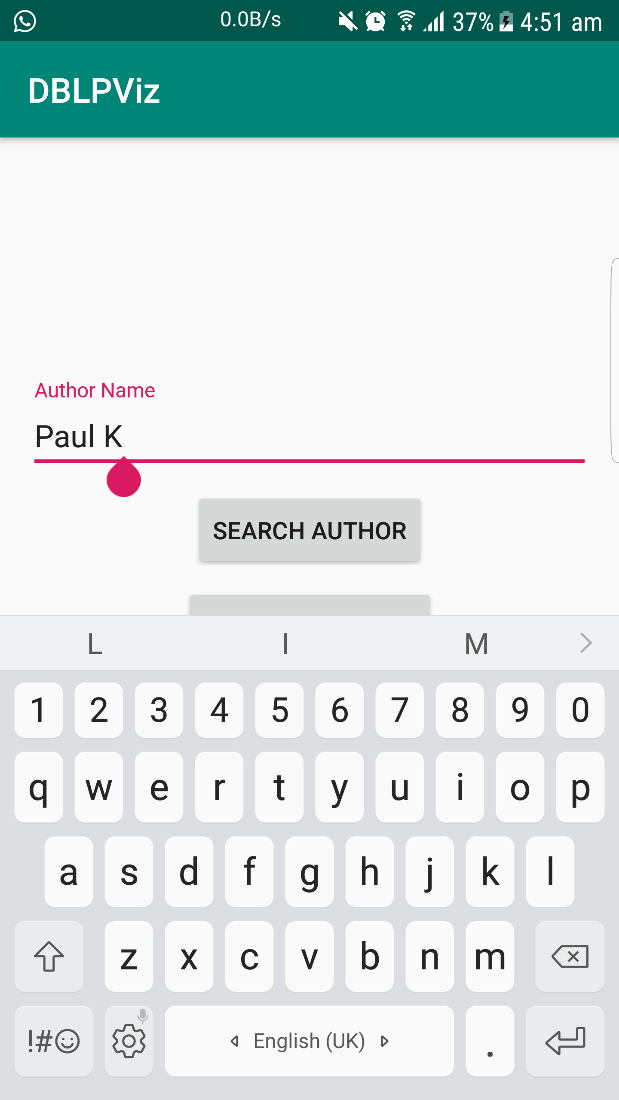
**This is done selecting all publications whose keys start with conf/ then aggregating based on conference acronym will give their count. Acronym is found in the same key after conf/.**

1. **Identify The FoRs for each of those conferences and journals and aggregate the count of journals and conferences into count of FoR.**
2. **The FoR with the largest count will become the FoR of the author.**

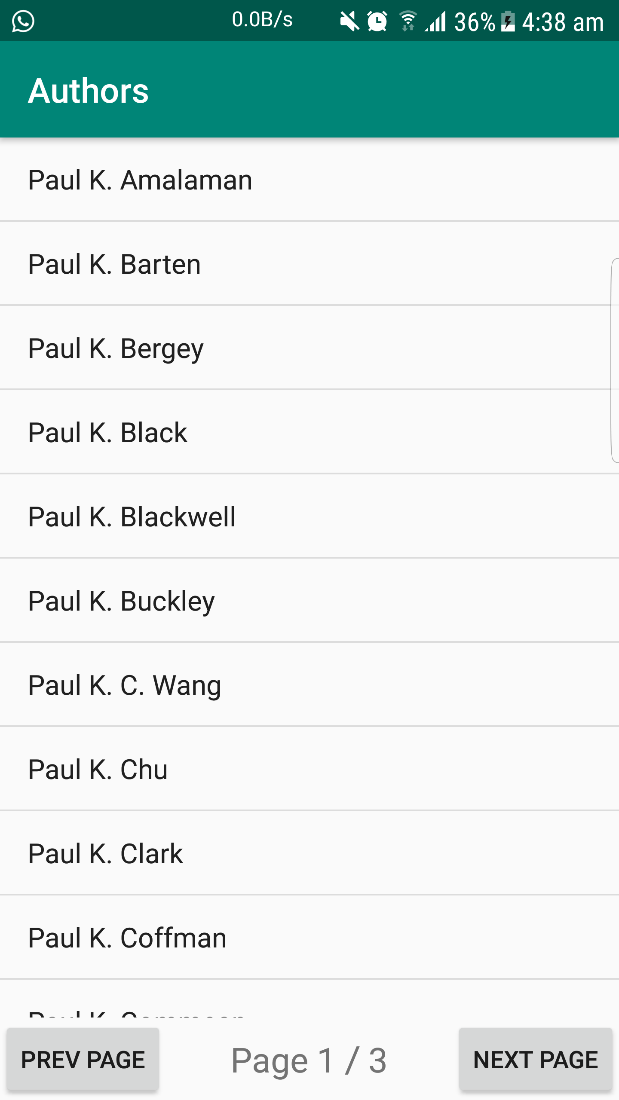
**Screenshots of the application**



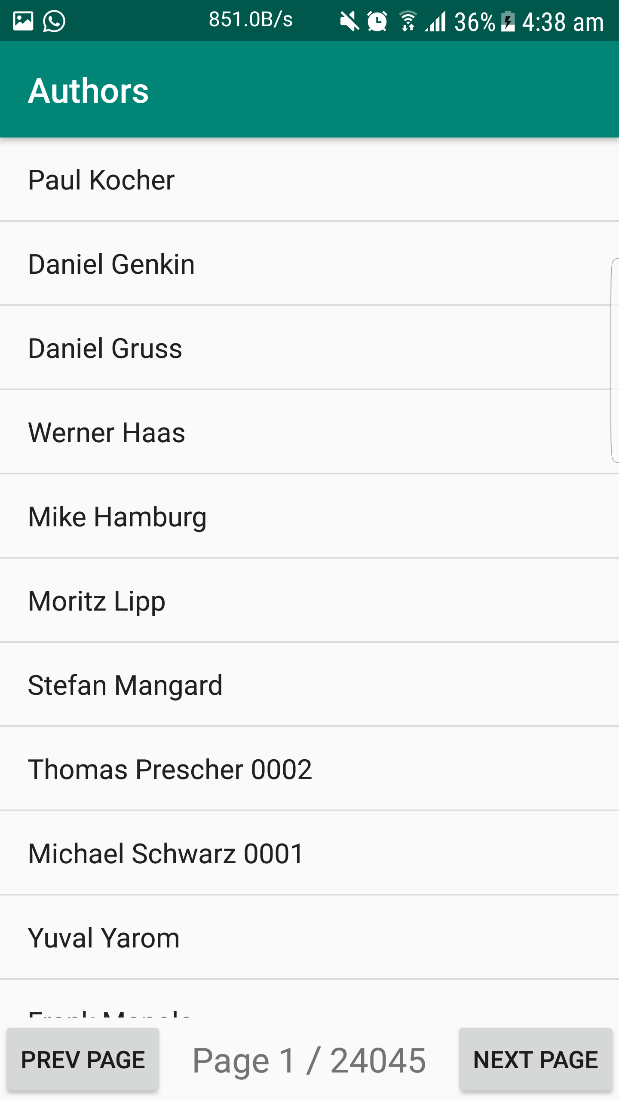
**Figure 1: Main Screen**



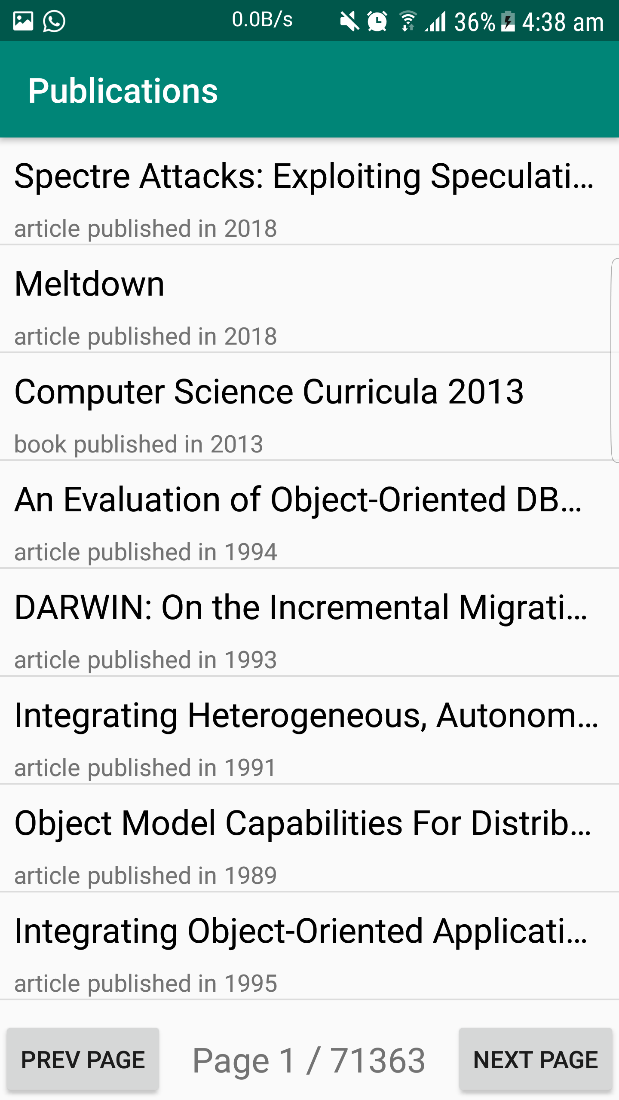
**Figure 2: Searching for an Author**



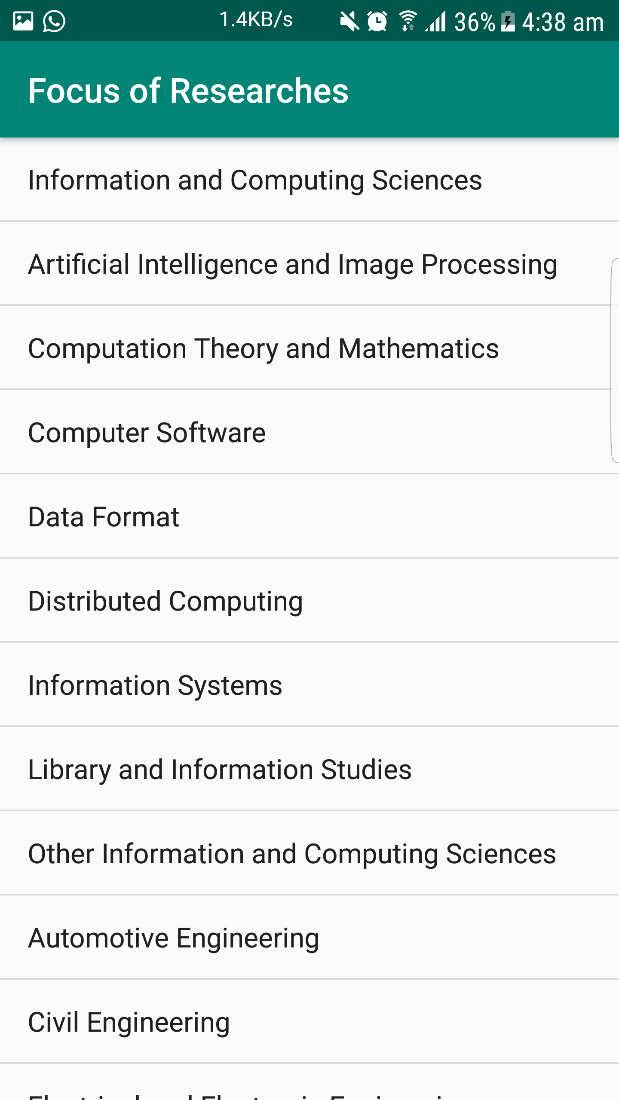
**Figure 3: Search Query Result**



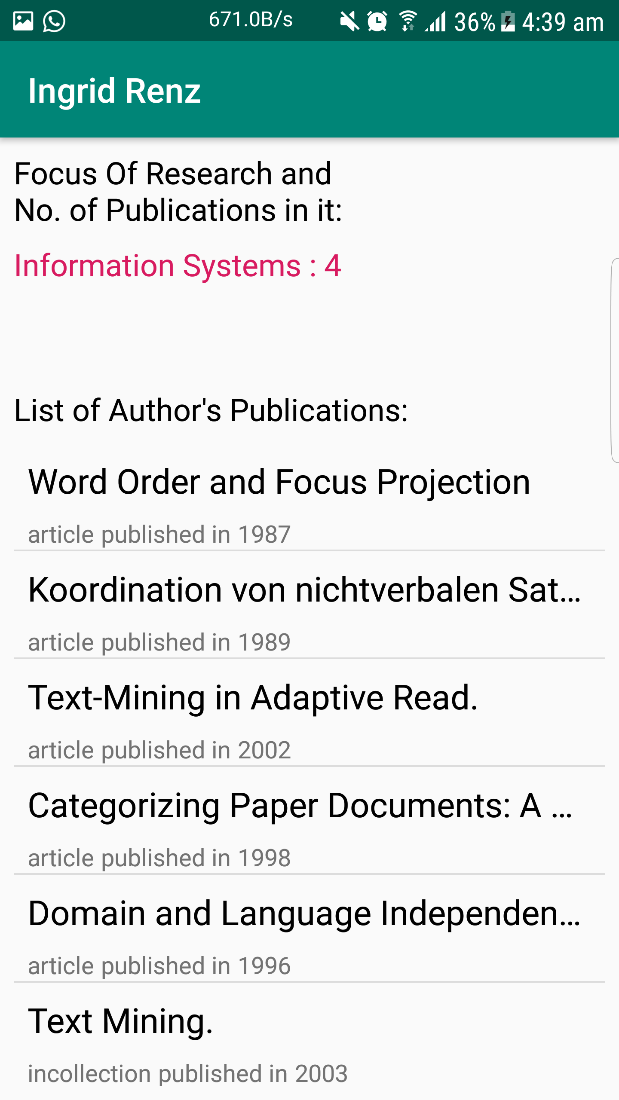
**Figure 4: List of all Authors page by page**



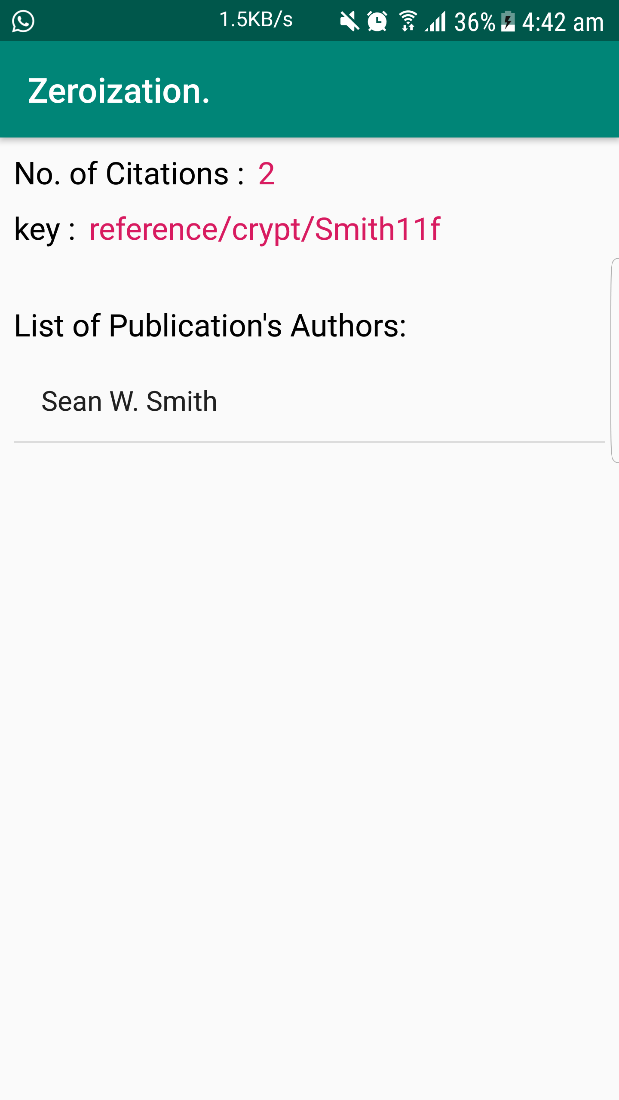
**Figure 5: List of all publications page by page**



**Figure 6: List of all FoRs**



**Figure 7: An Authors detailed info**



**Figure 8: A publication’s detailed info**