Web Crawler
Master's Project

Purpose of the system

Our aim of the project was to assist the research done by a sub-project of the CRC1342, the B06. CRC1342 (The Collaborative Research Centre 1342) analyses the global dynamics of public social policy rather than just focusing on the assumed 'main' factors which shapes country's social policies like domestic circumstances, international relations and networks, CRC1342 also take other factors like trade relations, migration, war and colonialism as well as the worldwide dissemination of ideas and legal regulations. And the sub-project B06 analyses the influence of sharply rising state revenues caused by a natural resource boom on the inclusiveness and generosity of state social policy in authoritarian regimes of the post-Soviet region. As part of that research, our project aims to look at and find articles from the web about social policies created by 3 particular Eastern European countries, Ukraine, Azerbaijan, and Georgia regarding internally displaced people (IDP). With the help of our team, we collectively aim to gather articles and information available on the internet and structure it based on keywords with relation to IDPs; e.g. IDP and Healthcare.

Success criteria and objectives of the project

- Success Criteria
  The success criteria for our project were to build a system that successfully scrapes, and stores articles from the given websites from the researchers and also sorts those articles based on priority.
  Our aim was to build a user-friendly interface with which researchers can filter those articles based on date of publishing, language and category.

- Objectives
  Data Scraping:
  Building the web crawler for scraping articles and the date of publishing from the provided websites of Azerbaijan, Ukraine and Georgia. Scrapped files to be stored in a database.

  Prioritization:
  As talked about earlier that the relevancy of the scrapped article is really important for multiple reasons. One being that there would be thousands for articles containing the given keywords and out of them we need to figure out which ones are really important to the researchers as they cannot read all of those articles manually.
  So, we tried to prioritize all the gathered articles. The prioritization is determined by how many times the keywords are mentioned in the article and how uniformly they are mentioned. The prioritization algorithm will assign specific weights to (sub-)keywords and ensure that the keyword IDP is at least mentioned once. The IDP keyword will have more weight than the search keyword. Based on this method we will give each article a prioritization score and sort them on descending order while showing the results.

Main Goal:

Our main goal was to build a local web-based interface, with all the files in a grid. We choose the grid-based interface to display this relevant information because we have several meta data to show along with article title and therefore gird would be the efficient choice as every meta data related to the article can be displayed in one line, which would make things very simple to read and relate.
The grid will be structured the following: one row will be one article and different columns indicate the meta data, such as source, date of publication, keywords, priority score, mentioned actors. The system will provide options for sorting the articles after keywords, date of publication and language (and source).

My Role:

UI design, User Research

Tools:

Figma.

User Research

According to User, they want a Folder grid system with sorting: The system should be able to portray the prior crawled articles in a grid system, and sort the articles after “priorities” and let the users download them.

- Functional Requirements
  - The system should be connected to a database holding all the articles and meta data
  - The system should be able to have an option to choose the country/language.
  - The system should be able to search for “IDP”, in the respective chosen language with all its grammatically correct versions and synonyms, plus keywords in a given list of websites.
  - The system should be able to portray these articles in a grid
  - The files should hold specific meta data such as:
  - Date of publication
  - Language/country
  - Headline of the article
  - Prioritization score
  - Source
  - Keywords mentioned
  - Actors and organizations mentioned
  - The system should be able to sort the articles after the meta data above
  - The system should provide an option to select these articles for download

- Nonfunctional Requirements
  - Usability
  - The System should be designed intuitively and usable by our group of users. It should follow basic design principles (e.g., Material Design), to ensure simplicity. Further to ensure the usability for our user group, we will co-create the system with our stakeholders and do user studies/interviews and ask about their point of view.

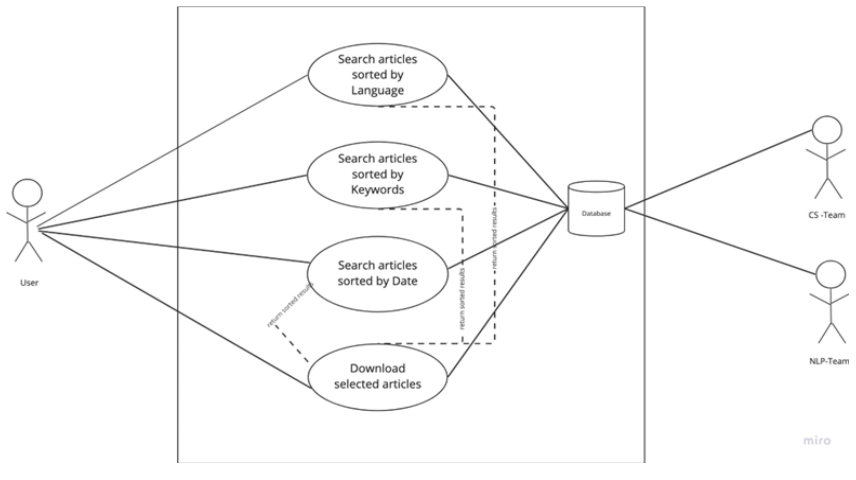
- Implementation
  - Every single website has a different structure and another (web architecture, class name etc.) so, each website has a separate crawling program. Injecting the keyword combination can be done automatically through looping and at the end of the running cycle the program should output a list of articles according to the relevant keywords in separate folders.

- Interface
  - The Interface would be web based and local, and will be deployed in the researcher’s computer. All the scraped files will be stored locally in the computer. The interface should have a normal search feature with filtering options for countries, keywords, and metadata like actors, time, and other keywords mentioned. After the search the grid view of search results should be displayed with columns like name of the article, date of publication, source, etc.

System models

- Scenarios
  - Scenario 1: User wants to search specific list of articles of one single country and one specific keyword (e.g. Education)
  - Scenario 2: User wants to search articles of a specific time- frame, of a specific country with a specific keyword.
  - Scenario 3: User wants to see relevant actors and organizations associated with specific articles.

Use case model



User Persona

Due to privacy issue original picture has not been used.

Name: Teacher. Dr. Heiko Pleines

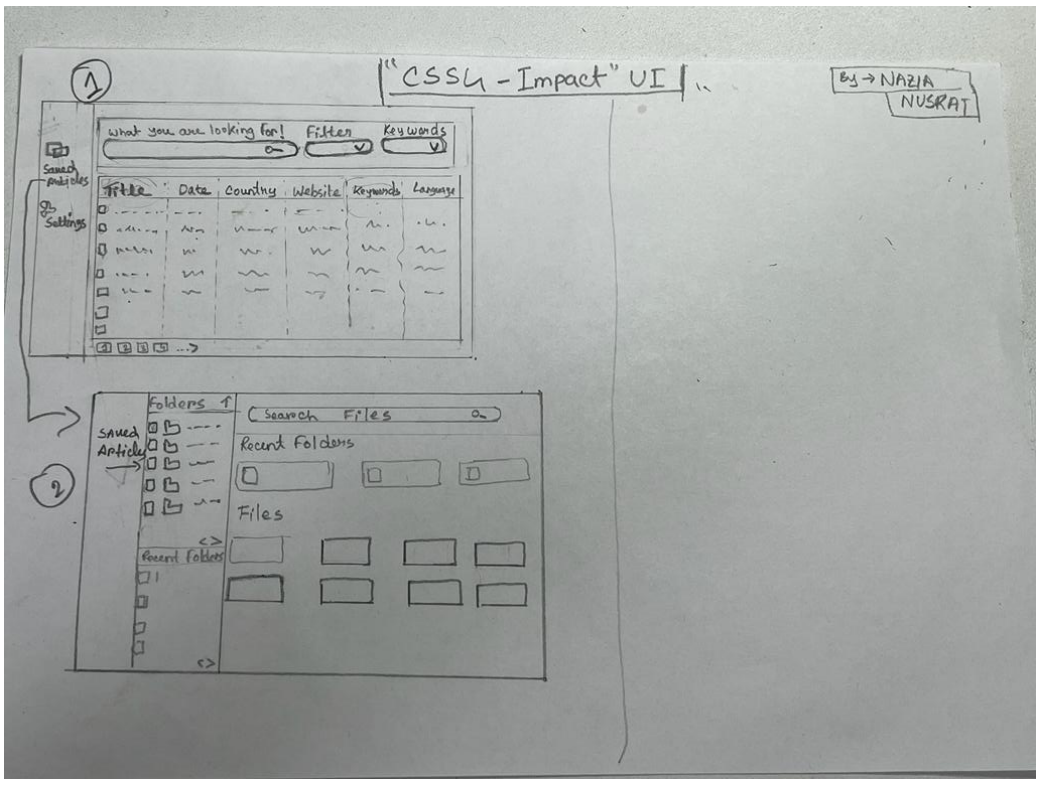
Occupation: Head of the Politics and Economics Department,
Deputy Director.
Member of the Institute for Political Science at the University of Bremen

Persona: Currently, the problem he faces:

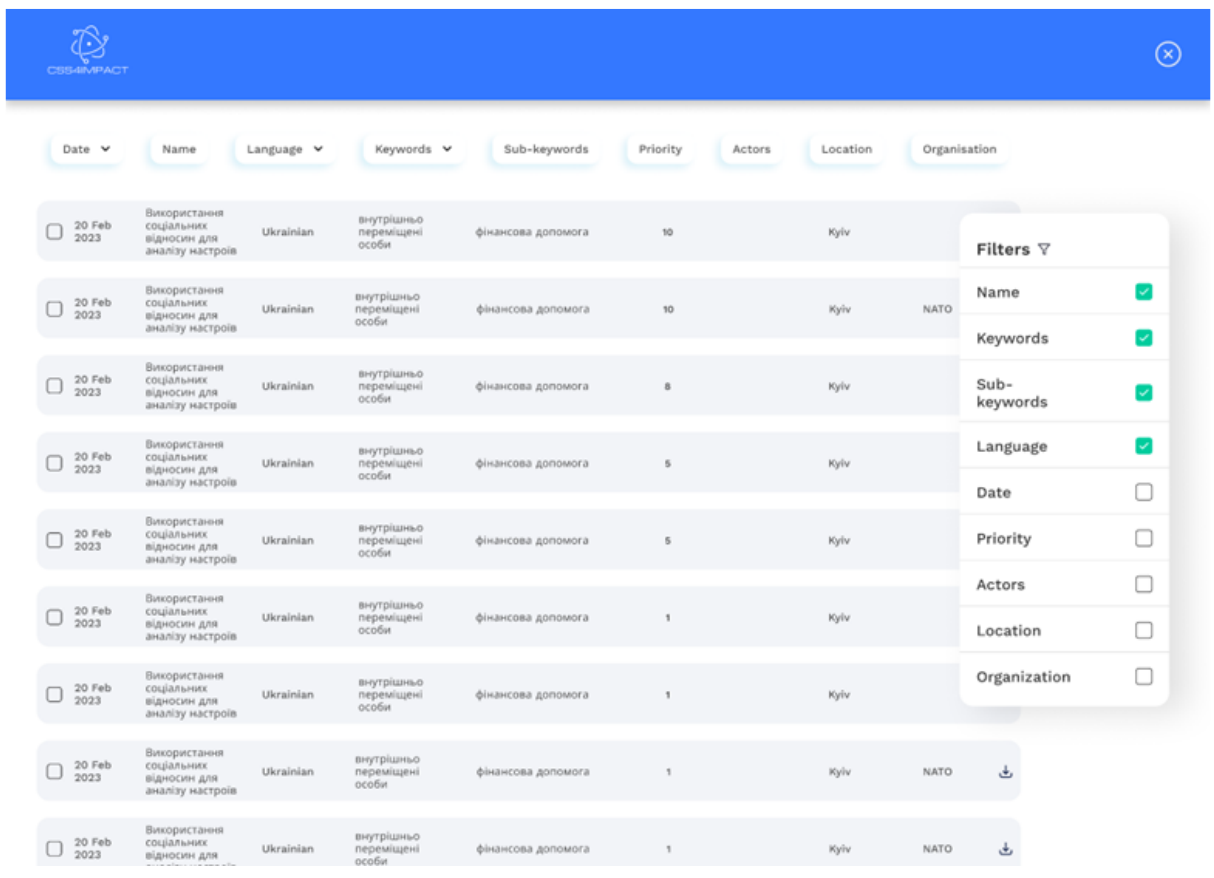
- He had to search for relevant articles manually. He and his team use the same keywords provided to us, and look them up on the same list of websites.
- He had to download the articles individually
- He and his team read the articles from oldest to newest or in the time frame of the relevant events that happened, it is highly dependent on the country and its political and sociological situation

Website: https://www.forschungsstelle.unibremen.de/de/3/20110606111616/20110620171822/Leiter_Arbeitsbereich_Politik_Wirtschaft_stellvertretender_Direktor.html

Low-Fi Wireframe:



Design:



The grid-based UI model allows the process of sorting articles effortless. User can sort articles by Languages, Keywords, and Date of publication. Additionally, users can see meta data like sub keywords existing in the article, priority score, actors, location and organizations mentioned in the article. In the end user can download selected articles.

Definitions, acronyms, and abbreviations

IDP: Internally Displaced People
NLP: Natural Language Processing
SNA: Social Network Analysis

Keywords:

The provided keywords are listed in here: https://seafile.zfn.uni-bremen.de/d/b9d535059597474d8829/?p=%2F

Front End:

The front end of the system will be implemented with React Native.

Back End:

The back end of the system will be implemented with (MongoDB).

References

Keywords, websites and social media accounts can be found under this link: https://seafile.zfn.uni-bremen.de/d/b9d535059597474d8829/