**FACULTY OF SCIENCE, ENGINEERING AND COMPUTING**

**School of *Computer Science & Mathematics***

**BSc DEGREE**

**IN**

***Computer Science***

**PROJECT PROPOSAL**

Name: Tomasz Przybylski

ID: K1602155

Title: A News Outlet, Journal and Magazine Search Tool for Investigative Journalists

Project Type: Build

Date: 11/10/2019

Supervisor: James Orwell

KU London Logo

Did you discuss and agree the viability of your project idea with your supervisor? Yes

Did you submit a draft of your proposal to your supervisor? Yes

Did you receive feedback from your supervisor on any submitted draft? Yes or No

[• Abstract 3](#_Toc21552726)

[• Aims and Objectives 3](#_Toc21552727)

[Aims: 3](#_Toc21552728)

[Objectives: 3](#_Toc21552729)

[• Technologies and Resources 4](#_Toc21552730)

[• Method and Workplan 4](#_Toc21552731)

[MoSCoW: 4](#_Toc21552732)

[• Legal, ethical, societal, security 5](#_Toc21552733)

[Appendix 5](#_Toc21552734)

• Abstract

As it stands, there is no tool for Investigative Journalists to quickly search through a large amount of different news outlets or magazines or scientific/literary journals. According to my customer, the traditional search engines such as Google, Bing etc. are not suited to the needs of investigative journalism and could greatly be improved for researching. This project will not aim to replace the large search engines but rather build a tool that will perform the job better for this use case.

• Introduction and Background

Researching a topic through search engines is tiresome and time consuming. Current search engines are not designed for this purpose, and as such one can often come across difficulties such as repetitive articles or irrelevant websites. This process also requires one to view tens or hundreds of articles, with each article having to be accessed individually, meaning lots of network traffic and bandwidth usage. This is a problem for individuals who are in places where internet is not readily available or where it is not at a standard as to accommodate such traffic. This project hopes to alleviate some of these problems for Investigative Journalists, Independent Researchers, Students and Individuals (henceforth will be referred to as “Users”). It will also allow authors of articles, journals and other reports to share their creations through another medium, increasing its reach.

The importance of free speech and journalism is paramount. It is a basic human right, the pinnacle of modern society and the crucible for all progress and inventions. This project is an extension of that and aims to aid anyone who wishes to pursue investigative journalism and research. It is important enough to pursue for this final year project and it is important enough to my customer to seek someone to create this tool.

As mentioned before, search engines can be successfully used to perform this task, although they leave a lot to be desired. A tool very alike the aims of this project is LexisNexis, an American “Computer-Assisted Legal Research” (CALR) tool. According to LexisNexis, “*LexisNexis gives you access to some of the highest-quality legal research and intelligence on the market. Our solutions help lawyers sift through varied sources of information to quickly get to the right information for the work they’re doing and provide practical guidance in a wide range of practice areas.*”[1]. They boast the largest legal and public-domain knowledge database in the world. Another similar tool that works on the same premise is Google Scholar, a free CALR tool. However, both LexisNexis and Google Scholar only works for Scholarly Articles and Case Laws and is only useful for legal research.   
There are a half dozen subscription-based CALR web services (such as LexisNexis) mostly based in but not limited to the US and the UK, creating a global “commercial market surpassing $8b” [2].

• Aims and Objectives

Aims:

The aim for this project is to create a tool that professional Investigative Journalists, Researchers, Students and Individuals (“Users”) could use in order to aid and speed up the process of researching relevant topics or information, and to aid Users who do not have enough bandwidth to facilitate regular browsing. The project aims to have a prototype ready by mid-December and the completed project by late-April

Objectives:

To achieve my aim, the following will need to be achieved:  
- Delivery of an Android GUI or webservice containing a home page, search page, results page, basic document manipulation page, document printout page, profile page that reacts to user input almost instantaneously.  
- Finding or manipulation of 2 or more existing databases of documents, journals, news articles or any other applicable media   
- [Major Contingency] If no applicable databases exist or if none are accessible to the project, there will be a need to create 2 or more databases and populate them with data (articles, documents, journals etc.) of at least 500 each entries  
- Adaptation of existing or designing and creation of 3 new searching and sorting algorithms for the project to use on the document databases.  
- [Major Contingency] A webserver or web service may be necessary to enable full functionality.

• Technologies and Resources

So far the following technologies and resources have been identified as necessary:  
- A database management and creation tool (such as phpMyAdmin)  
- Java, JDK11, Java-compatible IDE  
- AndroidSDK, Android Virtual Device Manager, Android 7.1.1  
- Possibly a server, a file transfer manager (such as WinSCP), a domain name  
- Git & Jira  
- Co-operation from a news source, scientific journal provider or independent authors/publishers  
- Access to journal, news, article databases (if they exist)  
- Windows Hypervisor Platform for AMD Ryzen compatibility  
- Sorting & searching Algorithms

• Method and Workplan

The project will have 5 milestones, with the first milestone being completed by Christmas 2019 and the fifth and last milestone being completed by May 2020 and consisting mainly of testing and final stretch goals. The full list of milestones with rough dates are as follows:

\*GANTT\*

An App GUI or webservice that reacts to user input in under 40ms – Prototype by milestone 1, completed by milestone 3.  
A searching algorithm that accepts up to 2 user inputs of keywords and search terms and chooses relevant content based on both of inputs – Completed by milestone 2.  
App accesses content stored on a database in under 40ms – Completed by milestone 1.  
Accessed content displayed on App GUI/webservice – Completed by milestone 2.  
App is available on Android store or via download link on website – Completed by milestone 5.  
User can download up to 3 files from database onto device – Completed by milestone 3.  
A searching algorithm that accepts up to 5 user inputs of keywords and search terms and chooses relevant content based on both of inputs – Completed by milestone 2.  
App automatically modifies files chosen for download by highlighting searched key terms. – Completed by milestone 4.  
User can download up to 10 files from database onto device – Completed by milestone 3.  
User can manipulate their version of files in the app (i.e. delete sections, merge file content together etc.) – Completed by milestone 4.  
User history and saved search terms.  
Storing basic user settings such as profiles (username & password), frequently searched terms, related topics/articles

\*Not Finished Yet - How will you use the technologies and resources to achieve your aims? Show the major phases of the project, milestones and deliverables. Consider major contingencies. Generate a schedule using a Gantt chart or similar form.

• Legal, ethical, societal, security

As the project does not use user data or user inputs and does not store them/use them internally, ethical issues are practically non-existent. However, if the ‘Could Have’ profile objective is achieved, user personal data such as email address, password, name and browsing information will have to be kept according to UK ethical and legal laws, for example GDPR 2016[4]. Another aspect of legal issues is the usage and storage of public-domain documents and published articles. The project will have to abide by UK publishing and referencing law. These are typically along the lines of correct citation and acknowledgement of the sources [3].

Societal issues include ensuring that all documents accessible by the app abide by UK law and do not promote discrimination, racism, hate etc. Also means that all documents must be checked to confirm the author. Another issue is to ensure that the app and all documents are accessible to all people.

Security issues relate to the storing of user personal data if the corresponding “Could Have” objective is met. This includes usernames, real names, addresses, passwords, browsing data, device permissions and anything else that classifies as user data under the General Data Protection Regulation or any other applicable privacy laws. Another issue is the security of the documents themselves and prevention of tampering through the app or on the database side. Access should only exist for authorised persons.

Under the current scope of the project, I do not believe there is a need for ethical approval or oversight.

Appendix

1. LexisNexis (2019) *About us* [online] Available at : <https://www.lexisnexis.co.uk/about-us/about-us> [Accessed 06.10.19]
2. New York Times (2019) *Harvard Law Library Readies Trove of Decision for Digital Age* [online] Available at: <https://www.nytimes.com/2015/10/29/us/harvard-law-library-sacrifices-a-trove-for-the-sake-of-a-free-database.html> [Accessed 06.10.19]
3. NCBI (2019) *PMC Copyright Notice* [online] Available at: <https://www.ncbi.nlm.nih.gov/pmc/about/copyright/> [Accessed 07.10.19]
4. Investopedia, (2019). *All you Need to Know About GDPR, the New Data Law*. [online] Available at: <https://www.investopedia.com/investing/all-you-need-know-about-gdpr-new-data-law/> [Accessed 11.10.19]
5. Website name, (Year published). Page title. [online] Available at: URL [Accessed Day Mo. Year].