**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

[• Abstract 3](#_Toc22247803)

[• Introduction and Background 3](#_Toc22247804)

[• Aims and Objectives 4](#_Toc22247805)

[Aims: 4](#_Toc22247806)

[Objectives: 4](#_Toc22247807)

[• Technologies and Resources 4](#_Toc22247808)

[• Method and Workplan 4](#_Toc22247809)

[Gantt: 5](#_Toc22247810)

[Deliverables: 5](#_Toc22247811)

[• Legal, ethical, societal, security 6](#_Toc22247812)

[Appendix 6](#_Toc22247813)

**•** Abstract

Currently, there is no tool for Investigative Journalists to quickly search through a large amount of different news outlets or magazines or scientific/literary journals. An individual has agreed to be a client for this project: they state that 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. This tool will be evaluated by the client upon completion, and prospects for this tool will be investigated. The project will be undertaken through weekly sprints and the use of Crystal-Clear Agile methodology, with an increased focus on the frequent delivery of working software and interaction with the client.

• **Introduction and Background**

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.

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.

Scholarly journals and academic articles have a formal network of connections & references (‘citations’) that show a clear-cut path to the original data or information; however, such a standard does not exist for most news articles, magazines & many other instances of public-domain information.

As mentioned before, search engines can be successfully used to perform the task of extensive research into a topic, 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 work for Scholarly Articles and Case Laws and are only useful for legal research.   
LexisNexis makes up one of half a dozen subscription-based CALR web services mostly based in but not limited to the US and the UK, creating a global “commercial market surpassing $8b” [2]. There may be other proprietary solutions, however there is no way to evaluate these without access.

• Aims and Objectives

Aims:

An 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 intensive browsing.

This project will explore the possibility of creating a standard for news articles, magazines & public-domain information, like the citations and referencing used for scholarly articles and will use this standard for determining which articles build on or relate to others.

Objectives:

To achieve the aims of the project, the following will need to be achieved:  
- Thorough research of similar industry solutions & their shortcomings  
- Definition of deliverables and objectives into MoSCoW (Must have, should have, could have, wont have this time)  
- Creation of some hand-constructed examples of the project’s designs and proof of concept  
- 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.  
- Utilizing 2 or more existing databases of documents, journals, news articles or any other applicable media   
- Adaptation of existing or designing and creation and evaluation of 3 new searching and sorting algorithms for the project to use on the document databases.  
- Close cooperation with potential stakeholders, such as the Users, article authors, database providers and the client.  
- Evaluation of design and implementation and presentation of completed product to the client  
- Support for the product post-deployment

• 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 early December 2019 and the fifth and last milestone being completed by June 2020 and consisting mainly of testing and final stretch goals. The full list of milestones with rough dates are as follows:

- Milestone 1: 08.10.19 – 03.12.19  
- Milestone 2: 03.12.19 – 04.02.20  
- Milestone 3: 04.02.20 – 07.04.20  
- Milestone 4: 07.05.20 – 05.05.20  
- Milestone 5: 05.05.20 – 02.06.20

Gantt:

Figure 1 shows an early Gantt chart of this project and the approximate durations and dates of main deliverables and milestones. This Gantt chart is prone to changes during the project life due to the Agile methodology in use for the purpose of this project.

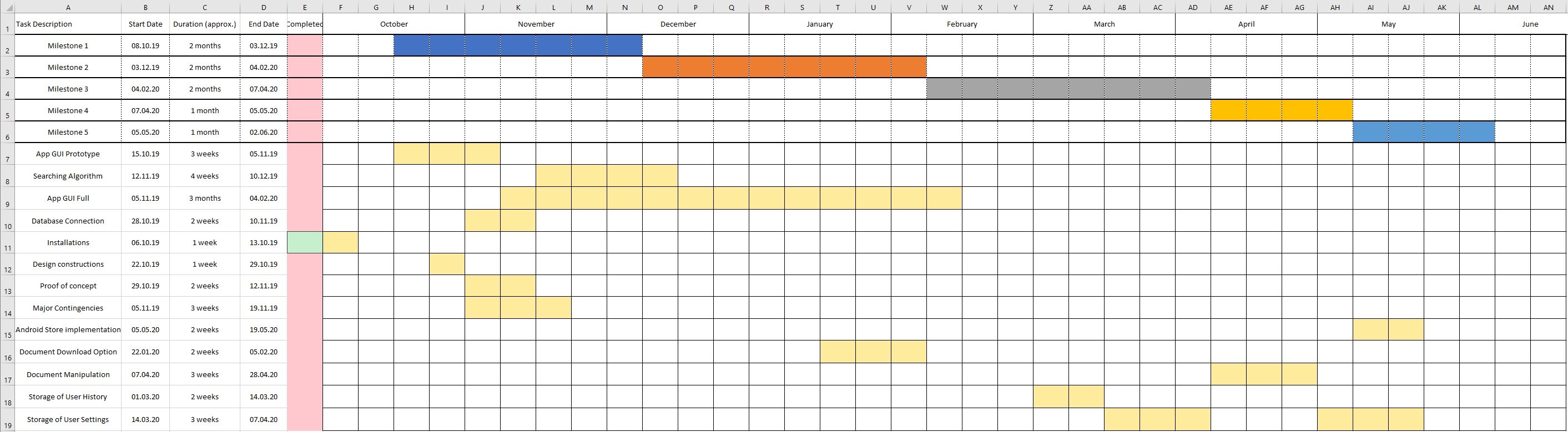


Figure 2 – Gantt chart Key

Figure 1 – Gantt chart at Project start



Deliverables:

- 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 4.  
- 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 – Completed by milestone 3.  
- Storing basic user settings such as profiles (username & password), frequently searched terms, related topics/articles – Completed by milestone 4

Major Contingencies:  
- A webserver or web service may be necessary to enable full functionality.  
- If no applicable databases exist or if none are accessible to the project, create 2 or more databases and populate them with data (articles, documents, journals etc.) – 500 entries each

Version control will be done through Gitlab and Sprints will be managed through Jira. This project will be undertaken using Crystal-Clear Agile methodology with the above set out schedule being adjusted as the need arises. The Crystal-Clear Agile methodology will be adapted to suit the needs of this project. The methodology will have an increased focus on delivery of working software and continuous integration, and on constant cooperation with the client of the project for guidance on the project’s direction and design.

• 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 Profile setting storage 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 2016-2018[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 objectives are met. This includes usernames, real names, addresses, passwords, browsing data, device permissions and anything else classified 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 back end. Access should only exist for authorised persons.

Under the current scope of the project, 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].