Creating an Advanced Research Workbench

Hugh Bedford
bdfhug001@myuct.ac.za
University of Cape Town
Rondebosch, Cape Town, South Africa

Dean Williams
wlldea003@myuct.ac.za
University of Cape Town
Rondebosch, Cape Town, South Africa

KEYWORDS

user requirements, digital libraries, NDLTD, research workbench, visualization, reading lists

1 PROJECT DESCRIPTION

The invention and innovation of the Internet over the past 25 years has led to the digitalization of publishing and a large increase in the amount of online content [12]. As a result, there was, and still is, an ever increasing need to manage the vast quantity of electronic data [9]. Digital libraries provided a method to address this need by allowing materials to be stored in an electronic format and allowing users to manipulate large collections of data [18]. There are now a large number of digital libraries available to use, several of which store academic literature such as electronic theses and dissertations. In his 1945 essay "As We May Think", Bush explained how our methods of reviewing results of research are generations old and are inadequate for their purpose [5]. This has largely remained unchanged, with most digital libraries having extensive resources but lacking powerful services that allow users to interact with those resources effectively [19]. Users struggle to find resources related to their informational needs and there is a lack of services to make the processes easier [19].

In order for society to progress, the efficient dissemination of knowledge across the world is paramount. As discussed by Bush, the ability to access and share such knowledge allows for progress in every facet of society [5]. Digital libraries, especially the Networked Digital Library of Theses and Dissertations (NDLTD), may be able to improve research efficiency by implementing enhanced user services. Lombardi explained that helping users find resources effectively and easily online with several other disorganised resources is the main priority of an academic library [22].

When conducting an analysis of literature in related areas, it was found that users wanted to have the ability to see most or all of the resources from the same subject area and were not able to [10]. The vast number of ETDs available on NDLTD makes it difficult to determine which to read [16]. Furthermore, studies have found that users prefer more visual-based interfaces for finding information [20]. Currently, NDLTD does not provide such features to its users.

2 RESEARCH OBJECTIVES

There are several academic search engines, like Google Scholar and Arnetminer, that provide tools that help researchers with searching for papers. However, their functionality and effectiveness leave much to be desired. When considering our primary focus, NDLTD, it has even less functionality than the aforementioned sites. Users of NDLTD are able to make a search on the library database for electronic theses and dissertations, but are provided with few tools to aid their searches.

The goal of this project is to provide functionality to the NDLTD global search that will allow researchers to access useful and relevant information more efficiently than the current academic search engines available and to expose users to more resources than they would be with a simple search. To accomplish this, we aim to implement two types of organisation of resources. The first is automatic organisation of content and the second is user organised content. Automatic organisation will consist of two central features. Firstly, resources will be categorised automatically based on their research topic and, secondly, users will be provided with a visualization of this organization. Similarly, user organisation of resources will consist of two central features. Firstly, users will have the ability to create public reading lists that can be viewed and accessed by all users. Users will also have the option of creating private reading lists, which are only accessible by the users who created them. This will effectively allow the user to bookmark content for later reference. In order to investigate these new organizational structures, the following research questions are proposed:

(1) Does automatic categorization of research topics provide a meaningful and useful organization?
(2) Will the new visualization of resources provide a more useful interface to users than the ones currently provided?
(3) Will users have a more positive user experience using the new visualization as opposed to the NDLTD Global Search?
(4) Will public reading lists provide a useful recommendation feature to help users find relevant content?
(5) Will private reading lists provide a better user experience than the current bookmarking tools users have at their disposal?

3 PROCEDURES AND METHODS

3.1 System Design

THETOOL will be developed in a layered architecture with the intention of layers being able to be developed and adjusted independently. The layers will follow the Model-View-Controller (MVC) design pattern. A diagram illustrating the architecture is shown in Figure 1.

3.1.1 View. The view of the system will be accessible via its webpage. The view will present the user services to the user.

3.1.2 Controller. The controller layer will handle all the user services. The controller will connect the user interface with the database objects, which, in our case, will be the electronic theses and dissertations. This layer will allow the user to interact with the user services that the system will provide (described in section 2) and manipulate the model (database) depending on user behaviour. The Controller was also be responsible for interfacing with the existing NDLTD services.

3.1.3 Model. This layer involves the storage and organisation of the digital objects. NDLTD metadata will be accessed from the currently existing database and search engine. A new database will be manipulated based on the user profile.

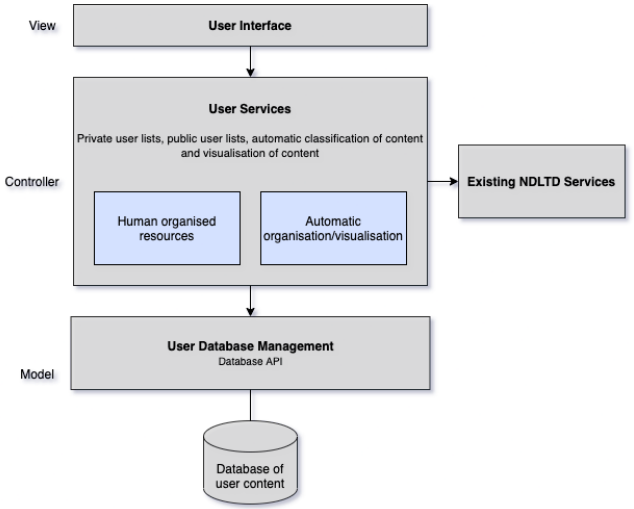


Figure 1: The architecture design on the system

3.2 Development Platform

THETOOL will be designed as a web-based application. This approach will be taken for several reasons. Firstly, the NDLTD Global Search already existed as a web-based application. Secondly, we believe researchers will not want to download an extra application. Lastly, if we develop local applications, the architectures of researchers' machines would need to be considered.

The View of the application will be developed in HTML, CSS and JavaScript. The front end will be dynamic and responsive. We could also make use of available frameworks such as D3.js to create responsive visualizations [7].

The Controller will be developed using Java and used to interface with the databases. The user database will be a MySQL database, just like the NDLTD metadata database. NDLTD metadata is stored as XML-formatted data [8]. Java libraries including Java DOM Parser and JDBC will be used to handle database queries and parse the XML metadata stored in the NDLTD database [14, 15].

3.3 Implementation Strategy

The development of THETOOL will follow an iterative development methodology to allow for evolving requirements and challenges we may encounter. We aim to loosely follow the SCRUM framework with weekly sprints and scrum meetings. SCRUM is a framework that helps the team develop a system through adaptive solutions for complex problems, embracing change and promoting an environment where all team members share an equal voice [13, 17].

This development approach will allow us to cater for possible changes to the requirements or unforeseen challenges that we may encounter.

We will develop our system from the ground up, starting with the Model, then Controller and finally the View. Once the system has been developed and tested, we will evaluate the system with UCT postgraduates to determine whether the system answers the research questions.

The first step of the development process involves gathering the existing NDLTD metadata database and search algorithms to run locally on our laptops. This will allow for easier development, and mitigate risks of connectivity issues that will occur during load shedding. The correct frameworks to query the database will also need to be installed. Once this has been set up correctly, the user services will be developed. The user services will be developed concurrently while waiting for ethical clearance from UCT before the user evaluation can begin. Private user lists and automatic generation of resources will be the first two user services that will be developed for the system. Upon completion of development of those services, the developers will move onto developing the public user lists and the visualization of resources.

Once the back-end of the user services have been developed, we will work together to design and develop the user interface of the system. The user interface will present all the user services to the user, so it will need to be developed in collaboration. Part of the user interface development process will be developing the system that handles user login. When we test the system internally and deploy the system onto a web hosting platform, we will concurrently prepare our usability questionnaire for the user survey phase of the implementation.

When developing the system, we will conduct regular simple unit tests to ensure the development of the services goes as planned. These tests will be conducted throughout the development process. Since we are testing usability and user preference, the system provided to users needs to be fully fleshed out and thus we cannot conduct surveys using early prototypes.

3.4 Experimental Design

There are two initial steps to setting up our user evaluations.

The first step is drafting up the system evaluation questionnaire. Our questionnaire will draw heavily from existing software evaluation questionnaires such as the System Usability Scale, Computer System Usability Questionnaire and the Questionnaire for User Interaction Satisfaction [3, 6, 11].The questionnaire will also contain several questions specifically designed to answer our research questions. Questions will focus on user experience of the system as a whole as well as the separate features implemented. Additionally, many questions will be focused on comparing our new features to the features (or lack of features) of the NDLTD Global Search web page. This will be done in order to answer the 2nd, 3rd and 5th research questions.

The second step is to draft up a list of instructions for the users to answer the questionnaires. These instructions will tell the users what tasks they must complete (or what actions they must perform) on both the NDTLD Global Search as well as our new web page. Instructions could include general tasks such as "Create a private reading list" to test user experience, as well as specific instructions such as "Enter 'Machine Learning' in the search bar" in order to

guide users to specific states in order to perform more instructions and tasks.

Since we will be comparing two separate web pages, our user evaluations will have to take this into account to keep the evaluations fair. For example, we will need to split the users up into two equal sized groups. The first group will use the NDLTD Global Search before using our web page, while the second group will use the two systems in the reverse order. This will help reduce bias. Other than this, we will make use of several other techniques expected of a scientific experiment to keep the evaluation as unbiased and fair as possible.

Once both documents have been drafted, we will then begin conducting the user evaluation. With the aid of our supervisor, we will gather a large enough sample size of Postgraduate students who are willing to partake in our experiment. Users will be provided the instructions document as a PDF as well as a link to the questionnaire, all via email.

We will make use of Lime Survey for the questionnaire. The tool allows us to use tokens to remove the association of a name with a participant response. Once users have finished answering the questionnaires on Lime Survey, we will gather up the results and start analysing them, with the primary goal of answering our research questions.

4 EXPECTED CHALLENGES

There are several challenges that we expect to encounter with the development of this research tool. We are yet to receive the NDLTD metadata database. After receiving the current NDLTD system, we will have a better understanding of how long it will take us to learn and understand the system. Before developing services that interact with the current NDLTD system, we need to understand how the system operates.

Physical meetings are often more successful when trying to communicate with team members. Since physical meetings are not advised during the current pandemic, all communication will need to take place online. This makes it harder to communicate, especially due to factors such as load shedding affecting connectivity.

Recruiting users to participate in our user survey is also an expected challenge. Postgraduates are busy individuals, so getting them to participate in an experiment that will take up their time may prove to be a challenge.

5 ETHICAL AND PROFESSIONAL ISSUES

Since we need to survey users to determine the overall success of our system, we will require ethical clearance. We will obtain ethical clearance from the UCT Human Research Ethics Committee. We will communicate the purpose of the research clearly to our participants before we ask them to participate in the study. Participants will be informed that their individual responses will not be made public or seen by anyone outside the project team. We will be clear that the project team will at no point make any physical contact with survey participants. If we need to contact a participant outside of the questionnaire, it will be done via online meetings services such as Microsoft Teams to observe social distancing guidelines.

Since we believe the spread of knowledge across the world to be essential to humanity's progression, we shall be keeping the project free and open source.

6 RELATED WORK

This section aims to highlight the need for certain user services by looking at related work. Overall, the use of a digital library should be easy to learn [10, 20]. The easier the digital library is to learn, the quicker researchers are able to find the work they are looking for. This highlights the importance of having a user interface with a focus on ease of use and user experience. To use digital libraries effectively and get users to extract the maximum benefit, support structures should be added to the service to aid users [2, 4, 20]. As mentioned, user features should be easy to understand, however, if they are slightly more complex, clear instructions will be provided to the user.

The literature reviewed presents evidence that there is a user need for digital libraries and Electronic Theses and Dissertations (ETDs) to recommend other literature that may be of interest to the user based on what the user is currently looking at [2].

Agosti and Orio found that professional researchers wanted to consider resources from other online collections or similar resources within the same collection [2]. It was also found that researchers often need to see most of or all the resources from the same subject area and are not able to [10]. This illustrates how there is a gap for tools that provide services such as these. Kani-Zabihi et. al. found users also want to be able to list the most important resources [10].

Agosti et al. found that if query results are presented in a clear and more visual format, it might appeal to and stimulate novice and non-domain users of the digital library [1]. This again is a feature we aim to develop and the literature shows it is a desired feature. Not only would it be appealing to novice users but a clearer presentation of results would also make digital libraries more efficient for all users overall [21]. Sweetnam et al. also found that there was a preference amongst the studied users for a visual-based interface for finding information [20]. We therefore conclude that there is an overwhelming amount of evidence suggesting that visualization of content will enhance user experience.

There are several tools that aim to aid research by offering a few services similar to the ones we aim to develop. An application called Zotero has features that aid with the organisation of research into different collections [24]. Secondly, VOSviewer offers text mining functionality that can be used to construct and visualize networks of important terms extracted from a body of scientific literature [23].

7 ANTICIPATED OUTCOMES

7.1 System

7.1.1 User lists. Users will be able to create lists of papers they are interested in. This could be papers in similar subjects, papers they are currently reviewing, or simply papers they have enjoyed. Users who search for a paper that exists in another users' list will be recommended that user list, but only if the user makes their list public. This will help with recommending papers to users by making use of other research already done.

7.1.2 Automatic organisation. Papers will be categorized automatically based on their metadata and displayed accordingly. Such categorization shall be similar to that of a library, where books are categorized based on their genre. Analysis of the metadata will be used to automatically create "Genres" of ETDs. ETDs will be grouped into these categories.

7.1.3 Visualization of resources. ETDs will be displayed in a visually pleasing manner, allowing users to understand and navigate search results easily. Elements of physical libraries, such as the grouping of books into genres, shall be incorporated into the visualization. Thus when searching for ETDs, users will be provided an easy to use and navigate visualization of the automatically categorized ETDs.

7.2 Expected impact of the project

We hope to create a set of services that will greatly impact the methods for research. We hope that these services will be a useful and efficient means of retrieving information from research papers. Ultimately, these tools will greatly aid the spread of information, allowing for efficient dissemination of knowledge across the world.

7.3 Key Success Factors

The clearest metric of success will be answering the research questions after conducting user evaluation. The project features will be deemed a success if:

- Automatic categorisation of research topics provides a meaningful and useful organisation of resources.
- The visualization was useful for finding relevant information.
- Users showed a preference for the new visualization over older systems.
- Public reading lists are a useful method of recommending resources to the user.
- Private reading lists are an effective way of allowing users to save resources in particular lists to refer back to them.

8 PROJECT PLAN

8.1 Risks

Risks were identified and put into a risk matrix. The matrix can be seen in Appendix A.

8.2 Timeline

The project timeline starts from 3 May and runs until 18 October when the final project deliverable is due. The detailed breakdown of the project schedule can be seen in Appendix B attached.

8.3 Resources Required

Several resources will be required to develop the system and then evaluate whether it is effective in what it's intended to do. Here is a list of the resources we will require:

- Personal computers for system development
- NDLTD metadata
- User interface frameworks, such as d3.js
- Lime Survey hosted at survey.cs.uct.ac.za
- Web hosting services, either Amazon Web Services or Digital Ocean
- OpenProject project management software

8.4 Deliverables

The following is the list of deliverables that will be produced during the project timeline:

- Project proposal presentation
- Initial software feasibility demonstration
- Final paper draft
- Project final papers
- Final project code
- Final software demonstration
- Project poster
- Project website

8.5 Milestones

The following will make up the project milestones:

Milestone	Date
Initial software demonstration	10 August 2021
Final software complete	17 August 2021
System testing complete	21 August 2021
User study questionnaires completed	29 August 2021
Final project paper submission	17 September 2021
Final code submission	20 September 2021
Final software demonstration	4 October 2021

8.6 Work Allocation

Hugh will develop the services that provide human organised content, which involves the public and private user lists. This includes developing a system to allow sharing of reading lists between users. Hugh will also develop an algorithm that promotes public reading lists to users based on relevance. Dean will develop the algorithm for automatic categorization of the ETDs and a visualization for the search results. A diagram illustrating the work allocation can be seen in Appendix C

REFERENCES

[1] Maristella Agosti, Franco Crivellari, Giorgio Maria Di Nunzio, and Silvia Gabrielli. 2010. Understanding User Requirements and Preferences for a Digital Library Web Portal. *Int. J. Digit. Libr.* 11, 4 (Dec. 2010), 225–238. https://doi.org/10.1007/s00799-011-0075-7

[2] Maristella Agosti and Nicola Orio. 2012. User Requirements for Effective Access to Digital Archives of Manuscripts. *Journal of Multimedia* 7 (04 2012). https://doi.org/10.4304/jmm.7.2.217-222

[3] John Brooke. 1995. System Usability Scale (SUS): A quick and dirty usability scale. *Usability Eval. Ind.* 189 (11 1995).

[4] George Buchanan, Sally Cunningham, Ann Blandford, Jon Rimmer, and Claire Warwick. 2005. Information Seeking by Humanities Scholars. *Research and Advanced Technology for Digital Libraries: 9th European Conference* 3652, 218–229. https://doi.org/10.1007/11551362_20

[5] Vannevar Bush. 1945. As We May Think. *The Atlantic Monthly* 176, 1 (July 1945), 101–108. https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/

[6] John P. Chin, Virginia A. Diehl, and Kent L. Norman. 1988. Development of an Instrument Measuring User Satisfaction of the Human-Computer Interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Washington, D.C., USA) *(CHI '88)*. Association for Computing Machinery, New York, NY, USA, 213–218. https://doi.org/10.1145/57167.57203

[7] D3.js. 2021. *Data-Driven Documents.* Retrieved June 18, 2021 from https://d3js.org/

[8] W. Flannery. 2008. Digital Libraries: Policy, Planning and Practice. *Library Management* 29 (2008), 452–453.

[9] Edward Fox, Hussein Suleman, Remesh Gaur, and D Madalli. 2003. Design Architecture: An Introduction and Overview, Design and Usability of Digital Libraries: Case Studies in the Asia Pacific. *Information Science Publishing* (2003), 22–37.

[10] E Kani-Zabihi, G Ghinea, and S Chen. 2006. Digital libraries: what do users want? Digital libraries: what do users want? Digital libraries: what do users want? Digital libraries: what do users want? Digital Libraries: what do users want? *Online Information Review* 30, 4 (2006), 395–412.

[11] James R. Lewis. 1995. IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human–Computer Interaction* 7, 1 (1995), 57–78. https://doi.org/10.1080/10447319509526110 arXiv:https://doi.org/10.1080/10447319509526110

[12] Norbert Lossau. 2004. Search engine technology and digital libraries-libraries need to discover the academic internet. *D-Lib magazine* 6 (2004).

[13] Jim Medlock. 2021. *A Short Introduction to the Scrum Framework.* Retrieved June 18, 2021 from https://medium.com/chingu/a-short-introduction-to-the-scrum-methodology-7a23431b9f17

[14] Oracle. 2021. *Java Documentation - Processing SQL Statements with JDBC.* Retrieved June 18, 2021 from https://docs.oracle.com/javase/tutorial/jdbc/basics/processingsqlstatements.html

[15] Oracle. 2021. *Java Documentation - Reading XML Data into a DOM.* Retrieved June 18, 2021 from https://docs.oracle.com/javase/tutorial/jaxp/dom/readingXML.html

[16] Ryan Richardson, Venkat Srinivasan, and Edward Fox. 2008. Knowledge discovery in digital libraries of electronic theses and dissertations: An NDLTD case study. *Int. J. on Digital Libraries* 9 (11 2008), 163–171. https://doi.org/10.1007/s00799-008-0046-9

[17] Scrum.org. 2021. *What is SCRUM?* Retrieved June 18, 2021 from https://www.scrum.org/resources/what-is-scrum

[18] Michael Seadle and Elke Greifeneder. 2007. Defining a digital library. *Library Hi Tech* 25 (06 2007), 169–173. https://doi.org/10.1108/07378830710754938

[19] Hussein Suleman and Edward Fox. 2001. The Open Archives Initiative. *Journal of Library Administration* 35, 1-2 (2001), 125–145. https://doi.org/10.1300/J111v35n01_08 arXiv:https://doi.org/10.1300/J111v35n01$_0$8

[20] Mark S. Sweetnam, Maristella Agosti, Nicola Orio, Chiara Ponchia, Christina M. Steiner, Eva-Catherine Hillemann, Micheál Ó Siochrú, and Séamus Lawless. 2012. User Needs for Enhanced Engagement with Cultural Heritage Collections. In *Theory and Practice of Digital Libraries*, Panayiotis Zaphiris, George Buchanan, Edie Rasmussen, and Fernando Loizides (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 64–75.

[21] Yin Leng Theng, Norliza Mohd-Nasir, and Harold Thimbleby. 2000. Purpose and Usability of Digital Libraries. In *Proceedings of the Fifth ACM Conference on Digital Libraries* (San Antonio, Texas, USA) *(DL '00)*. Association for Computing Machinery, 238–239. https://doi.org/10.1145/336597.336674

[22] John V.Lombardi. 2000. Academic Libraries in a Digital Age. *D-Lib Magazine* 6, 10 (2000), 1–7.

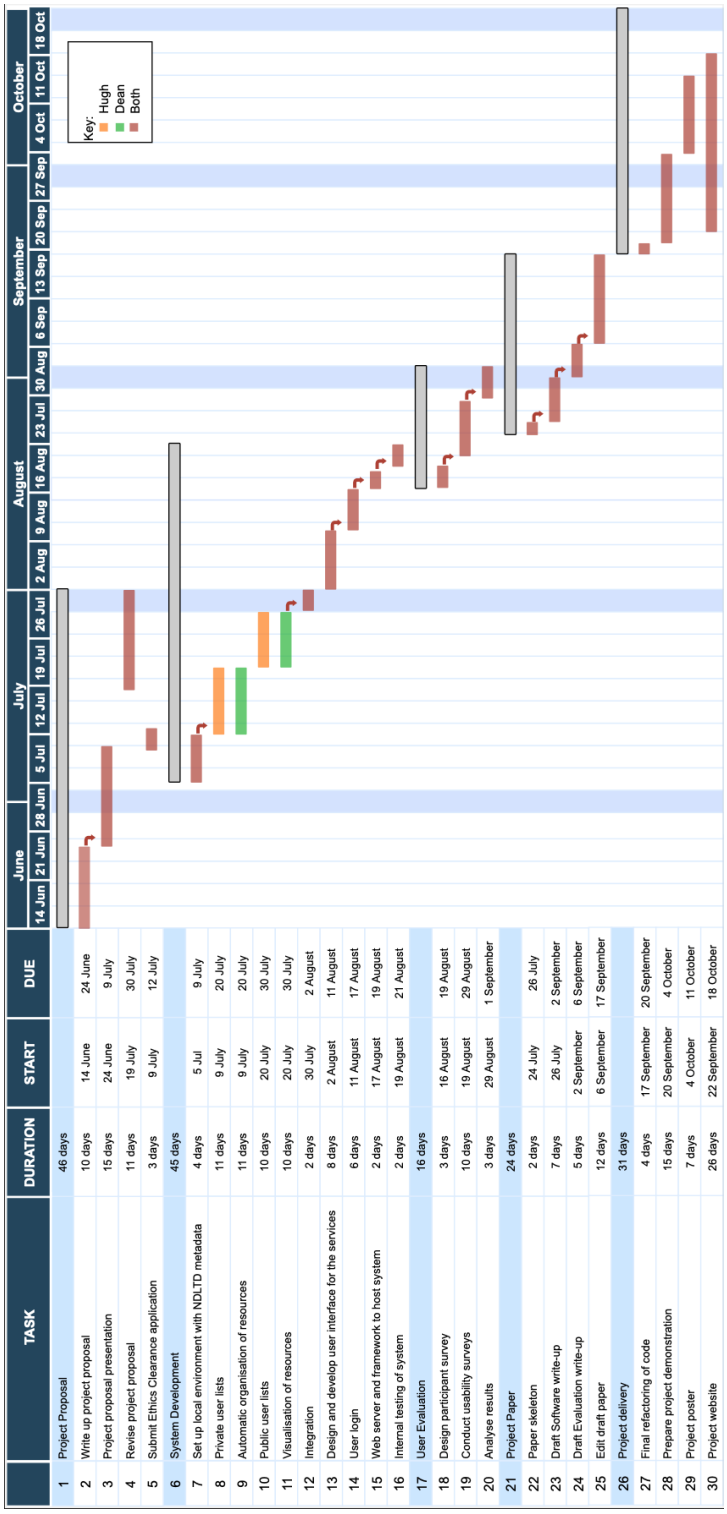
[23] VOSviewer. 2021. *Visualizing scientific landscapes.* Retrieved June 18, 2021 from https://www.vosviewer.com

[24] Zotero. 2021. *Your personal research assistant.* Retrieved June 18, 2021 from https://www.zotero.org

A RISK MATRIX

Risk	Consequence	Probability	Impact	Mitigation	Monitoring	Management
Scope creep	Impair overall quality and, thus, hurt user experience	Low	Critical	Having a clear vision of the system and sticking to the implementation strategy	Have weekly meetings to discuss with group members and supervisors.	Shed unnecessary features if scope creep is identified
Taking time to understand the software framework being used	Delay the development process	High	Marginal	Discuss implementation strategy and frameworks early	Clear communication between team in weekly meetings	Receiving help from team members or project supervisor if needed
Not finding enough users to participate in the survey	Poor accuracy of results	Low	Catastrophic	Ask project supervisor to help to gather participants in the study	Pay attention to the rate at which we gather participants	Increase recruitment efforts
Load shedding disrupting workflow/ reducing available time	Delay the development process	Medium	Marginal (Critical if stage 4 or above)	Regularly backup work	Pay attention to load shedding schedule	Travel to a location which is not currently being loadshed
Slow information retrieval	Negatively affects user experience	Low	Critical	Ensure the search is sufficiently quick before developing other features	Compare system retrieval times with current available system - NDLTD	Add UI features to distract the user from the retrieval times
Partner drops out of honours	Significantly higher workload for remaining user.	Low	Critical	Maintain positive energy and attitude within team	Clear team communications in weekly meetings	Ensure both parts of the project are clearly separable

B TIMELINE



	TASK	DURATION	START	DUE
1	Project Proposal	46 days		
2	Write up project proposal	10 days	14 June	24 June
3	Project proposal presentation	15 days	24 June	9 July
4	Revise project proposal	11 days	19 July	30 July
5	Submit Ethics Clearance application	3 days	9 July	12 July
6	System Development	45 days		
7	Set up local environment with NDLTD metadata	4 days	5 Jul	9 July
8	Private user lists	11 days	9 July	20 July
9	Automatic organisation of resources	11 days	9 July	20 July
10	Public user lists	10 days	20 July	30 July
11	Visualisation of resources	10 days	20 July	30 July
12	Integration	2 days	30 July	2 August
13	Design and develop user interface for the services	8 days	2 August	11 August
14	User login	6 days	11 August	17 August
15	Web server and framework to host system	2 days	17 August	19 August
16	Internal testing of system	2 days	19 August	21 August
17	User Evaluation	16 days		
18	Design participant survey	3 days	16 August	19 August
19	Conduct usability surveys	10 days	19 August	29 August
20	Analyse results	3 days	29 August	1 September
21	Project Paper	24 days		
22	Paper skeleton	2 days	24 July	26 July
23	Draft Software write-up	7 days	26 July	2 September
24	Draft Evaluation write-up	5 days	2 September	6 September
25	Edit draft paper	12 days	6 September	17 September
26	Project delivery	31 days		
27	Final refactoring of code	4 days	17 September	20 September
28	Prepare project demonstration	15 days	20 September	4 October
29	Project poster	7 days	4 October	11 October
30	Project website	26 days	22 September	18 October

Figure 2: Gantt Chart showing the project timeline

C WORK ALLOCATION

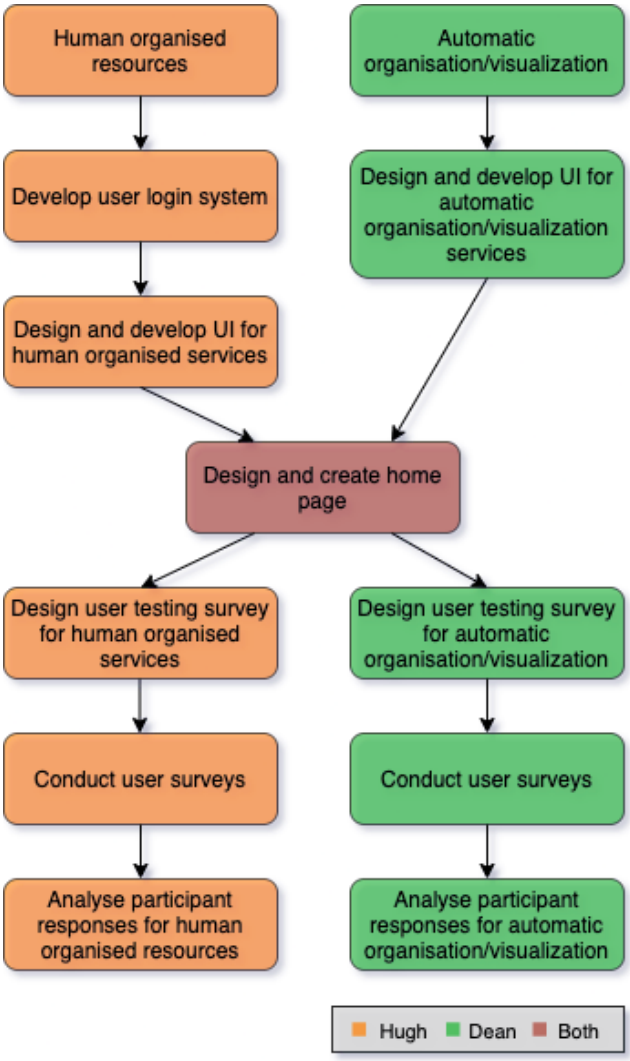


Figure 3: Diagram showing the work allocation for the project