CARDIFF UNIVERSITY

INITIAL PLAN

FINAL YEAR PROJECT
CM3203 – 40 CREDITS

BibTeX Web Database

Author
Mr Miles BUDDEN

Supervisor
Prof. Frank LANGBEIN

February 2, 2020

Contents

1 Project Description

BibTeX is a tool used to insert metadata about references stored in .bib files [3] into the LaTeX typesetting system [5]. It is commonly used by groups of collaborators on a project to track their references. An issue arises when multiple collaborators wish to update the .bib file at the same time as the file must be synced between. This can often lead to errors in the file when collaborators do not do this correctly. Although there are existing solutions for this problem, such as Mendeley, the majority of them are closed source and cannot be self-hosted. Therefore the need for an open source and free alternative is evident.

This project is a tool to store and manage the aforementioned references online. With the help of this tool, collaborators will be able to concurrently add, view, and edit references whilst keeping them synchronised (see subsection 2.1). Whilst adding references to the database, the user will have varying methods of adding references such as uploading existing BibTeX files, manually adding fields, and uploading PDFs to be parsed for information (see section 2.2). Users will be able to download the references in the form of BibTeX files in order to interface with LaTeX. As well as these features for interacting with the references themselves, there is need for a feature to search the text of the PDFs as well as to annotate them (see section 2.3).

The aims for this project are to implement all of the features described above in order to make this project valuable to a researcher in need of an open source alternative for their reference management system. Time permitting more features could be added but I will consider this an extension and therefore not part of the initial plan.

2 Aims and Objectives

2.1 Implementation of a BibTeX Database

My first aim for this project is to allow users to store BibTeX data in a database that is accessible via a web interface. Through this interface, the user should be able to add, view, and edit data stored. This data should not be public and only accessible by the user and any other users the primary user has granted access to. The core objectives of this aim are as follows:

1. The project should be accessible via a web browser.
2. The user should be able to enter data about the reference including files associated. For example, a PDF.
3. Once entered, the user should be able to view this data either in a human readable form or as BibTeX data.
4. This data should be editable on a field by field basis.

2.2 Trivial Data Entry

My second aim is for the entry of data into the database to be as trivial as possible. For example, say the user wishes to enter the data about a PDF copy of a paper, then they can simply upload the PDF and all possible metadata is extracted out and the BibTeX fields are automatically populated. The core objectives of this aim are as follows:

1. The user should be able to enter data about the reference in a form.
2. The user should be able to enter data about the reference by providing a link to the webpage that provides information about the reference.
3. The user should be able to enter data about the reference by uploading a PDF copy of

the reference and having the contents of the PDF automatically parsed and the metadata fields populated appropriately.

4. Messy data entry should be cleaned to match the existing data.

2.3 Searchable Data

My third aim is for all of the data stored on the database to be searchable. This includes all of the data stored in the BibTeX entries as well as the content of the PDF files and any metadata associated with the reference entry. The core objectives of this aim are as follows:

1. There is a search box presented to the user.
2. Upon entering their query, they should be presented with all of the stored references that contain the search query.
3. An exact match is not required.
4. The matched string can be from either the metadata or the reference (PDF) its self.

3 Work Plan

3.1 Work Items

Below are the individual work items I have identified as well as an estimate of how long they will take to complete.

Initial Report – The initial report containing the project description as well as the aims and timeline. I estimate this section to take the first week. This is due to the short length as well as the deadline after the first week limiting the time I can work on it.

Project scaffolding – The core code of the project to allow the web application to be started. I estimate this to take considerably less than a week as the majority of this code can be generated by a scaffolding tool such as Yeoman [6] or a framework specific tool. In this section I am also including other tasks such as the initialisation of a database but the time taken for these tasks is also negligible.

User management – The basic infrastructure to allow a user to register, login, logout, add others to a group. These are fairly standard tasks for a web application and therefore I do not expect these to take longer than a week for full implementation.

Database design – Determine how data is going to be stored in the database and implement said design. Due the unstructured form of BibTeX data, the BibTeX data its self will not require much planning to store as this will be stored in a NoSQL database. The rest of the data (such as user information) will be stored in a relational database. As previously stated, this is very standard data for a web application to store and therefore I do not think this section will take longer than a week.

BibTeX user interface – The interface that will allow the user to manually add fields to the metadata of an article and for said data to be converted to BibTeX data to be stored in the database.

Converting BibTeX data into a format storable by the database – NoSQL databases store data as JSON-like [4] so it would be advantageous to convert the BibTeX data to JSON. I estimate this to take 2 weeks as it is relatively complex to build a parser for one format and to convert in into a new format.

PDF data extraction – Extracting the data out of the PDF including the text it contains as well as any metadata that it also holds. I estimate that this section will take a week as this is a task that has been implemented in other projects and so will not be too challenging.

Metadata acquisition from existing data – Using the information extracted from the PDF, locate its source online as well as any metadata associated with it. I estimate this section to take 2 weeks as there are many possible flows depending on what data can be extracted from the PDF. For example, given the DOI, the complete BibTeX entry can be easily obtained [2]. However, if the DOI is not present, this could be significantly more challenging.

Search Interface – Allow for the user to enter search queries. I estimate this to take a week or less at it is a simple interface element.

Search Implementation – Functionality to allow the user to search all of the data stored in the database. I estimate this to take 2 weeks as

I know relatively little about searching data such as PDF files.

Final Report – This is the final report for the project. I expect this to take 4 weeks if done concurrently with any final outstanding coding tasks.

3.2 Milestones

I have identified 4 milestones for this project. These are as follows:

Milestone 1 – The base implementation is complete. Although lacking in extra features, the code database as well as BibTeX input, user management and groups are complete.

Milestone 2 – The user can now trivially input data. There are multiple options for the user to input data with including scraping of metadata from a PDF.

Milestone 3 – The user can now search all of their references including metadata data as well as the references themselves.

Milestone 4 – The final report is complete and the project is ready for submission.

3.3 Supervisor Meetings

I have agreed to meet once a week with my supervisor. I plan on keeping these meetings shorter and then to have a longer review meeting after each previously stated milestone is reached.

3.4 Timeline

I have represented the timeline of the work items as well as the milestones in a Gantt chart which can be seen in Figure 1. The column dates are the week beginnings that I intend to carry out the corresponding activities in.

4 Ethics

In order to determine whether this project has met its aims, it will be necessary to carry out user testing to check that, for example, data is indeed trivial to enter. This research will require collecting data from test participants about what they thought about the interface and their interactions with it.

In order to align with the University policy on ethics [1], I, as well as my supervisor, will need to undertake the "Research Integrity Online Training Programme" and to notify the Computer Science Ethics Committee.

References

[1] Richard Booth, Alexia Zoumpoulaki, and Liam Turner. *Computer Science and Informatics Ethics.* URL: `https://www.cs.cf.ac.uk/ethics/` (visited on 02/02/2020).

[2] Crossref. *DOI Content Negotiation.* URL: `https://citation.crosscite.org/docs.html` (visited on 11/03/2019).

[3] Alexander Feder. *BibTeX Format Description.* 2006. URL: `http://www.bibtex.org/Format/` (visited on 01/29/2020).

[4] MongoDB. *Documents — MongoDB Manual.* URL: `https://docs.mongodb.com/manual/core/document/` (visited on 02/02/2020).

[5] The LaTeX Project. *LaTeX - A document preparation system.* URL: `https://www.latex-project.org/` (visited on 01/29/2020).

[6] Yeoman. *The web's scaffolding tool for modern webapps — Yeoman.* URL: `https://yeoman.io/` (visited on 02/02/2020).

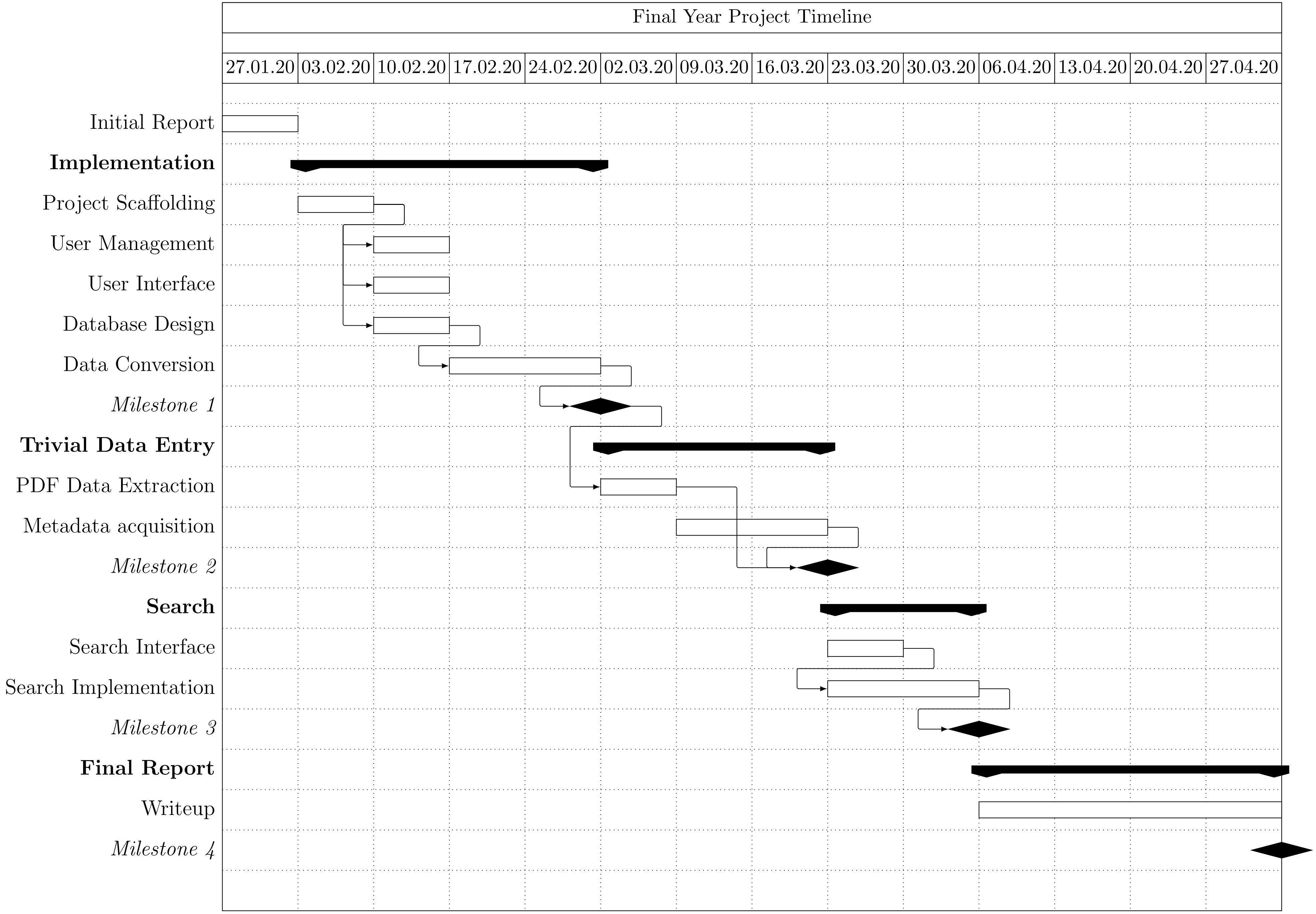

Figure 1: Gantt Chart