Keeping track of electronic reading material

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE

REQUIREMENT FOR THE DEGREE

Of

Master of Science in Computer Science

At

The University of Waikato

By

Mohammed Subhi Alanazi

_____

The University of Waikato

2013

Abstract

Because reading electronically has become widespread, keeping track of the wide range of material one reads has become a problem. To explore this issue, a user study was conducted concerning how people keep track of the materials they have been reading, are currently reading, and are planning to read. We observed that people use different methods for these three tasks, and that there is a discontinuity in the processing of reading materials. Moreover, people have no effective common strategy for keeping track of what they read. As a result of this study, we derived four requirements for developing a system that helps people keep track of their reading materials.

Instead of developing a completely new system, seven available systems were reviewed to find out which one can be extended and cover the requirements. Two open source software projects were chosen to be analysed more deeply (Zotero and Greenstone). As a result of this analysis, we decided to combine the two software systems to work together. Tracking Reading Material System (TRMS) was developed as a solution for the problem. TRMS combines Greenstone and Zotero software, using Greenstone to store and organise documents, and Zotero to capture the documents and metadata. Four participants used TRMS, and evaluated it over a week's use, and completed a diary. Each participant was interviewed individually pre- and post-use of the TRMS. This data showed that the TRMS was partially successful, but modifications are needed for a more satisfactory user experience.

Acknowledgements

In the name of Allah, all praise to Allah for the strength and blessings he has given me to complete this thesis.

I would like to take this opportunity to offer my sincerest gratitude to my supervisor, Dr Annika Hinze, for her consistent support, encouragement, kindness, guidance, and patience as I was carrying out my project and obtaining the needed knowledge. This thesis would not have been possible without her extreme support and close supervision. I thank my co-supervisor, Sally Jo, for her support, encouragement, and help in sometimes.

I also wish to thank my parents and family for their support and encouragement during my study in New Zealand. I would like to tell them that their tears of farewell were not wasted, and they were an incentive for me to do my best to finish my study and to come back to them having been successful.

I would also like to thank the people who contributed to my study and provided assistance during my research. Finally, I would like to thank my friends in New Zealand, especially Mumrize, Amay, Saad, Muteb, and Hassin, as well as those back in Saudi for supporting my best effort on my project.

Table of Contents

List of Figures

List of Tables

Chapter 1

1 Introduction

Reading scientific publications in an electronic form has become common in the last decade (Boyce, King, Montgomery, & Tenopir, 2004). Boyce et al. observed that according to *Ulrich's Periodicals Directory*, more than 80% of the active reviewed journals were provided in digital form. Even though a few of these were available in digital form only, the majority of journals are provided in both print and electronic versions. Often the electronic copy provides additional content. A recent study on book access found that 'people want major reference works, serials, and books online' (Bunkell & Dyas-Correia, 2009). Bunkell and Dyas-Correia observed that, by the time of their study (the study was not general but, rather, was concentrated on a distinct subset of monographs on ScienceDirect[1]), books available online were used more often than print books. Access to electronic materials has increased significantly, supported by the presence of such digital libraries as Scopus, IEEE Xplore, DBLP, ACM, and Google Scholar. These kinds of digital libraries provide access to articles, papers, and books in different disciplines (Hull, Pettifer, & Kell, 2008).

Access via typical bibliographic software is usually impersonal and restricted. In particular, no distinction is made among documents that users are already reading, planning to read, or considering for reading. Because people prefer to have easy and portable access to their reading material (see Chapter 2), many download all the electronic files of articles or books that they want onto their computers, thus ensuring full control of the materials. As a consequence of the number of files, users have difficulties in maintaining a well-defined folder structure and in keeping track of their reading in general. As a result, they sometimes duplicate material and spend much time looking for information or specific papers.

---

[1] http://www.sciencedirect.com/

1.1 Focus

This project aims explore how people keep track of their reading materials. It further aims to develop a system that will help people in keeping track of their reading material. For example, users of the system will know what material they have read, are currently reading, or plan to read. In preparation for developing this system, we conducted a user study to investigate how people deal with their reading material. To explore the effect of the system, we also conducted a short study after people had begun to use our system.

1.2 Structure

The structure of this thesis is as follows:

Chapter 2 describes this study and explores how people keep track of their reading material, including what kind of methods and tools they use. It details the process of collecting and analysing the data, and it presents the findings and discusses them.

Chapter 3 outlines the requirements derived from the findings of the study. It reports how these requirements can compose a system for tracking reading material. The systems that help people to organise digital material will be reviewed and compared to the requirements.

Chapter 4 explains the process of system design, which is presented based on the revealed requirements. It will show how the proposed system (TRMS) was developed and how the process ended with combining Zotero and Greenstone into one system.

Chapter 5 describes the components involve in the implemented system and how they will work together to reach the goal of the study. It explains the problems and solutions encountered in the implementation stage. It will discuss the limitations that might affect the process.

Chapter 6 describes the results of the participants' evaluation of the TRMS for tracking reading materials. It analyses and discusses the participant's feedbacks.

Chapter 7 summarizes the whole thesis and discusses the potential for system development in the future.

Chapter 2

2 User Study: Exploring the Tracking of Reading Material

This chapter reports on the execution and results of a user study in which we explored how people keep track of their reading materials. It includes a description of the study goal, explains how the study was conducted (Sections 2.1 and 2.2), and provides background on the people who participated in the study. The results of the study will be demonstrated in Section 2.3 and the findings will be presented and discussed in Section 2.4.

2.1 The Purpose of the Study

The study aims to explore what kind of materials people keep track of, as well as what software or methods they use for tracking. This knowledge will help enable people to recognize their needs and overcome the issues they face. The results will help them with knowing how to develop a system for the easy tracking of reading material.

2.2 Study Methodology

2.2.1 Procedure

Two methods were chosen for gaining a better understanding of what needed to be explored for the study: an initial questionnaire and an interview. The first step was to obtain an Ethical Approval application, which is required because humans are involved in the study. The approval for this study came from the Human Research Ethics Committee at the Department of Computer Science, University of Waikato (see Appendix A). All the participants were invited by the researcher, via either email or mobile phone, to take part in the study. The participants chose the location for the interview and their preferred time. Some participants were interviewed at their homes and others at their offices.

Prior to each session, the participants were informed about the study and its goals. They were also told that at the end of the study, they would have the opportunity to receive a summary concerning its outcome. Then, each participant was given a

participant information sheet for obtaining confirmation to participate in our study.

In the first part of the study involving the initial questionnaire, participants were given the form and asked to fill it out. The form contained questions about the participants' background: age, job, reading hours, and computer use (see Appendix A). In the interview part, the researcher asked the participants some questions related to their experiences in tracking their reading material. This interview was audio recorded with the participants' consent (see Appendix A).

2.2.2 Participants

Twenty people participated in this study (14 males and six females). Of the participants, 12 were between 18 and 38 years old, and the rest more than 39 years old (see Figure 2.1). Fourteen of the participants were students enrolled in either the Master's or PhD program at Waikato University.
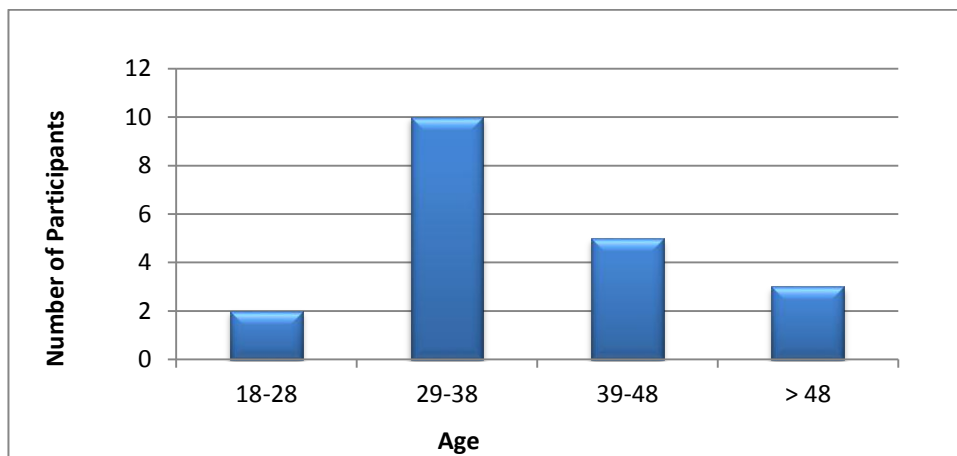
Figure 2.1: Participants' age groups

Half of them were studying in the computer science department and the rest in other departments. Of those who were studying, seven were also working at the university, and the other six were academic staff from the computer science department at Waikato University.

All the participants were highly experienced with using a computer: 18 of them had been using the computer for more than 10 years and the other two, for 5 to 10 years. On average, they were spending more than 4 hours a day on the computer. All of the participants also had experience in writing documents, and about half of

them indicated that they had been writing for approximately 1 to 5 years (see Figure 2.2).
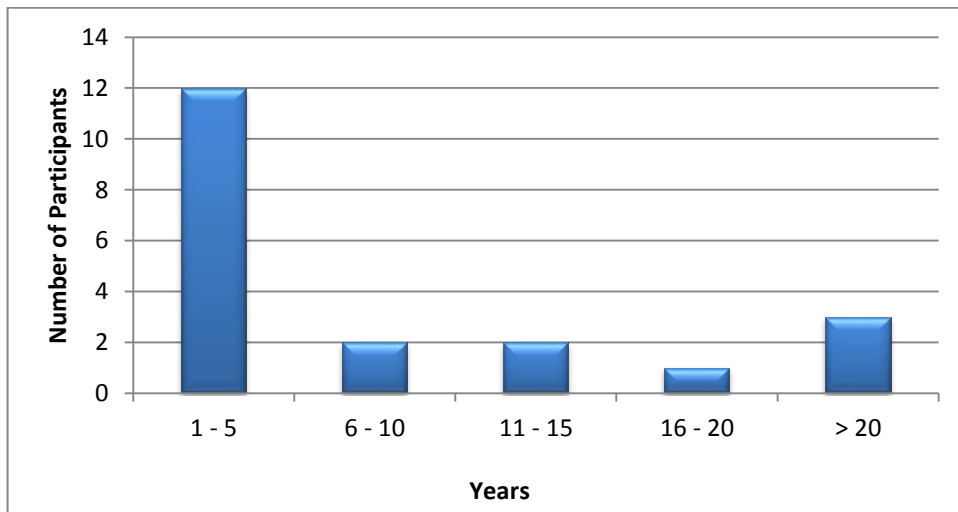
Figure 2.2: Participants' experiences in writing documents

2.3 Data Collection and Analysis

The participants' reading interests were examined to determine whether they were
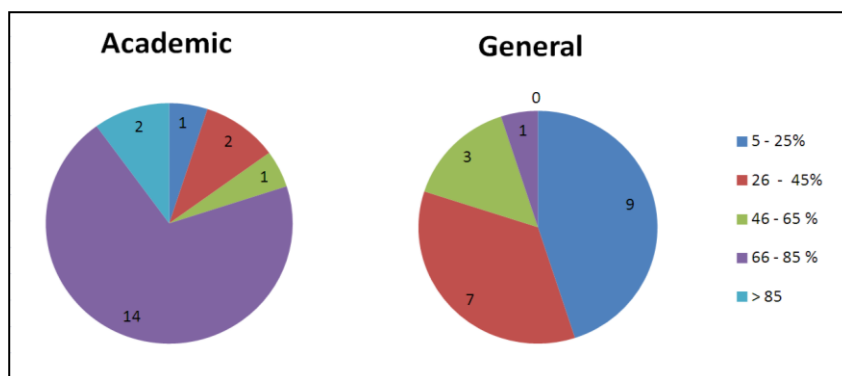
Figure 2.3: Percentages of academic and general reading

academic or general and to discover what kind of material they read in each category. It was also determined whether the participants read the material in printed or digital copies, what kind of devices they used for reading digitally, and finally, how they tracked their materials.

Figure 2.3 shows the results obtained from an analysis of both academic and general reading. Fourteen of the participants indicated that academic reading occupies about 66% to 85% of their reading materials. However, for only one participant did his general reading take up about the same percentage (see Figure

2.3). Two participants indicated that 85% or more of their reading materials were academic, whereas no one mentioned this percentage in his general reading.

Of the 20 participants, about half indicated that less than 26% of their reading is general reading, and only one participant indicated this percentage for his academic materials. Seven participants said that from 26% to 45% of their reading is general, and only two participants gave about the same percentages for academic reading (see Figure 2.3).

2.3.1 Academic Reading

The results for types of academic reading are presented here, as well as preferences for electronic and paper-based reading.

2.3.1.1 Types of Academic Reading

Figure 2.4 shows the participants' responses among the five types of academic reading that were mentioned. From the graph below, we can see that most participants   read   articles   and
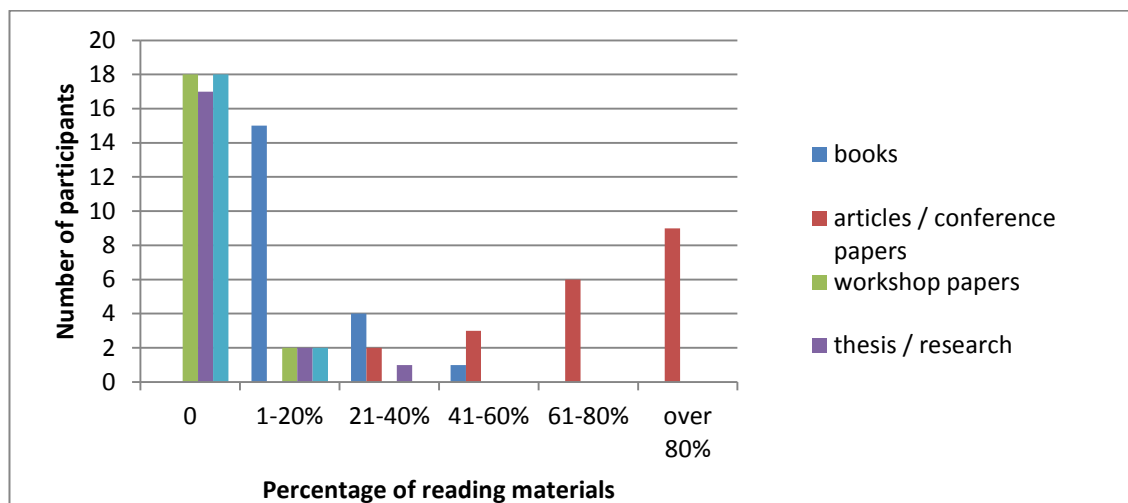
Figure 2.4: Percentages of academic reading among five types

conference papers more than other materials (15 participants read more than 60% articles and conference papers). Three-fourths of the participants indicated that reading books constitutes the lowest percentage of their academic reading (15 of 20 indicated 1-20% of book reading). The other types of reading—workshop

papers, theses, research, and reports—were indicated by only seven participants and made up only about a quarter of their academic reading.

2.3.1.2 Electronic Versus Hardcopy

Table 2.1: How participants usually read their academic reading (online/hardcopy)

		Articles & conference papers	Books	Workshop papers	Thesis & research	Reports
A	Mostly electronic	11	4	1	2	1
B	Mostly electronic, but I print if I want to focus	3	0	0	0	0
C	Electronic/hardcopy (electronic more)	4	0	0	1	0
D	Electronic/hardcopy (50% each)	0	2	0	0	0
E	Electronic/hardcopy (hardcopy more)	2	4	1	0	1
F	Mostly hardcopy	0	10	0	0	0
G	Total number of participants	20	20	2	3	2

Table 2.1 shows how often the participants read various kinds of materials. It is immediately obvious that all participants read articles, conference papers, and books, whereas only a few participants read workshop papers, theses and other research, and reports (see Table 2.1, row G). Thus, a few main points may be derived from the preliminary analysis of the data in Table 2.1:

- Eighteen participants read their articles and conference papers digitally much more than in print (see Table 2.1, rows A, B, and C). Only two participants prefer to read in hardcopy (see Table 2.1, row E).

- About 14 participants read their books mostly in hardcopy (see Table 2.1, rows E and F); however, only four participants read them in digital copies (see Table 2.1, row A).

- The other reading materials (workshop papers, thesis, research, and reports) are read digitally; no one reads them mostly in hardcopy (see Table 2.1, rows A, C, and E).

2.3.2 General Reading

We here present results for the types of general reading as well as preferences for electronic and paper-based reading.

 Types of General Reading Figure 2.5 represents the personal reading of participants among three types of materials, which are newspapers, magazines, and novels.

The most striking result to emerge from the data is that for 15 participants, at least two-thirds of their personal reading goes to newspapers, for eight of whom newspapers constitute more than 80% of their personal reading (six of the 20 indicated 61-80%). However, the other six mentioned that newspapers make up less than half of their personal reading.

A significant number of participants read magazines. However, none of them indicated that magazines represent more than 30% of their personal reading (12 of 14 indicated 1-20% and two of 14, 21-40%).

Approximately half of the participants who read novels indicated that reading novels does not comprise more than 20% of their personal reading (seven of 13 chose 1-20% for novel reading). The number of participants who focus on reading novels was not significant (see Figure 2.5).
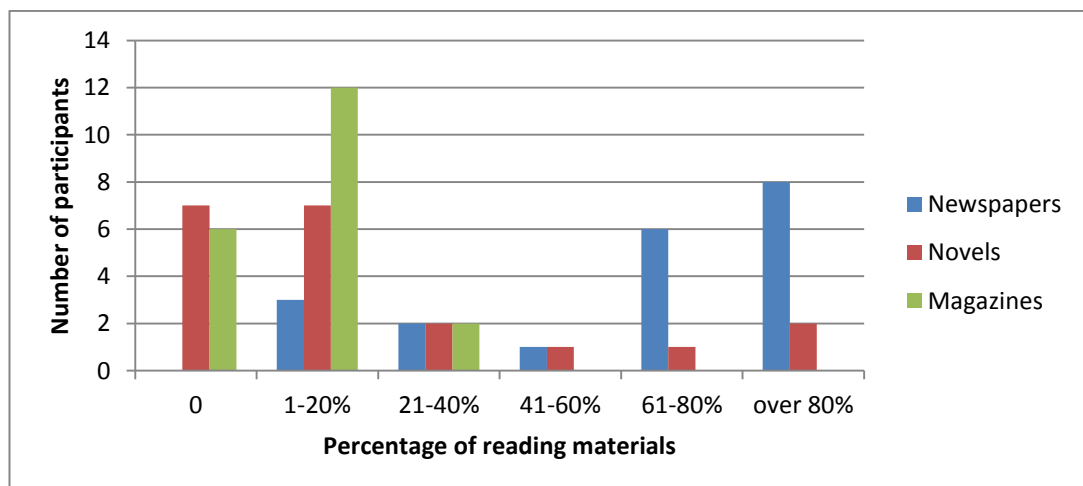
Figure 2.5: Percentages of the three types of general reading

2.3.2.1 Electronic Versus Hardcopy

The results obtained from the preliminary analysis of how often the participants read general materials can be compared in Table 2.2. All of the participants read newspapers, whereas only 14 and 13 participants read magazines and novels, respectively (see Table 2.2, row G). The main points from the initial analysis of data in Table 2.2 are explained below:

- Over half the participants read newspapers online, whereas only about a quarter of them read mostly in hardcopy (see Table 2.2, rows A, C, and F). In contrast, over half of the participants who read novels use mostly hardcopy, with only five participants reading more online (see Table 2.2, rows A, C, E, and F).

- Of the 14 participants who read magazines, nine read them online more than in hardcopy (see Table 2.2, rows A and C).

- Only two participants read newspapers and magazines in both hardcopy and online equally (see Table 2.2, row D). All other participants seem to have a clear preference for either the hardcopy or electronic version.

Table 2.2: How participants usually do their general reading (online/hardcopy)

		Newspapers	Novels	Magazines
A	Mostly electronic	12	3	4
B	Mostly electronic, but I print if I want to focus	0	0	0
C	Electronic/hardcopy (electronic more)	1	2	5
D	Electronic/hardcopy (50% each)	1	0	1
E	Electronic /hardcopy (hardcopy more)	0	1	0
F	Mostly hardcopy	6	7	4
G	Total number of participants	20	13	14

2.3.3 Devices Used for Reading

All participants prefer to use electronic methods for academic reading. However, the devices used for general reading differ from one participant to another. Some participants stay with one device, whereas some use more (see Figure 2.6). The types of devices used are tablet, mobile, laptop, and PC (see Figure 2.6).

Of the 20 participants, 18 use devices for general reading and over half indicated that they use tablets for reading online (see Figure 2.6; 10 of 18 use tablet). The laptop comes in the second rank of the devices most used, with seven participants (see Figure 2.6).
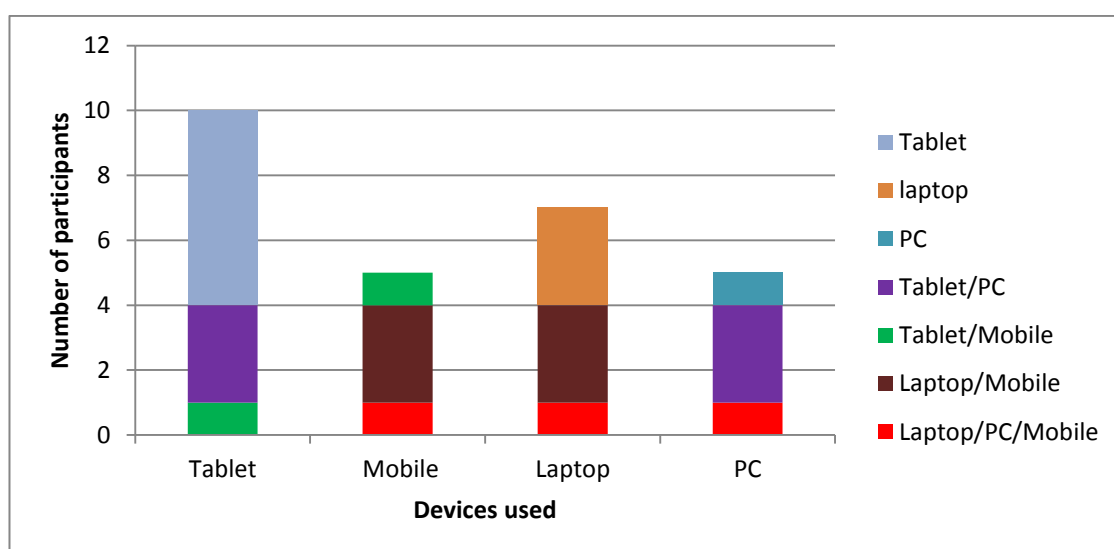
Figure 2.6: The types of devices used for personal reading

The PC and mobile are used by only five participants each (see Figure 2.6). Only participants 7 and 19 did not use any devices because they prefer to read hardcopy.

2.3.4 How Participants Track their Reading

All the participants were asked if they keep a list of the materials they have read, are currently reading, and are planning to read. All participants indicated that they do not care about tracking the general reading that they have either read or are reading (either electronic and printed); they are interested only in tracking personal material *to be* read. Seven participants pay attention to their electronic

11

general reading that they want to read in future by sending the material to their emails or bookmarking websites.

Although the question concerned lists, all the participants explained that rather than keeping a real list, they follow a certain procedure. This method assists them in recognising and keeping track of their materials, but even so, all of them have faced some problems with their own procedures.

For academic reading, all participants keep the electronic materials in their computers, and only 10 participants print the materials in various circumstances (see Table 2.3). Participants 10, 13, 15, and 20 mentioned that the reason for keeping the materials electronically is that sometimes printed materials get lost.

Table 2.3: Circumstances participants prefer for printing documents

Participant No	Participants' Responses
2	'I prefer printed copy: When the PDF file has many pages, I print out the pages that I want and highlight the points I need, [that is,] when the PDF is image and it is not searchable and highlightable'.
6	'The reason [for] printing out is to keep them in my repository. I sometimes like to read them again and again, and I do not want to use my laptop for reading, and then I can highlight them and make notes'.
8	'I read in paper only when I cannot get it electronic[ally]'.
9	'Sometimes I just print the electronic ones because sometimes it is easier to read on a piece of paper than through screen'.
10	'I have to have the print copy because that is how I organise the project that I am working on. I get a collection of print documents; then, I read them through; I annotate them and circle things. The circling and annotating are the important stuff for me'.
13	'Any materials I want to concentrate on, I print [them]. When I print, I can highlight and write some notes. Later on, if something pops up, I can go through only the highlights instead of reading all the pages. Also, I can reference when I am

writing'.

15 'I use printed copies more because I can do highlight. People who have a problem in their eyesight will prefer reading through printed copy. For electronic [reading], you need to check the battery of the device you use if you want to go somewhere to read. In terms of safety, you have to take care of everything; for example, if you leave your laptop or tablet anywhere definitely people will go to [meaning steal] it'.

16 'If it is related to sociology, I prefer to print it out because I need to make some comments. If it is about computer science, I will use the electronic [device]. Probably I do not need extra comments; I need only highlight the main points'.

17 'If I find any interesting paper, I print it out to read it one day in future. Write the URL line on the paper I print'.

20 'If any material is important, I print it out'.

None of the participants who usually print out some of the papers have a procedure to follow but rather keep many stacks in their offices, and the important ones are left on the desk. Participant 10 talked about this difficult situation and said, 'One of the big problems is after 6 months when I need one point from some papers that were placed in stacks, I know that the papers I want are somewhere in my office, but I have no idea where they are'. In addition, participants indicated that all the materials which can be available only in hardcopy, such as books, are placed on a bookshelf. Participants 14, 15, and 16 organise the materials printed out in folders according to the same order as on their PCs.

Some participants pointed out what motivates them to read electronic copy. Participants 1, 2, 6, and 7 mentioned that instead of reading all of the article or paper, they can search for the exact information they want in a few seconds. Those participants added that they can also annotate and highlight any information found through the electronic materials by using a tool like a PDF reader. Four participants indicated that electronic material can be easily organised and used (3, 6, 7, and 9).

Figure 2.7 shows that all the participants prefer to keep track of the electronic/online materials that they have read or are planning to read, whereas only 11 participants prefer to track what they are currently reading.
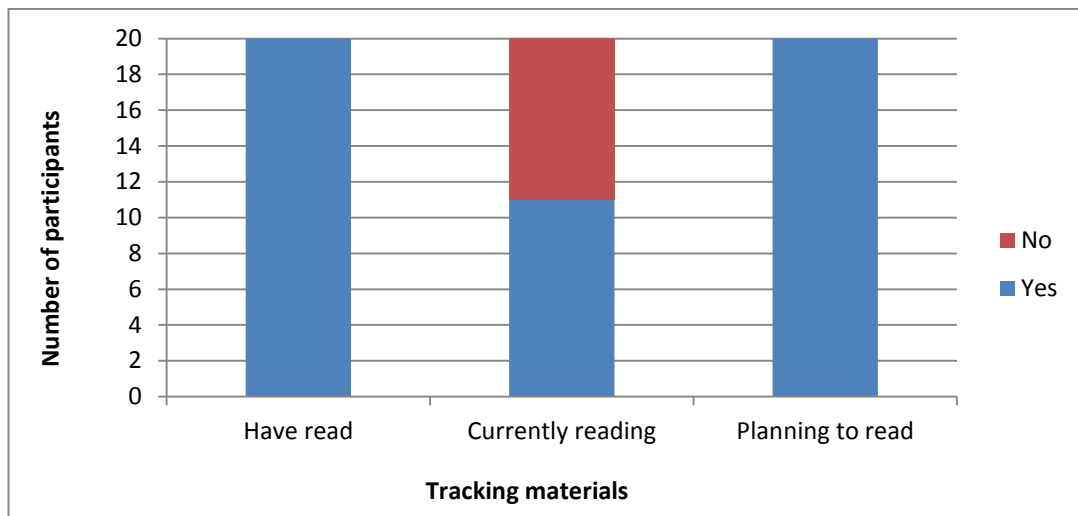
Figure 2.7: Participants' responses concerning the tracking of electronic materials they have read, are currently reading, and are planning to read

2.3.4.1 Materials That Have Been Read

All the participants reported that they tend to track their academic digital materials only when they do a project, research, write a paper, or something similar. For other material, they do not think they need to be entirely aware of what they have, so they simply organise it by subject.

Participants use three main methods to track the materials that they have read (see Figure 2.8), some using just one method and some more than one. Eleven of the participants rely on only one method (see Figure 2.8, which shows that 10 use only folders and subfolders, and one uses only software). The others use two methods, except for one participant, who uses all the methods.
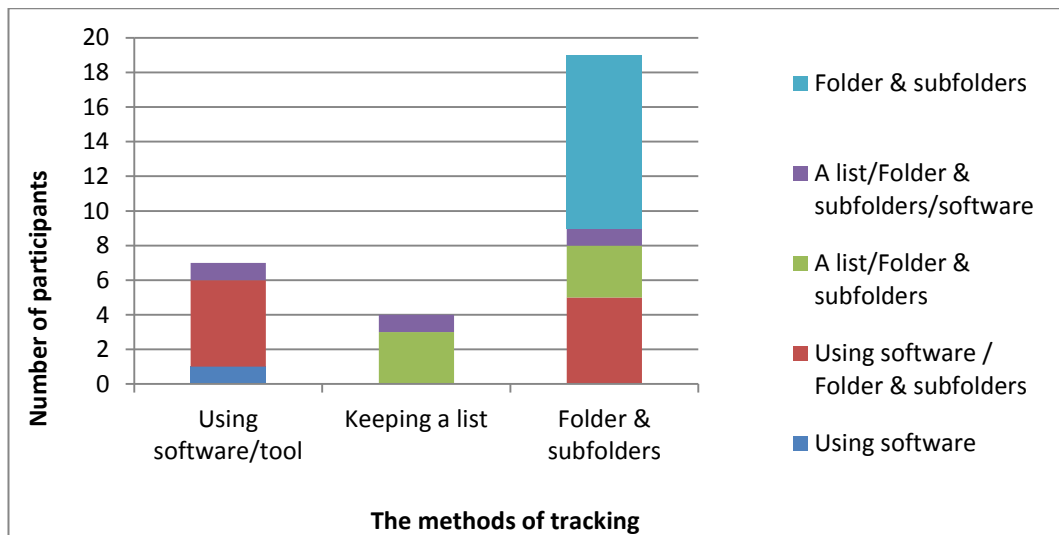
Figure 2.8: How the participants keep track of the materials they have read

The majority of participants use folders and subfolders (see Figure 2.8 showing that 19 of 20 use folders and subfolders) for tracking, and only nine of those use another method besides. Three and five participants use a list and software, respectively (see Figure 2.8). Only one participant uses both a list and software. Each method is explained below.

- *Using software/tool*

  Of the 20 participants, only six use software for tracking their materials, and one participant uses the Safari bookmark tool. All participants who use software/a tool prefer to use additional methods for tracking their reading materials as well, except for one participant. He uses only Mendeley[2] for tracking and referencing purposes.

  Zotero,[3] NVivo,[4] and Safari bookmark are preferred by only one participant each. Mendeley and Endnote[5] have two participants each. The Endnote and Zotero users indicated that they use the software for doing their references in writing as well (see Figure 2.9).

---

[2] See www.mendeley.com

[3] See www.zotero.org

[4] See www.alfasoft.com/en/menu-products-en/statistics-and-analysis/132-nvivo-en.html

[5] See www.endnote.com
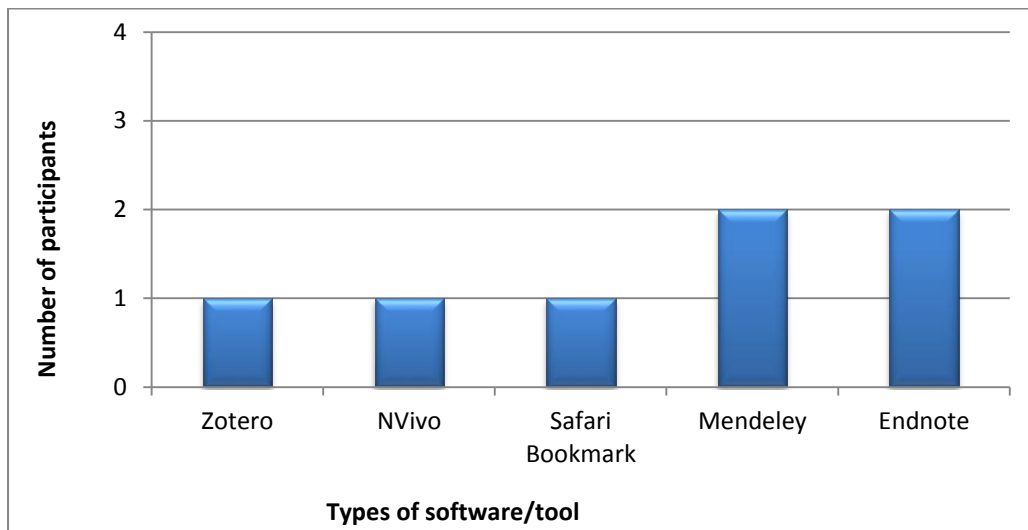
Figure 2.9: Types of software/tools used for tracking

- ***Folders and subfolders***

All participants who use this method create a folder for the project or paper and then create subfolders based on the outline. Afterwards, each item is placed in the folder for the part to which it may belong. For example, the main folder could be a PhD or master's thesis, and then inside that folder will be additional folders for introduction, background, and methodology.

Some participants mentioned the importance level of the materials. Inside the subfolders, they have folders called 'less important', 'important', and 'very important'. In addition, they indicated that not all the materials they place in their folders should be read or used. To remember them, they rely on their memories from reading the title or abstract.

The participants who use software for referencing indicated that all the materials added to referencing software are usually read and used in their writing projects.

- ***Keeping a list***

Only four participants indicated that they keep a list, and they pointed out that the list has only the materials used in a project or paper or those they might refer to later (see Figure 2.8). Participants had their individual techniques, depending on what they want from the list.

- Participant 7 said, 'I keep a list for each paper or project, and in the list I write all the titles [of the] materials I used (books, articles, and conference papers). A list is written by word document and placed in the project's folder. Each item in the list will be explained [by] which part [it] was used [in]: for instance, introduction, background, or discussion'.

- Participant 8 said, 'It depends actually where I am; if I am at my desk, basically I write notes in Emacs[6]; I put them in a file, and if it is a paper, I give it a number. I keep a copy of that paper. If it is a book, I just put the title of the book on it. Then, I keep all of those in a directory of notes. [I use] ordinary folders. If I am in my car, I use [a] phone app called EverNote,[7] and then I transfer that to my directory'.

- Participant 15 said, 'I use Endnote, and sometimes I do it manually by Excel[8] because sometimes I could not find it in Endnote, which I think … is very crucial. For example, like [for] disclosure, I do have a list of the authors, titles, other findings, and what method they use'.

- Participant 19 said, 'If I read something electronically, then I make note in my database. It is something called "all references or quotations", and it does not really work because it [is] on Word file. At least it helps me because it has got all [the] information there, and I can go back to any info whenever I want. My database

---

[6] See http://www.cs.colostate.edu/helpdocs/emacs.html

[7] See http://evernote.com

[8] See http://office.microsoft.com/en-us/excel/excel-2010-features-and-benefits-HA101806958.aspx

in my PC is like (title- author-publication year – a page no. if I do quotation)'.

2.3.4.2 Materials That Are Being Read

All the participants keep track of the materials they have read and are planning to read, whereas only about half keep track of the materials currently being read—that is, have read, 20 participants; currently reading, 11 participants; and planning to read, 20 participants (see Figure 2.7).

Table 2.4 presents all the responses of the participants who indicated that they prefer to keep track of their current reading. Over half have their own methods for tracking and they are different from each other (see Table 2.4, rows B, C, D, E, F, and G). However, they follow different methods; four of them use software for tracking (Table 2.4, rows D, E, F, and G). The remaining five leave the materials open on the screen (see Table 2.4, row A).

Table 2.4: Participants' responses for how they keep track of current reading materials

	Participant No.	Method used
A	1, 7, 8, 9, 20	They leave them open on the screen until they finish them [paraphrased by the researcher]
B	2	'I highlight the last point I have read and leave a sticky note on the monitor about the material that I am reading'.
C	3	'Leave some sticky notes/notes in my office'.
D	11	'The reading list in Safari is the only way that I follow'.
E	15	'I use Excel sheets and Endnote'.
F	16	'I use Foxit Phantom PDF Reader[9] on my computer and EZ PDF Reader[10] on my tablet'.

[9] http://www.foxitsoftware.com/

[10] http://download.pandaapp.com/android-app/ezpdfreader_1.5.0.0-id4479.html

| G | 18 | 'I put them in Endnote or leave them open'. |

2.3.4.3   Materials to be Read

Most participants follow one of two methods for materials they want to read in the future (see Figure 2.10, which shows that five use software/tools, and nine download the materials). Only two participants use two methods for tracking reading materials for the future. Participant 12 said, 'If I find something interesting, I either send it to my email or save it somewhere on my computer to read it later'. Participant 16 said, 'If I can download the paper I will do [it], and then I use software called Remember The Milk[11] to organise when I want to read it'.

Two participants indicated that they plan for reading materials, but there is no specific way to follow because, they said, ' it is pretty rare to do that' (see Figure 2.10, 2 participants, None).

Each participant who uses the software/tools method has a different type of software, and only two of those use Remember The Milk. The others use APA Reference, reading lists in Safari, and bookmarks.
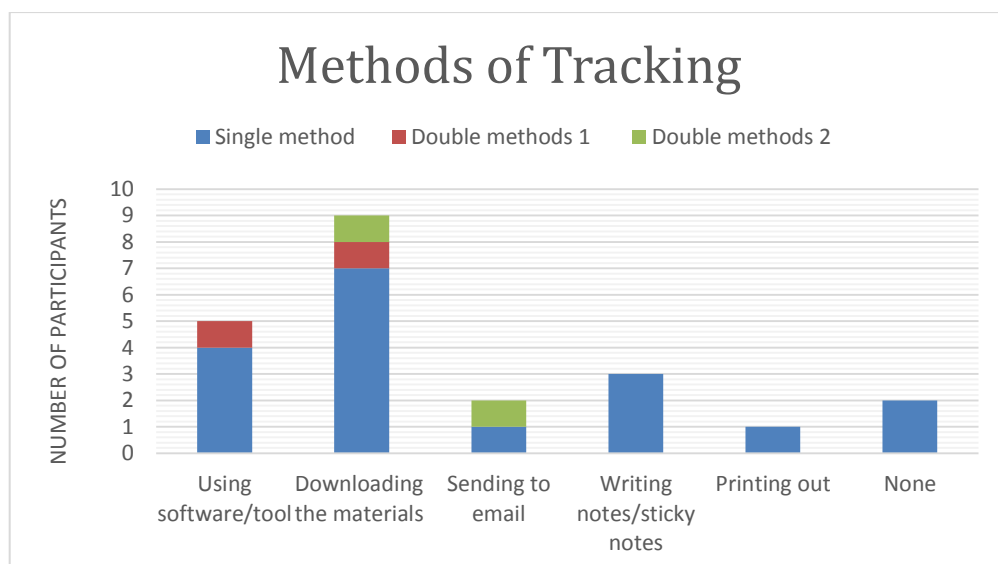
Figure 2.10: How participants keep track of materials they are going to read

---

[11] http://www.rememberthemilk.com/

In terms of the downloading method, none of the participants who download documents have a specific place or location on their PCs for storing the documents. Some of them have a folder called 'Want to read' and some place new documents with the materials they have read. One participant uses his download folder as a default location for all the materials. Participant 6 downloads his materials on his desktop and when he was asked why, he said, 'I leave them on my desktop in my computer because it is like—look at these things in front of you; you need to read them if you want to'.

The participants who send emails to themselves send either the titles of materials or the materials themselves if they can. The others who write notes or print out the material mentioned that their method makes it easy to remind themselves all the time about the materials.

2.3.4.4 Problems Encountered

The majority of issues that were indicated by the participants were gathered and analysed for materials previously read, currently reading, or to read (hereafter may be referred to as 'reading status'). The issues are divided into two parts: electronic/online and hardcopy.

Problems relating to electronic/online materials:

- They spend much time looking for a specific paper or little information is used (participants 5, 6, 8, 9, 10, 11, 13, 17, 18, and 19).
- They read a paper more than once because they do not know if it has been read or not, or they do not remember which parts are in that paper (1, 2, 5, 20).
- They need to use one paper in two or three different projects, so they need to place one copy in the folder for each one, and each copy has different highlights and comments. Moreover, sometimes in one project they need to use one paper in different parts of the same project, with individual highlights and comments (6, 7, 14, and 20).
- Sometimes they save a hyperlink to an article list, and when they return to it later, the hyperlink does not work (4, 12).

- They cannot tell what downloaded files are about because the name of file is usually numbers and letters (2, 10).
- They have lost files and could not find them again (18, 19).
- They download a paper more than once because they don't know if it has been downloaded or not (2).
- They cannot always know whether the material is important or not (14).

The problems relating to hardcopy materials are as follows:

- After several months, when they need just one point from some papers that are placed in stacks, they know that the papers are in the office, but they have no idea where (10, 13, 15, 18, and 20).
- Sometimes, they have so much printed material that they do not know where they placed the papers, and then they need to print them again. As a result, they also lose all their comments and the highlights they made (15, 18, and 20).
- Sometimes they request a book that is not available in the library, and when the book comes in, they have trouble remembering why they requested it (1, 4, and 20).
- They spend a lot of time looking for specific information among the papers (18 and 20).
- The folder is not enough to keep all the stacks of papers (15).
- Sometimes it is annoying to have one's office filled with a lot of papers (20).

2.3.5 **Reading Hours**

Figure 2.11 shows how many hours participants spend reading in a day. The graph demonstrates that most participants spend more than 4 hours reading every day, and about a quarter of participants spend less than 4 hours (see Figure 2.11). The number of hours that was given sometimes increased or decreased, depending on what they are working on at the time.
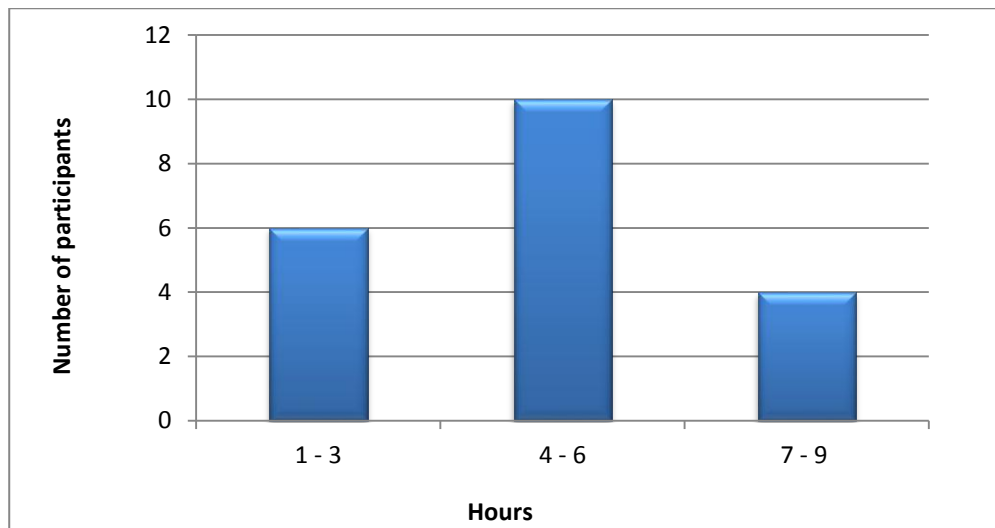
Figure 2.11: Reading hours spent in a day

2.3.6 **Library/archives**

Well over half (13) of the participants have a collection of books and some printed copy at their homes, whereas seven participants have nothing at home. Four of those who have nothing have been in New Zealand for only 2 to 3 years and are keeping all their books at their offices because of moving from one house to another. The other three who have nothing have shipped some books to their own countries and have sold some because they intend to leave New Zealand.

The amount of physical materials that participants have ranges from a few up to 10,000 items. Three and four participants indicated that they have several thousands and hundreds of printed items at their homes, respectively. In addition, six participants indicated that they have less than 200 items. All of those who keep physical materials at home indicated that they are organised mainly by subject. Most also have digital materials, but they are not as extensive as the physical ones.

Participants 8 and 10 indicated that they have several hundreds of digital files in their Kindle software. All participants mentioned that they usually keep all the digital and hardcopies related to their academic work at their offices. Moreover, they sometimes have recreational materials.

2.3.7 Materials That Have Been Written

All the participants indicated that they have written academic documents, and each one has written more than one type of document, such as articles and conference papers. Figure 2.12 shows what kinds of materials the participants have written. From the graph below, it can be seen that most participants have written conference papers, articles, and theses more than anything else. Only three, five, and five of the participants have written books, academic papers, and dissertations, respectively.
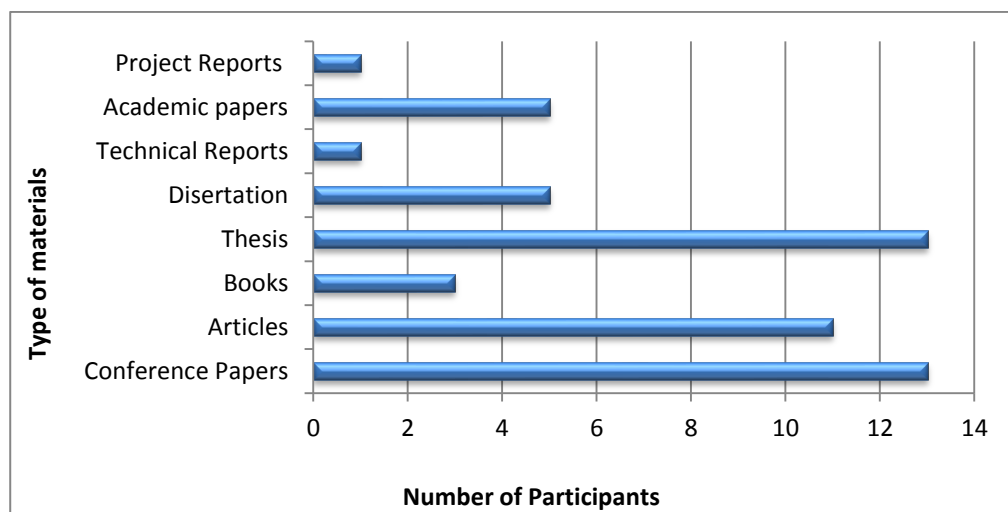
Figure 2.12: The types of materials previously written by the participants

2.3.8 Tools/software Used for Writing Documents

Significantly, all the participants normally use Microsoft Word for writing documents except participant 8 (see Figure 2.13). When he was asked why, he answered, 'Mostly I use only LaTex, and I do not have a Windows machine. Sometimes I use Open Office'. Ten of the 19 participants indicated that they use only Word, whereas five of those use other software/tools besides Word. The other four participants use two other software packages/tools with Word.
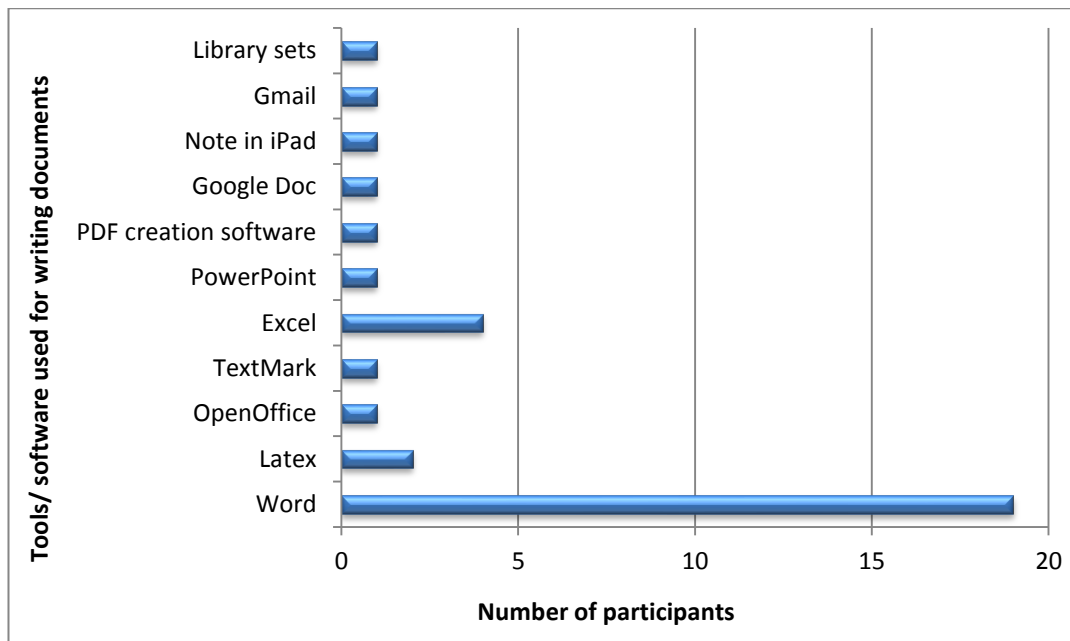
Figure 2.13: Tools/software the participants use in writing documents

2.3.9 Software System Used for Referencing

Fourteen of the participants indicated that they use software and tools for doing their referencing (see Figure 2.14). Endnote is used by a majority—eight of 12 participants (see Figure 2.14). Only two participants use a Microsoft Word tool and one participant each uses Bibtex, Jabref, Mendeley, and Zotero (see Figure 2.14).

Six of the 20 participants do not use any software, choosing rather to do their referencing manually (see Figure 2.14). Besides the manual referencing, only participant 20 uses some sources that may help with referencing. When he was asked, 'Which software system do you use for referencing?' he answered, 'I do it manually, and sometimes, I search in Google and see how the others did the reference for that article or book, and I use some websites, such as the library's web site at Waikato University'.
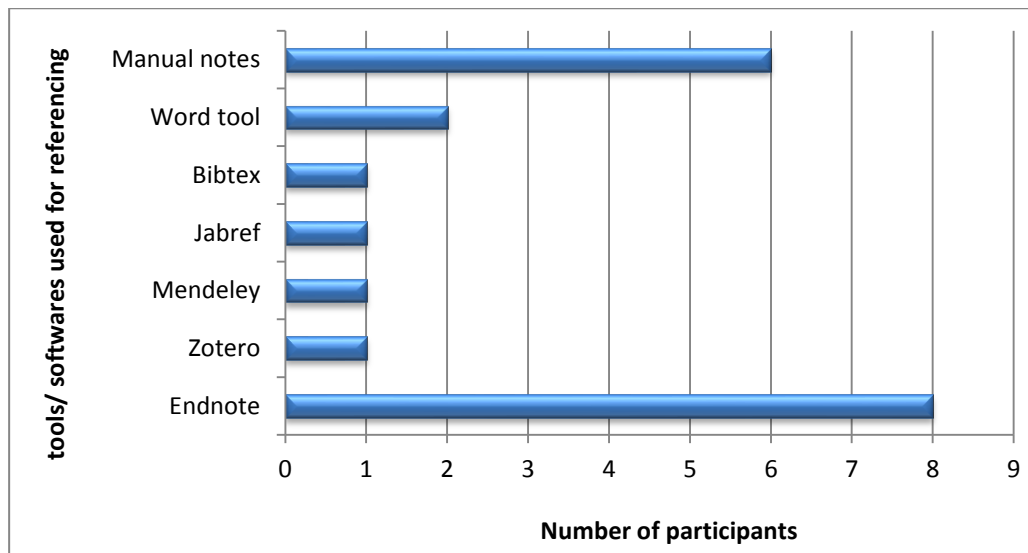
Figure 2.14 : Tools/software the participants use for referencing

2.4 Findings and Discussion

2.4.1 Reading Focus of Participants

The results of this study show that most study participants do more academic than personal reading (see Figure 2.3, Chapter 4). The sample for the study may have affected this finding because 14 participants are graduate students, and the other six are academic staff at Waikato University. Most of the participants read articles and conference papers more than other materials, and these materials comprise more than two-thirds of their academic reading (see Figure 2.4, Chapter 4). By comparing the data from that figure with the data in Table 2.1, we conclude that most participants read academic articles and papers electronically (see Table 2.1, rows A, B, and C). In consequence, we observe that most participants' reading is electronic. Interestingly, for all electronic academic reading, all participants prefer to read on a PC screen, whereas for personal reading they prefer several other devices, such as tablets and laptops. For academic reading, some participants expressed that preference because of the bigger screen and the useful tools available on a PC. For personal reading, the tablet was participants' first preference. A possible explanation may be its ease of mobility.

25

2.4.2 **Participants' Strategies for Keeping Track of Materials**

On the question of keeping a list of the reading status of their materials, this study found that all participants keep track of their reading materials, but they use different strategies for tracking the three different types of material.

- Materials they have read: Even though participants mentioned that they use different kinds of techniques, an overwhelming number—19 out of 20—use folders and subfolders to keep track of what they have read, perhaps in addition to other methods. As a consequence, there is no particular order to the documents, or they may not be a physical order that matches the electronic one. They have no annotations that permit the users to recognise their materials or to capture the context and purpose of the reading (beyond naming the folders and documents).

- Materials they are currently reading: Only 11 of the 20 participants keep track of the materials they are currently reading, and five of those follow the strategy of leaving the materials open on the browser. The other six have their individual techniques that help them remember the materials they have not finished yet. This finding might lead us to conclude that a step could be missed between the 'currently reading' and 'have read' because the material is not tracked properly from the beginning; thus, it may be difficult to keep track of it after it has been read.

- Materials they are planning to read: Participants use a variety of techniques, though there is no uniformity in this kind of tracking. There is, however, certainly a need to keep track of the materials one is planning to read because most of the participants have some method for doing so, or more than one method. Though everyone is doing it, even the same people have different strategies for keeping track of where they are in their reading, and there are further differences in strategies among different people.

A discontinuity can be observed in the processing of reading materials because the participants use different methods. For materials that have been read, most participants use a technique involving folders and subfolders; however, for other

types ('currently reading' and 'planning to read'), no one uses this method. It seems obvious that although participants try different methods, there is no effective common strategy. As a consequence, two issues indicated by 10 participants are that they spend too much time looking for a specific paper, and sometimes little of the information is used. Moreover, they end up reading one paper more than once because they either do not know if it has been read or not, or which parts of the paper they have read. This finding may have important implications for developing a single strategy that can help people keep track of any reading material easily and effectively.

2.4.3 Referencing

The results of the study reveal that there are two groups of participants. One group of six participants does the references manually, and the other group of 14 participants uses software. Thus, it seems that the majority uses software to format references when writing. We assume that for referencing, people prefer to use the software that goes with the word processing software they use to create documents. In this study, a significant number of participants (19 of 20) use Microsoft Word for their documents. Because of that, most of the participants who use bibliographic software (eight of 14) use Endnote, which is compatible with Word. All the other software used for referencing can be used with Word as well, except for the one participant who uses Latex for writing, and he uses BibTex.

2.5 Summary

This chapter reported on the analysis and the findings of the user study. The aim of the study was to investigate what kind of materials people track and what tools or methods they use for tracking. There were 20 participants in the study (14 males and 6 females). The results of the study show that most participants read academic material more than general material (see Table 2.5). More than 50 percent of participants' academic reading is articles and conference papers, which they prefer to have in digital form. In addition, the participants use different techniques to track their reading of the material (see Table 2.5). Table 2.5 shows that no one strategy or software exists that can be used for all three phases of reading.

Table 2.5: The techniques participants use for each reading phase

Reading phases	Techniques participants use				
To read (20P)	Using tool/software (5P)	Downloading the materials in a folder (9P)	Sending to email (2P)	Sticky notes (3P)	Printing out (1P)
Currently reading (11P)	Leaving the screen open (5P)	Each one use different strategy (6P)			
Have read (20P)	Structure the material in folder and subfolders (19P)	Keep list (4P)	Using a tool/software(7P)*		

*Note: The numbers in line 3 do not add up to 20 because some participants use more than one method.

Participants mentioned three problems: spending too much time looking for information or a paper, reading a paper more than once, and having to duplicate files that have different highlights and comments.

In the next chapter, we will extract requirements from the findings of the study that match people's need for tracking material. In addition, we will explore the number of existing systems that can be matched to the revealed requirements.

Chapter 3

3 Requirements and Related Work

This chapter discusses the requirements that should be fulfilled, based on the analysis of the data collected from the reader participants (as reported in Chapter 2). Furthermore in this chapter, four bibliographic types of software are reviewed, as well as two social library networks and one digital library. The last section offers a comparison among these, and they are also reviewed in light of the requirements revealed from the study.

3.1 Requirements

In the user study, four aspects were discovered that must be considered in developing a system for tracking reading materials: localisation, organisation and management, indication, and annotation. Each one has its own requirements.

- **Localisation of materials**

  1. **Downloading the reading material and metadata with the system:** the system should be able to download the user's material and to provide the metadata (see Section 2.3.4).

  2. **Storing the material itself (such as papers, images, and websites) and in addition, the material's metadata:** the system should allow the user to keep the material (i.e. PDF, webpages, and image) and its metadata. All the participants in this study (see Chapter 2, Section 2.3.4.1) indicated that they intend to keep real material for any source they have read, and most of them use bibliographic software for doing their references. The metadata of the item help the user with the bibliography.

  3. **Browsing the material through the system:** the system should enable the user to browse the materials already downloaded in the library through a window that is part of the system. The participants mentioned that they intend to download their reading items; therefore, they need to display the material for reading.

  4. **Search feature:** searching was one of the problems participants indicated in the study (see Section 2.3.4.4, Problems). The system should provide a search function that can search in the metadata

and contents of the items downloaded in the library. It would be a helpful tool for reducing users' efforts in looking for items or information.

- **Organisation and management**

    The system should support a flexible structure for the materials because users want to group them according to the current task.

- **Indication**

    1. **Indication for the reading status of the material:** the system should provide an indication for each item that shows the state of material—that is, whether it is read, unread, or to read (see Section 2.3.4.4, Problems). This will help the users to identify which items they have finished and those they have not.

    2. **Indication for material used in writing papers:** the system should assist the user in knowing where each item of his or her collection is referenced—that is, in which project and in which sections in a project the paper is used as a reference (see Section 2.3.4.4, Problems).

- **Annotation:** The system should enable users to write a description about each item in the collection. This might assist them in recognizing and remembering the materials more easily. Users need to be able to write a short description about each paper because the participants mentioned in Section 2.3.4.4 (Problems that they cannot recognise their materials easily, that they need to see and read what is there to remember. The system can facilitate this process if it enables users to write a description about each item they have read in their own words.

In summary, users should be able to employ the system to fulfil the following requirements:

- R.1   Download the reading material and metadata with the system.
- R.2   Store the material itself as well as its metadata.
- R.3   Browse the material through the system.
- R.4   Search the materials.
- R.5   Organise and manage materials.
- R.6   Indicate the reading status of materials.

R.7 Indicate material to be used in writing papers.

R.8 Annotate.

We will use this list of requirements to analyse related work, and later to design our own system. In the implementation done in this project, we focus on requirements R1 to R6, leaving R7 and R8 for future work.

3.2 Related work

Here we review four popular citation management software systems: Endnote, RefWorks, Zotero and Mendeley. The main differences among these software packages is that RefWorks and Endnote were designed to manage citation, whereas Zotero and Mendeley were developed to manage papers (Butros & Taylor, 2010). Zotero and Mendeley are easier to use than Endnote and RefWorks (Barsky, 2010). Moreover, there are two social network digital libraries, LibraryThing and Goodreads, and one digital library, Greenstone. This presentation looks at the different aspects of each one, such as its strengths and how it works. In addition, any requirements from Section 3.1 that are present in those systems will be highlighted in this section.

3.2.1 **Endnote**

Endnote[12] is a commercial tool for organising and managing references. It has desktop software and a web-based interface. Endnote has some abilities that facilitate organising citations for thesis and papers. In addition, it provides many options for customization and formatting; thus, it may be convenient a wide variety of research projects. It suits researchers and postgraduate students because it can manage and maintain large libraries (Trinoskey, Brahmi, & Gall, 2009; see Figure 3.1).

---

[12] http://endnote.com/
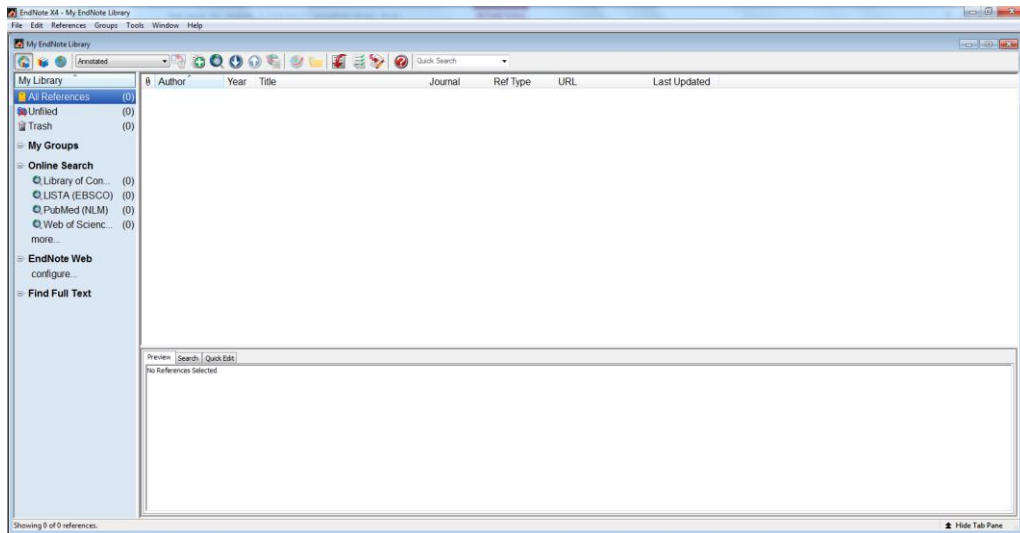
Figure 3.1: Endnote desktop window.

Sources can be added into Endnote four ways: typing the metadata in a template, exporting from appropriate databases, using a search engine provided in Endnote, or downloading the sources to a device (PC/laptop) and then importing them (Butros & Taylor, 2010). Endnote can deal with books, articles, and PDFs; however, the records may not be shared with other users. Nonetheless, the records can be exported to other citation software, such as Word, Open Office, and LaTex. The greatest disadvantage of Endnote is that the records cannot be backed up. Anyone who uses Endnote must have an account and must purchase the license for use (Mathew Willmott, n.d.).

Endnote allows the users to search for references by metadata fields, such as author and year. The users should specify the library they want to search in and then they can add the logic for their search action ("EndNote Online User Guide - Getting Started," n.d.).

The same report shows that Endnote allows users to attach PDF files to their references. Moreover, these files can be displayed, highlighted, and annotated in Endnote.

Endnote fully supports four of the identified requirements: R1, storing documents and metadata; R3, browsing materials; R5, organising and management; and R7, annotation. However, it also partially supports the R1 (downloading) and R4 (search) requirements because it deals with only metadata for both functions. None of the indication features are provided in Endnote.

3.2.2 **RefWorks**

'RefWorks is dedicated to providing a high quality web-based research management, writing, and collaboration tool for the academic, government and corporate research communities' ("RefWorks: About Us," 2009). Some of its features are similar to those of Endnote, and they have been produced for the same purpose.
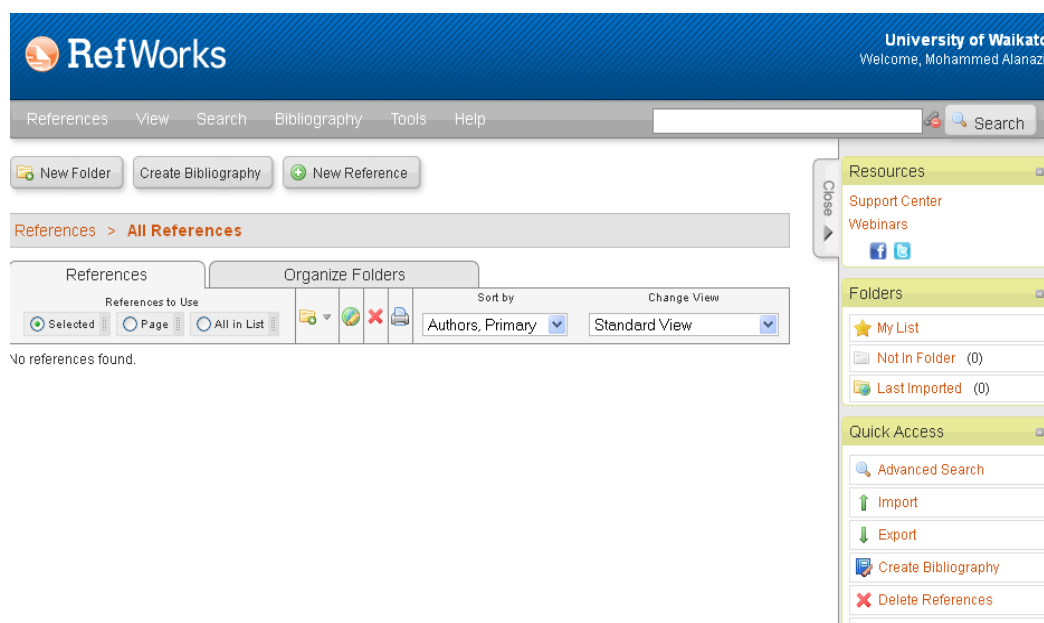
Figure 3.2: RefWorks' main screen.

RefWorks is web based only; thus, it can work with any machine, and there is no installation. In RefWorks, all the references can be easily exported from any appropriate database and citation software. In addition, the user can share citations with other users who have an account in RefWorks. There are many popular bibliographic formats and styles, and the user can easily import articles and books. However, it is not convenient for large libraries because it can become unwieldy. It works with Word and LaTex through Write-N-Cite and Bib Tex, respectively. All the records in RefWorks can be backed up.

RefWorks has two ways to add references to its database/library. First, from *References* (see Figure 3.3, on the top left-hand corner) the user chooses *Add New*. In this way, the user needs to specify the style of bibliography and the reference type. The user enters all the fields manually, and he or she may also attach a file to that reference (see Figure 3.4 and Figure 3.6).
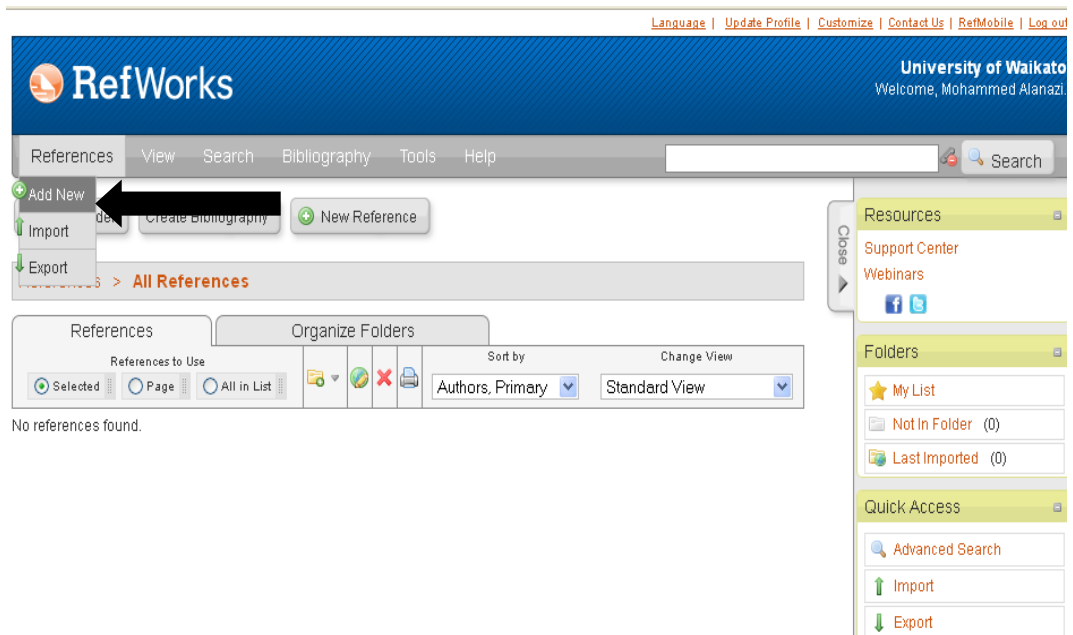
Figure 3.3: Adding and importing new sources in RefWorks
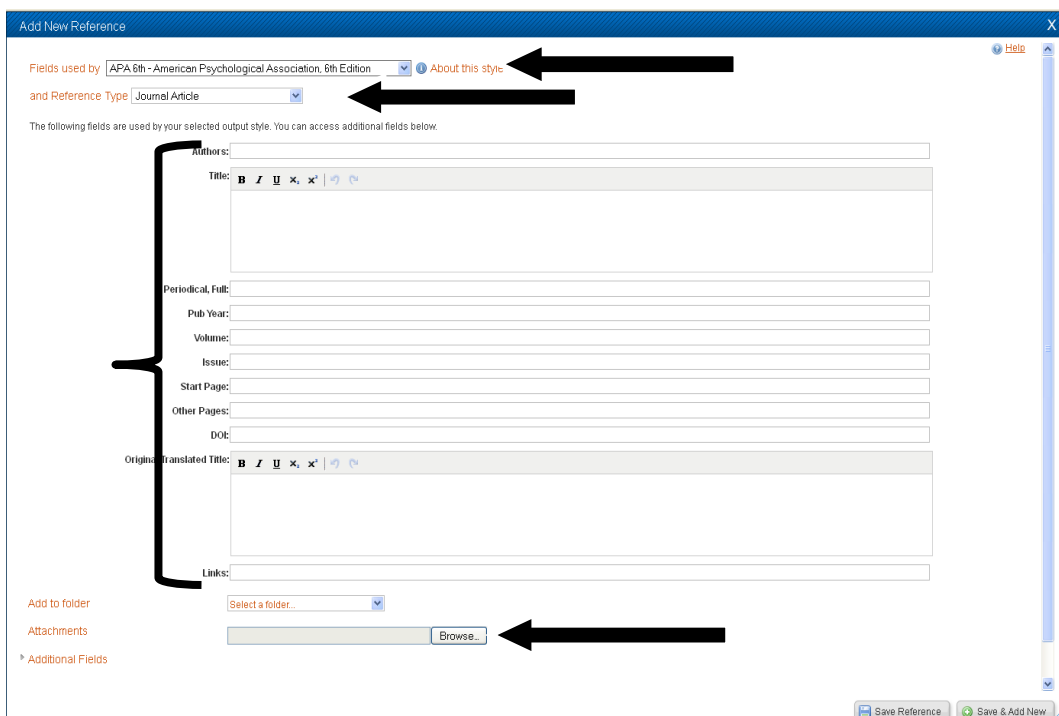
Figure 3.4: The fields of the source added

Second, from *References* (see Figure 3.3 at the top left-hand corner), the user chooses *Import.* When users specify the data source and database, they can then import the file that contains the references (see Figure 3.5). The references will be added to the library. Users can next attach a file and annotation for each record in their library by editing the reference and then attaching the file they want (see

Figure 3.6). The attachment can be browsed by external software, such as a PDF reader or web browser.
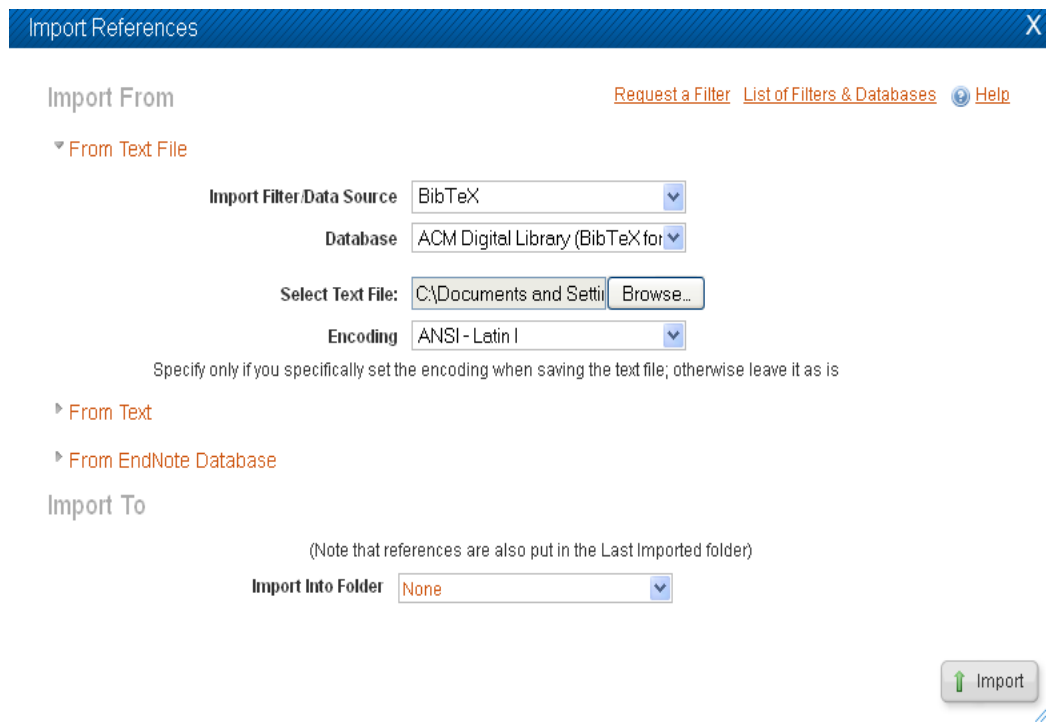
Figure 3.5: Importing data from an external source
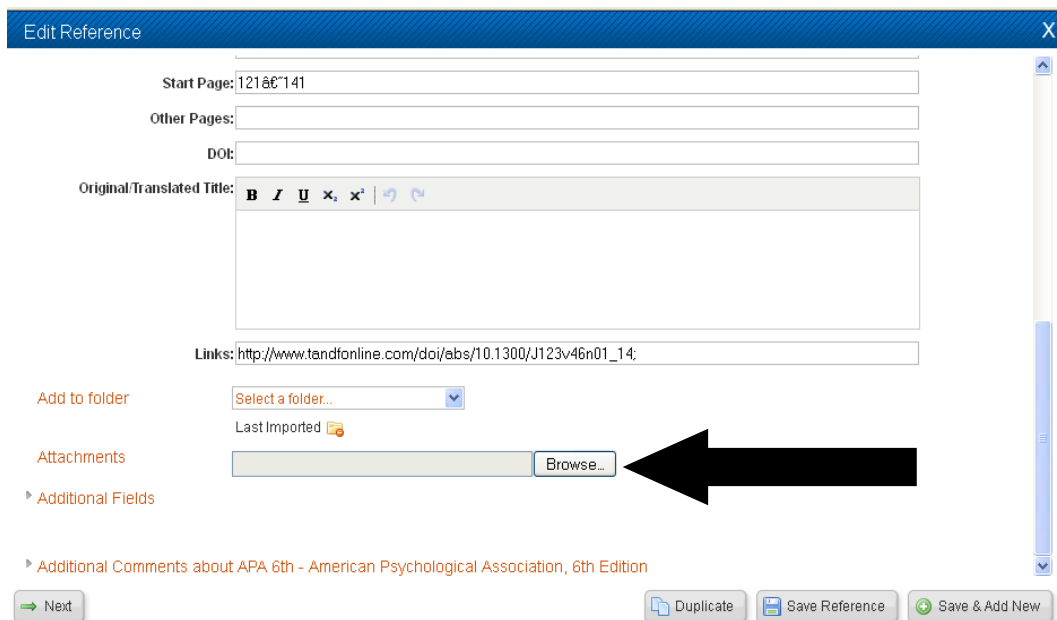
Figure 3.6: Attaching a file for metadata

All the metadata and attachments are included in the search function, and the user can specify the range of search: for instance, author, title, and attachment. RefWorks enables users to organise their references in folders and subfolders.

RefWorks supports R2 (storing metadata and documents) and, subsequently, R5 (organisation and management) as well. In addition, it supports R8 (annotation), which is under the additional fields. It partially supports R4 (search) as it is able to search through metadata fields only. The other requirements are not satisfied in RefWorks.

3.2.3 Zotero

Zotero is a free open source code that combines features of both Endnote and RefWorks. There are two ways to use Zotero: the first is to install the desktop software which has all the control functionalities (Puckett, 2011; see Figure 3.7).
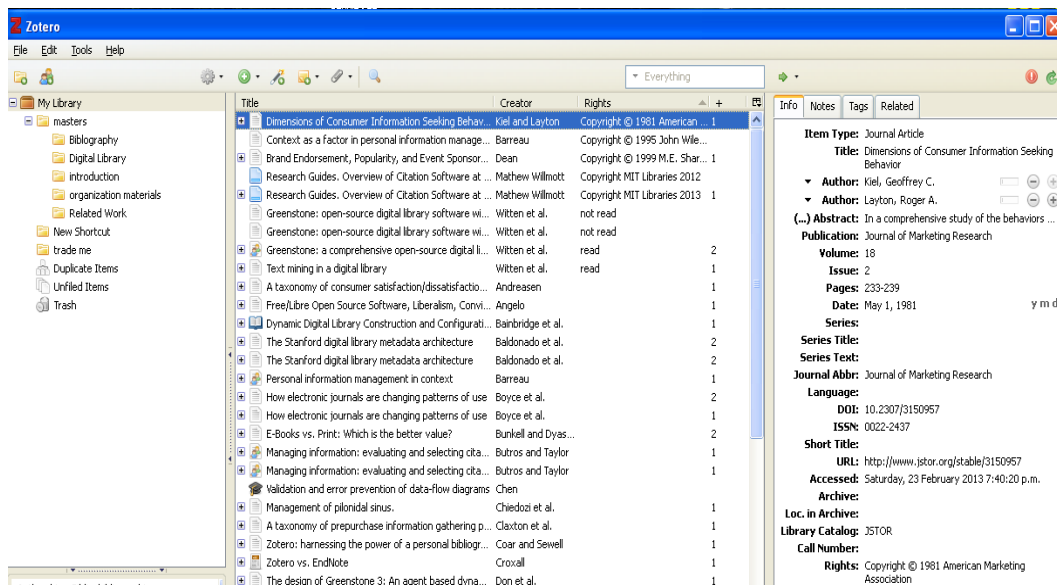
Figure 3.7: The Zotero standalone window

The second way to use Zotero is through a Firefox extension, which enables users to add any items they want to their library while they browse the Internet (see Figure 3.8).
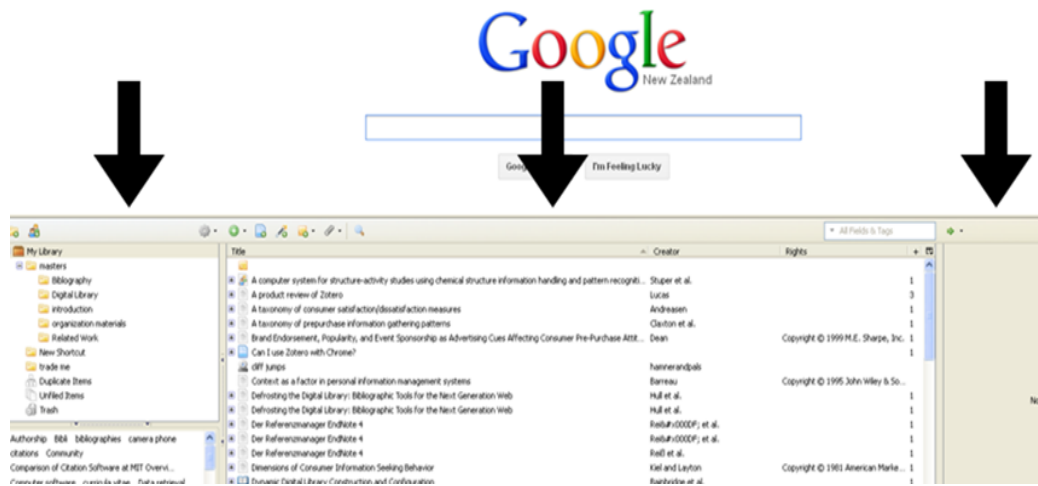
Figure 3.8: Zotero Firefox extension window

Users can control, organise, and manage their collections by using either the desktop software or the version on Firefox because the two are connected to each other. Each Zotero location has its own purpose: the desktop software, for example, is to be used for citation while writing a document. It can be used in Word, Open Office, LaTex, and Google Docs through installation of the plug-in. The Firefox extension, then, is for importing/downloading the items to the Zotero library. All the records and their attachments can be downloaded easily from many databases in one step (the user needs to click on the icon that appears in the URL bar (see Figure 3.9). The most exciting feature is that Zotero can automatically recognise the types of records added, such as 'conference papers and books'. The collection can be synchronized over multiple computers (Puckett, 2011), thus greatly contributing to overall organisation and management of materials.
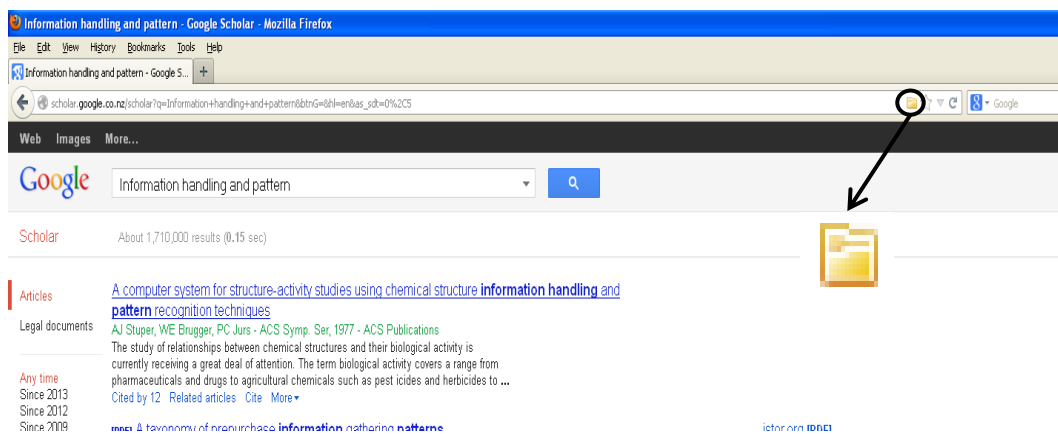
Figure 3.9: The icon adds the metadata and files to the library

The search function in Zotero is beneficial because it enables the user to search through the metadata and the attachments of the collection (rmzelle, 2012).

Zotero works with some databases that RefWorks and Endnote do not—for instance, USPTO and Factiva. Moreover, users can share their records by setting up a group profile. As with Endnote and RefWorks, the records can be exported to any citation software. Zotero has difficulty in dealing with duplicates, so it takes a long time to verify (Butros & Taylor, 2010; Puckett, 2011).

Users have the advantage of some features that may assist in organising records, like tagging records, highlighting text, and taking notes on a page (Butros & Taylor, 2010; Puckett, 2011). From the review, it can be seen that Zotero supports fully five requirements: R1, downloading; R2, storing; R4, searching; R5, organising; and R8, annotating the metadata and documents. In addition, it partially supports R3 (browsing), but one cannot browse the document through Zotero itself. Zotero does not support either R6 (reading) or R7 (citing materials).

3.2.4 Mendeley

Mendeley is free software that both has desktop software and is web-based, and having an account is essential for using any one of them. An interesting feature ("Overview | Mendeley," 2012) is that synchronization for both versions is provided as iTunes and MP3 music (Butros & Taylor, 2010).

Mendeley imports and organises two types of records: PDFs and bibliographic citations only. There are two methods for importing references into the Mendeley library. First, PDF files may be added to the desktop by dragging them. The

metadata will be extracted automatically from the PDF itself. Second, only the metadata of references may be imported from the web (Butros & Taylor, 2010; see Figure 3.10).
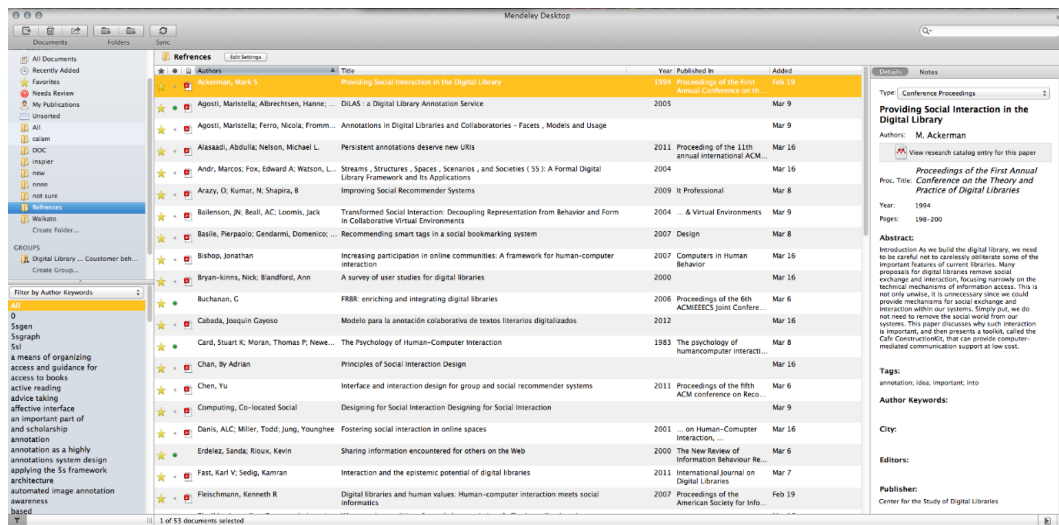
Figure 3.10: Adding files to Mendeley.

Mendeley has certain features that may assist users with both organising papers and reading them. Files may be organised in folder and subfolders, and the 'Using tags' feature can help with identification. All PDFs in Mendeley are searchable and can be filtered by publication, author, or tag. Each PDF can be indicated as read, unread, or favourite. The PDFs can be browsed in Mendeley itself because it has a PDF reader that enables the user to highlight and annotate (Butros & Taylor, 2010).

Users are able to share references by setting up groups, but they can also control their own libraries, choosing whether to make them viewable to others or not. Mendeley has many popular styles and bibliographic formats, and it works with two word processors, Word and Open Office (Butros & Taylor, 2010b; Mathew Willmott, n.d.).

Mendeley fully supports five requirements: R2, storing; R3, browsing; R4, searching; R5, organising; and R8, annotating the metadata and documents. Moreover, it partially supports requirements R1 (downloading) and R6 (reading status). With regard to the former, it only downloads metadata from the internet, and in terms of the latter, it provides only two indications: read and unread. The indication for citation is not provided in Mendeley.

3.2.5 **LibraryThing**

'LibraryThing is an online service to help people catalog their books easily' ('About LibraryThing,' 2013). LibraryThing is described as a social website so that users can share their book collections with their friends who have an account in LibraryThing. It consists of two parts, the users' collections and the social part (see Figure 3.11). Users can add books to the collection by searching through the Library of Congress[13] and over 695 world libraries. Any item in the collection can be tagged, reviewed, annotated, and rated, and can be shared with friends.
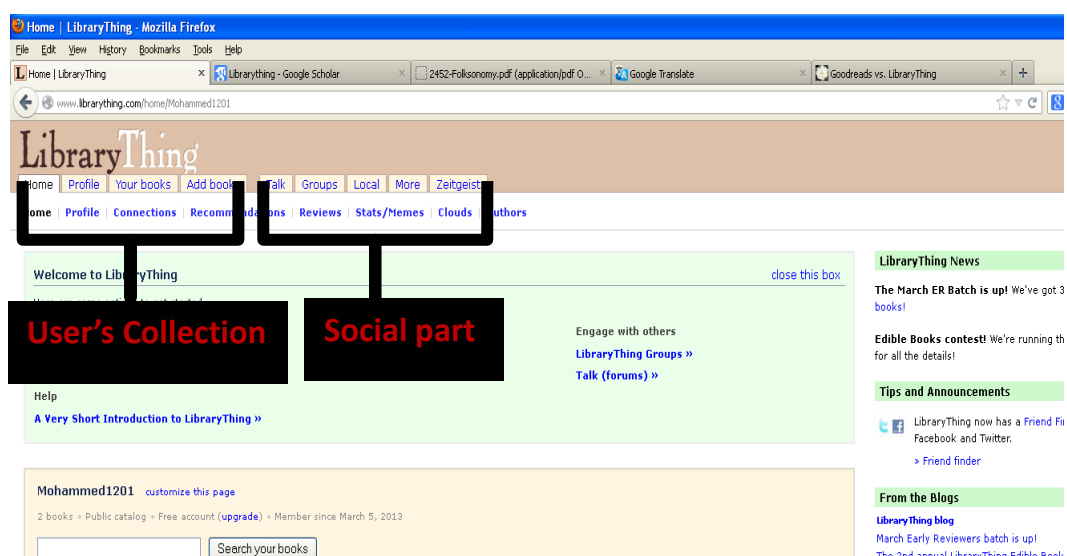
Figure 3.11: LibraryThing main window

Users can organise their collections in folders and sub folders, such as 'To read' and 'Currently reading' (see Figure 3.12). When users add a book, they get only its picture and information, such as title, author, and publication date.

---

[13] See http://www.loc.gov/index.html

Figure 3.12: LibraryThing collections

LibraryThing supports only three requirements: R5, organisation and management; R6, reading phase; and R8, annotation. It partially supports R1 (downloading), R2 (storing), and R4 (searching), providing those functions only for metadata. The other three requirements are not provided in LibraryThing.

3.2.6 Goodreads

'Goodreads is the largest site for readers and book recommendations in the world' (About Goodreads, 2013). Goodreads is web-based software that requires a free account, and it helps people to organise their reading lists and creates library catalogues. Users can add books to their bookshelves by searching for them (see Figure 3.13). Users have options in adding the books they want. They may first add the book directly to the 'Want to read' shelf or they may specify the shelf they want: 'Read', 'Currently reading', or 'Want to read'.
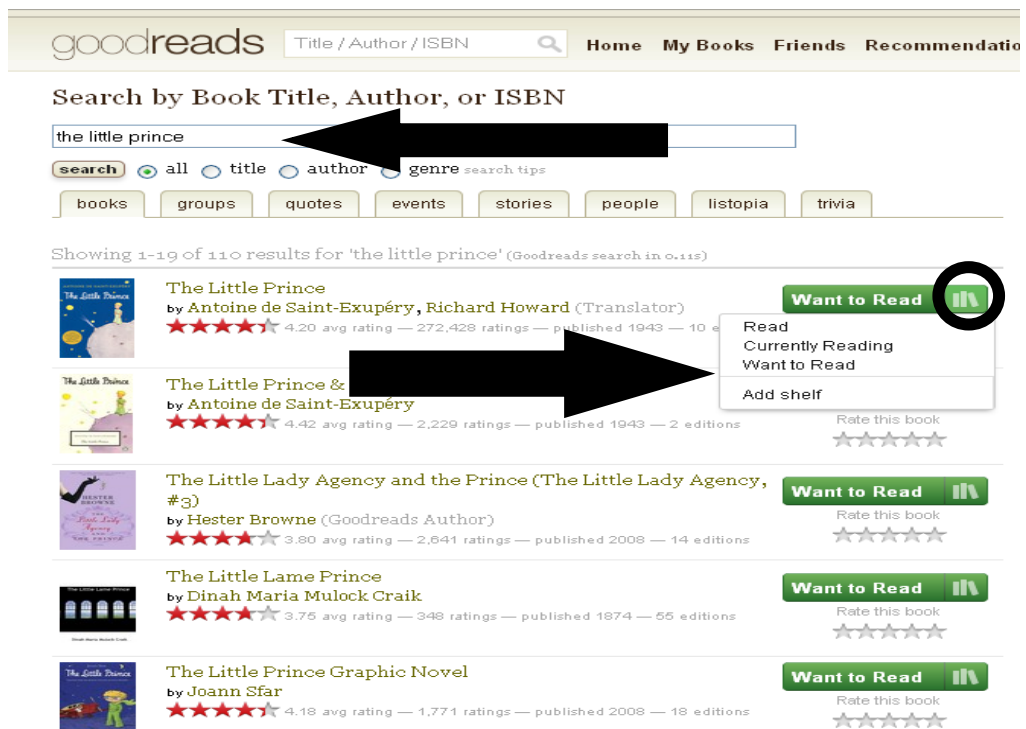
Figure 3.13: Search and add items to the library

Goodreads provides three default shelves: namely, 'Read', 'Currently reading', and 'Want to read', but users are able to create other shelves. Moreover, they can annotate and rate the items in the collection. Adding a friend in the friends' list gives users the opportunity to share their collections, reviews, discussions, and suggestions. The website also allows the user to buy a copy of the item added (see Figure 3.14).

Figure 3.14: Purchase a digital copy feature in Goodreads

Because Goodreads is a social network digital library like LibraryThing, it supports some of the same features (R5, organisation and management; R6, reading phase; and R8, annotation). It partially supports R1 (downloading), R2 (storing), and R4 (searching) but these functions are available only for metadata.

3.2.7 **Greenstone**

Greenstone is open source software for building one's own digital library, which may contain books, images, audio, video and PDFs. It is able to gather, organise, and build those items automatically (Witten, Bainbridge, & Boddie, 2001). Moreover, it provides three main functions: browse, search, and index. To have a digital library in Greenstone, users need to install the free software, which consists of two main parts: building and browsing. For building, users import a collection, add the metadata for each file, and then design search and index styles (see Figure 3.15).
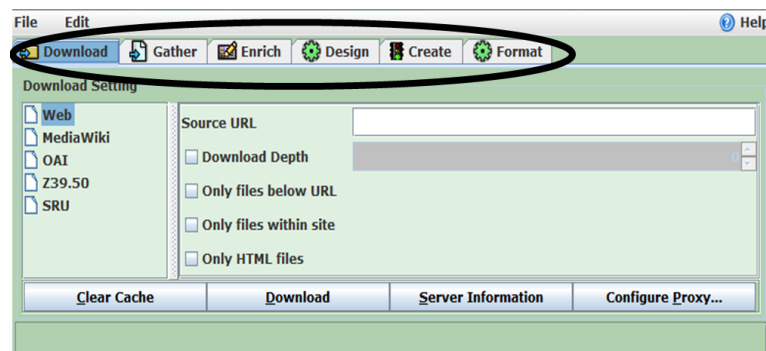
Figure 3.15: Greenstone librarian interface (GLI)

After the library is built, users can browse the material and the metadata included in the collection by a web browser. Greenstone provides a search function which enables a search of the content of each item and the metadata (see Figure 3.16). While browsing, users can edit and add metadata elements so it is possible for them to add comments to any item in the collection while they are reading (Witten, Bainbridge, & Nichols, 2009).
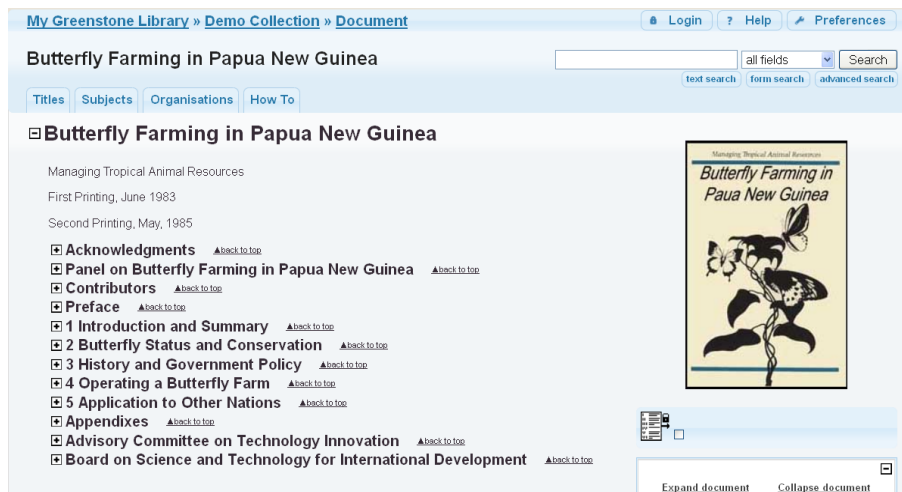
Figure 3.16: Reader's interface

Greenstone supports four features of the requirements: R2, storing; R3, browsing; R4, searching; and R5, organising the documents and metadata. It does not support the rest of the requirements (R1, downloading; R6, reading phase; R7, material used in writing paper; and R8, annotating).

3.3   Summary

In this chapter, we derived requirements that match the findings from Section 2.4. The developed system should have four aspects for tracking reading materials: localisation, organisation and management, indication, and annotation. For localisation, the reading materials should reside on storage that is always accessible to the user (e.g., local hard disk or personal space in a data cloud). The system should be able to download the materials (e.g., PDF, image, Web page). However, it also needs to store provenance information for citing or accessing the original files. In addition, the system must provide search functionalities that allow quick access to contents, comments, and metadata. In terms of organisation and management, the system should support a flexible structure for the materials because users need to be able to group them according to the current task. For indication, the system should visualize the reading status of the materials— indicating, for example, if it has been (partly) read, is unread, or is cited. For annotation, the system should enable users to write a description about each item in the collection, thus assisting with easy recognition and providing an aid for remembering the materials more easily.

After an explication of the requirements, seven systems were reviewed that claim to help people with organising and managing their reading material. Table 3.1 shows how these systems support our requirements. It is immediately obvious that whereas all the systems provide one aspect of the requirements—that is, R5 organisation and management—none supports all the requirements.

Table 3.1: Comparison of the systems reviewed

		Software reviewed						
		Endnote	**RefWorks**	**Zotero**	**Mendeley**	**LibraryThing**	**Goodreads**	**Greenstone**
System Requirements	**R1-Downloading documents and metadata**	Metadata	———	Supported	Metadata	Metadata	Metadata	———
	R2-Storing documents and metadata	Supported	Supported	Supported	Supported	Metadata	Metadata	Supported
	R3-Browsing materials	Supported	Not through software	Not through software	Supported	———	———	Supported
	R4-Searching	Metadata	Metadata	Supported	Supported	Metadata	Metadata	Supported
	R5-Organising and managing	Supported	Supported	Supported	Supported	Supported	Supported	Supported
	R6-Indicating reading status	———	———	———	Read and unread	Supported	Supported	———
	R7-Indicating whether used for citation	———	———	———	———	———	———	———
	R8-Annotating	Supported	Supported	Supported	Supported	Supported	Supported	———

▮ means supported   ▮ means partially supported   ——— means not supported

LibraryThing and Goodreads provide the ability to indicate materials in the three reading phases—read, to read, and currently reading—but no system so that users can track where the item has been cited.

Four systems out of the seven cover more than three aspects of our requirements: Endnote, Zotero, Mendeley, and Greenstone. Mendeley supports the same requirements that are supported by Greenstone (R2, R3, R4, and R5) in addition to R8. However, Endnote and Zotero resemble one another in supporting three requirements: R2, R5, and R8. The only software that provides R1 (downloading documents and metadata) is Zotero.

Chapter 4

4 System Design

This chapter focuses on a system design for managing materials, from making the decision to getting the system ready. It discusses which systems are both convenient for our requirements and extendable. It also explains each one in detail and analyses its functions to demonstrate which one is better for extension.

4.1 Design Decision

The main consideration here is extending one of the available software systems that meets some of the proposed system's requirements. From the variety of options, the Greenstone digital library software and Zotero were chosen, mainly for the two reasons that they are freely available and that they use open source code. In this section, each of the two systems is examined in greater detail, and an analysis offered concerning which requirements are met and for which ones the software would need to be extended.

4.1.1 Detailed Overview of Zotero

Zotero is open-source bibliographic software that allows researchers and students to collect, store, organise, and manage their resources. Moreover, Zotero can do in-text citation and bibliography, which benefits Microsoft Word, Open Office, and Latex users by means of a plug-in tool (Puckett, 2011; Trinoskey, Brahmi, & Gall, 2009).

Zotero has a browser extension that helps users to collect and manage their resources easily and all the information for those materials (i.e. the metadata) while they browse the Internet. It works only with Mozilla Firefox (George Mason University, n.d.; Kessler, 2007). However, for other browsers, like Chrome and Safari, there are separate plug-in and standalone applications (Puckett, 2011).[14]

---

[14] available from the website: http://www.zotero.org/download/
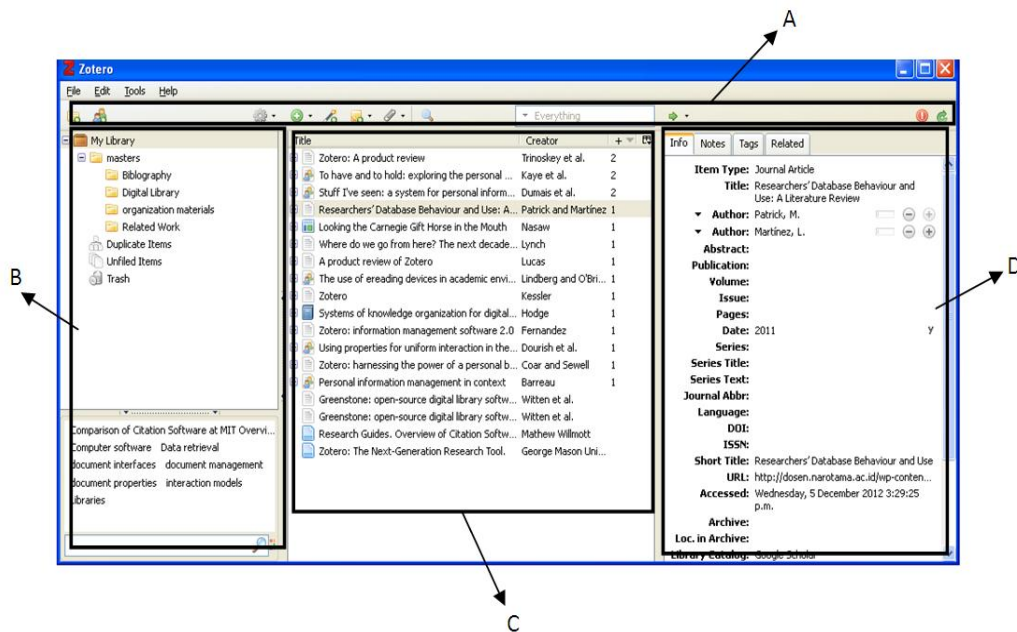
Figure 4.1: The main window for Zotero

Zotero's main window has five parts (see Figure 4.1). Each part is explained below:

- Part A: this bar contains Zotero's main functions, such as *Create collection, Add source, Attach file*, and *Search* (see Figure 4.1, Part A).

- Part B: this part, called the collection panel, allows users to organise and manage all their sources. A default collection, called 'My Library', contains all the sources (see Figure 4.1, part B). In this collection, users may create further collections and sub collections and may also control all collections by renaming and moving the sources among them.

- Part C: this part, the title panel, displays the title of the source and its author. Beside each source is an icon which represents the type of source, such as books or web pages (see Figure 4.1, part C). It displays only the sources in the collection that are selected in the collection panel.

- Part D: this part displays the information for any selected source from the title panel (see Figure 4.1, part D). This information includes bibliographic metadata—such as title, author, and publication company--notes, tags, and any related item in the library (Kessler, 2007).

Zotero has many features that can help the researchers and students in terms of bibliographic information, but here Zotero's main features are highlighted along

with their implications for organising sources, adding sources, adding notes and tags, and searching sources.

- **Organising sources:** users in Zotero are able to create a collection under any of the collections in their libraries (see Figure 4.2, which shows how the folders are structured in Zotero—that is, the 'Master's Folder' has five subfolders: 'Bibliography', 'Digital library', 'Introduction', 'Organization materials', and 'Related work'). Users can move any item in their library by dragging and dropping the items they want.
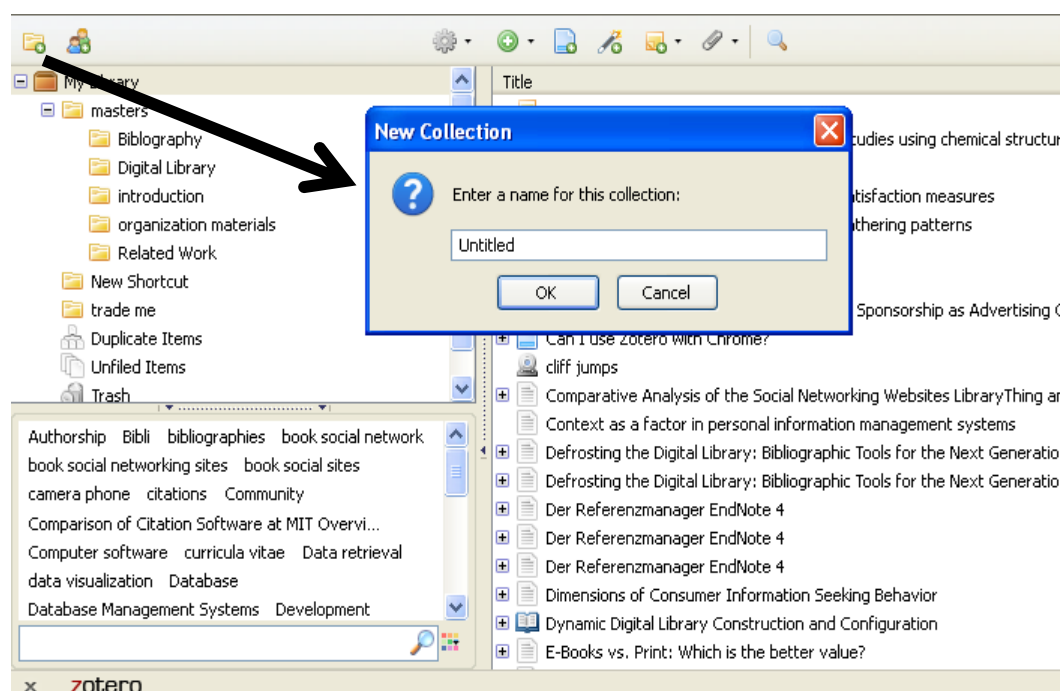
Figure 4.2: Adding new collections

- **Adding sources:** There are two ways of adding sources to Zotero. First, they may be added manually: the user can add the source by clicking on the green button (see Figure 4.1, part A) and then choosing the type of source (book, journal article, webpage, etc.). All the metadata or information for the source (author, title, date of publication, and other fields) must be entered by the user as well (see Figure 4.1, part D).

The second method is automatic, as Zotero has an automatic source sensor that can recognise the type of item on the webpage (such as a book, journal article, or newspaper article; see Figure 4.3). The user needs only to click on

the icon which is located on the right side of the URL bar by which either single items or several items can be added (see Figure 4.3). With just a click on this icon, the item will be stored in the user's library and all the bibliographic information for that item will automatically be added to the library, as well. This option requires users to install the browser extension, by which means they can import automatically from several websites.
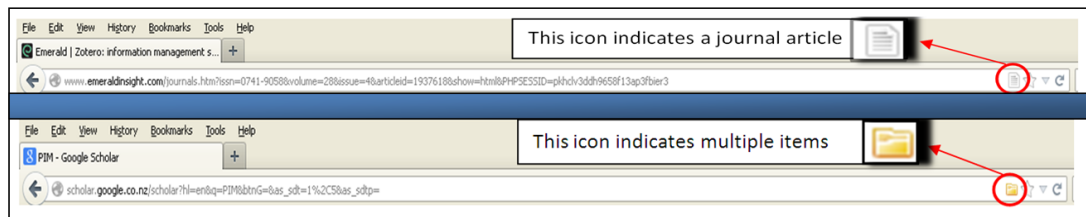
Figure 4.3: Zotero's recognition feature according to the type of item

For each source, users are able to add files (such as pictures, PDFs, and web pages) that they may need, or such a file can be placed as an additional file to that source (see Figure 4.4).
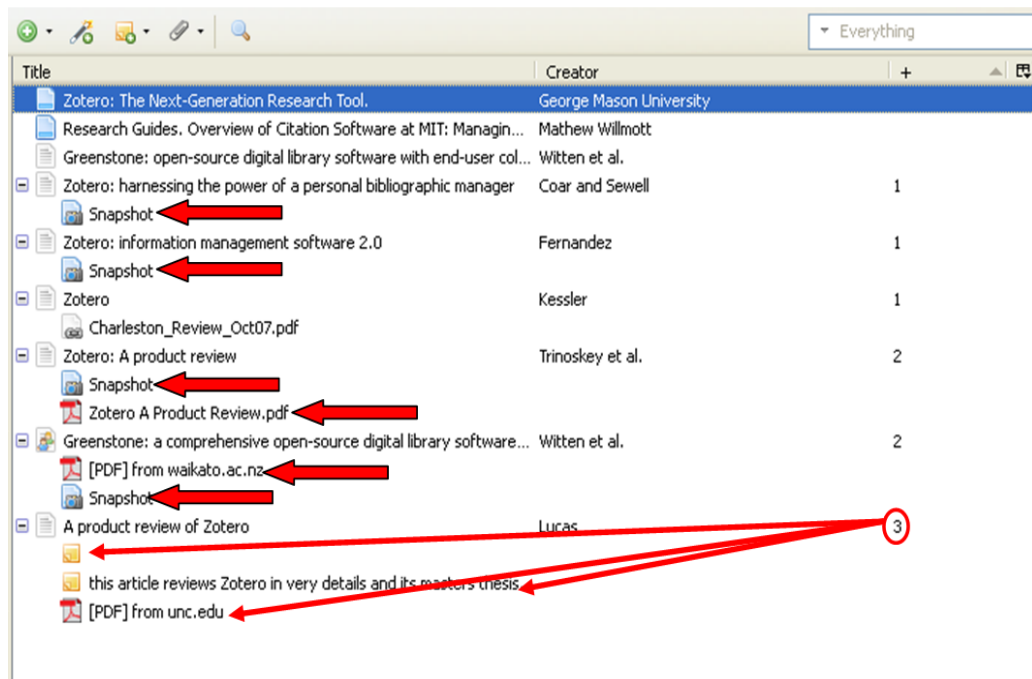
Figure 4.4: Zotero can keep each item for any source

- **Adding notes and tags:** Zotero allows users to add annotations to any source in their library; each note will be listed under the source that has been annotated, and the note can stand alone as well (see Figure 4.5). As shown in Figure 4.5, users have several tools and enough space for writing their annotations. The user may easily add not only notes but also many tags for any source (see Figure 4.6).
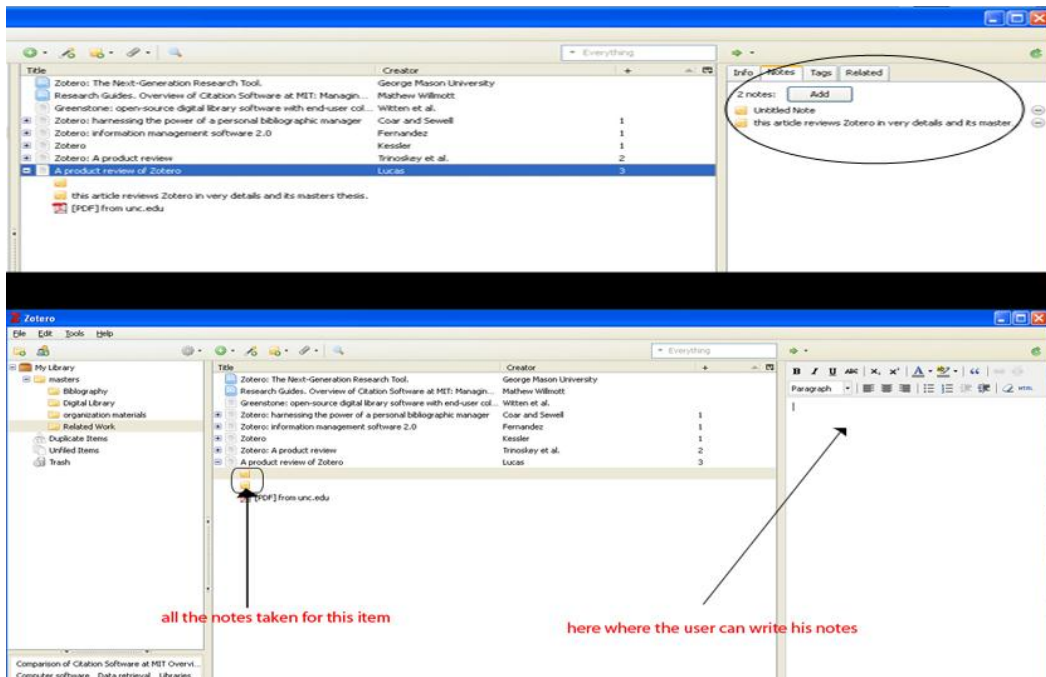
Figure 4.5: Adding notes for any source
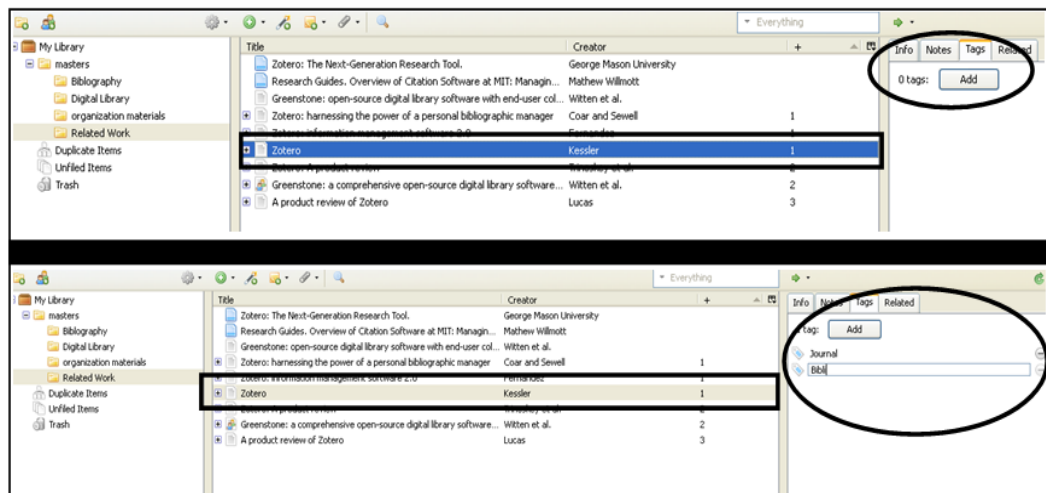
Figure 4.6: Adding tags

- **Searching sources:** In a review of the features of Zotero, Kessler (2007) pointed out the benefits of Zotero for both researchers and students. She ranked the search function in Zotero with five stars out of five, which means it is excellent. All the information in Zotero is searchable, including the title panel, the content of such sources as PDFs and web pages, the information for the source (metadata), the tags, and the notes taken (see Figure 4.7).
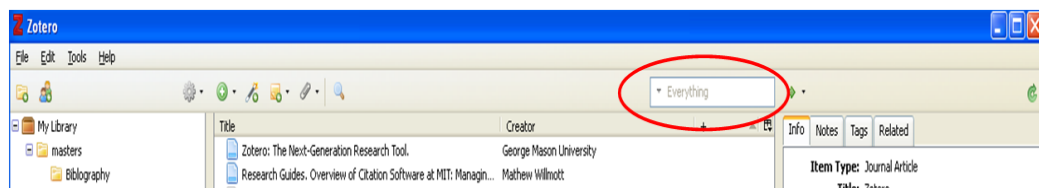
Figure 4.7: Search feature

Another option in Zotero is the advanced search (see Figure 4.8). In this option, users can control and specify the search scope, and can name and save their searches.
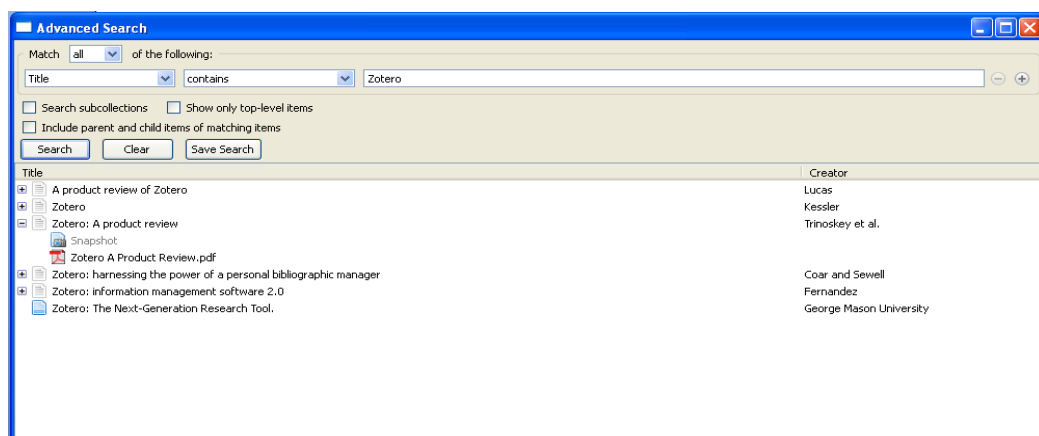
Figure 4.8: Advanced search window

4.1.1.1 Analysis of Zotero

Table 4.1 shows the main features that Zotero offers along with the features of the proposed system that it does not have. Zotero already supports five of the features required by our system (see Table 4.1; functions are R1, R2, R4, R5, and R8):

- **R1-Downloading the material through the software**: may be done by means of the two methods already explained in this chapter, manually and automatically (see Figure 4.1, part E, and Figure 4.3).

- **R2- Storing the material itself as well as its metadata:** Figure 4.4 in this chapter shows the attachments for each source in the library and how they can be attached easily with several types of items. The information for each source can be displayed as it is described in Figure 4.1, part D.

- **R4- Searching feature**: as described in this chapter in Section 4.1.1, there are two options for searching, standard and advanced (see Figure 4.7 and Figure 4.8).

- **R5-Organisation and management**: The part on organising sources in Section 4.1.1 of this chapter demonstrates that Zotero allows its users to create and separate their collections and add items to any collection or sub-collection by dragging and dropping the item.

- **R8- Writing a short description about each paper**: Section 4.1.1 also explains and demonstrates in Figure 4.5 how easily users can annotate their items.

Zotero supports further functions not mentioned in our requirements list, which may, nevertheless, be useful for our system:

1. Zotero can recognise and describe the available types of sources from any webpage.

2. The metadata of sources are automatically attached to the library.

3. The system for adding sources is very flexible.

4. Citation.

Zotero is missing three features required in our proposed system:

- **R3-Browsing the material using the software.**
- **R6-Indicating which material has been read, is currently being read, or is designated to read.**
- **R7-Indicating the material used as references in documents.**

Table 4.1: Existing functions that Zotero offers and others that are needed

Functions offered by Zotero	Functions needing to be developed in Zotero
R1-Downloading the material through the software	R3-Browsing materials through the software.
R2-Storing the material itself in addition to a list of the material's metadata	R6-Indication for material read, currently reading, to read
R4-Searching feature	R7-Indication for material used references in writing documents.
R5-Organisation and management	
R8-Writing a short description of each paper	

- The metadata of sources are attached automatically to the library.
- -Zotero can recognise the type of sources (this feature does not work with all websites).
- The system for adding sources is very flexible.
- Tagging
- Citation

4.1.2 Detailed overview of Greenstone

Greenstone is open-source, digital library software that constructs and presents collections of information. Two main features can help users in Greenstone: ease of finding information and good browsing facilities. Moreover, the collections in Greenstone, which are maintainable and extendable, can be automatically rebuilt (Witten, Boddie, Bainbridge, & McNab, 2000). Greenstone has two important processes. The first is building the collection, which is done by the Greenstone librarian interface (GLI), an independent Java application. It involves gathering materials and metadata, and designing the library activities. The second is the reader's interface, which is the Greenstone home page, for browsing and searching the collection. Both of them are run on the same computer, using the

Greenstone server. Therefore, the collection can be accessed by a standard web browser through either local or remote access (Witten et al., 2009).

4.1.2.1 Greenstone librarian interface (GLI)

Building a digital library in Greenstone has five main steps: gathering documents, enriching the documents with metadata, and designing, building, and formatting the collection.

1. Gathering documents: the items can be collected by either downloading them from the Internet or attaching them from the computer (see Figure 4.9 and Figure 4.10). For the former, if the sources are external and need to be in the collection, users must employ the Download feature (see Figure 4.9). They need protocols, like the web page HTTP. For the latter, if the sources are saved on the computer, the Gather step would be used (see Figure 4.10). Adding items from the left panel (computer files) to the right panel (collection files) is easily done by dragging and dropping either individual files or whole folders.
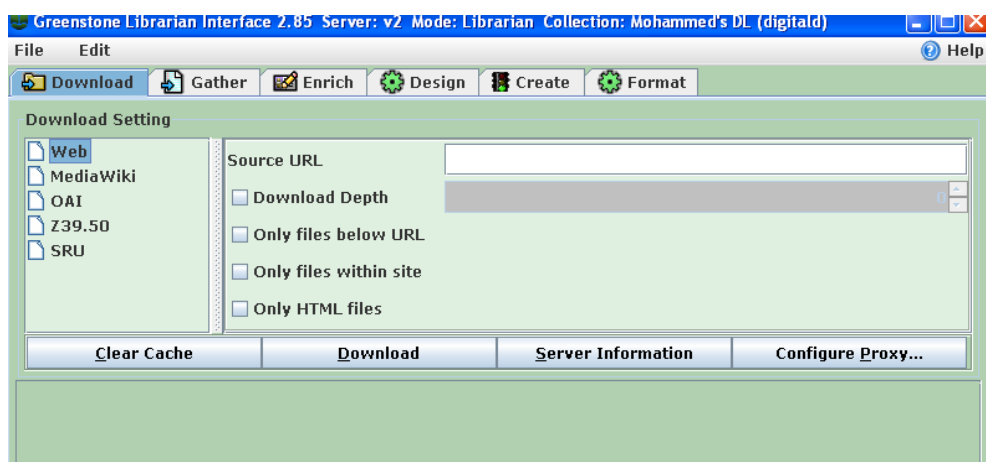
Figure 4.9: Downloading files from the Internet in Greenstone
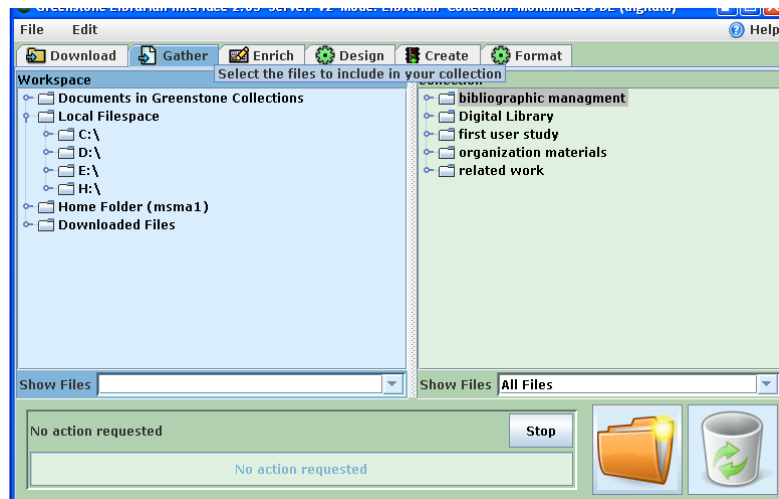
Figure 4.10: Attaching files from the computer in Greenstone

2. Enriching the documents with metadata: in Figure 4.11, the documents
    collection shows in the left panel, and the metadata shows in the right
    panel. Users can assign each item in the metadata by adding it manually to
    each document, or one item can be assigned to several documents. If the
    metadata are assigned to a folder, all the items in this folder will inherit the
    assigned metadata (see Figure 4.11). At this phase, the user can edit or
    update the metadata. Different metadata sets can be assigned to one
    collection, such as the Dublin Core Metadata Element Set. Moreover,
    users can create their own metadata at any time by using either the
    Manage Metadata Sets button (it is located down on the left in the Enrich
    tab; see Figure 4.11) or the Greenstone Editor For Metadata Sets (in the
    Windows platform, *Start > All programs > Greenstone-3.05 >
    Greenstone Editor For Metadata Sets*).
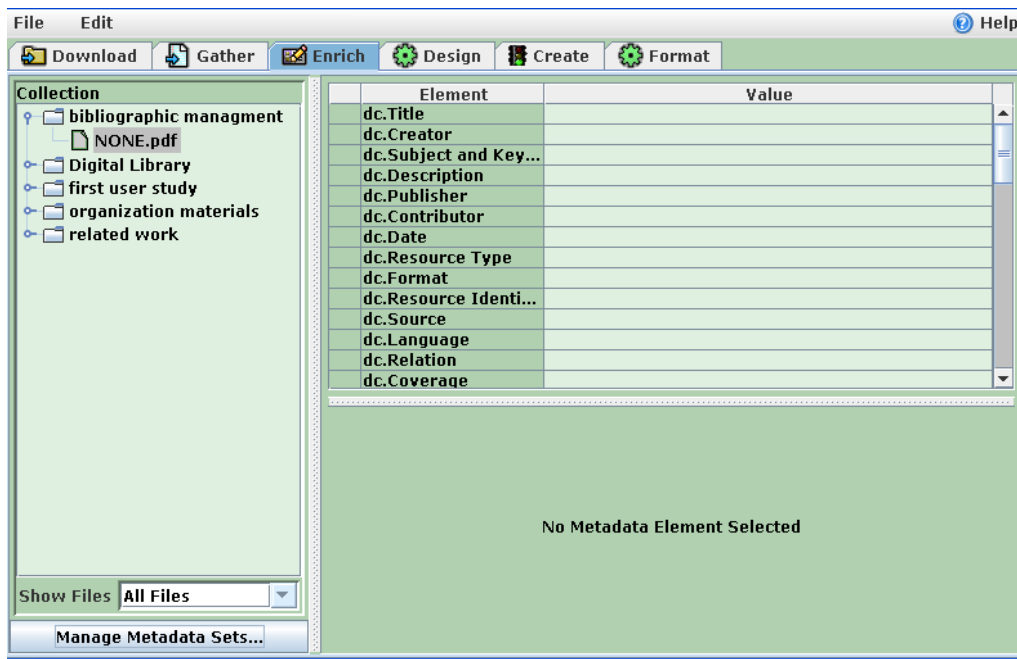
Figure 4.11: Enriching metadata in Greenstone

3. Design the collection: this is the phase in which users organise and structure their collection. It involves a series of four different actions, which can be seen on the left panel in Figure 4.12. Each one can be controlled by clicking on the functions that appear on the right side in Figure 4.12. The series of actions is explained below:

- **Document Plug-ins:** plug-ins accomplish the major work of the import process (Witten, Bainbridge, & Nichols, 2009). A plug-in can be defined as an ID that lets Greenstone identify the type of imported file. Each imported file has to be processed by one of the plug-ins provided; hence, when it is imported, it is passed to the plug-ins to find one that can process it. For example, *moo.pdf* is passed as a PDF file and *moo.txt* is passed as a Text file. At that panel, the user can do three operations: add, configure, and remove plug-ins.

- **Search Indexes:** Greenstone can search in full text and it has three indexers: MG, MGPP, and Lucene. MG performs the search operation using three separate indexers—paragraph, section, or document level. MG has been re-implemented and restructured into MGPP with some improvements. MGPP performs the search operation at the word level to handle the field, phrase, and

proximity searching efficiently. Lucene performs the searching at the single level by handling field and proximity searching. It was developed to facilitate building the collection incrementally. After specifying the indexer type for searching, the user can assign an index for title or author to be more specific.

- **Partition Indexes:** these divide the document into several sections so that each section can be researched individually. Partitioning indexes can help to control the search space. For example, the indexes can be partitioned based the language, such as English, Spanish, or Arabic.

- **Browsing Classifiers:** they are used for browsing functions and are based on the metadata. Browsing classifiers are provided to organise and structure the materials in the collections. Greenstone has many types of classifiers, such as *Lists*, which is the default, and *Hierarchies*. Each classifier has different features for browsing the materials. The user needs to specify the type of classifier and then add it by choosing an item of the metadata, such as title, file name, or author.
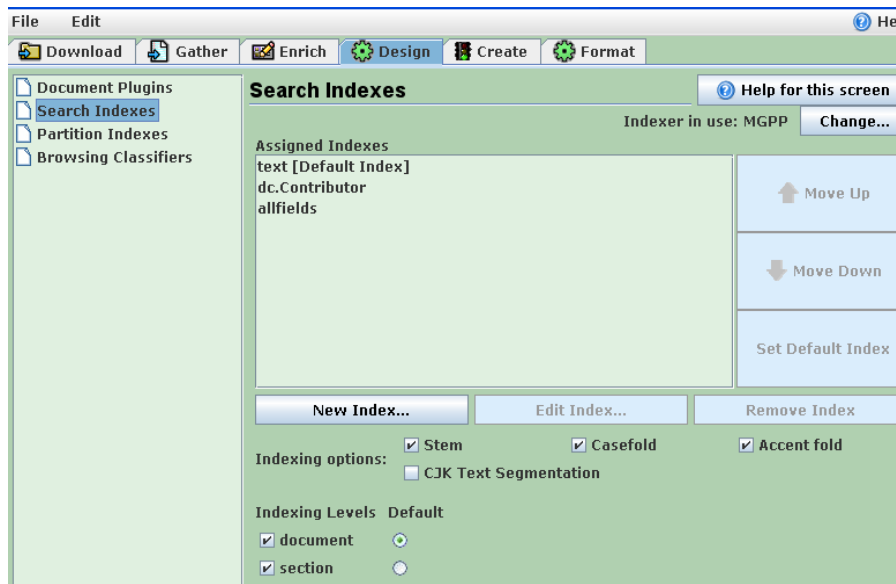
Figure 4.12: Designing the library in Greenstone

4. Building the library: when the library is ready for building, the user needs only to click on the button *Build Collection*, and the process of generating

59

will take a few minutes (see Figure 4.13). The library may then be opened through the web browser by following the *Preview Collection* link.
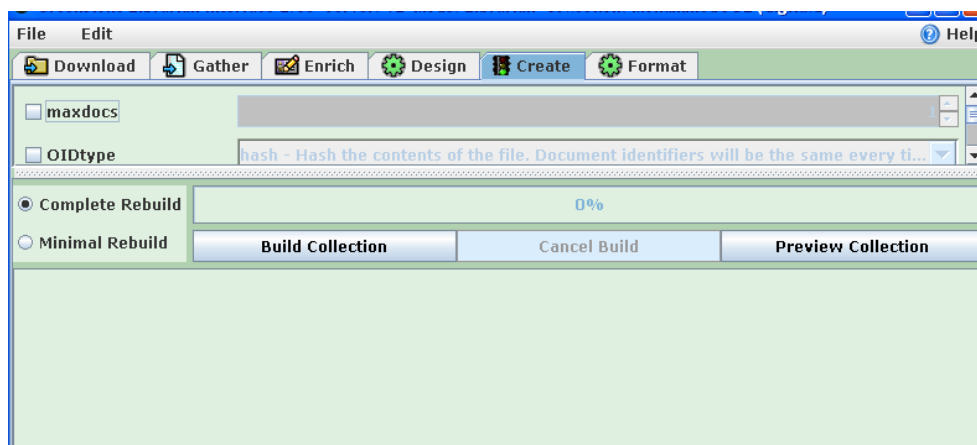
Figure 4.13: The final step in building a library in Greenstone

Formatting the collection: 'the formatting language is powerful but mysterious. And, in truth it is not very interesting'(Witten, Bainbridge, & Nichols, 2009). Because of that we will go through its parts quickly. Formatting consists of four parts: General, Search, Translate Text, Cross-Collection Search. The general part allows users to write information about the collection, such as title and description. The search part is for specifying names for search indexes. The translate text part allows users to provide some of the parts in their collection's interface in another language. Cross-collection search gives users the opportunity to search in several different collections.

4.1.2.2 Reader's Interface

After the library is built, the collection will be shown on a webpage called the 'Reader's Interface'. Some of its most useful features are searching, browsing the materials, and editing the metadata.

- Searching: The user can search for any information in his collections including full text, metadata, and all other information related to the collections. The search depends on the design of the collection, so some of the collections have indexes of chapters, paragraphs, or titles. For instance, Figure 4.14 shows how the search function is used in the Greenstone demo

collections, and we can see how Greenstone retrieves the information. The search can be limited by title, author, and so on.
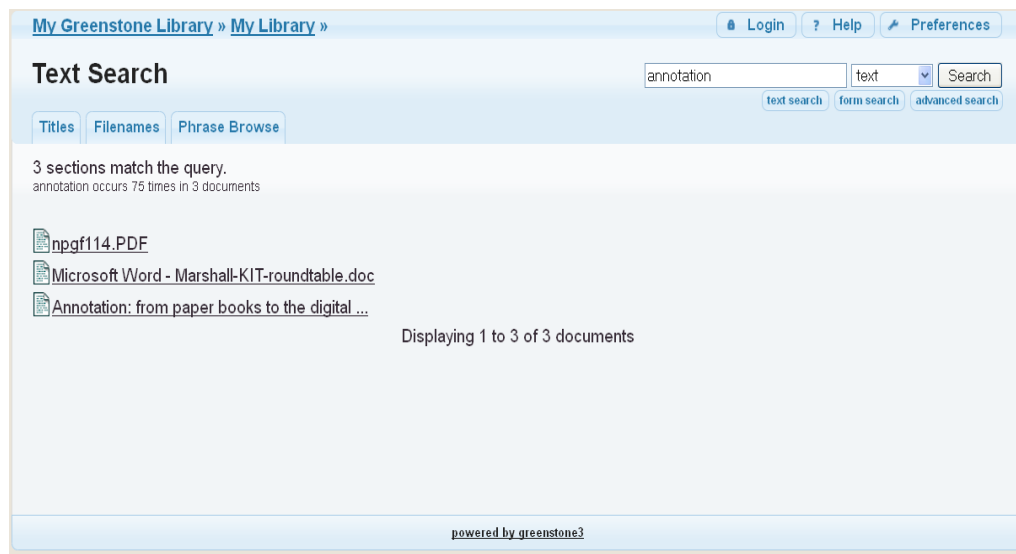
Figure 4.14: Search function in Greenstone

- Browsing materials: the browsing function involves the structures of data which are generated from the metadata, so it may help users to display their collections by lists of titles, authors, subjects, and so on (see Figure 4.15 where all the materials are shown by subjects). Users can browse any material in their collections through Greenstone itself (see Figure 4.16). As Figure 4.16 shows, the material is displayed by sections so that users need only click on the chapter or section they want. Greenstone can deal with seven types of files: HTML, images, office documents, PDF, text files, audio, and XML.
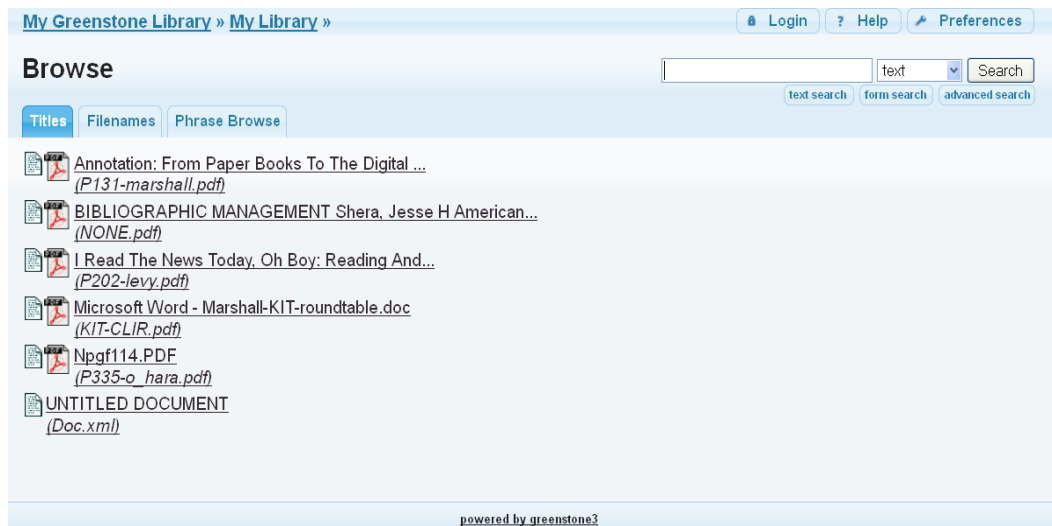
Figure 4.15: Browsing materials by subject
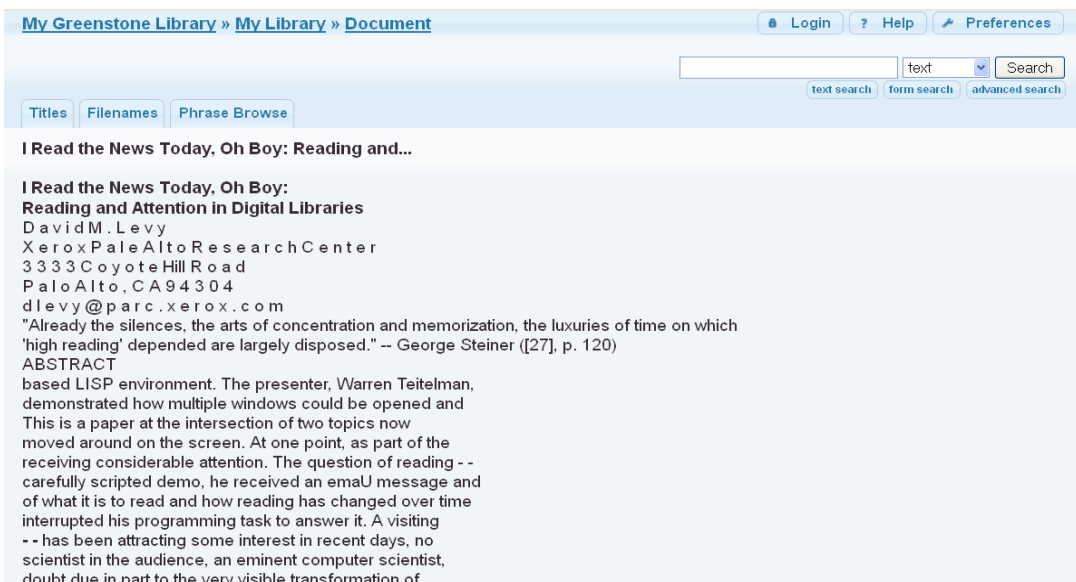
Figure 4.16: Greenstone browsing the content of the material

- Modifying the metadata through the reader's interface: by logging in as an admin, users can edit any metadata in their collection by choosing the file and then clicking on edit metadata (see Figure 4.17). The user can modify and add new elements to the metadata and then save the changes (see Figure 4.18).
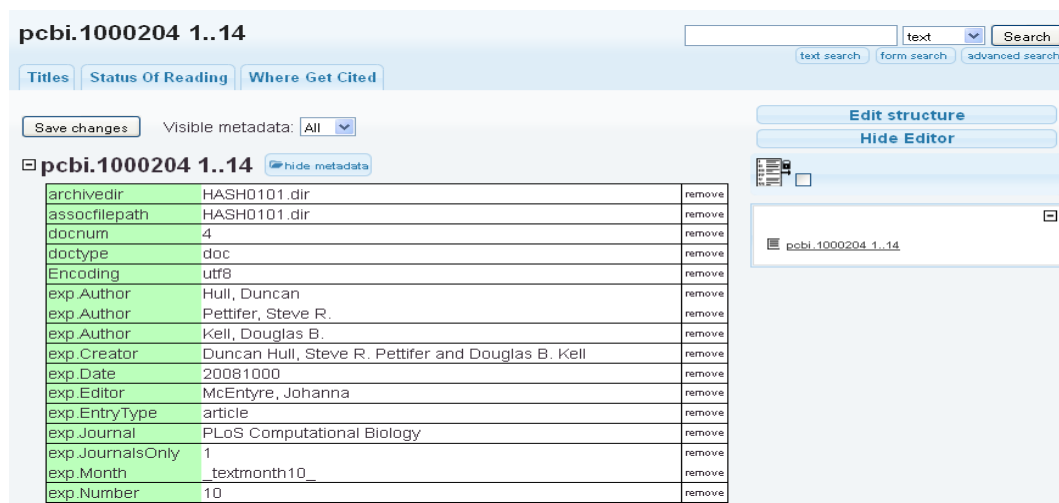
Figure 4.17: Modifying the metadata

Figure 4.18: Adding new metadata

4.1.2.3 Analysis of Greenstone

Table 4.2 shows the main features that Greenstone offers and the features of the proposed system that it does not have. Greenstone already supports five of the features required by our system (see Table 4.2, functions R1, R2, R3, R4, and R5), and the other three features can be adapted (R6, R7, and R8):

1. **R1-Downloading the material through the software:** as explained in Section 4.1.3.1, for the first step (gathering documents) of building a library, there are two options for attaching the materials, either from the computer's hard drive or by downloading from the Internet (see Figure 4.9 and Figure 4.10).

2. **R2-Storing the materials themselves in addition to a list of their metadata:** as described in Section 4.1.2, the first and second steps (enriching the documents by metadata) in building a library in Greenstone are enriching each file or material by metadata, as shown in Figure 4.11.

3. **R3-Browsing the material through the software:** as explained in Section 4.1.2, and shown in Figure 4.15 and Figure 4.16, the materials can be displayed through Greenstone.

4. **R4-Searching feature:** this feature in Greenstone is as described in Section 4.1.2, step 3, and as shown in Figure 4.14.

5. **R5-Organisation and management:** the collection can be structured and organised in Greenstone as described in Section 4.1.2 in step 3.

6. **R6-Indication for the material's reading status:** although there is no function that supports this requirement, it can be adapted by creating a metadata set called 'Tracking material' (see Section 4.1.2, step 2) and assigned in the browsing classifiers (see Section 4.1.2, step 3). Thus, instead of an indication that shows the reading phase, there will be a browsing classifier for each phase.

7. **R7-Users can indicate the material used as references in writing documents:** it can be adapted as shown for requirement 6.

8. **R8-Users can write a short description about each paper**: in Section 4.1.2, step 2, users can employ the element called 'Description', which enables them to write individual notes.

Table 4.2: The functions Greenstone offers and others that need to be developed

Functions offered by Greenstone	Functions that need to be developed in Greenstone
R1-Downloading the material through the software.	R6-Indication of material's reading status .
R2-Storing the material itself as well as its metadata.	R7-Indication of the materials used as references in writing documents.
R3-Browsing the material through the software.	R8-User can write a short description about each paper.
R4-Search feature.	
R5-Organisation and management.	

4.1.3 **Comparison between Zotero and Greenstone**

This section reviews all the requirements of the system in comparison to the features available in Zotero and Greenstone. In addition, we compare the efficiency of each requirement between Zotero and Greenstone. Table 4.3 shows the requirements that are supported, not supported, or partially supported by either Zotero or Greenstone ((+) means supported, (-) means not supported, and (-+) means partially supported).

Table 4.3: Comparison of Zotero and Greenstone

Requirements	Zotero	Greenstone
R1-Users can download materials through the software.	+	-+
R2-Users can keep the materials themselves in addition to a list of the metadata.	+	+
R3-Users can browse materials using the software.	-+	+
R4- Search feature	+	+
R5-Organization and management	+	+
R6- Indication of material's reading status	-	-+
R7- Indication of materials used as references in writing.	-	-+
R8- Users can write a short description of each paper.	+	-+

A comparison of the efficiency by way of a description of each requirement in terms of functional performance follows for both systems.

R1-Download the reading material through the software: in Zotero, users can download a document and its metadata automatically by one click while browsing the Internet. However, in Greenstone, the user needs to import the documents and metadata manually.

R2-Storing the reading material as well as a list of its metadata: both Zotero and Greenstone provide this service. In Zotero, any file can be imported to the library or collection as a source, or it can be specified as belonging to any source: for example, one paper can have three or four sub-files, such as PDF, image,

video, and webpage. Similarly in Greenstone, multiple items can be grouped in one folder to which metadata has been assigned, and then all the items will inherit this metadata.

R3-Browsing the reading material through the software: users in Zotero can browse all the types of files that have been added to their libraries, but the browsing process does not occur through Zotero itself. The operator of each type must be downloaded onto the computer; for example, a PDF file needs a PDF reader, a webpage needs a web browser, and so on. However, in Greenstone all the files are displayed through either the browser of Greenstone or those of their operators.

R4-Search feature: in both Zotero and Greenstone, users can search through the metadata and content files, a basic of our requirements.

R5-Organization and management: Zotero uses folders and subfolders, and Greenstone uses partition indexes and classifier indexes. Although they are different in the way they organise material, they achieve the user requirement for being able to organise and manage materials.

R6-Indication of material's reading status: neither Zotero nor Greenstone supports an indication for the reading status of the material. Greenstone, however, provides an alternative option that might fulfil the requirement. The solution was described in Section 4.1.2.3.

R7- Indication of material used as references in written documents: as for R6, neither Zotero nor Greenstone supports such an indication, but Greenstone has a feature that can be adapted, as described in Section 4.1.2.3.

R8- Writing a short description about each paper: this feature is available in both Zotero and Greenstone, but adding notes in Zotero is much easier than in Greenstone.

4.1.4   Conclusion

In conclusion, we should note, as a reminder, that we have here addressed requirements R1 to R6. Options for how to incorporate R7 and R8 will be discussed in a future chapter. It is clear that Greenstone supports the same requirements as Zotero does. In addition, Greenstone is more flexible then Zotero

because two requirements of our system which are certainly not supported in Zotero could be fulfilled through adaptations in Greenstone and thus serve the purpose. Although Greenstone seems to have all the requirements, it does not provide as efficient a procedure for collecting materials and metadata as Zotero does. An analysis of both Zotero and Greenstone has shown that neither of the two systems is a clear favourite as a foundation software that can be extended. Another option may be to combine the two systems and extend them to work together to reach the goal of this project. The next section explains how the two systems will be combined.

4.2   Combining Zotero and Greenstone

In this combined system, Greenstone serves as a document management system and Zotero as a metadata and document provider.

From the overview shown in Section 4.1.2, it seems clear that Greenstone can assist people in identifying which materials have been read, are being read, or will be read. Moreover, they can also browse, index, and search their materials easily. However, Greenstone cannot provide the documents' metadata for helping people prepare their citations as they write. As bibliographic data management, Zotero can offer a solution for this issue because, as shown in Section 4.1.1, Zotero automatically imports documents and metadata from websites. The idea of this system is to use Greenstone for storing and organising the documents and Zotero for capturing document metadata. For a complete, beneficial system, two steps must be carried out:

1. Design and implement software (i.e. a Tracking Reading Material System [TRMS]) to transfer the data automatically from Zotero into Greenstone (shown in blue in Figure 4.19).
2. Extend Greenstone to capture information about the reading status—have read, currently reading, and to be read.

Figure 4.19 shows a high level design for system processing. The PDF and its metadata usually reside in a website and the user downloads these data using Zotero. TRMS will import the data downloaded into the Greenstone collection.
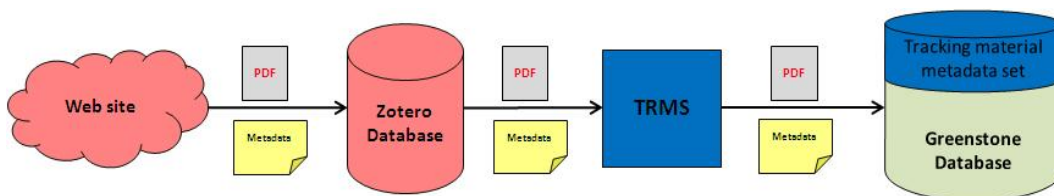
Figure 4.19: High level design of TRMS

4.3   Transferring Data from Zotero into Greenstone

This section describes how the process works manually in the TRMS, starting with the importation of PDFs and metadata from Zotero and ending with the incorporation of these data into Greenstone. This will allow the system to address requirements R1 to R6.

4.3.1   Importing the Items to Zotero

After the Zotero extension is installed in Firefox, an icon appears at the lower left-hand corner of the web browser (see Figure 4.20). From this icon, we can access, organise, and control all the records and their metadata that we have imported.
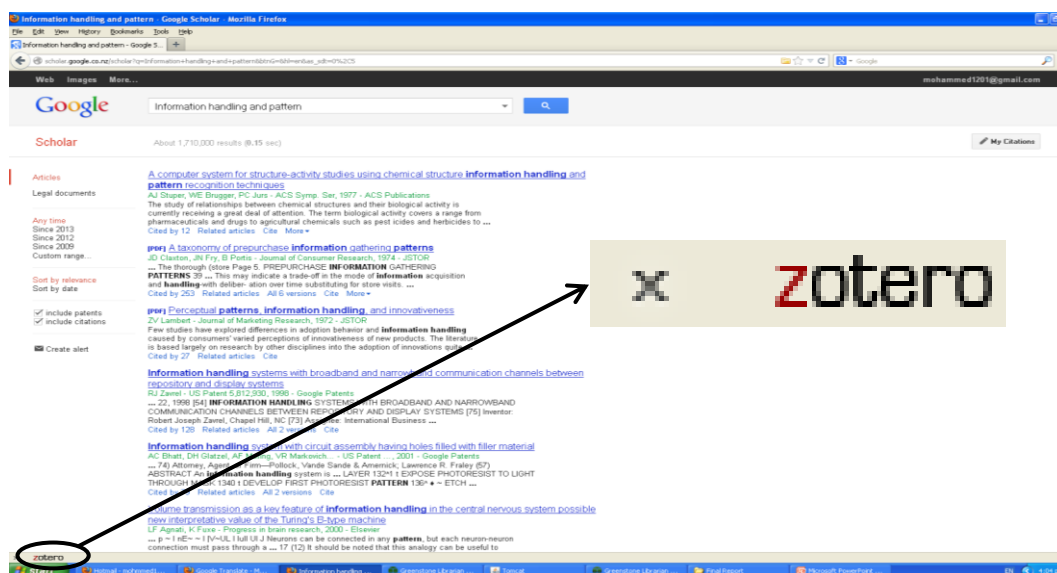
Figure 4.20: Zotero Firefox extension

The items may now be manually imported in several steps into the user's library in Zotero. A Google Scholar source serves as an example here:

1- The user types 'information handling and pattern' in the Google Scholar search textbox (see Figure 4.22) and then clicks on the Zotero icon (lower left-hand corner) to access the collection. A panel comes up in the lower part of the web browser, giving access to the existing Zotero collection. There are two ways of importing metadata and items:

A. The user may download a single item with the following steps:

    A.1   Choose the source desired.

    A.2   Click on the *Add* icon (see Figure 4.21).

    A.3   The metadata and the PDF file will be downloaded into the Zotero library (see Figure 4.21).
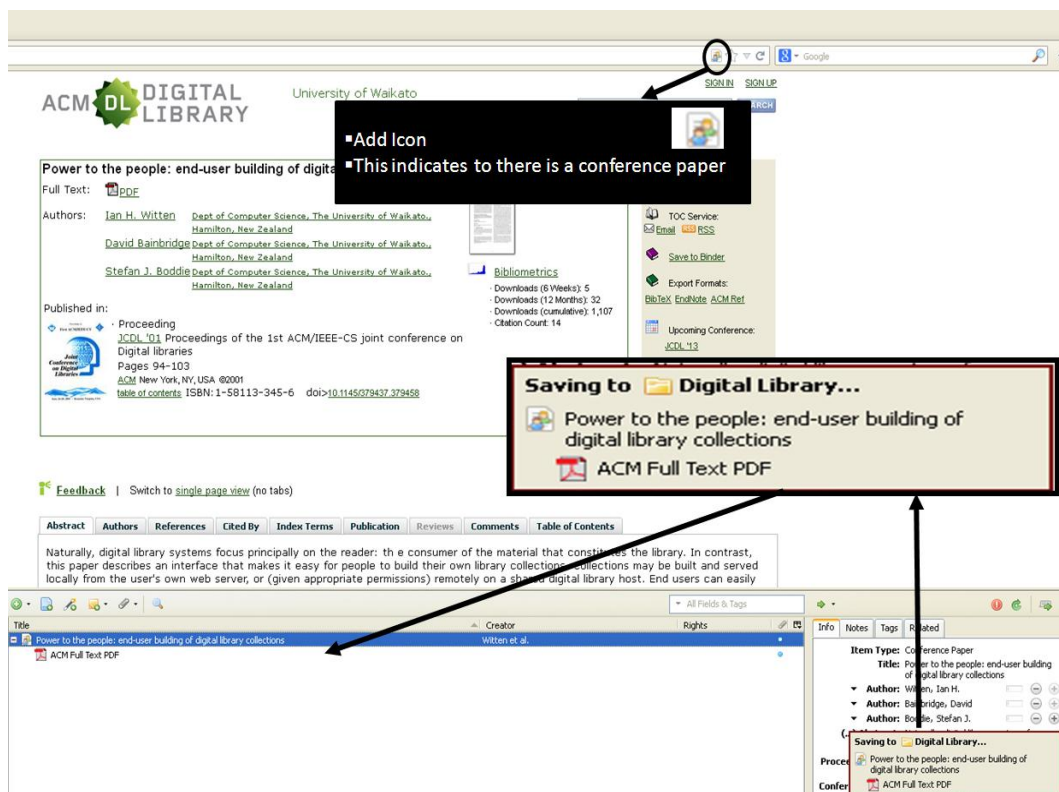
Figure 4.21: Process for downloading a single item

B. Alternatively, users can download several items at one time, first, by following step 1 of the procedure above. At the right-hand side of the URL bar, an *Add* icon indicates that Zotero detected material that may be included in Zotero (e.g., articles or conference papers).
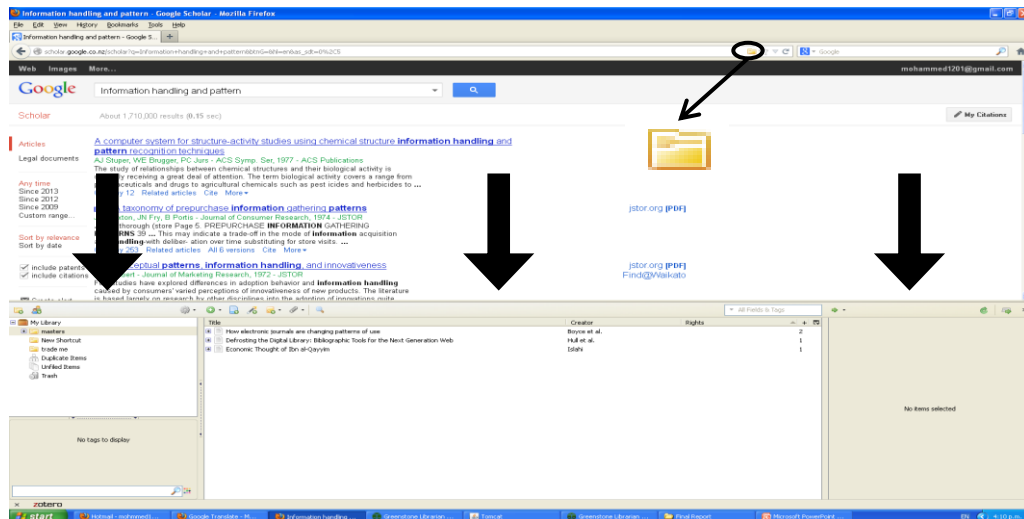
Figure 4.22: First step for searching and adding sources

B.1 When the user clicks on the *Add* icon (see Figure 4.22), a window pops up with all the items detected by Zotero. Then any item(s) desired for download may be checked (see Figure 4.23), and the user can click *OK*.
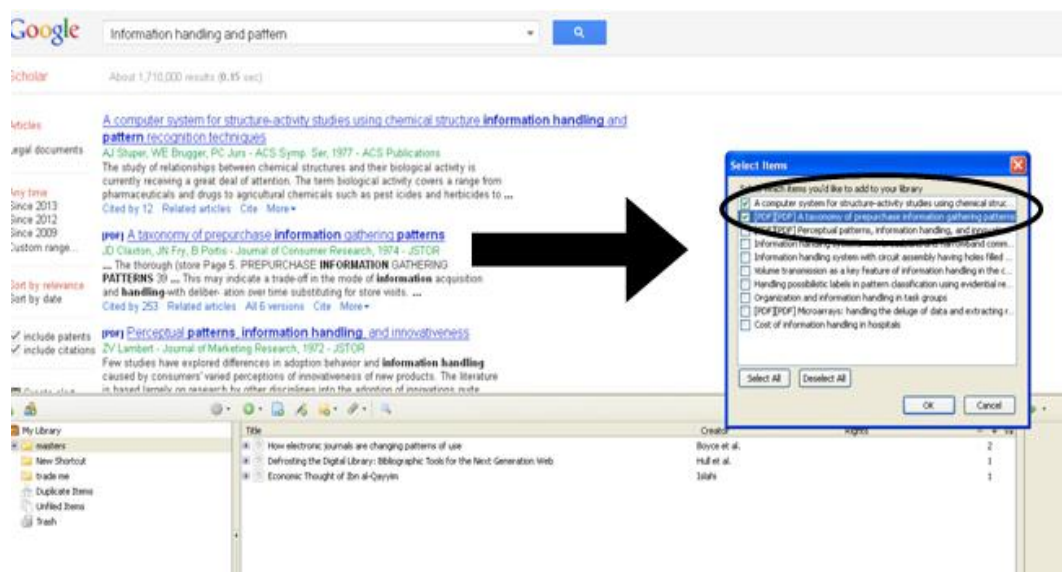
Figure 4.23: Second step for searching and adding sources

B.2 The item and metadata are then downloaded and saved in the Zotero library, and Zotero will automatically recognise the types of items—for instance, article, webpage, or book (see Figure 4.24).
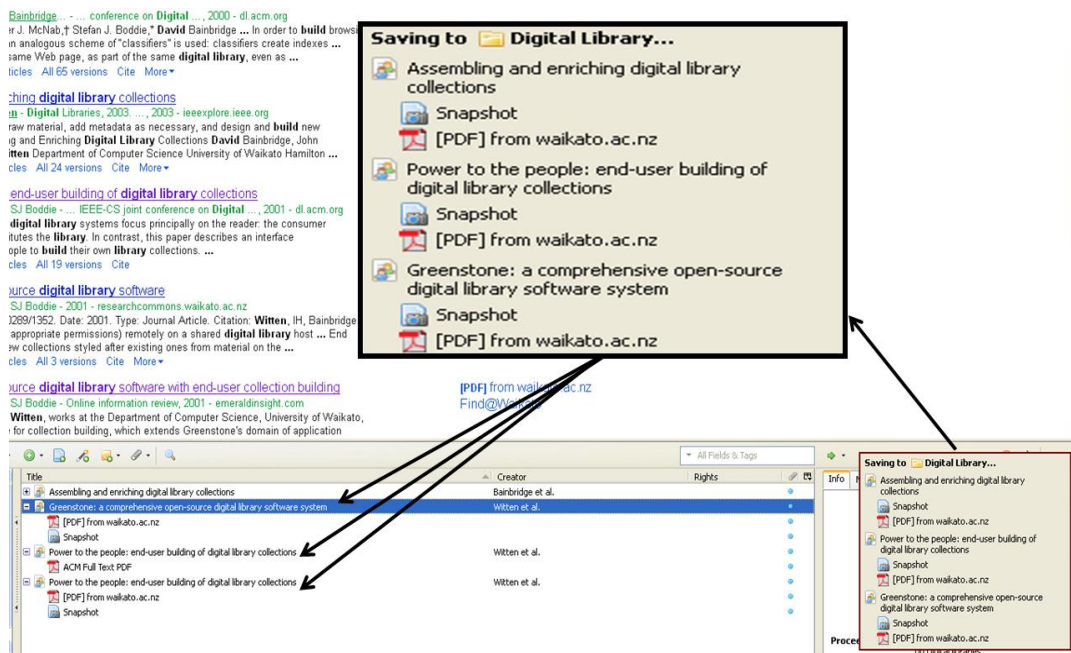
Figure 4.24: Pop-up message showing the items saved

4.3.2   Exporting a Bibtex File from Zotero

Items may be exported from Zotero as a Bibtex file in five steps:

1- The items desired for export are selected (Shift + Left-click; see Figure 4.25).
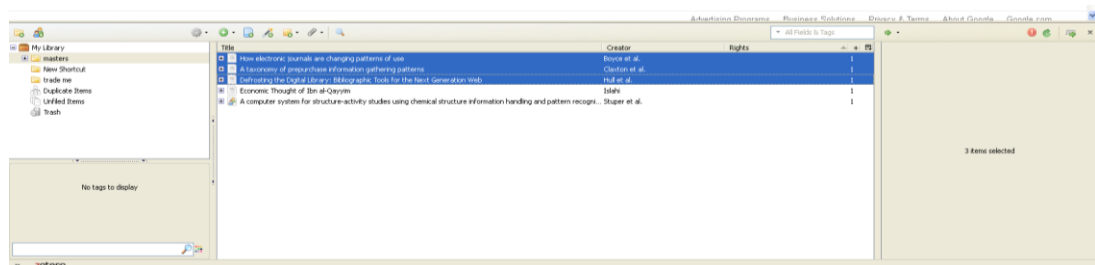
Figure 4.25: Select the sources for exporting

2- The user right-clicks on the items and then goes to *Export Selected Items* (see Figure 4.26).
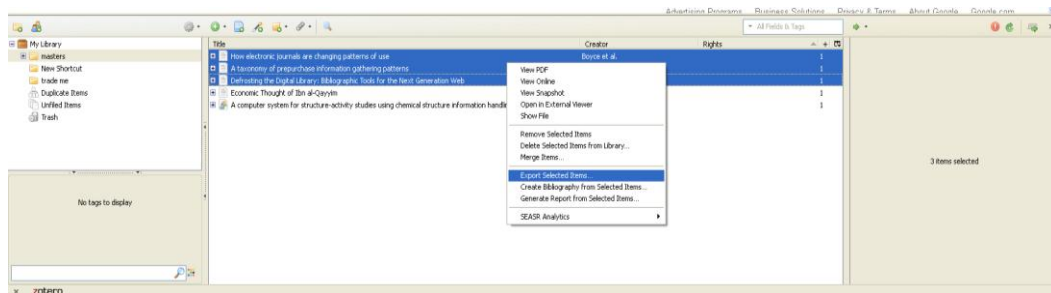
Figure 4.26: Select the export action

3- Users choose the format they want (here, Bibtex format; see Figure 4.27).
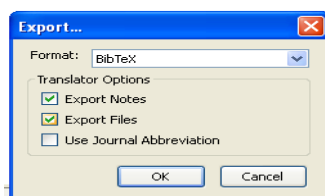
Figure 4.27: Formatting the export action

4- Next, they choose the folder in which to save the library (see Figure 4.28).
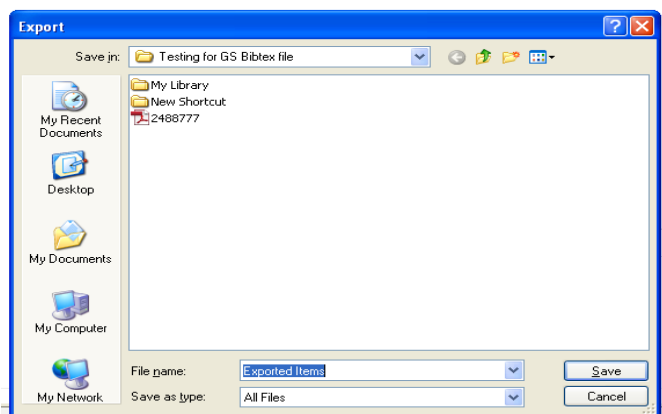
Figure 4.28: Specify where the exported items will be saved

5- The files will be exported to a folder called 'Exported Items'. In that folder there are Bibtex files and a folder containing all the files related to the exported metadata (see Figure 4.29).

Figure 4.29: The sources and metadata exported

4.3.3 Greenstone Extension

Greenstone has a folder called 'Metadata' which contains sets of metadata. When users create a collection, they can choose the metadata they want from any available sets (the available sets are shown in Figure 4.30).

Development Library Subset Example Metadata (dls)
Dublin Core Metadata Element Set, Version 1.1: Reference Description (dc)
New Zealand Government Locator Service Metadata Standard version 2.1 (nzgls)
Qualified Dublin Core Metadata Element Set, Version 1.1: Reference Description (dc)

Figure 4.30: Metadata sets provided by Greenstone

Each one of these sets has an abbreviation called a 'namespace' that helps the users to recognise its attributes while they enrich the metadata fields. Figure 4.31 shows an example of (dc) metadata.

Element	Value
dc.Publisher	
dc.Contributor	
dc.Date	
dc.Resource Type	
dc.Format	
dc.Resource Identifier	

Figure 4.31: An example of a dc.set

Other metadata sets are automatically assigned to the collection, such as *Exploded Metadata Set (exp)* and *Extracted Greenstone Metadata 1.1 (ex). Exp* is assigned if there is a file like Bibtex which, because it contains group metadata records, is

73

extracted out (see Section 4.3.4, step 3) and *exp* metadata is assigned when any item is imported to the collection.

Therefore, because Greenstone allows users to create their own metadata sets, a set can be established that is called 'Tracking Material'. It contains one element: the reading phase. This will help users to distinguish between the material they have read, are currently reading, or are planning to read. This set can be established in seven steps, as follows:

1. In the beginning, the user goes to *Start > All Programs > Greenstone-3.05 > Greenstone Editor for Metadata sets (GEMS)* (see Figure 4.32).
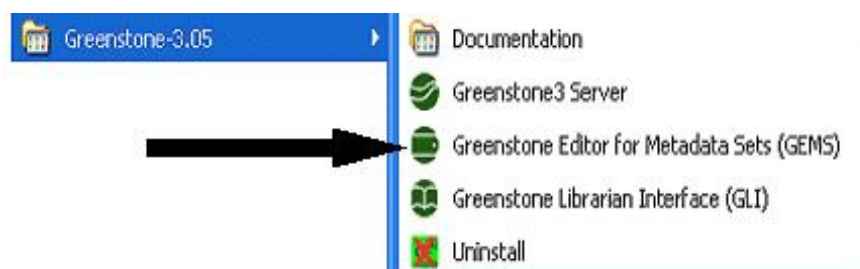
Figure 4.32: GEMS in the Greenstone package

2. As Figure 4.33 shows, a main window comes up. To create a new set, the user goes to *File > New*.
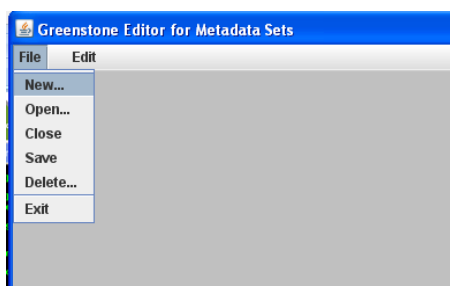
Figure 4.33: GEMS window

3. A *New Metadata Set* window pops up; the user fills the fields title and namespace, and then clicks OK (here, 'Tracking Material' was chosen for the title and 'TM' for the namespace; see Figure 4.34).
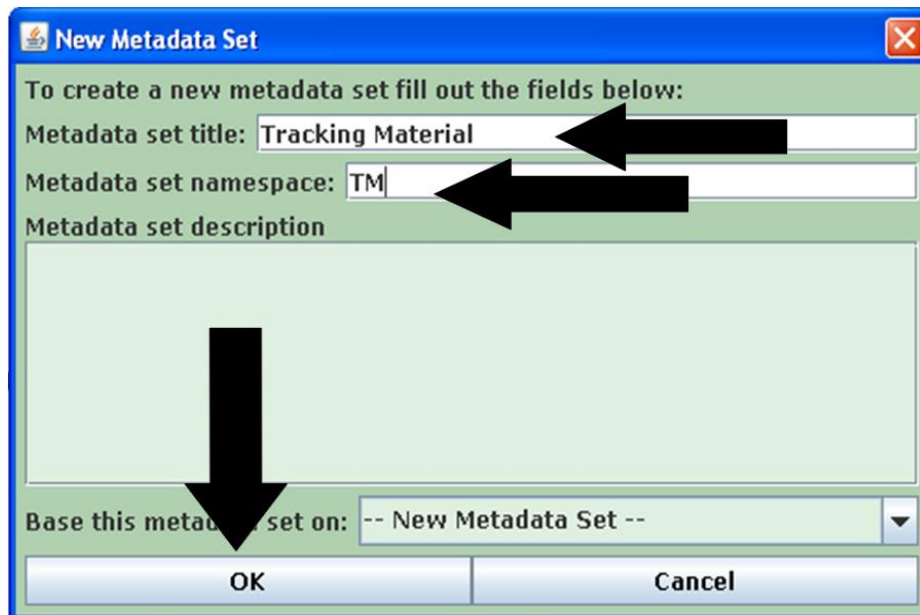
Figure 4.34: Creating a new metadata set

4. Next, the user adds the element ('reading phase') and right-clicks on the title of the set, choosing *Add Element* (see Figure 4.35).
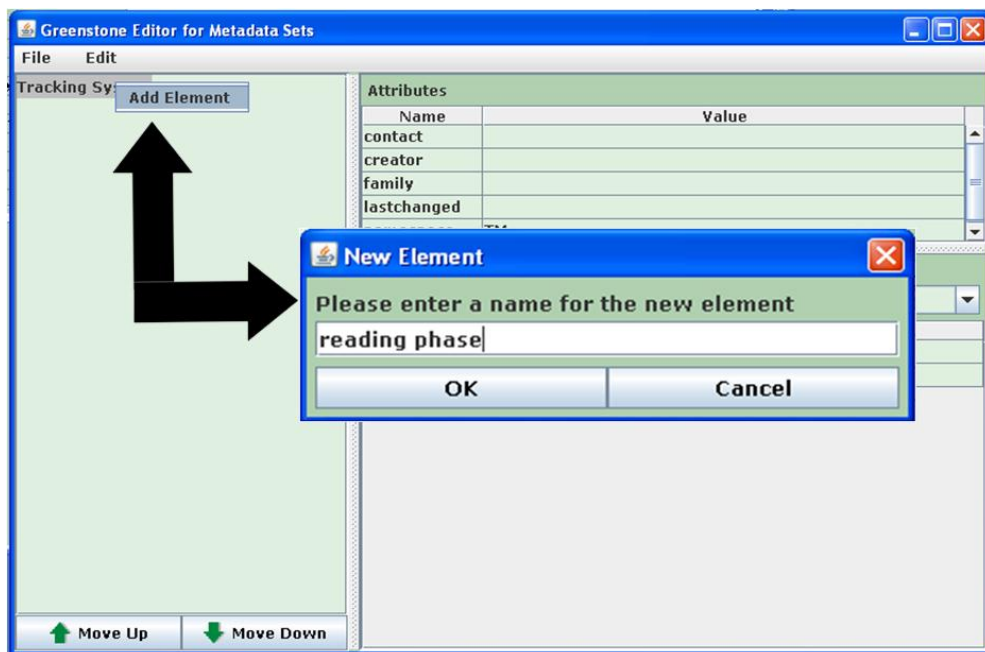
Figure 4.35: Adding element to metadata set

5. When the *New Element* window pops up, the user writes the name of the element (here, *reading phase*) and then clicks OK (see Figure 4.35).

6. Finally, the user saves the metadata set by going to *File > Save* (see Figure 4.36).
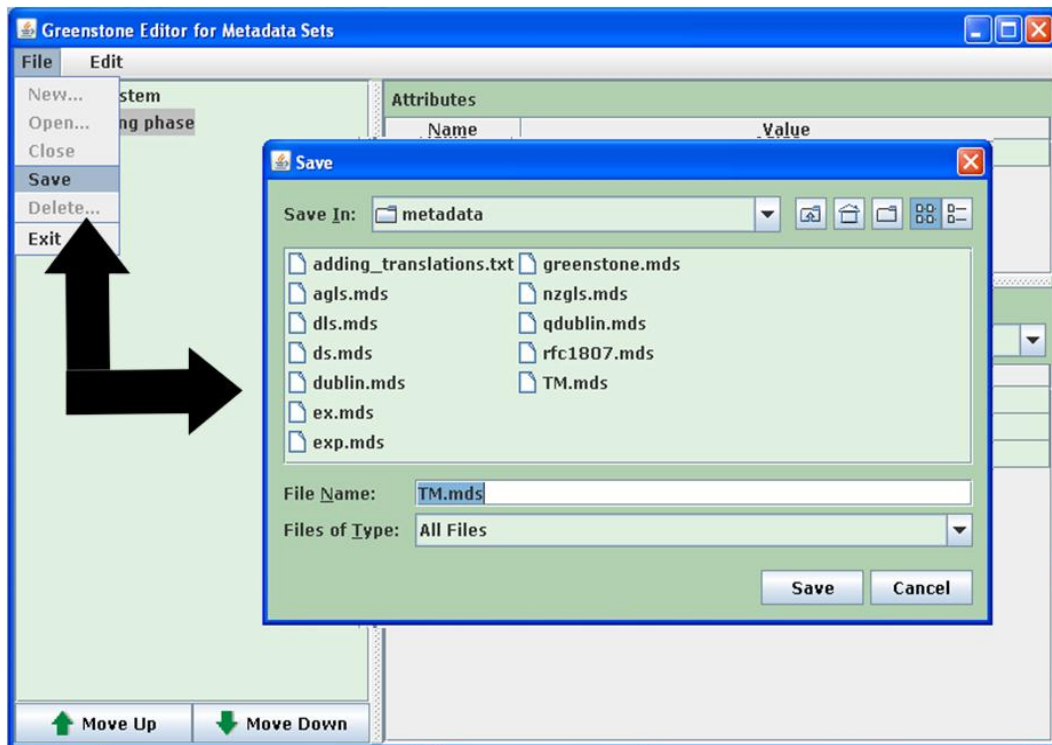
Figure 4.36: Save the metadata set in Greenstone's metadata folder

7. The set is saved in the metadata folder which contains all the Greenstone metadata sets (see Figure 4.36).

4.3.4   Importing a Bibtex File and Designing The Collection

This section explains the process of building a library in Greenstone by using the Bibtex file that was exported from Zotero (see Section 4.3.2). First, it should be noted that Greenstone 3.05 was installed on a Windows platform (see Figure 4.37). In the beginning, the user goes to *Start > All Programs > Greenstone-3.05 > Greenstone Librarian Interface (GLI)*.
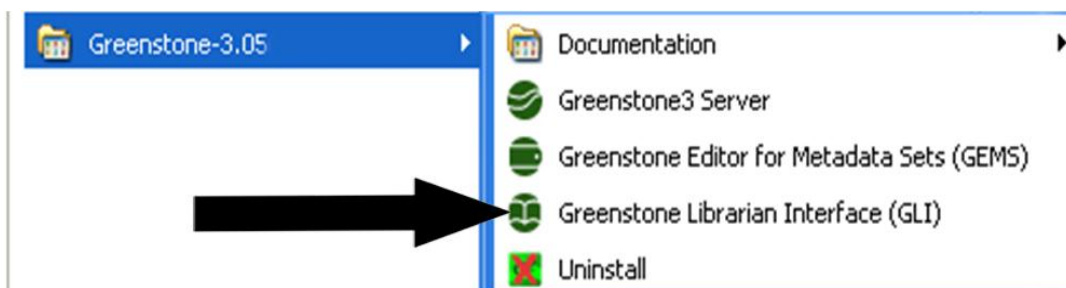
Figure 4.37: Package of Greenstone

1. To create a new collection the user follows *File > New* (see Figure 4.38). Then, he or she writes the collection title in the window that pops up (see Figure 4.38).
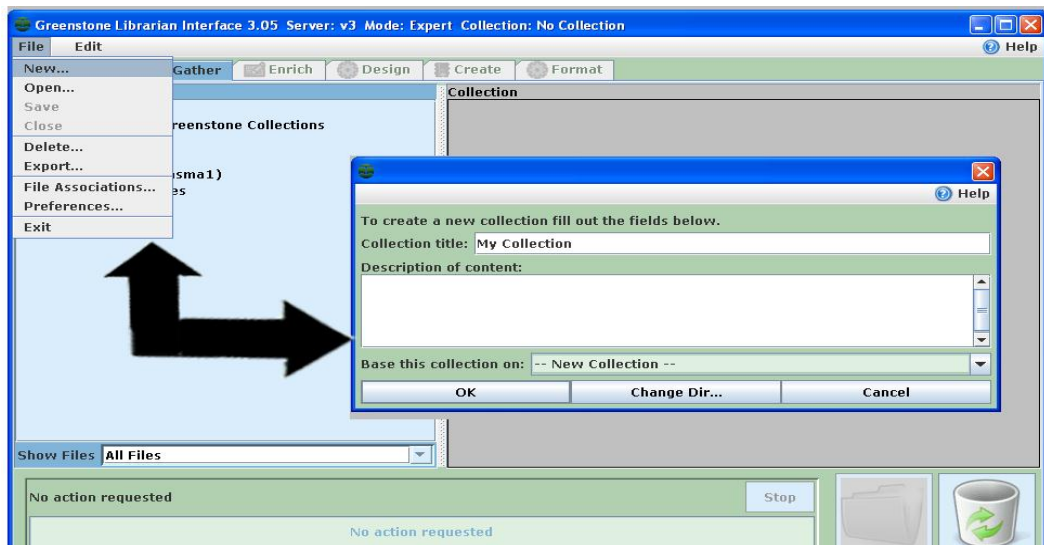
Figure 4.38: Creating a new collection

2. The user imports the Bibtex file ('Exported Items') to the empty library by dragging and dropping from the left to the right panel (see Figure 4.39).
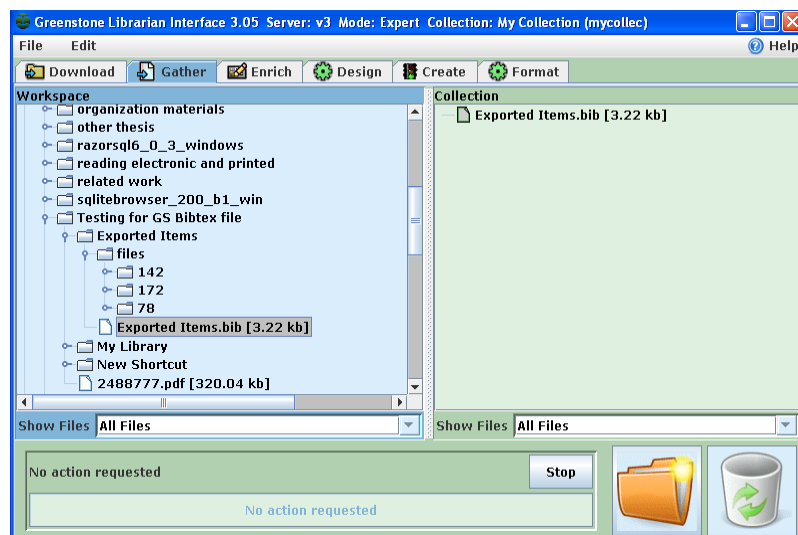
Figure 4.39: Importing the Bibtex file to My Collection

3. To access each source individually, the user explodes the file and extracts the entire metadata database. Next, the user can right-click on the Bibtex file and then *Explode* the metadata database. When the window pops up, the user clicks on *Explode* (see Figure 4.40).
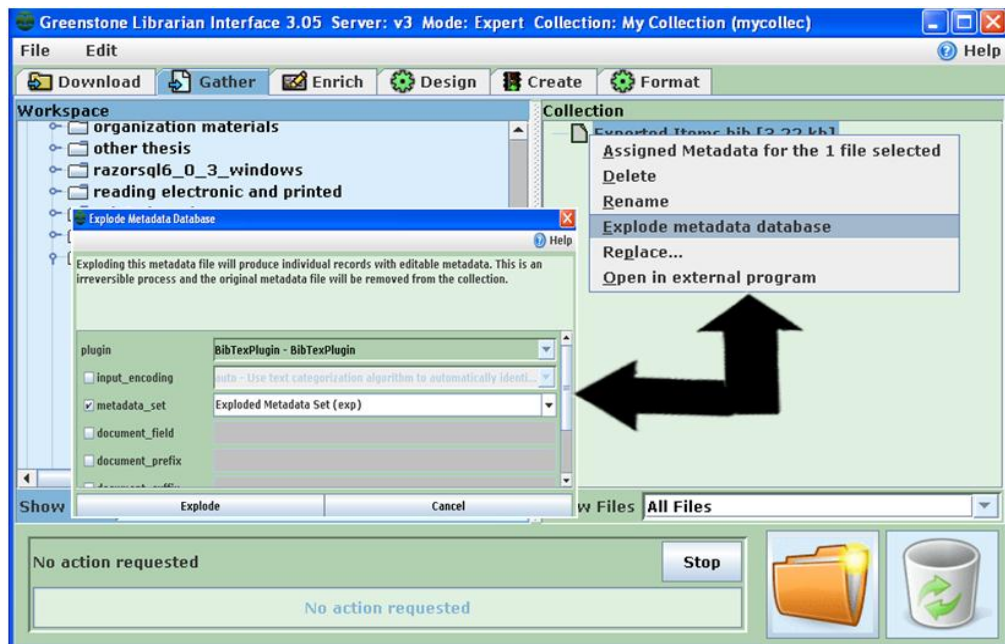
Figure 4.40: Extracting all sources from the Bibtex file

When the metadata is exploded, an XML file and the sources are obtained (see Figure 4.41).
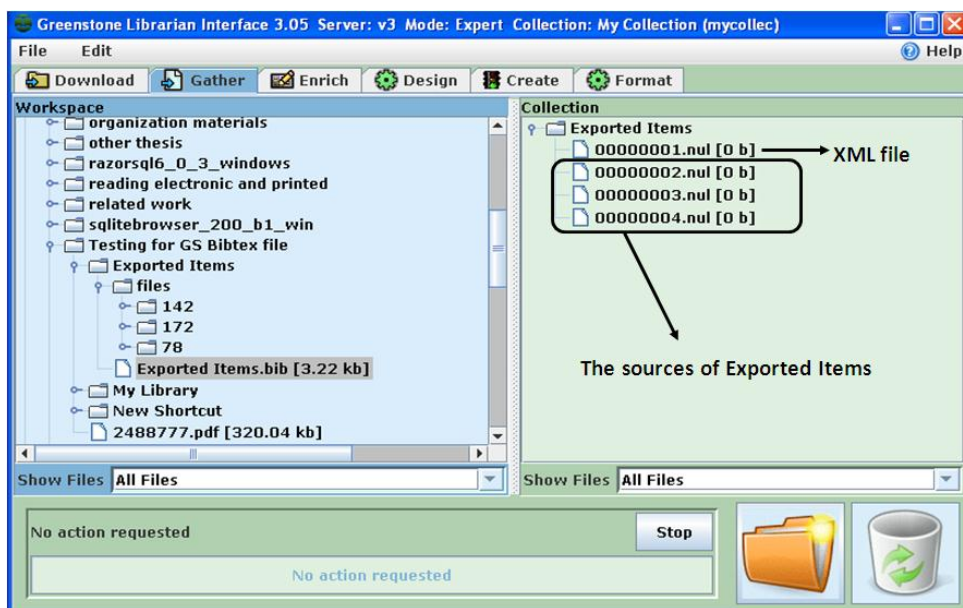
Figure 4.41: The sources after being extracted

4. Next, a PDF file can be assigned for each metadata by right-clicking on the item and choosing *Replace;* the metadata will then be merged with the file chosen (see Figure 4.42).
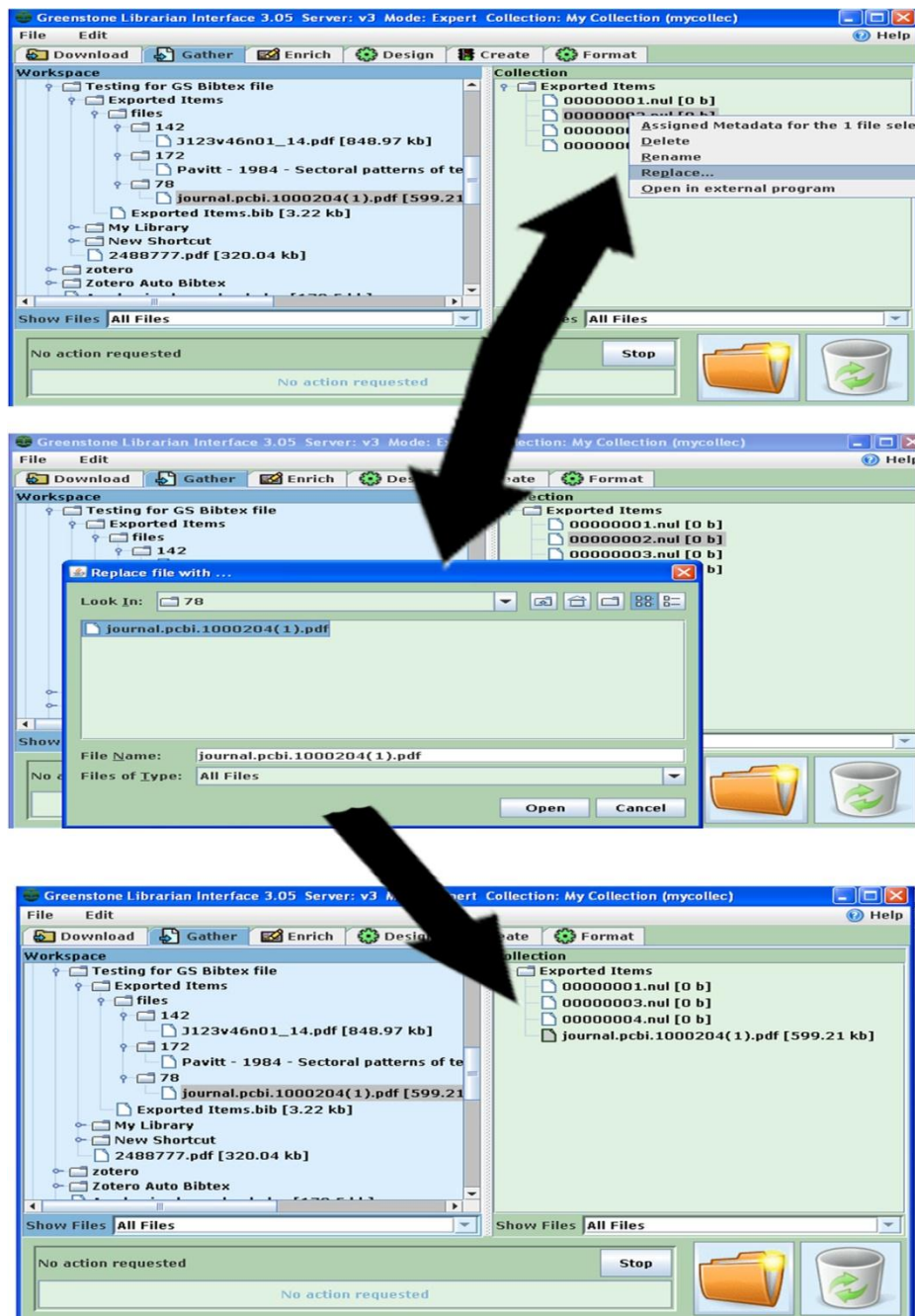
Figure 4.42: Assigning a file for each metadata exported

5. The user repeats step 4 for the other metadata (see Figure 4.43).
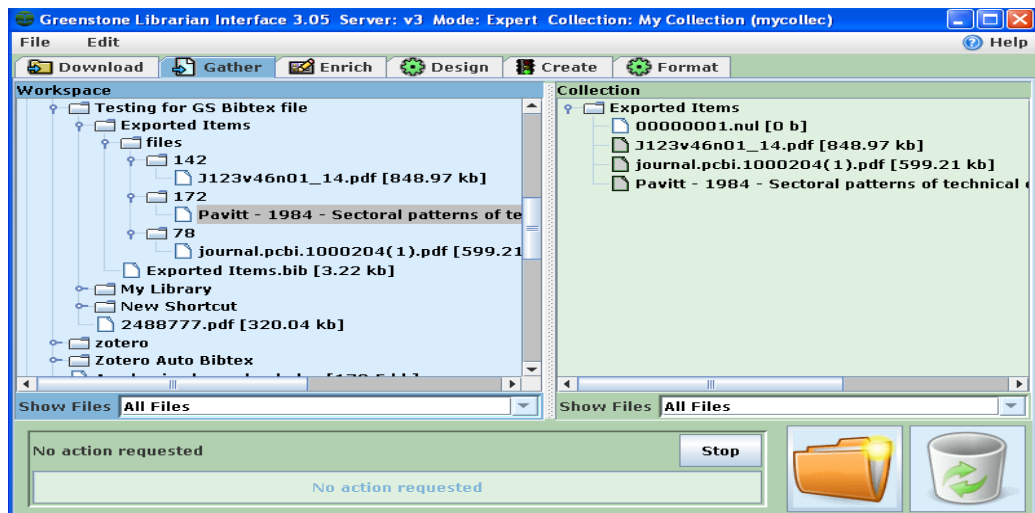
Figure 4.43: All the metadata assigned to files

6. The next step is enriching the metadata. The metadata has already been obtained for each item (see Figure 4.44). However, the 'Tracking Material' metadata set which we created in Section 4.3.3 must be added and enriched.

7. To add the TM set, the user goes to *Manage Metadata Sets,* which is located on the lower left-hand corner of the Greenstone Librarian Interface (see Figure 4.44).
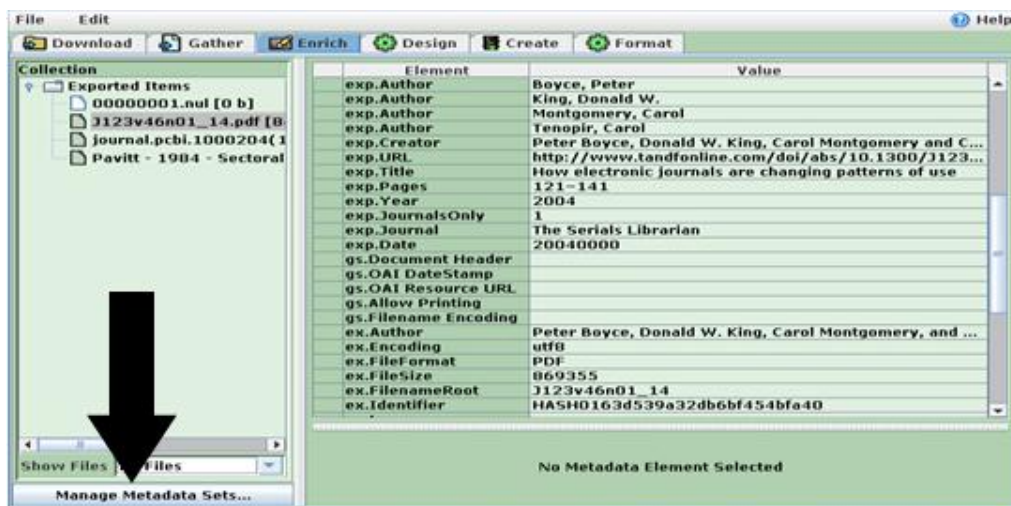
Figure 4.44: Enrich tab, managing and enriching metadata

8. A window pops up showing all the metadata assigned to the collection (see Figure 4.45, step 1). The user clicks on *Add* to see all the available metadata sets, chooses 'Tracking Material' (TM), and then clicks *Add*. TM metadata is

assigned to the collection (see Figure 4.45, step 2). After ensuring that TM is added, the user clicks on *Close* (see Figure 4.45, step 3).
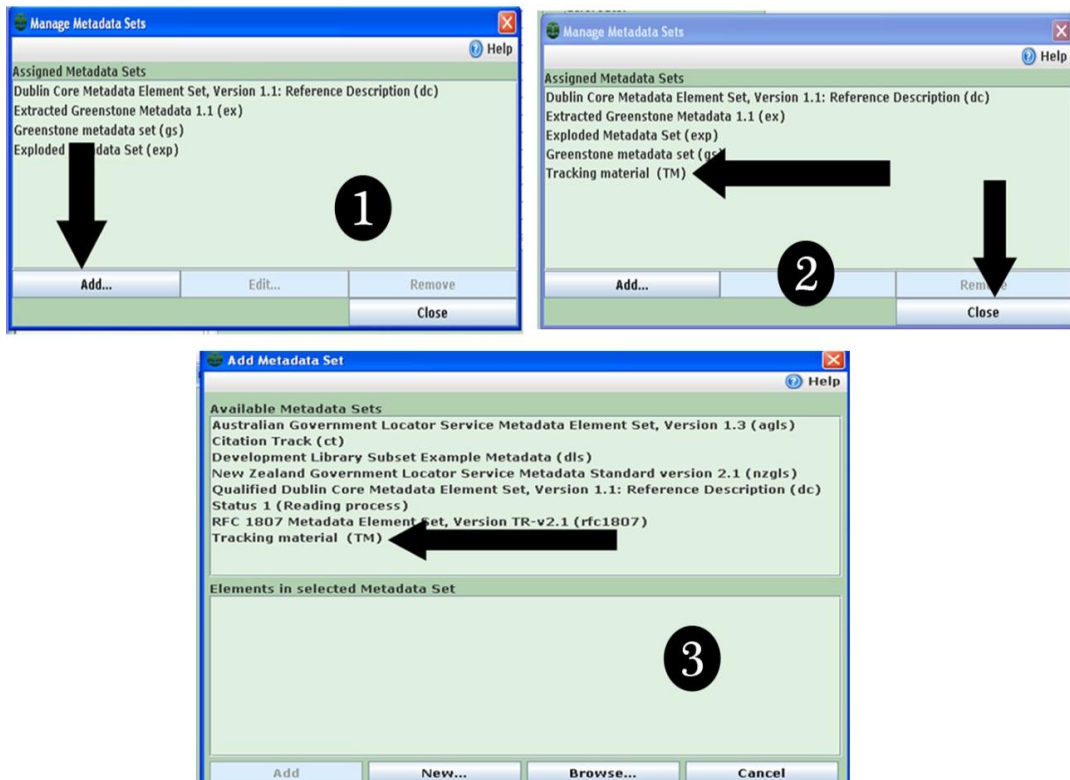
Figure 4.45: Adding the TM metadata set

9. Next, the user can specify the reading phase (have read, currently reading, or to read) of each item in the collection (see Figure 4.46).
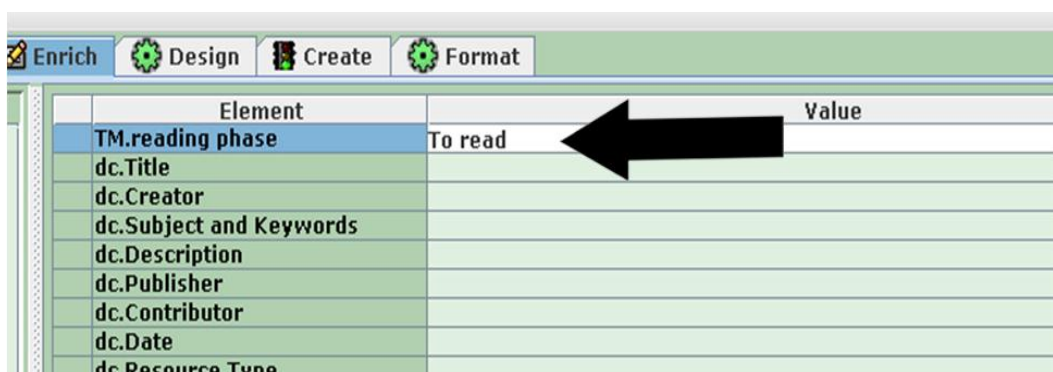
Figure 4.46: Adding values to the added metadata set

10. The second-to-last step is the library design. It has four parts: document plug-ins, search indexes, partition indexes, and browsing classifiers.

10.1   The document plug-ins are to help users browse the materials they have added to the collection. Greenstone can identify the type of each item added and what kind of plug-in it needs. For instance, if the item is a Word document, it needs the Word plug-in (see Section 4.1.2). In this example, Bibtex and PDF plugins are needed, which are already assigned (see Figure 4.47).
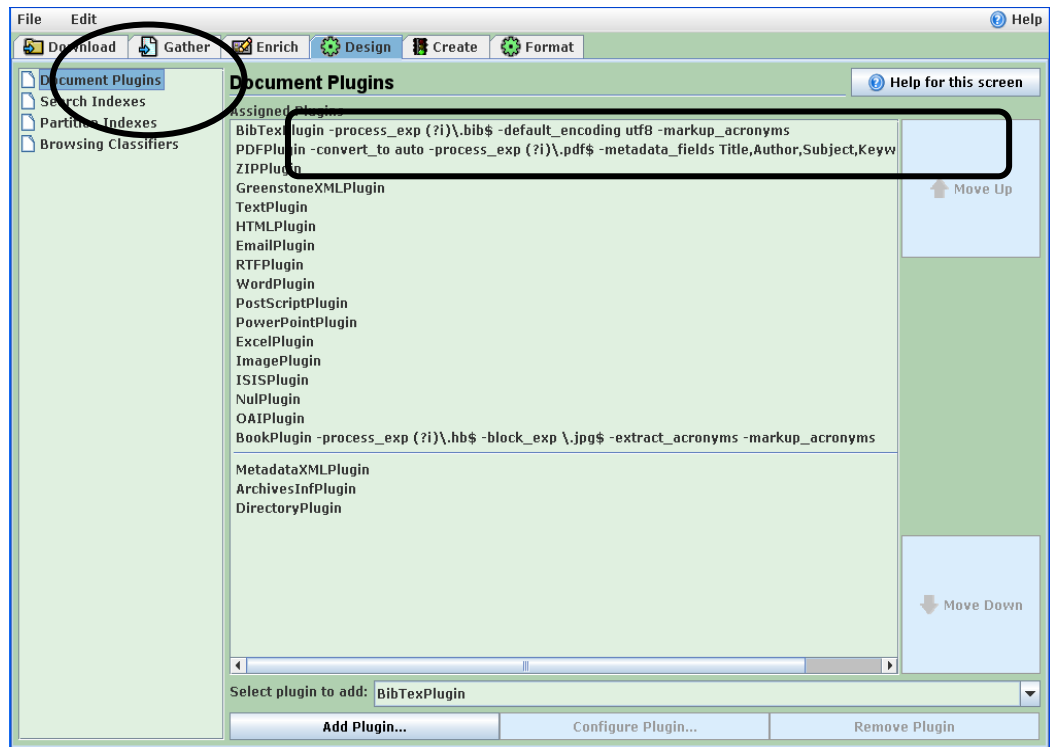
Figure 4.47: Adding document plug-ins

10.2   For the search indexes step, the search text is the default index (see Figure 4.48). Add another index for all the metadata exploded by clicking on *New Index*.
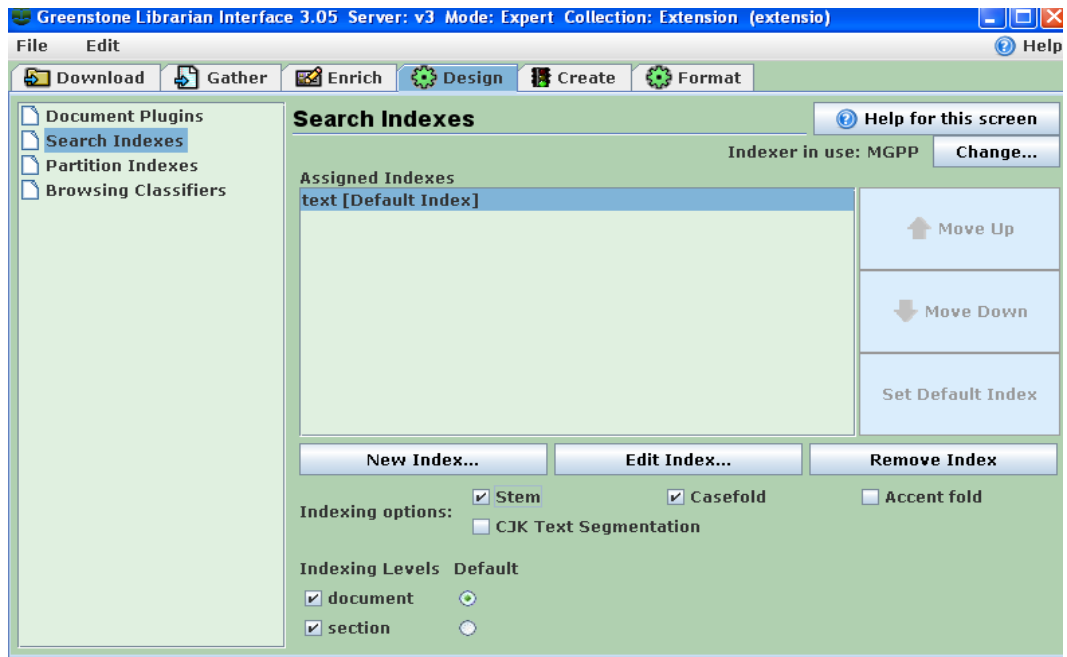
Figure 4.48: Search indexes

When a window pops up, the user checks all the exploded metadata, 'Track Materials' metadata, and full text, and then clicks on *Add Index* (see Figure 4.49; the metadata start with *exp.*).
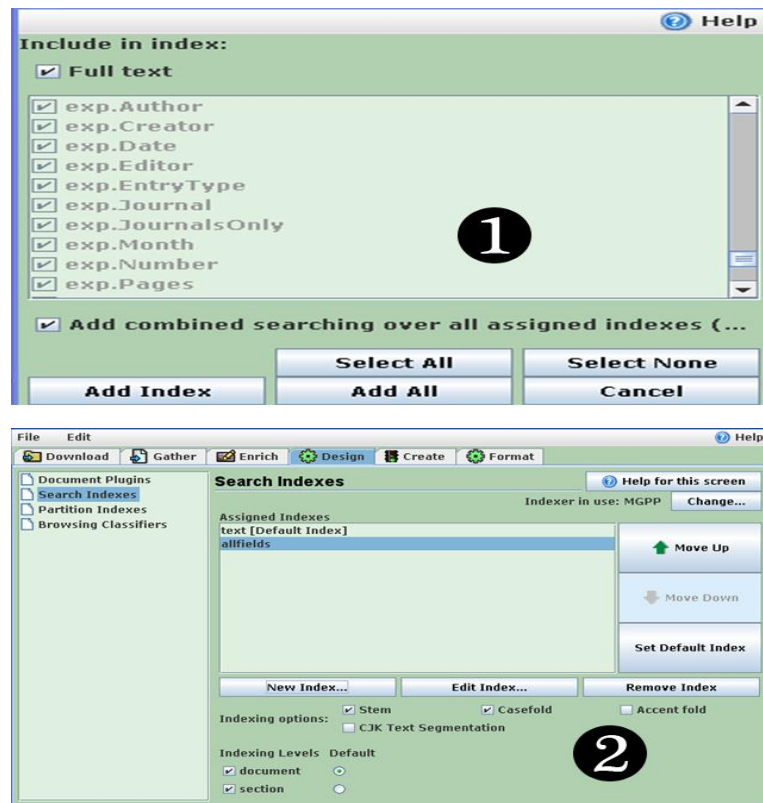
Figure 4.49: Assigning the elements in the search index

10.3 Partition indexes: because dividing the document into sections is not required in this system, this part is not needed.

10.4 For browsing classifiers, the items can be categorised into groups based on the metadata for browsing them. The user divides the items into two groups: title and reading phase.

10.4.1 The user selects the classifier type (*List* was chosen; see Figure 4.50), and then clicks on *Add Classifier*.

Figure 4.50: List classifier

10.4.2 From the window that pops up, the user chooses the metadata and the partition types (see Figure 4.51) and then clicks *OK* (*exp.Title* was chosen, which is the title from the exploded metadata set; see Figure 4.51).
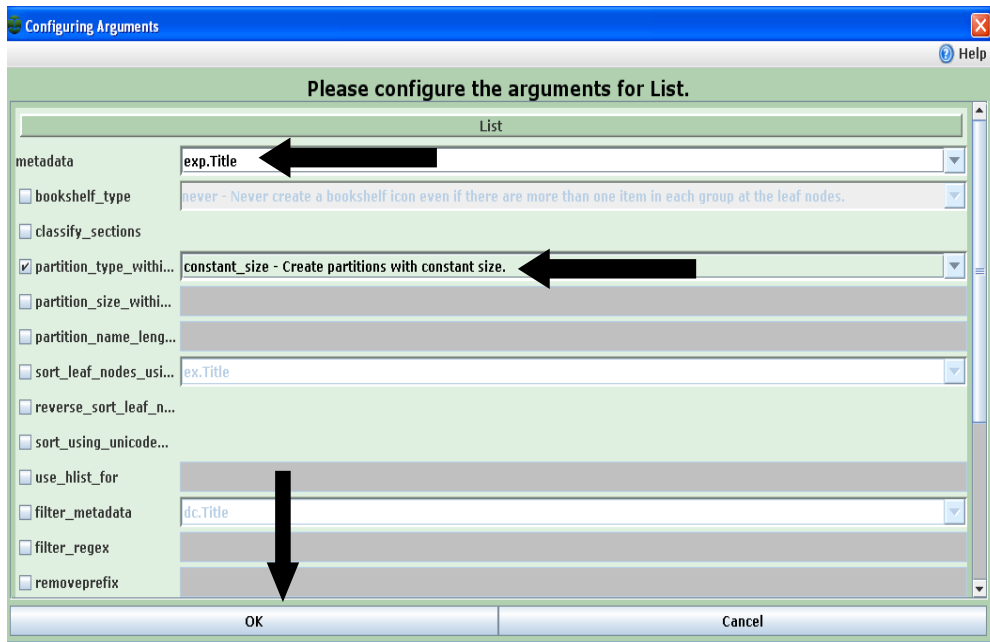
Figure 4.51: Adding a list classifier for *exp.Title*

10.4.3 For the *TM.reading phase*, the user selects the classifier (*Hierarchy* was chosen; see Figure 4.52) and then clicks on *Add Classifier*.
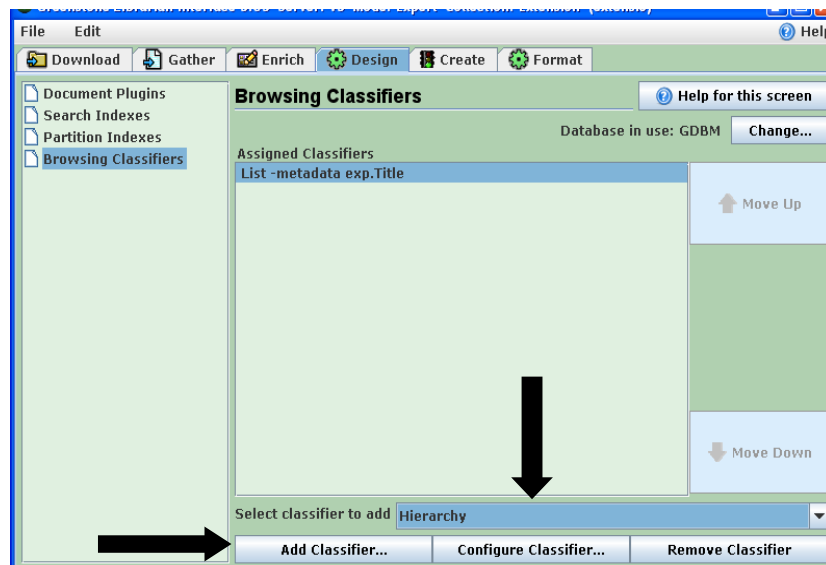
Figure 4.52: Hierarchy classifier

10.4.4 From the window that pops up, the user chooses the metadata and then clicks *OK* (here, *TM.reading phase* is chosen; see Figure 4.53).
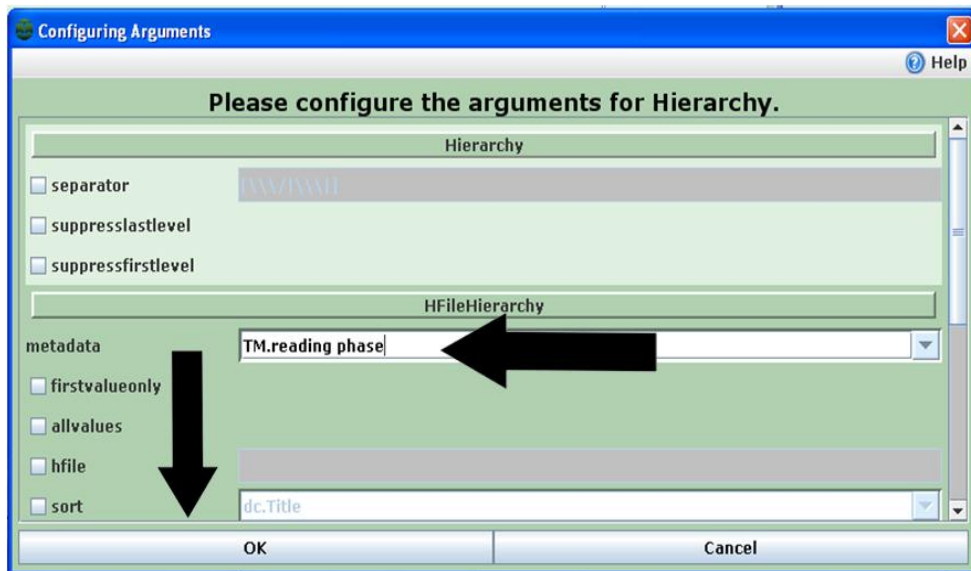
Figure 4.53: Adding a hierarchy classifier for TM.reading phase

11. Building the library is the last step (see Figure 4.54).
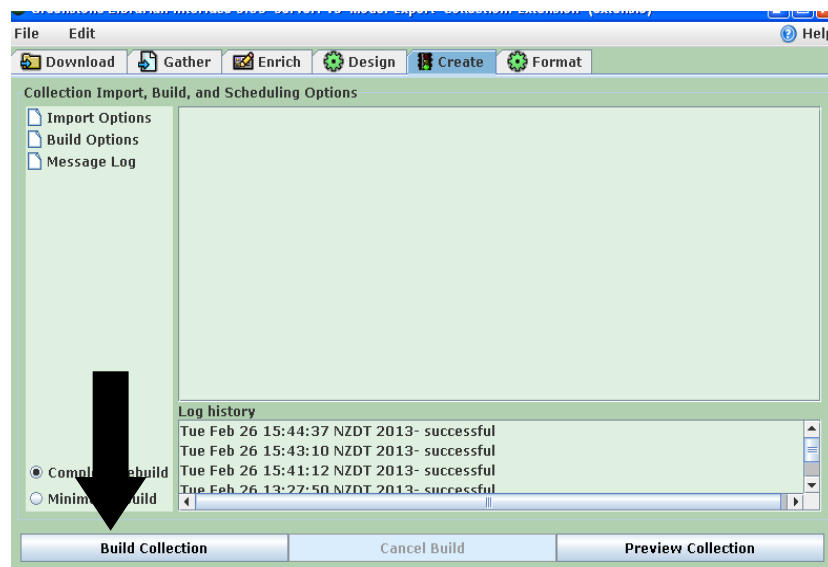
Figure 4.54: Building the library

12. When the message shows that the library has been built, the user clicks on *Preview Collection* (see Figure 4.55).
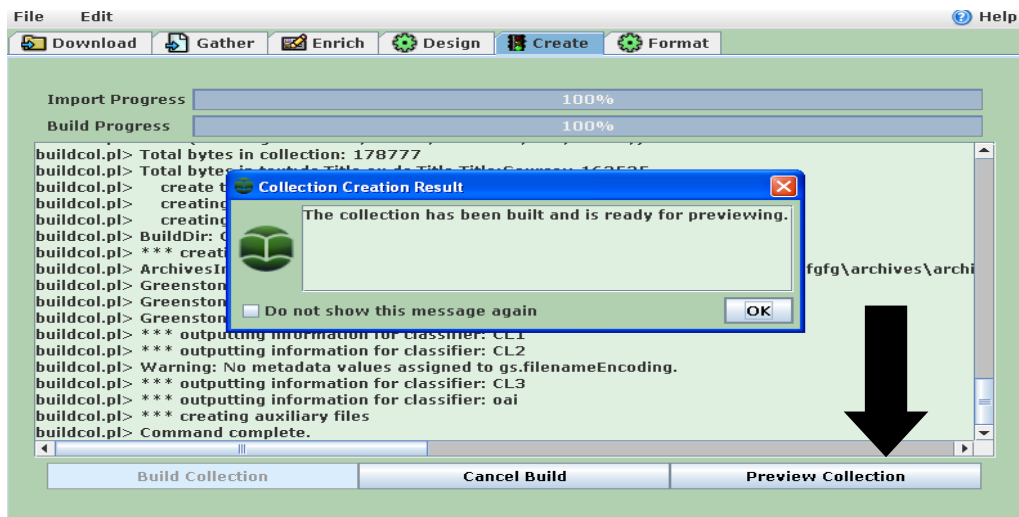
Figure 4.55: The message shows that the collection is ready for browsing

4.3.5 **Browsing the Library**

After the library has been set up, browsing the collection is next, the second part of the Greenstone Digital library. The last step in Section 4.3.4 was *Preview Collection*. The collection displays on a web page (see Figure 4.56).
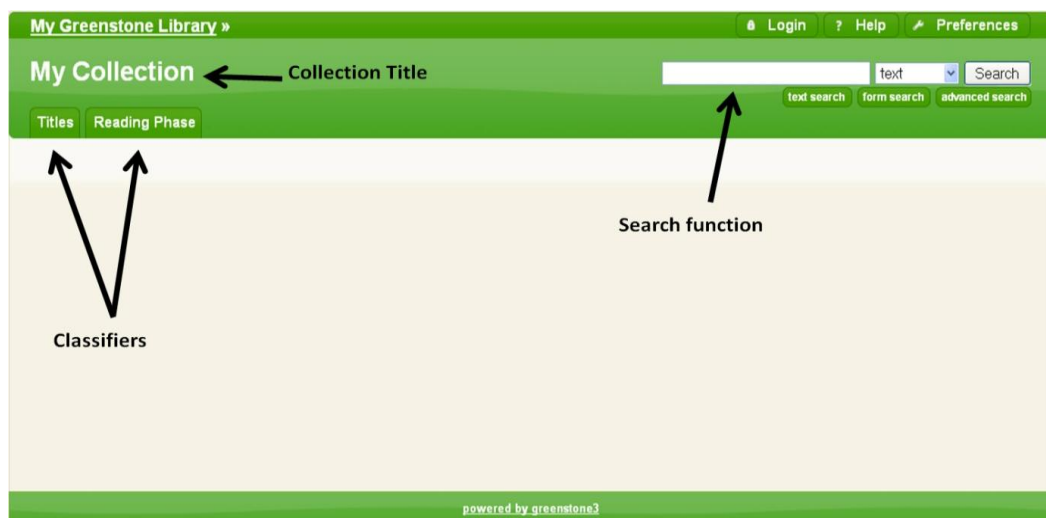
Figure 4.56: Main window of the reader's interface

- To list all the items by their titles, users click on Titles (see Figure 4.57). Either a PDF reader or Greenstone may be used to view any item (by means of the plug-in added in Section 4.3.4, step 10.1, Document Plugins; see Figure 4.58).
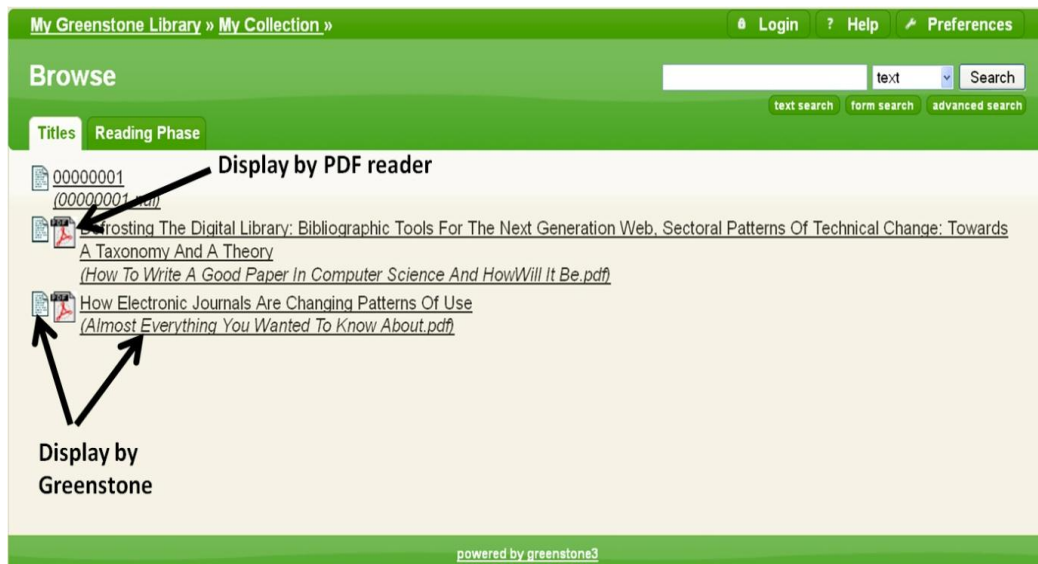
Figure 4.57: Display all the items by their titles

Figure 4.58 shows an item displayed with a PDF reader and Greenstone.
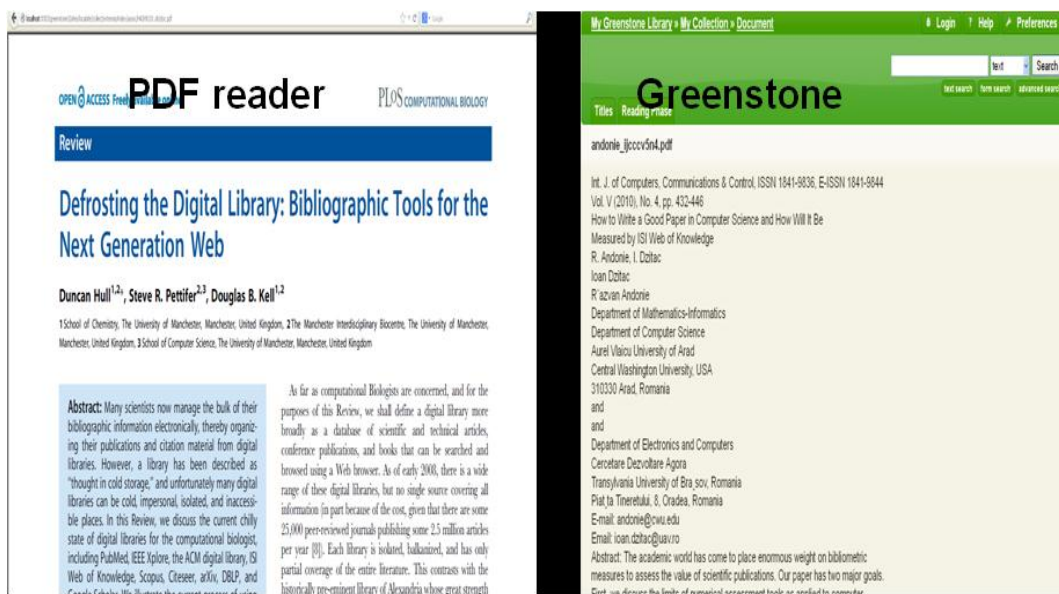
Figure 4.58: Example of one item displayed by both the PDF reader and Greenstone

- To browse the items by reading phase, click on 'reading phase', which is the second browsing classifier (as explained in Section 4.3.4, step 10.4). The items are grouped based on their TM.reading phase metadata (see Section 4.3.4, step 9, and Figure 4.59).
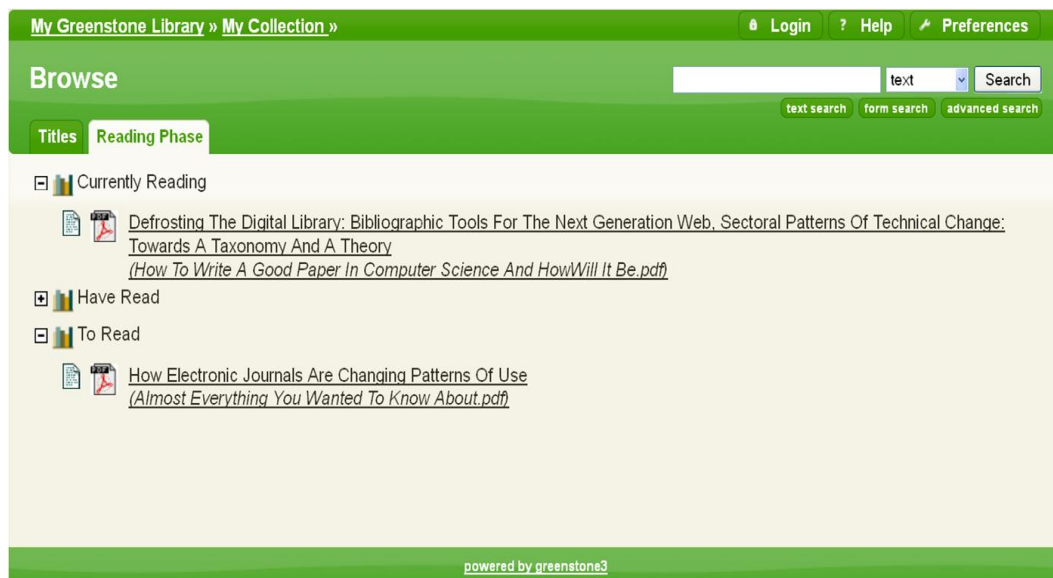
Figure 4.59: Displaying the items by their reading phases

4.3.6 **Shifting the Items among the 'Shelves'**

This section shows how the items can be moved from one phase, or shelf, to another. For example, when users are reading items they have not finished yet, they will be placed on the 'Currently reading' shelf, but if users have finished reading an item, it should be placed on the 'Have read' shelf. These things can be done by using the metadata. To shift items, users employ the following steps:

1. Log in as an admin (the user name and password should be assigned during the installation of Greenstone; see Figure 4.60).
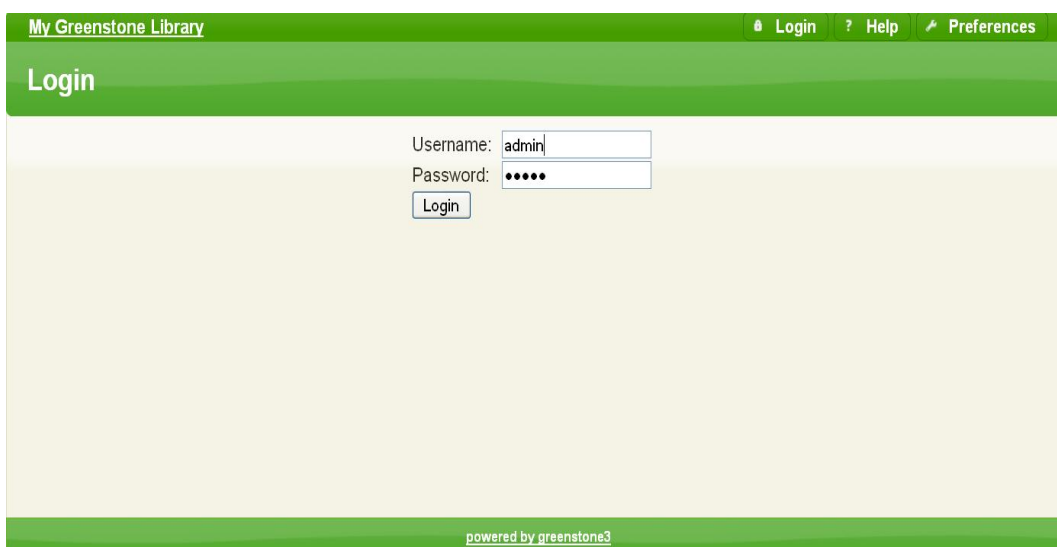
Figure 4.60: Admin login window

2. They choose the item they wish to move (the second item is chosen; see Figure 4.61).
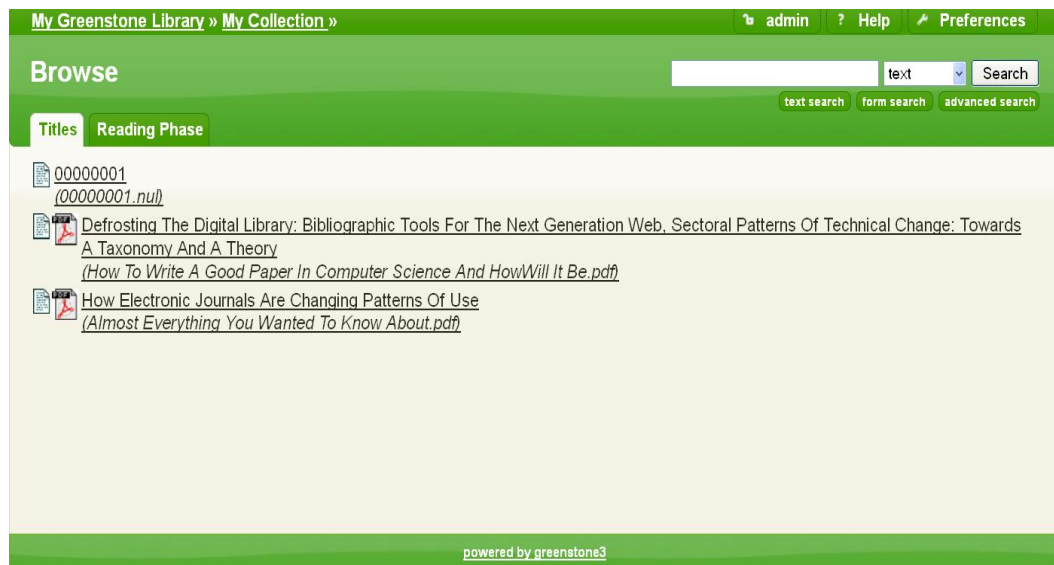
Figure 4.61: The window after the user has logged in as an admin

3. The item is viewed through Greenstone, and the user then clicks *Edit Content* (see Figure 4.62).
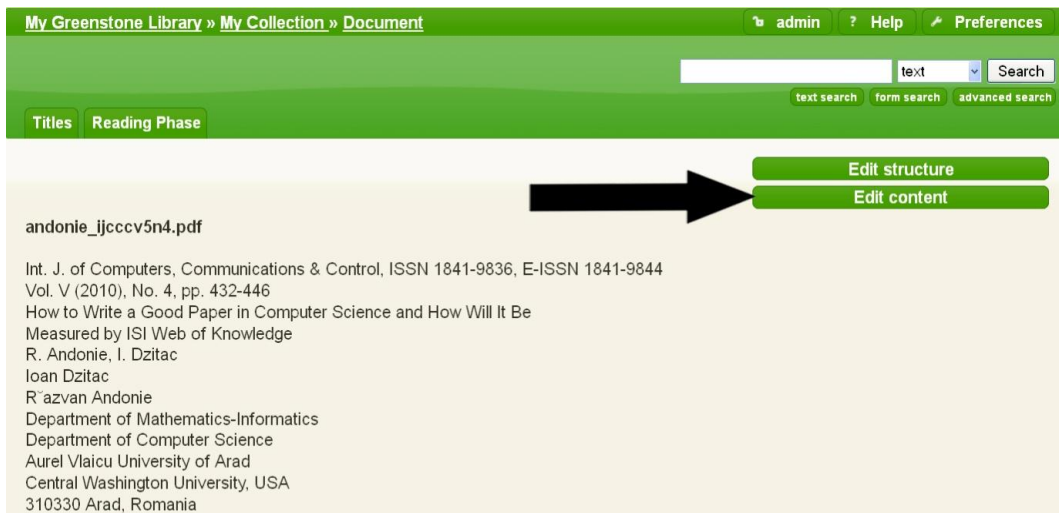
Figure 4.62: Browsing items through Greenstone

4. The user clicks on *Edit Metadata* (see Figure 4.63).

Figure 4.63: Edit the metadata

5. The metadata will be shown (see Figure 4.64).

Figure 4.64: Displaying all the metadata

6. The reading phase is modified to 'Currently reading' or 'Have read' (see Figure 4.65).

Figure 4.65: Modifying the reading phase

7. Finally, the user clicks on *Save Changes* (see Figure 4.66). The collection will be rebuilt so that the item will be placed on the chosen shelf.
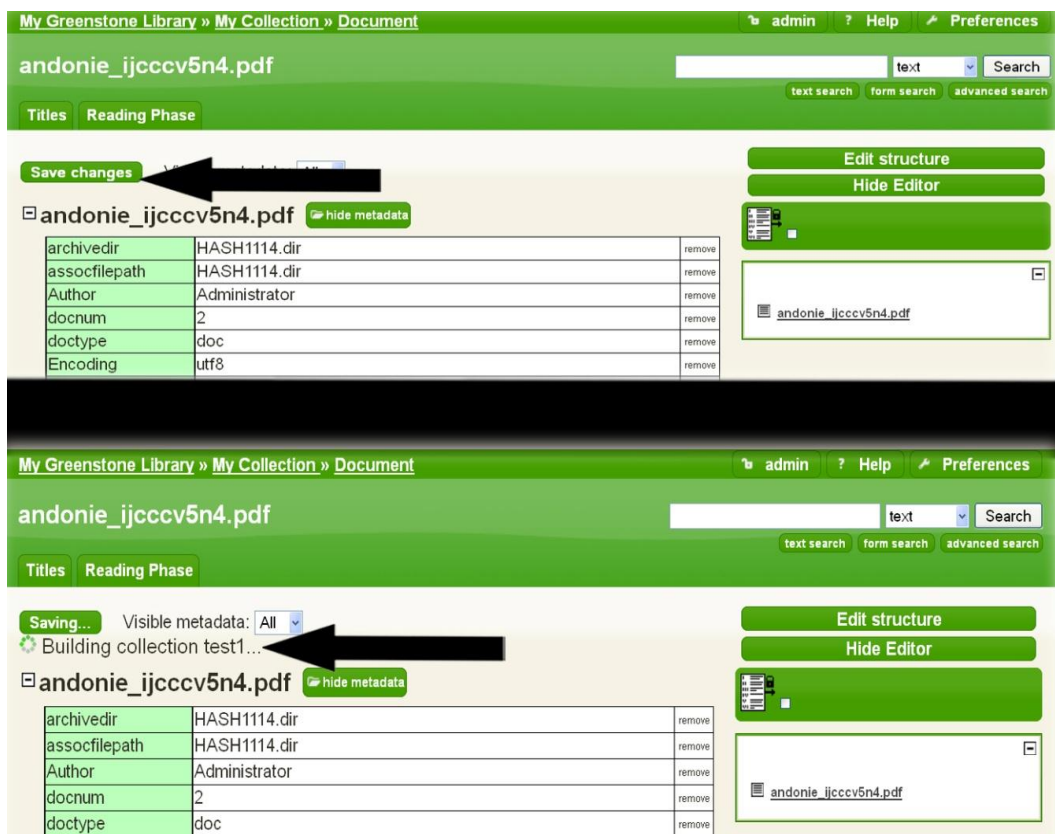
Figure 4.66: Saving and building the collection

4.4 System Design

Subsequent to the explanation of how the system works manually, this section provides an overview of the entire process, from obtaining the items to browsing. Figure 4.67 illustrates briefly the steps of the system (TRMS) as described below:
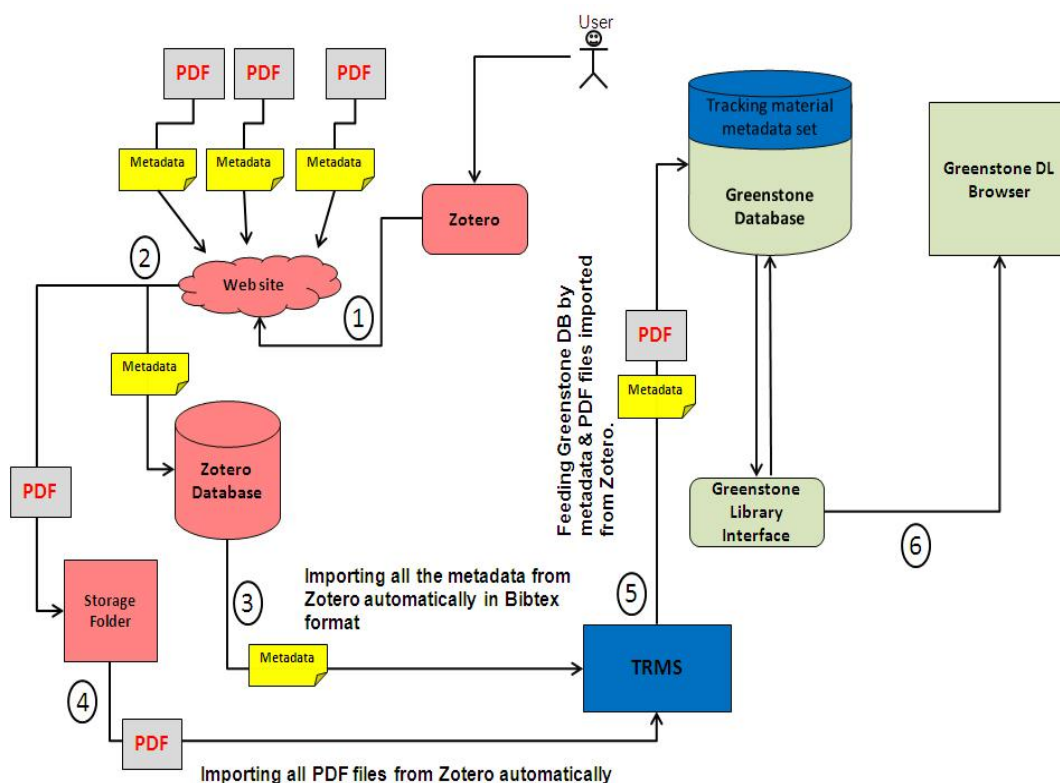
Figure 4.67: The initial architecture of TRMS

1   Using the Zotero extension Firefox tool, the user adds records from websites (equivalent to steps 1-3 in Section 4.3.1).

2   Zotero stores the metadata of the records added (type of source, title, author...), and if there are any files like PDFs or Web pages, Zotero stores them in a folder called 'Storage Folder' (equivalent to steps 1-3 in Section 4.3.1).

3   TRMS imports these metadata automatically in Bibtex format (equivalent to steps 1–4 in Section 4.3.2) from Zotero.

4   TRMS imports the metadata files automatically from Zotero.

5   TRMS exports the metadata and PDFs to Greenstone database.

6   Greenstone builds the library by organising the metadata and its files, which will be shown in the browser.

4.5 Summary

This chapter details the process of system design in three stages: system decision, exploration of functionality and the architecture of the system. In the first stage, two software packages, Zotero and Greenstone, were chosen with the possibility of extending one of them and making it congruent to our requirements. This software was chosen for two main reasons: both are freely available and they are open source code. Each one is reviewed in detail to show how each one works, as well as to examine the functions that support the requirements of our system.

Although Greenstone seems to cover five out of the six requirements (R2, R3, R4, R5, and R6) for the system, it does not support R1 as efficiently as Zotero does. In Greenstone, items that have already been downloaded to the computer can be imported manually. However, Zotero downloads items and their metadata automatically from such websites as Google Scholar and ACM. This leads us to combine the two systems so that they can work together. Zotero can act as a bibliographic and document provider, and Greenstone as a digital library organiser.

The second stage is to demonstrate how this concept works, showing that it consists of three steps. In the first step, the metadata and the documents are imported into the Zotero database and then exported as Bibtex files. In the second step, the Greenstone database is extended through the creation of a metadata set which contains a reading status element. The final step is to import the Bibtex file and the exported documents into Greenstone, and then design and build the library. Afterwards, the library will have an index for the reading status consisting of three shelves: 'Have read', 'To read', and 'Currently reading'.

The last stage is to combine the three steps in the second stage into one system called the Tracking Reading Materials System (TRMS). The TRMS will transfer the data from Zotero into Greenstone automatically and will be able to capture the information about reading status.

Chapter 5

5   Implementation

This chapter explains how Zotero and Greenstone are combined and extended in our system (TRMS). First, Section 5.1 introduces the general design of the system by summarising all of the steps in the TRMS architecture. Section 5.2 investigates how the collection is displayed in Greenstone and where Zotero stores the data. Section 5.3 describes our method for obtaining data from Zotero and organizing it according to Greenstone standards. In addition, we highlight the problems encountered during each step of the process. Section 5.4 explains the limitations that might affect the processing of the system. The chapter concludes with a summary of the TRMS process.

5.1   Introduction

This section provides an overview of the entire process, which ranges from obtaining the items to browsing them in a digital library. The system has several components (see Figure 5.1). Papers located on websites have two components, i.e., PDFs and metadata, which are indicated by grey and yellow, respectively, in Figure 5.1. When Zotero accesses websites, it downloads and stores the metadata in its database and the PDFs in its storage folder (red in Figure 5.1). The components of the Greenstone digital library are indicated in green. The TRMS components (TRMS Bridge, metadata set, and tracking material in the web browser) were developed in this project and are indicated by blue. Before explaining the TRMS process, we provide a brief description of the eight elements used by our software (TRMS), as follows (these elements are present in the architecture shown in Figure 5.1).

1- Storage folder in Zotero: this folder contains the PDF files in Zotero (see Section 5.2.2, *storage* folder).

2- TRMS folder: this folder receives the copied pdfs, and it needs to be created and named as TRMS by the user.

3- Xmlfile: the location of the XML file. This file holds all of the metadata imported from Zotero. This file should be located with the Greenstone collection in the TRMS system.

4- Extension file: the file types transferred from Zotero to Greenstone. The type is set to PDF by default but it can be changed.

5- Zotero database: the user needs to specify/assign Zotero database where the Zotero collection is located (see Section 5.2.2).

6- Copied Zotero database location: the location where the Zotero database will be copied. This can be anywhere but the TRMS folder is the best location.

7- Greenstone collection location: a location from which Greenstone can display files (usually located in the Greenstone installation: "Greenstone3/web/sites/localsite /collect/<collection's name> / import").

8- Bash file for building the GS library: this file (build.sh) must be installed with the TRMS software and placed in the TRMS folder. It contains the commands required to build the Greenstone collection automatically.

Figure 5.1 is an extension of the architecture shown in Figure 4.67 in Chapter 4 and it provides a brief illustration of the steps used by the proposed system (TRMS), as follows.
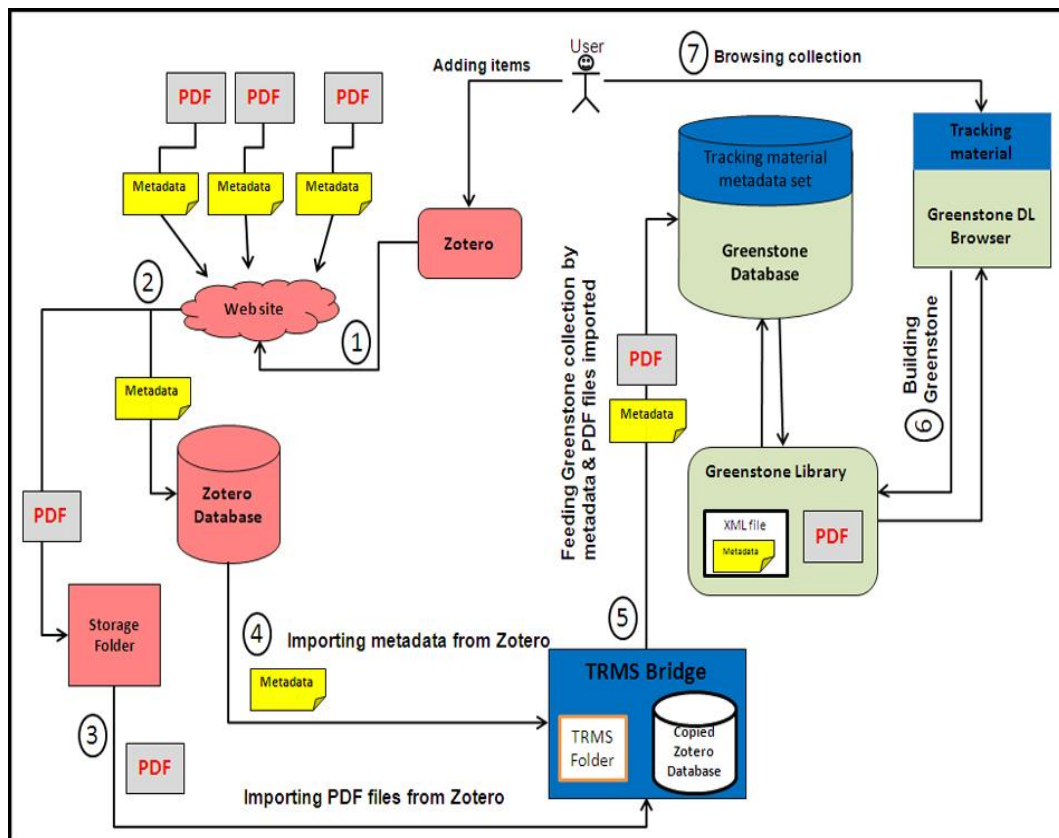
Figure 5.1: Details of the TRMS architecture.

1   A user adds records from websites with the Zotero Firefox extension tool.

2   Zotero stores the metadata of the added records (type of source, title, author...), while files such as PDFs or web pages are in a folder called "Storage Folder".

3   TRMS Bridge imports the PDF files automatically from Zotero.

4   TRMS Bridge imports the metadata automatically from Zotero.

5   TRMS Bridge exports the metadata and PDFs to the Greenstone database.

6   TRMS Bridge builds and activates the library that is available to the user.

7   A user moves items from the reading phase to another phase using the Greenstone reader interface.

There are seven steps in the proposed system. The first and second steps are performed by Zotero and the sixth step is performed via Greenstone, although the TRMS Bridge automates the building process. Thus, we focus on the third, fourth, fifth and sixth (building process) steps during the transfer of PDFs and metadata to Greenstone. In Section 5.2, we explain how the data are stored in Zotero and Greenstone. In Section 5.3, we introduce the TRMS process that transfers the data from Zotero to Greenstone. Finally, in step 7, we show how the user can change the metadata of items to move them to the reading phase.

5.2   Data storage in existing software elements

Before describing the implementation of steps 3–6, we need to understand how Greenstone displays the collection (see Section 5.2.1) and how Zotero stores the metadata and PDFs (see Section 5.2.2). This allows us to find ways of obtaining items from Zotero and preparing them to conform to Greenstone's standards.

5.2.1   How Greenstone stores items and metadata

According to Witten, Bainbridge and Nichols, all of the metadata of a collection are stored in an XML file called metadata.xml. Greenstone reads the metadata from this file, which is located in a folder called ***Import*** (this is usually located in the Greenstone installation; refer to the seventh element in Section 5.1). In addition, documents such as PDFs or txt should be placed in the ***Import*** folder with the XML file that needs to be displayed (Witten, Bainbridge and Nichols 2009).

Figure 5.2 shows an example of an XML file, which contains three metadata items. The file name element is the key that indicates the file to which the

97

metadata applies. In the example, the metadata of item 1 applies to a file called "The power of human.pdf" and the metadata of item 2 applies to a file called "001.nul" (see Figure 5.2). The metadata will be displayed when these files are placed with the XML file (Witten, Bainbridge and Nichols 2009).

Figure 5.2: Example showing the XML file format.

This investigation helped us to recognize where metadata and PDFs should be transferred and organised in Greenstone. It also helped us to determine how the metadata should be structured in the XML file that allows the Greenstone Reader's Interface to display the collection correctly.

5.2.2 Location where Zotero stores the metadata and PDFs

In general, the Zotero collection is stored in a folder called Zotero. This folder is located either in the Firefox profile (when using the Zotero Firefox extension) or in the Zotero profile (when using Zotero Standalone). If the Zotero Standalone and Firefox extension are both installed, the user will be asked to select one of these sources as the storage collection for adding records or they can use them individually (Figure 5.3) (Zotero_data [Zotero Documentation] n.d.).

98
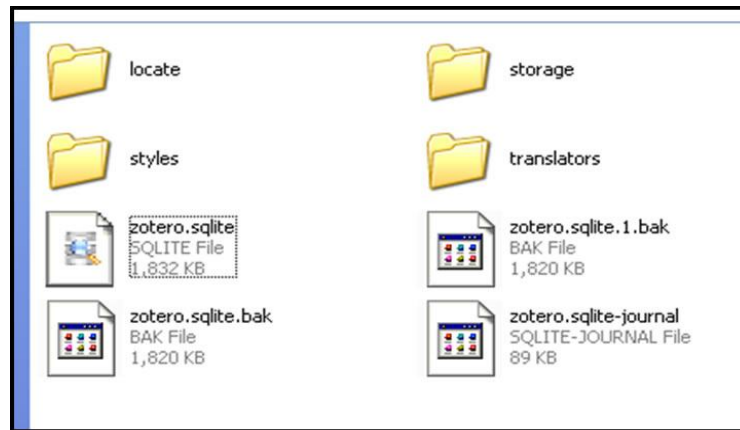
Figure 5.3: Zotero collection.

Two main file types from a Zotero collection are required by our system: metadata and PDF files. According to the Zotero documentation, all of the metadata for the collection are stored in a database called **Zotero.sqlite** while all the attachments, such as PDFs and web pages, are stored in the folder called **storage** (Zotero_data [Zotero Documentation] n.d.)**.**

After identifying the locations of the metadata and PDF files, we analyse the database to determine the required fields and how they can be extracted. Zotero uses SQLite, which is a small portable database system (Jeon, Bang, Lee 2012; Bi 2009).

To analyse the Zotero database, we installed SQLite Manager[15] in the Mozilla add-ons. Figure 5.4 shows the tables in the Zotero database.
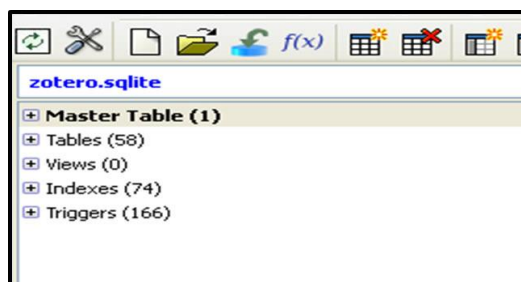
Figure 5.4: Zotero database content.

The database has one master table, 58 tables, 74 indexes and 166 triggers. We identified seven tables that contained data relevant to TRMS, because they refer directly to items and their metadata (i.e., tables items, itemAttachments, itemCreators, creatorData, itemDataValues, itemData and fields), while the other

---

[15] https://addons.mozilla.org/en-us/firefox/addon/sqlite-manager/

tables store extra information, such as user events, synchronisation and folder organisation. In these seven tables, we consider eight fields (seven were exported from Zotero as a Bibtex file in Section 4.3.2, along with the file name we need to merge the file with its metadata; see Section 5.2.1), which contain the information that needs to be displayed in Greenstone (file name, author (first and last name), URL, title, year, journal list, volume and date) (Figure 5.5).
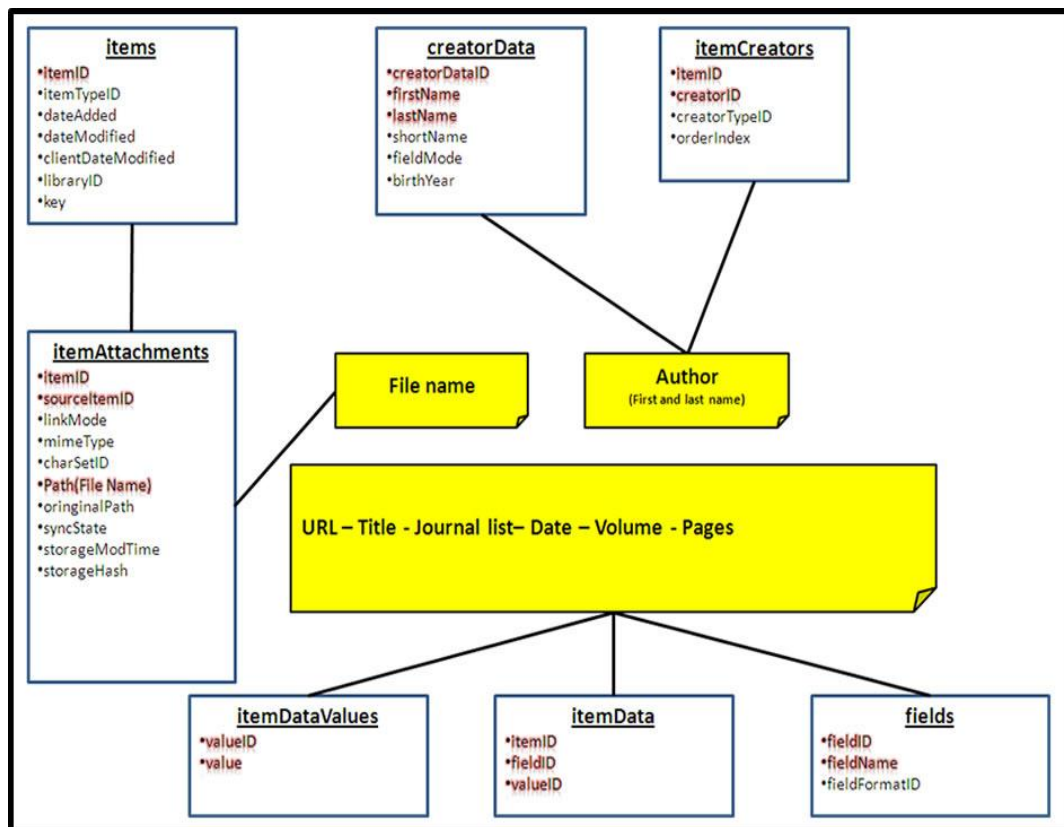
Figure 5.5: Fields retrieved from the Zotero database and their tables.

Figure 5.5 shows the tables and entities used to retrieve the required fields, where the entities are highlighted in red. This figure shows that the fields are distributed in three parts, according to their retrieval source, as follows. 1) The file name is retrieved from the table itemAttachments. 2) The author is retrieved from two tables: itemCreators and creatorData. 3) The other fields are retrieved from three tables: itemDataValues, itemData and fields.

Figure 5.6 shows the relationships among the tables and the keys that connect the tables. Three different queries are used to obtain the eight fields: (1) a complex query to obtain the file name field, (2) a simple query for the author field and (3) a simple query is used for each other field (the queries are clarified in Section 5.3.3).
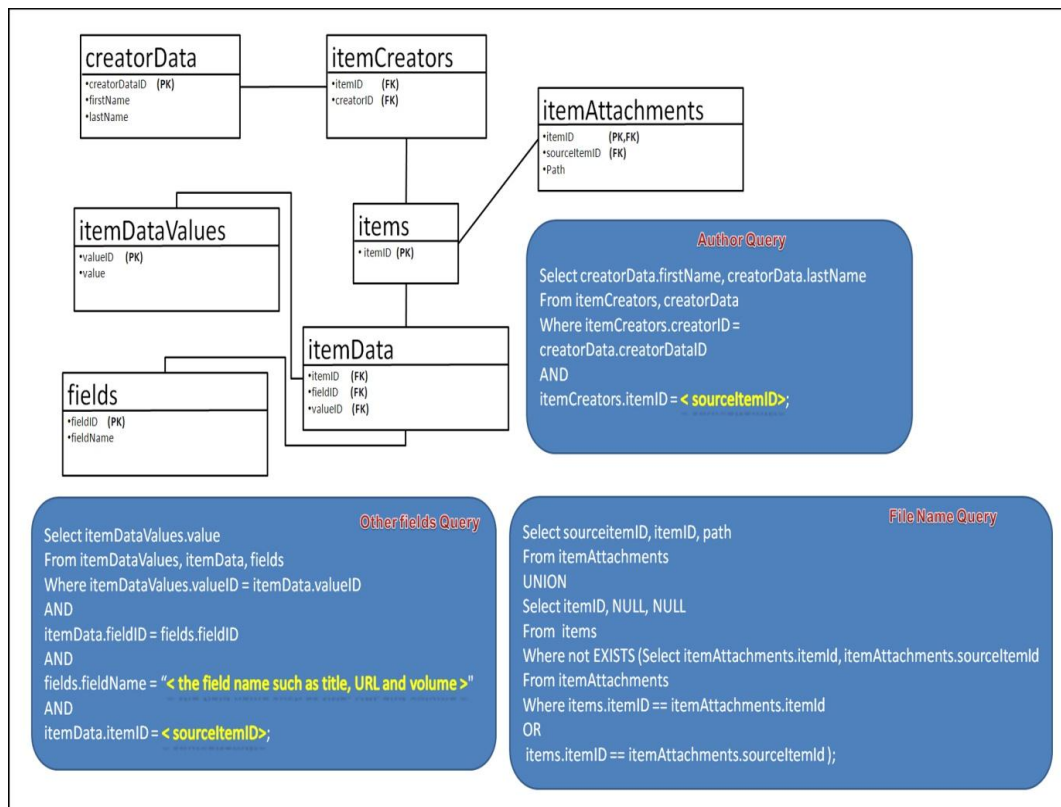
Figure 5.6: Relationships among the tables and the queries used for retrieval.

5.2.3 Connection with the Zotero database

To obtain the data from the Zotero database (Zotero.sqlite), TRMS needs to connect with the database, before running the queries shown in the previous section (see Figure 5.6).

As explained in Section 4.3.1, items and metadata are added by the user from a webpage using the Zotero icon located in the URL (see Section 4.3.1). Thus, the Zotero database is active and connected to the web browser. Therefore, the database is locked for any external connection, which means that TRMS cannot connect to the database until the web browser is closed (the database will be deactivated at this time). Zotero's documentation mentions that the database can be accessed by an external read-only connection, but this connection is for backup usage only (forums.Zotero.org, n.d).

The Zotero database is SQLite, which can manage huge volumes of data and work efficiently. It is also lightweight, fast, portable and developers can use it without copyright restrictions (Jeon, Bang, Lee 2012). Thus, we can import metadata if the problem can be solved by copying the entire Zotero database to another folder each time.

5.3   Implementation: new software elements

The information presented in the previous section suggests that it is possible to develop new software that combines these systems. This section provides details of the TRMS process, which comprises five main parts: importing PDFs, importing metadata, feeding the collection, building the Greenstone collection and browsing the collection. This software was implemented in the Java[16] programming language. In this section, we explain the steps of TRMS in detail (see Figure 5.1: Steps 3, 4, 5 and 6). We describe how TRMS operates during each step and how it handles the problems encountered. In Section 5.3.1, we introduce the TRMS process and describe the high level processing of steps 3–6. Each separate step is then explained in detail in independent sections. In Section 5.3.6, we show how TRMS can help the user to move items among reading phases in a simple manner.

5.3.1   Overview of the TRMS Process

Figure 5.7 shows the TRMS process cycle and the potential outcomes for users. Before TRMS processes these steps, it displays a confirmation message to ensure that the user actually wants to proceed with the action (see Figure 5.7). The process will be stopped if the user clicks "No". If the user selects "yes", TRMS will proceed through the steps as follows: importing PDFs and importing metadata. If any new items are detected, it continues the process as follows: writing the new items to the XML file, building the Greenstone Collection, displaying a message with the number of items and PDFs added to the collection (see Figure 5.7) and ending the process. Otherwise, a message is displayed to say that no new items have been added (see Figure 5.7).

---

[16] http://www.java.com/en/download/faq/whatis_java.xml
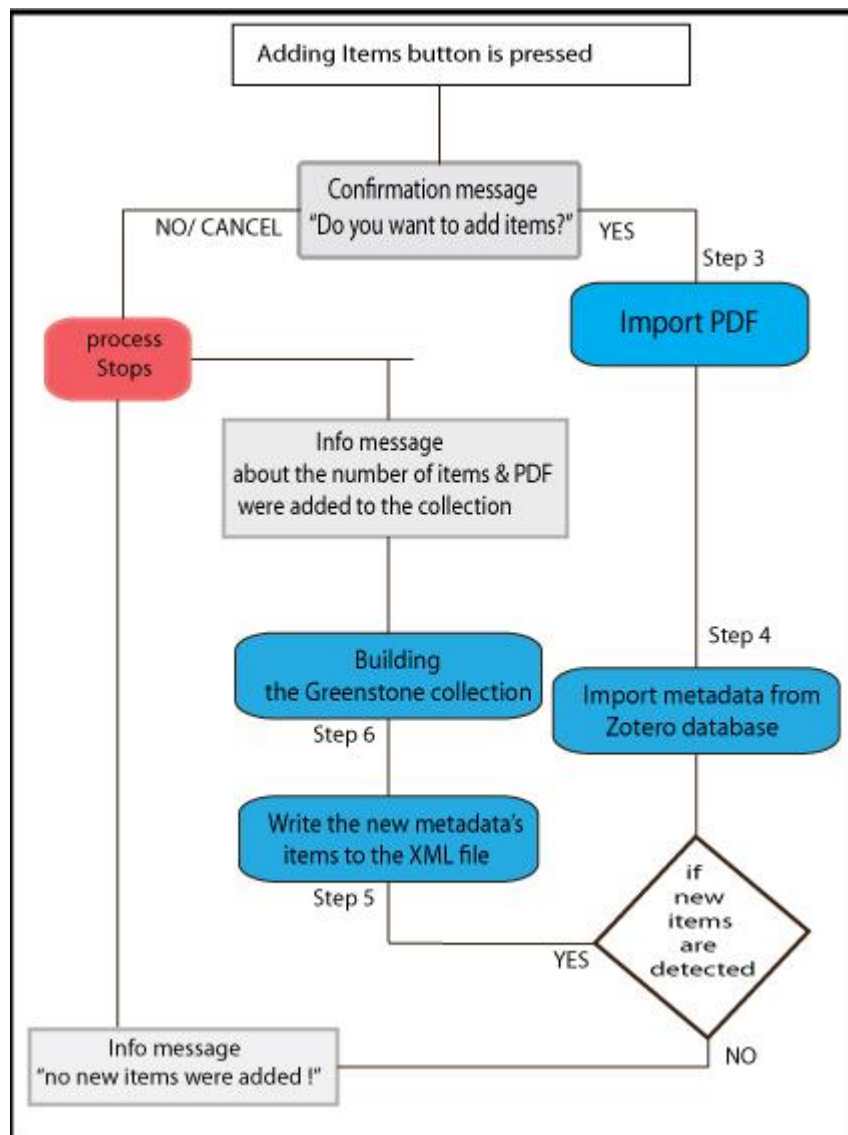
Figure 5.7: TRMS steps during each run.

As shown in Figure 5.8, the TRMS interface has eight elements, which are explained in Section Figure 5.1. The links to these eight elements are defined as input arguments for the software.

These elements must be set before using the software (but only once when it is first used). In addition, two buttons are located at the top of the window: the Hide button and the Adding Items button. The former is used to hide and show the eight elements, while the latter is use to add items as well as activating the TRMS process, which must be pressed each time the user wants to add items.
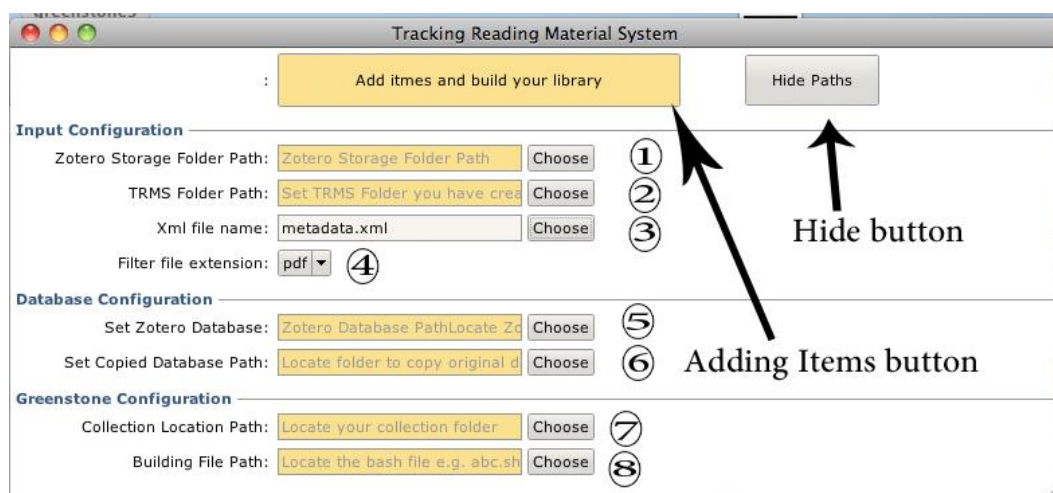
Figure 5.8: TRMS interface.

5.3.2 Importing PDF files (Step 3)

The first step in the TRMS process is importing the PDF files from the *storage* folder to the TRMS folder. Zotero can deal with multiple files for each entry because the *storage* folder may contain other files, such as css, gif and js files, in addition to PDFs. These files are components of HTML files and some might be attached to a source.

To prepare for the importing process, TRMS adds all of the names of the files located in the *storage* folder to a list called import. The content of the list is examined by passing it through two filters: extension and new item filters (see Figure 5.9). The former is used to provide PDF files only whereas the latter provides the new files that have not been imported to the TRMS folder previously. The extension filter checks the extension of each file in the import list, which are removed from the list if they are not PDFs. The new item filter checks each file name and they are removed from the list if they exist in TRMS folder because they have been imported before. After the list has been filtered, it only contains PDF files that have been downloaded recently (not in TRMS folder). The files in the import list are copied to the TRMS folder (see Figure 5.9).

In the last step, each file is copied to the TRMS folder and to the Greenstone collection (see Section 5.3.4, step 5). The reason for passing a copy to the collection is explained in Section 5.3.5.
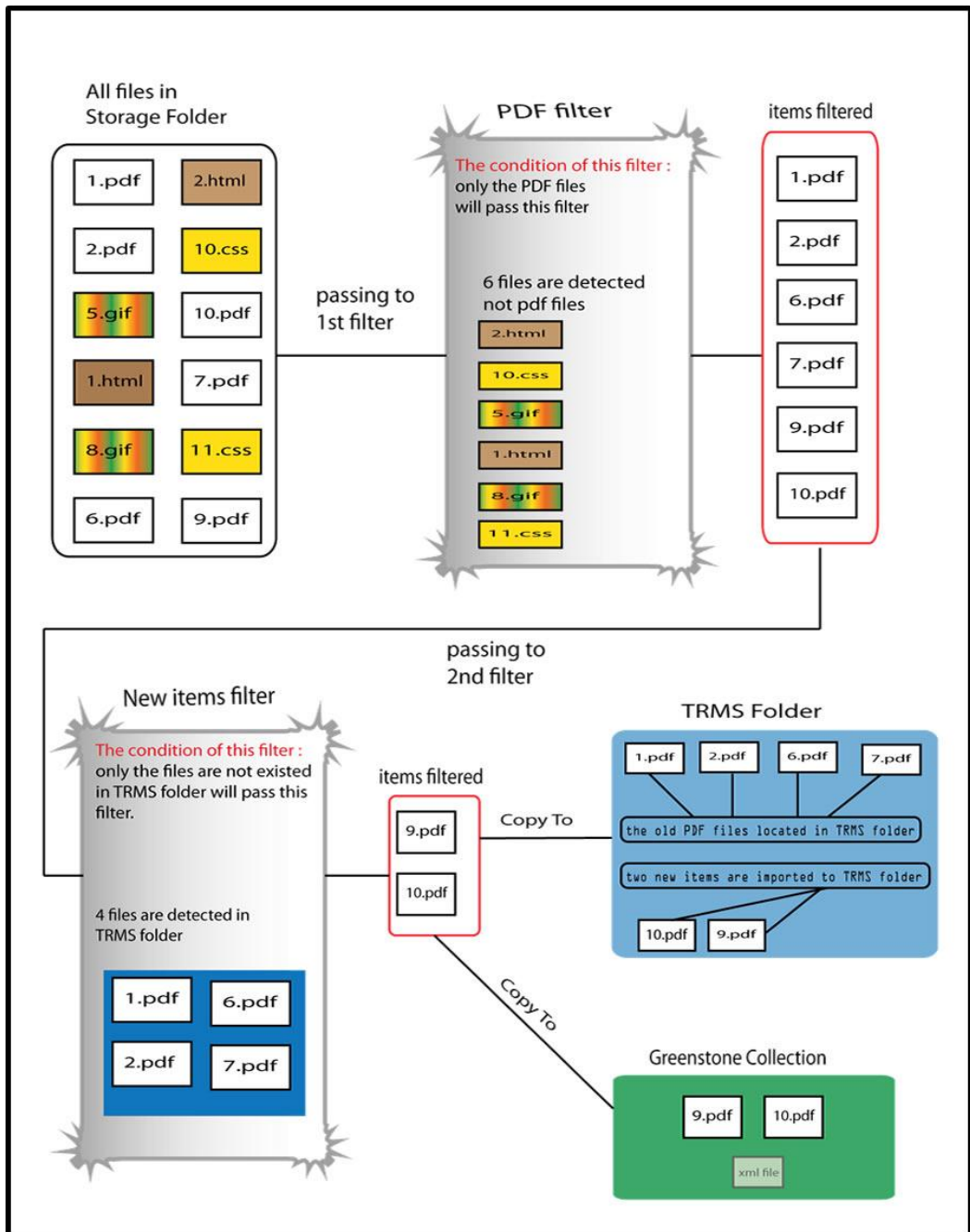
Figure 5.9: PDF import process.

5.3.3 Importing metadata (Step 4)

This section describes the process used to extract the eight fields (the fields are specified Section 5.2.2) from the Zotero database. The fields are structured in an XML file, as shown in Figure 5.2. Using the XML file structure, we obtain the file name of each attachment (the parent in the XML file) and all of its metadata (a child in the XML file) (the fields are specified in section 5.2.2). Thus, we need to deal with the attachments to obtain their file names. However, not all of the

sources have attachments in Zotero. Thus, we need to handle two types of sources: 1) sources with attachments, and 2) sources without attachments.

1) Sources with attachments

In this case, a simple query is used to obtain three attributes from the itemAttachments table (Figure 5.10) in the Zotero database: path (the file name of the attachment), itemID (the ID of the attachment) and sourceID (the ID of the source to which the attachment belongs).

```
Select sourceitemID, itemID, path
From itemAttachments;
```

Figure 5.10: Query used for sources with attachments.

A query for the author's field and queries for each other field are used to obtain the other fields (authors, title, etc.), where the sourceID is a condition key (see Figure 5.6 in Section 5.2.2). During the testing of queries, we identified three issues that may affect the process outcome.

1- Multiple attachments in one source lead to duplicated items in the XML file. Figure 5.11 shows an example where two items, i.e., a PDF and HTML, both have the same metadata. A single source may sometimes have two HTML attachments.

```
</Description>
</FileSet>
<FileSet>
  <FileName>SLEEP - 2006 - Behavioral treatment of bedtime problems and night.pdf</FileName
  <Description>
    <Metadata mode="accumulate" name="dc.Title">Behavioral treatment of bedtime problems an>
    <Metadata mode="accumulate" name="dc.Author">SLEEP , PEDIATRIC</Metadata>
    <Metadata mode="accumulate" name="dc.Creator">PEDIATRIC SLEEP</Metadata>
    <Metadata mode="accumulate" name="dc.Volume">29</Metadata>
    <Metadata mode="accumulate" name="dc.URL">http://nasms.org/app/download/6709694804/Revi
    <Metadata mode="accumulate" name="dc.Pages">1263</Metadata>
    <Metadata mode="accumulate" name="dc.Year">2006</Metadata>
    <Metadata mode="accumulate" name="dc.Journal">Sleep</Metadata>
    <Metadata mode="accumulate" name="dc.Date">20060000</Metadata>
  </Description>
</FileSet>
  <FileSet>
  <FileName>attachment.html</FileName>
  <Description>
    <Metadata mode="accumulate" name="dc.Title">Behavioral treatment of bedtime problems an>
    <Metadata mode="accumulate" name="dc.Author">SLEEP , PEDIATRIC</Metadata>
    <Metadata mode="accumulate" name="dc.Creator">PEDIATRIC SLEEP</Metadata>
    <Metadata mode="accumulate" name="dc.Volume">29</Metadata>
    <Metadata mode="accumulate" name="dc.URL">http://nasms.org/app/download/6709694804/Revi
    <Metadata mode="accumulate" name="dc.Pages">1263</Metadata>
    <Metadata mode="accumulate" name="dc.Year">2006</Metadata>
    <Metadata mode="accumulate" name="dc.Journal">Sleep</Metadata>
    <Metadata mode="accumulate" name="dc.Date">20060000</Metadata>
  </Description>
</FileSet>
<FileSet>
  <FileName>Sleep - 1992 - Technical note automatic sleepwake identification.pdf</FileName>
  <Description>
```

Figure 5.11: Example showing duplicated items in an XML file.

2- In Zotero, any source that has been downloaded several times will lead to duplicated items with the same file and metadata in the XML file.

3- TRMS only transfers the PDF files whereas Greenstone merges the files with their metadata based on their file names (see Section 5.2.1). Therefore, the HTML items will not be displayed because they do not have a file in the Greenstone collection where the XML file is located.

These three issues are addressed by passing all of the data through several filters.

o **Solution to the first and second issues:** for any source with a PDF and HTML attachments, a filter only passes the item with a PDF extension. In TRMS, only PDF files are transferred from Zotero so the PDF attachment's metadata is used as the first priority relative to other types of attachment. Figure 5.12 shows that we still have duplicated items (the items passed by the first filter). Another filter is created for sources with multiple HTML files or those that are already duplicated

107

in Zotero. All of the data are passed through and each item is compared with the items that have passed through the filter already. If any of the metadata match the metadata of items that have passed already, the filter will delete the item. Thus, this item will not be passed.
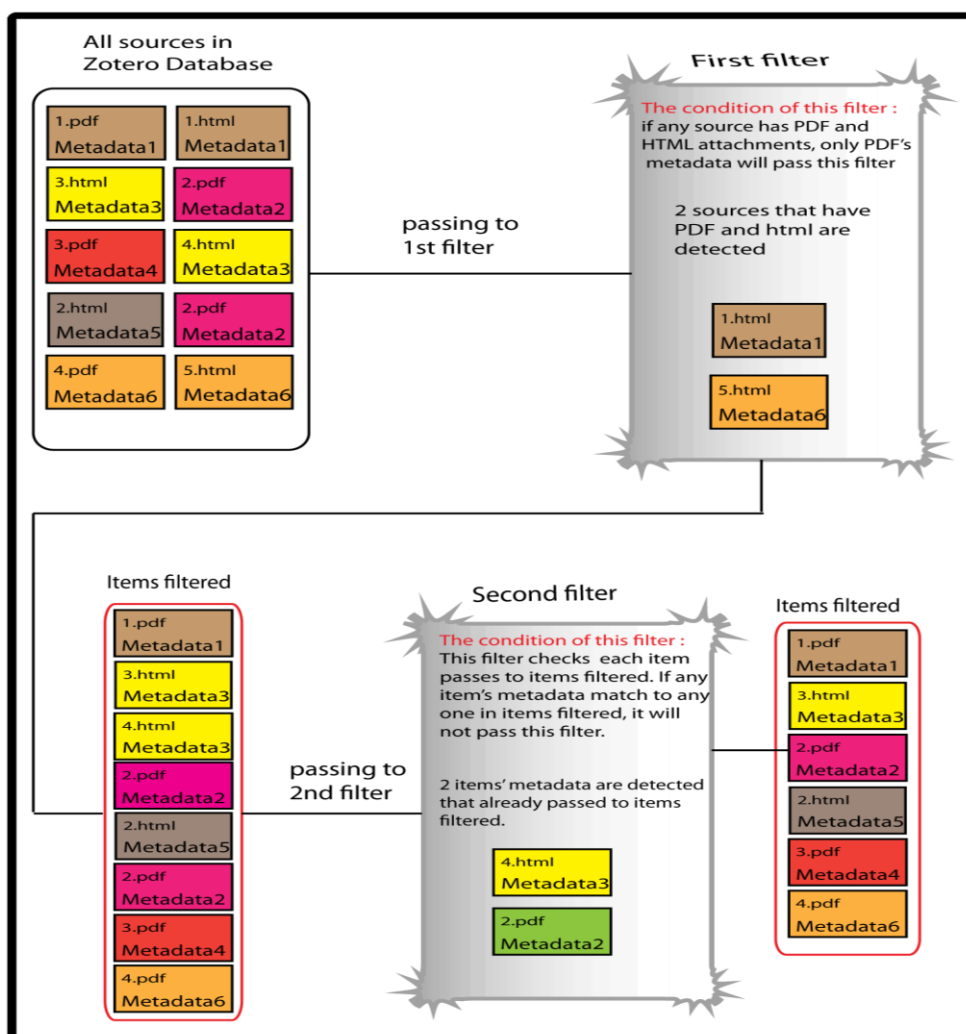
Figure 5.12: Data retrieval from the Zotero database – Part 1.

o **Solution to the third issue:** the items filtered by the second filter (see Figure 5.12) are grouped into two lists: one for PDF items and another for HTML items (see Figure 5.13). The file names of the items in the HTML list are renamed in numerical order with "nul" extensions. The two lists are then merged together to check whether each item exists in the XML file. If an item's name and metadata match an item in the XML file, it will be removed from the list. At this stage, we only have the new items that have not been written to the XML file before. These

items are added to the XML file and a file is created for each new nul file in the Greenstone collection (see Figure 5.13).
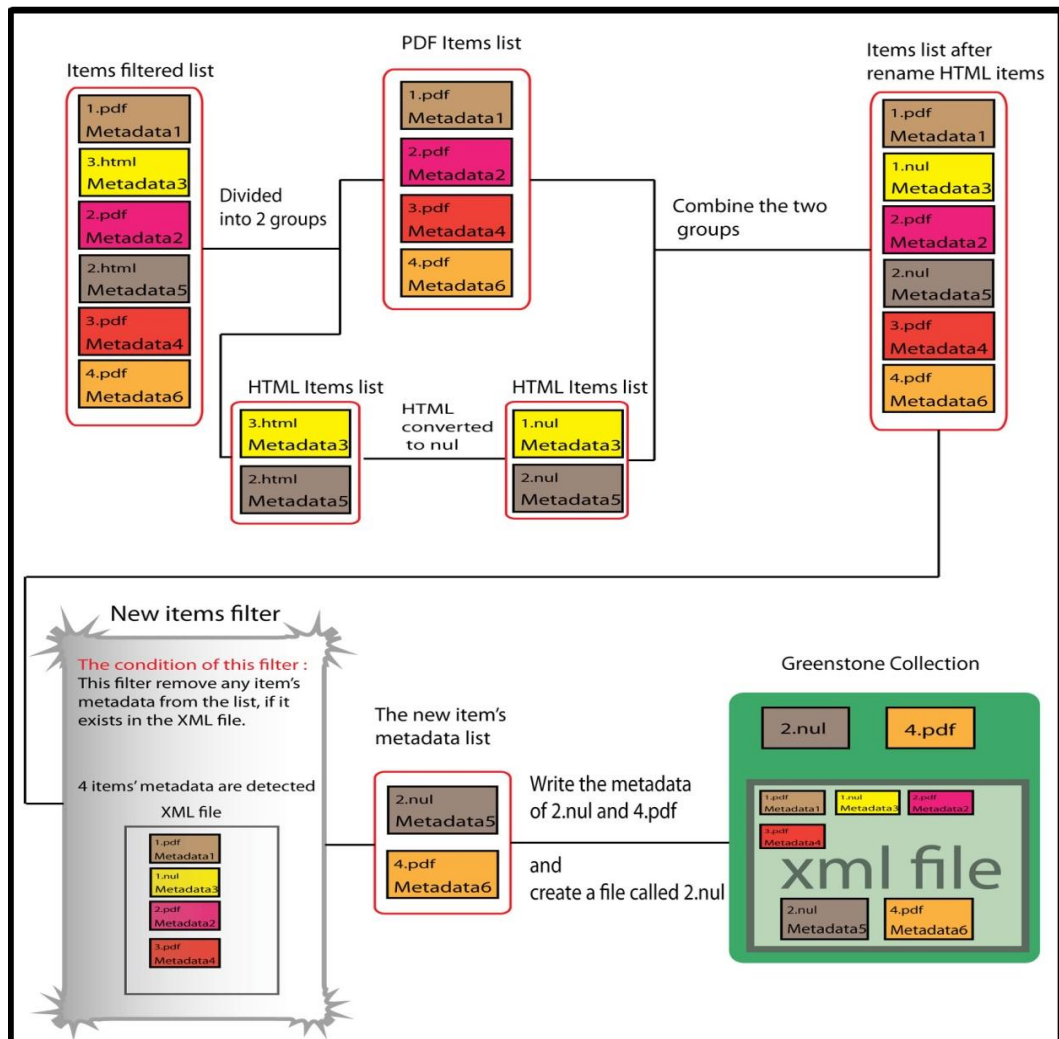
Figure 5.13: Data retrieval from the Zotero database – Part 2.

2) **Sources without attachments**

If the sources have no attachments, a query is used to obtain the sourceID only. This query selects the itemID (as sourceID) from the items table, provided that the item does not exist in the itemID and sourceID attributes in the itemAttachments table. To obtain other fields, we can use the query used for **sources with attachments.** These items are listed but the full-text is not available to users.

```
Select itemID
From items
Where not EXISTS (Select itemAttachments.itemId,
itemAttachments.sourceItemId
From itemAttachments
Where items.itemID == itemAttachments.itemId
OR
 items.itemID == itemAttachments.sourceItemId );
```

Figure 5.14: Query for sources without attachments.

In both cases (with and without attachments), the souceID is used to obtain all the required fields. Thus, the data retrieval process will be integrated by merging the first query for both types. The use of a UNION operator to combine the two queries might be the best way to integrate the process (Figure 5.15). However, the path (file name) and itemID columns should not be null values when passing the created filters (see **sources with attachments**). This can be solved by setting a default value "noAttachment.html" for any item that does not have a file name. Moreover, the itemID can be set as the sourceID.

```
Select sourceitemID, itemID, path
From itemAttachments
UNION
Select itemID, NULL, NULL
From items
Where not EXISTS (Select itemAttachments.itemId,
itemAttachments.sourceItemId
From itemAttachments
Where items.itemID == itemAttachments.itemId
OR
 items.itemID == itemAttachments.sourceItemId );
```

Figure 5.15: Query after integrating sources with and without attachments.

5.3.4 **Feeding collection (Step 5)**

As explained in the description of the TRMS architecture (Section 5.1), two elements are transferred from Zotero to Greenstone: PDFs and metadata. We have learned that Greenstone displays the files that are located in Import folder (Section 5.2.1) so the PDF and nul (nul sources do not have a PDF, see Section 5.3.3) files are placed in that folder. In addition, all of the imported metadata are written to an XML file located in the same folder.

Each item added to the XML file has a new element called a reading phase (Figure 5.16). This element has a default value of "To read" because each item should be in the first stage of the reading phase "To read" when it is first imported. The user can then move it to the second stage ("Currently reading") or the third stage ("Have read"). The method used to move items between reading phases is explained in the next section (Section 5.3.6).


```
</FileSet>
<FileSet>
  <FileName>SLEEP - 2006 - Behavioral treatment of bedtime problems and nig
  <Description>
    <Metadata mode="accumulate" name="dc.Title">Behavioral treatment of bed
    <Metadata mode="accumulate" name="dc.Author">SLEEP , PEDIATRIC</Metadat
    <Metadata mode="accumulate" name="dc.Creator">PEDIATRIC SLEEP</Metadate
    <Metadata mode="accumulate" name="dc.Volume">29</Metadata>
    <Metadata mode="accumulate" name="dc.URL">http://nasms.org/app/download
    <Metadata mode="accumulate" name="dc.Pages">1263</Metadata>
    <Metadata mode="accumulate" name="dc.Year">2006</Metadata>
    <Metadata mode="accumulate" name="dc.Journal">Sleep</Metadata>
    <Metadata mode="accumulate" name="dc.Date">20060000</Metadata>
    <Metadata mode="accumulate" name="TM.reading phase">To read</Metadata>
  </Description>
</FileSet>
```

Figure 5.16: Example showing a reading phase element.

5.3.5 Building the collection (Step 6)

The Greenstone building process proceeds through several steps in the command line (see Figure 5.17). A user must set two variables in the script before using TRMS: the collection path and the collection title. These steps are established in a bash file and they allow TRMS to be invoked each time we build the library. These steps are described as follows.

1- Clear the directory.
2- Locate the Greenstone installation to move through the steps.
3- A bash file is invoked in the Greenstone installation called setup.sh. This file contains the system variables that allow the system to recognise the next commands.

4- This command converts the source documents and metadata (metadata.xml) into Greenstone standards. The results are placed in an **archive** folder (located in the collection folder) and the collection is built from this folder.

5- Greenstone builds the collection during this step.

6- Two more three steps are required to operate the collection. First, to ensure that the **index** folder (located in the Greenstone installation) does not contain files, the command "*re –rf index*" is used to remove any files. Second, all the contents of the **building** folder (located in the collection folder) are moved to the **index** folder (see Figure 5.17, final command line), before activating the library by opening a URL. The final step in the building process is included in the TRMS software.

```
#!/bin/bash

# Please define the two variables below and save the file before closing it.

# please write the greenstone source for example, "/Users/MohammedAlanazi/Desktop/greenstone3"

GSLOCATION="/Users/hammadalshammari/Desktop/greenstone3"        The variables must be set

# Please write your collection or library, you have created in greenstone.For example "test", "test22".

COLLECTIONNAME="test2"

# now please save the changes you made and close the file

cd //                                          1

cd $GSLOCATION                                  2

source gs3-setup.sh                             3

import.pl -keepold -site localsite $COLLECTIONNAME    4
buildcol.pl -keepold -site localsite $COLLECTIONNAME  5

cd web/sites/localsite/collect/$COLLECTIONNAME
rm -rf index                                    6
mv building index
```

Figure 5.17: Contents of the bash file used to build the Greenstone library.

Note: As explained in step 3, the conversion results are placed in the **archive** folder each time the collection is built, which means there will be duplicate files in the collection if a new file is added to an old one. Thus, before activating the collection, TRMS deletes the PDF and null files from the **import** folder. Any new PDF or nul files detected by TRMS are added to the Greenstone collection (see sections 5.3.2 and 5.3.3).

5.3.6 Browsing the collection (Step 7)

At this point, the collection is ready for browsing. Greenstone provides two options for browsing the collection. The first option is to use the icon located in the Greenstone installation (see Figure 5.18). The second option is to use the command line via the terminal application.
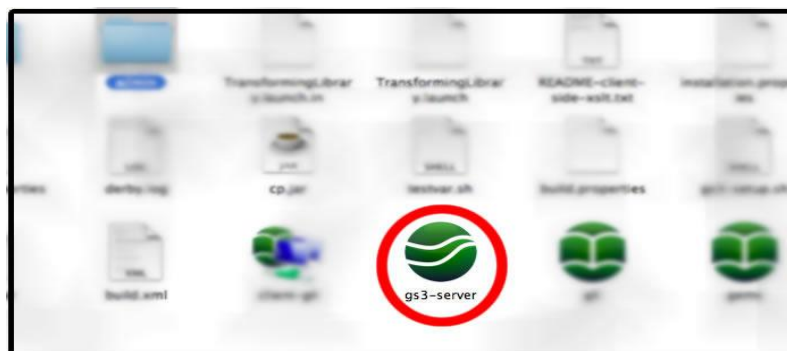
Figure 5.18: Server icon.

The main aim of this method is to help users to identify their reading material. As explained in Chapter 4 Section 4.1.2, Greenstone can be adapted to create an element in the metadata called the "TM.reading phase". This allows Greenstone to capture the value of this element from the XML file (as explained in Section 5.3.4, step 5). As shown in Section 4.3.6, the user can move any item between phases by changing the value of the TM.reading phase element and clicking on the save changes button. Changing the values manually moves items between the three phases successfully, but it might lead to issues that could affect the process efficiency. For example, the user might misspell the phase's name (by typing "hvae read" instead of "have read") or use mixtures of small and capital letters each time they type the value ("Have Read" is different from "have read" and "Have read").

To avoid these issues, three buttons are available for moving between the reading phases: "Have read", "Currently reading" and "To read". These buttons assign a value to the TM.reading phase element and rebuild the library. For example, if the user wants to move an item from the "To read" phase to the "Have read" phase, they go to the item (the user should be logged in as admin) and select the edit metadata button. The user only sees the "Have read" and "Currently reading" phase buttons because the item is already in the third phase "To read" (Figure

5.19). If the user clicks on the "Have read" button, the value of the TM.reading phase element changes to "Have read" and the library is rebuilt. At this point, the "Have read" button disappears and the "To read" button is visible.
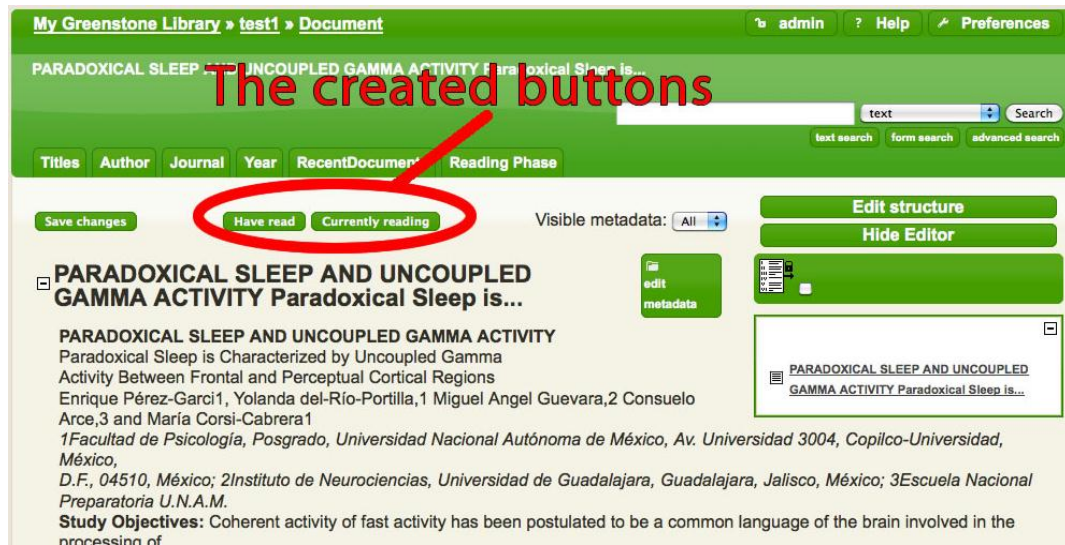
Figure 5.19: Greenstone reader interface and the created buttons.

This example demonstrates how the process of moving items among phases has been improved, which is more efficient and effective than typing manually.

5.4 Discussion of the Limitations of the Software

The implemented system combines two existing types of software to produce a hybrid system. This section discusses the limitations of each independent software system, i.e., Greenstone, Zotero, and the hybrid software system TRMS.

5.4.1 Greenstone

Changing the metadata and rebuilding the library via the Reader Interface browser is the main TRMS step conducted in Greenstone because we need to rebuild the library each time we add or edit items and metadata. Greenstone3 (windows binary version) did not support this process in an appropriate manner and it sometimes corrupted the collection. These technical issues with Greenstone3 (Windows binary version 1st of June) meant that TRMS could only be used on Mac machines (binary version 1st of June). However, the building process operates on Mac machines and the user only has to enter two setting before using Greenstone (but only once when it is first used). First, the search indexer must be changed from MGPP to Lucen while the database used by the browsing classifiers must be changed from GDBM to JDBM.

5.4.2 **Zotero**

Metadata cannot be imported until there is a connection with the database. Thus, connecting to the Zotero database is an essential step for obtaining the metadata. However, as explained in Section 5.2.2, the Zotero database does not allow any external connections while it is active. The Zotero database is usually active when the web browser is open. Thus, deactivating the database was not possible. The solution to this issue is to copy the entire database each time the importing process is invoked. Before copying the database and fetching the metadata, the user must ensure that all of the data and PDFs are downloaded into the database and the storage folder, respectively. If some data are not downloaded, they will not be transferred. Figure 5.20 shows the message displayed by Zotero when data is downloaded from websites (it usually appears in the right-down corner). The user must wait until the message disappears to ensure that the downloading process is complete.

Figure 5.20: Zotero message when downloading items.

5.4.3 **TRMS**

After items have been transferred from Zotero to Greenstone, Greenstone requires a few seconds to rebuild the collection (depending on the number of items added). New items cannot be added during the rebuilding period. Therefore, we decided to give the user control over *when* the items are incorporated into the collection, instead of rebuilding the collections repeatedly.

5.4.4 **HTML Files**

Occasionally, Zotero cannot download PDFs or the PDF is either restricted (some commercial digital libraries require payment or membership) or not available online, so Zotero provides either a snapshot of the website or the metadata alone.

115

we only transfer PDF files from Zotero and Greenstone that lack metadata referring to a file in the collection folder. We create null files for any sources without PDF files and display the metadata in Greenstone.

5.5 Summary

This chapter describes the development of TRMS and its architecture (Figure 5.1). The processes performed by the three components, i.e., Zotero, TRMS bridge and Greenstone, can be summarised in seven steps: (1) Zotero is used to add items from websites (PDFs and metadata); (2) Zotero stores the PDFs and metadata in a storage folder and the Zotero database, respectively; (3) TRMS imports the PDFs; (4) TRMS imports the metadata from Zotero; (5) TRMS feeds the imported PDFs and metadata to Greenstone; (6) TRMS builds the library to display the imported files; and (7) the collection can be browsed by a user.

Our implementation focused on steps 3–6. In step seven, we created the "To read", "Currently reading" and "Have read" buttons on the Greenstone Reader Interface which allows the user to move items in the library between reading phases.

Chapter 6

6 Evaluation TRMS

In the beginning of this research, a study was conducted to explore how people keep track of their reading materials. From the findings, eight requirements were considered for inclusion in a tracking material system. The system which was developed included six of these requirements (TRMS). All the steps were followed to help people in tracking their digital reading material. To determine whether TRMS reach the goal of the thesis, it needed to be evaluated. This chapter reports and discusses the results of an evaluation study conducted for TRMS.

6.1 Approach

Two methods were chosen for gaining better results in our evaluation: an interview and a diary study. The first step was to obtain the ethical approval required because humans were involved in the study. The approval for this study came from the Human Research Ethics Committee at the Department of Computer Science, University of Waikato (see Appendix B). The study involved three phases: an interview before using the software, a diary study while using the software, and an interview after using it.

6.2 Purpose

This project aims to develop a system that will help people to easily track their reading material. For this purpose, we designed and developed TRMS. This study was conducted to assess the TRMS process and evaluate the features developed for helping people to track material.

6.3 Procedure

Four participants were invited via email by the researcher. They were interviewed at their offices. Prior to each session, the participants were informed about the study and its goals. They were also told that, at the end of the study, they would have the opportunity to receive a summary concerning its outcome. Then, each participant was given a participant information sheet, then signed a consent form to participate in our study.

The study consisted of three phases: interview, use of the software, and a second interview.

In the first phase, the participants were interviewed before using the software to identify their background experience with Greenstone, Zotero, downloading papers and tracking materials (see Appendix B.2).

In the second phase, the participants were asked to use a machine supplied by Technical Support Group (TSG) and equipped with our software (the three components were Zotero, Greenstone and TRMS Bridge) for one week. The participants were provided with a diary study sheet to record any observations, notes or issues they faced while using the software (see Appendix B.3).

In the third phase, the participants were interviewed after using the software and asked about their experiences during the study (see Appendix B.4).

6.4 Participants

The study was designed to involve users of Zotero who are familiar with adding items from the internet. Finding participants who met the specific standards was one of the study's limitations that are described in Section 1.5. We found two participants who use both Zotero and a Mac computer (one of the limitations of Greenstone see Section 5.4.1). The other two, who were chosen randomly, use only a Mac. In the Zotero group, both are PhD students (one male and one female), and in the non-Zotero group, one is a Masters student (male) and the other is an associate professor (female). All participants are from the Computer Science Department at Waikato University.

6.5 Limitations

The machine supplied and Zotero user limitations may affect the study sample, and in turn affect the results.

Machine supplied: We have only one machine, which was supplied by TSG. Therefore, we could not run the study for more than one participant in a week. We invited four participants, so the study took one month (one week for each participant). We would have liked to have had more participants, but the time was not sufficient with this procedure.

Zotero users: It was difficult to find Zotero users who also use a Mac computer because one of the limitations in our system was to use the Greenstone Mac version.

118

6.6 Results

The following sections will present the results of the three phases. Phase one, is the interview before intervention including Greenstone, Zotero, Downloading Paper and Tracking Materials and diary study; phase two, diary study; and phase three, interview after intervention.

6.6.1 Interview before Intervention (Phase I)

To assess how well the participants know the components of the software, a before-intervention interview was used. In addition, this phase examined the experience of the participants in downloading papers and tracking materials. Knowing the participants' background in these areas may have assisted our understanding of their reactions and behaviours while conducting the experiment. This section reports the participants' experience in four fields: Greenstone, Zotero, Downloading Paper, and Tracking Materials.

6.6.1.1 Greenstone (GS)

The participants were asked if they had ever used Greenstone (Q1) and then asked two subquestions regarding the experience: Q1.1. Describe your experience. Q1.2.Things you like and dislike? Q1.3. Any issues you have ever encountered? Table 6.1 shows the participants' responses.

Table 6.1 Participant's responses to questions about Greenstone in first interview.

	Q1.	Q1.1.	Q1.2. & Q1.3.
P1	Yes, but not for building and managing. I have used a website built in GS.	Few times, 1 hour in total.	It was a bit slow.
P2	No	No	No comment
P3	Yes, but not for building and managing. I have used a collection	3 years from 1993–1996.	• It requires me to do too many steps such as what sort of metadata and how this is going to be organised.

	which was built in GS.		• I have to go through and build the whole collection when add new document.
			• It still too much trouble for building my own personal collection.
P4	Yes, I have tried to design and manage a collection.	About one month.	• It does not extract the metadata auto.
			• It does not have tags.
			• Reader interface is not convenient.
			• I would like view the collection for ex 2 or 3 dimensions.

The responses to the first question revealed that three of the participants had used Greenstone for a variety of periods.

For the first subquestion, only one positive point was given, by P4, who said, 'I can organise all my docs in one place'. However, all the participants who had used Greenstone noted several points that they were not satisfied with (see Table 6.1). None of the participants mentioned any issue in the last subquestion.

6.6.1.2 Zotero

The participants were asked if they had ever used Zotero (Q2.) and then asked subquestions regarding the experience: Q2.1. Describe your experience. Q2.2.Things you like and dislike? Q2.3.Any issues you have ever encountered? Table 6.2 shows the participants' responses.

Table 6.2 Participant's responses to questions about using Zotero in the first interview.

	Q2.	Q2.1.	Q2.2. & Q2.3.
P1	Yes.	About 9 years	• Integration into the web browser for one click capturing of a miss data.
			• The Word linking to the database.
			• The ability to link a downloaded paper within the interface.
			• Notes, tagging and groups features.
P2	Yes	About 5 years	• Integrated into the web browser.
			• Click a button to capture info.

		• Share folders or libraries with others. • Word Plug-in. • Sync all my computers.	
P3	No	None	No comment
P4	No	None	No comment

Table 6.2 shows how P1 & P2's answers to the first sub-question indicated their satisfaction with using Zotero. The responses to the first question revealed that two of the participants had been using Zotero for a long time.

For the things they dislike, they agreed on one point, which is the limit of storage data, and P1 added two more points. First, he would like the notes interface to be easier to use and the tagging system to be much better.

For the second sub-question, both had only one issue in their early use. P1's issue was about corrupting database and P2's issue was about additional set-up with Sync.

6.6.1.3 Downloading Papers

The participants were asked whether they downloaded digital papers related to their projects or research into their computers (Q3); this was followed by two sub-questions. The first sub-question asked how they do that (Q3.1). The second asked what kind of materials they usually download (Q3.2).

Table 6.3 Participant's responses to questions about downloading papers in first interview.

	Q3.	Q3.1.	Q3.2.
P1	Yes	Put the reference in Zotero and then download the file, open the PDF, copy the title and rename the PDF with title. Keep the file into the folder system (a folder structure 'tree structure' which matches tree structure's Zotero).	PDF
P2	Yes	Put the reference in Zotero and then download the file. I've got folders set up in my computer. The folders are related to the folders that I have in Zotero. Where I put the citation in Zotero folder, I	PDF

		will put the PDF in same folder in my computer.	
P3	Yes	I usually put the file in download folder which I never delete anything from. Then I put a copy in a folder that has to do with particular project that I'm working on.	PDF
P4	yes	I download the file into my projects' folder directly.	PDF

The response to the main question indicated that all the participants do usually download their materials into their computers. In response to the first sub-question, all the participants indicated that they download the papers from the internet without using any tool. However, each one has a different procedure for downloading digital papers (see Table 6.3). In response to the second sub-question, all the participants indicated that they usually download PDF files for their projects or research.

6.6.1.4 Tracking Materials

The participants were asked to describe their experiences regarding how they keep track of the material they have read, are currently reading and are planning to read, and what issues they have encountered.

- **Keeping track of material the participants have read**: When the participants were asked about the material they had read, they described the whole process they follow, from downloading the paper until they have finished reading it. They then commented on the issues they found.

Table 6.4 Participants' responses on tracking material they have read

	Material already read	
	Keeping track	Issues discovered
P1	• download a file into a folder (*To read and classify*) • Books to be kept filed in (*To retrieve*) • Once read, books moved to appropriate, related folder	• Tag needed to indicate the usefulness of the material waiting to be read. Knowing what I need to use these for. • Knowing whether I downloaded them.
P2	PhD readings organised in folders	I have many PDFs so I break down

	(unread) and subfolders (have read).	the PDFs into folders and having folder in each to track what materials I have read and unread.
	once a PDF is read, it is transferred to read subfolder	
	I only do that with my PhD stuff.	
P3	• Writing begins with really big broad searches in Google Scholar and ACM digital library - anything relevant downloaded until no more new stuff found or new stuff or a large number of files like 20 or 30 accumulated	• I have run in across a couple of cases where I thought I have new doc and then after a while I realised that I read this and it just layout with double of columns.
	• Check these online for relevance, mainly the abstract and introduction relevant material - small set of 10 to 15 copied to folder created for that project.	• Niggling feeling that there are some other things I used in the past that would be really useful here but I cannot remember enough of the title or the author to be able to go back and find.
	• From that folder, print them out and read them at once.	
	• Start reading in depth, starting to circle things and write notes.	
P4	• Download files into a folder related to my project.	
	• When read, files are moved to the relevant folder.	
	• Subfolders classified based on the subjects.	

- **Materials the participants are currently reading:** All participants indicated that they do not track what they are currently reading. However, only two participants raised issues in this phase (see Table 6.5).

Table 6.5 Participants' responses on tracking material they are currently reading.

	Material currently being read	
	Keeping track	Issues discovered
P1	No, I only open them on a tab or a PDF reader.	• It only works until the computer or the software crashes and then you lose them. • Having too many projects while writing.
P2	Normally just open.	When the computer is shut down.
P3	No	No
P4	No	No

- **Materials the participants are planning to read:** The responses to this question consisted of two groups. First, P1 and P2 answered this question by referring to their answers in the 'have read' question. However, the second group, which included P3 and P4, indicated that they do not keep track of what they are planning to read. Therefore, only P1 and P2 provided responses to the sub-question regarding issues they may have encountered (see Table 6.6).

Table 6.6 Participants' responses on tracking material they are currently reading.

	Material planned to be read	
	Keeping track	Issues discovered
P1	As same as the Have read question.	• It is quite useful if I have something to tell me read it now or leave it for six months. • A way to easily download something and almost classified or know why I downloaded it. • Knowing when and where I downloaded something.
P2	As have read method.	Downloading one item several times. Download one item and forget why I download that file and what I need it for.
P3	I don't have a way to order which one I should read next. I am really short term project base.	No
P4	No	No

6.6.2 Diary Study (Phase II)

While the participants were using the software, they were given a diary sheet that contained about 40 pages with four questions on each page: Q1.What are you trying to do? Q2.Did you observe any issue? Q3. How did you resolve the issue? Q4.Further notes or questions? The participants were asked to log any action they did while using the software (see Figure 6.1).
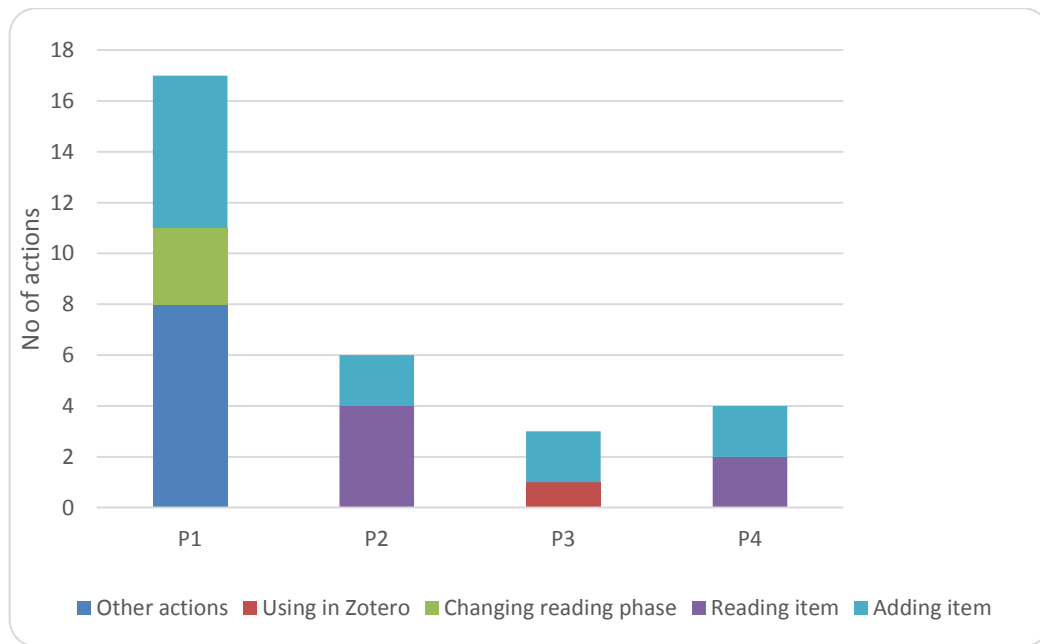
Figure 6.1 Common actions written in diary sheet

The actions were grouped into five categories: adding item, reading item, changing reading phase, using Zotero and other actions. The total number of actions taken by the participants in this study was 29, and a surprising finding was that most of them (17) were taken by P1 (see Figure 6.1)

The 'adding items' action was the most common action (used 12 times) among the participants. P2, P3 and P4 tried to add items twice during the study; however, P1 tried this action six times. Only P2 indicated that there were too many steps required to add an item and they were complex.

For the 'adding item' action, only two participants (P2 & P4) observed that the sources they added do not have PDF. Only these participants indicated that they read items and observed that the PDF formatting in Greenstone is not convenient and both of them used the other way of displaying items, which is through a PDF reader (through the web browser). P4 claimed that when he viewed an item on a PDF reader he was not able to highlight or annotate it. Only P2 observed that one PDF file content was not shown in Greenstone (it was holding lines instead of text). P3 did not have any observation regarding the software but it was about

Zotero software. She was trying to determine whether Zotero remembers saved documents across different pages and the issue of that action was that she could not open a new search window containing Zotero.

The other actions that P1 mentioned in the diary sheet were open app, use the app, log in to change a reading phase, what does 'save changes' button do?, 'edit structure' button, click on a journal, wanted to order the list and wanted to manage my list. The most important actions are summarised with the participant's observations in Table 6.7.

Table 6.7 Actions and observations for P1

| Action | Observations | | | |
	1st	2nd	3rd	4rd
Change reading phase	'Edit content' doesn't seem like the right name	Clicking 'edit content' then clicking 'currently reading' is too much overhead	It takes long time to build the collection	The system should understand and move the items among the phases auto
Log in to change a reading phase	There is no indicator that I need to log in to change a reading phase	No need to log in		
'Edit structure' button	What is this button?			
Click on an journal	I can't click the name of the journal			
Wanted to order the list	Can't order the list			
Wanted to manage my list	Tagging or grouping			

6.6.3 Interview after Intervention (Phase III)

When the participants finished using the software, they were interviewed about their experience. The questions were Q1. How long have you used TRMS? Q2 What did you like particularly? Q3 What did you dislike particularly? Q4 Have you faced any problems? If yes, please explain. Q5 Which features would you like to see in the future? All the questions and their responses are summarised in Table 6.8 below.

Table 6.8 Participant's responses to questions about their use of TRMS (second interview).

	Q1	Q2	Q3	Q4	Q5
P1	5 days	Knowing the item I started reading or completing it. Group in author, journal and year.	Not able to group or structure the items like my Zotero. The buttons need to be pressed to move the items among phases (log-in edit content). Expending the shelf does it work on clicking on its name. Classifying the items like a tree.	Forget to log in couple of times. New item added .... the items are not shown until I click on the any tab.	Group items. Combine the active message with and TRMS message. Adding notes. To see HTML converted to PDF.

P2	3 days	Able to mark something that you started reading it. Have several things started reading rather than have one.	Difficult to get a PDF into Greenstone. Formatting of PDF in Greenstone. The Greenstone was horrible. The topography was terrible. So many steps in the process in different buttons. Lines in PDF, it is annoying.	When I started I forgot which I icon opens the library. Not having formatting. The pictures were not there.	More straightforward. Not having many steps to get the article. Highlight and take notes. To view it as a PDF.
P3	2 days	Going to Google scholar and tick things and moving them over to digital collection. Having the metadata connected to the source. Being able to have a list of titles instead of list of file names.	The message box annoys me. Seems design for one PhD student .... working on one field for long time.	No, it seems to be fairly straightforward.	Adding notes. Remembering what search term I had tried. Come up with a way of distinguishing the sub collection.
P4	3days	Adding process faster than Greenstone. Instead of using GLI, I click only a button for adding items.	Browsing the doc in the webpage. PDF formatting. Annotation not supported. Evolution of importance. Bookmarking while reading inside the book or article.	Not able to do subcollection. Could not use tags for retrieval. I usually forgot clicking the buttons.	Links inside the doc should work as a hyperlink. Creating collections and subcollections. Tagging. Annotation. Bookmarking while reading.

Table 6.9 Items added by the participants

	source type	To read	Currently reading	Have read	Total items added	
P1	PDF	0	5	3	8	12
	Non-PDF	4	0	0	4	
P2	PDF	0	2	0	2	11
	Non-PDF	9	0	0	9	
P3	PDF	22	0	0	22	22
	Non-PDF	0	0	0	0	
P4	PDF	4	0	0	4	5
	Non-PDF	1	0	0	1	

Table 6.10 Issues with websites, PDF, non-PDF and comments

	Problems noticed		Comments or recommendations
	Websites	Papers with PDF and non-PDF	
P1	None	Message tells me no PDF associated with this. Showing all bibliography.	Some automated system might be cool. Group items in topics and have tags.
P2	None	Topography colour, info structure and some PDF holding lines.	No comment
P3	None	None	Being able to tag.
P4	Could not download papers from Waikato University website.	None	No comment

6.7   Analysis and Discussion

The results of this study show that managing downloaded items might be the first step in tracking material. The correlation observed between downloading items and tracking material can be explained from the participants' experiences. The results shown in Section 1.3.1 indicate that P1 and P2, who use Zotero, are more organised in downloading papers than P3 and P4, who download all items into one folder. The Zotero users (P1 & P2) follow a consistent procedure when adding any new item to their collection. However, they do not use Zotero or its functions to do this, because storage space is limited in Sync systems, and more

storage must be paid for. They keep track of their material in two phases: 'to read' and 'have read'.

When asked, each participant responded that what each one liked in TRMS was a feature that fulfilled a lack that they had felt in their own experience. Zotero users pointed out the reading phase feature (R6) and how it can be useful to know how many papers they have read or have started reading.

In contrast, P3 and P4 indicated the automated adding item and metadata feature (R1) and how the file is connected to its metadata. This rather contradictory result may be due to users' different background experiences.

Therefore, an implication of this evaluation is that there are two perspectives, and these can be combined. The first perspective comes from the people (P1 & P2) who are organised in downloading items and try to keep track of their reading material. They are Zotero users, but do not use the downloading function of that system. The second perspective comes from the people (P3 & P4) who are not organised (and do not use any software) and do not care about tracking material. By merging the evaluation feedback of these two viewpoints, we will obtain a comprehensive assessment for several aspects.

Each requirement of the TRMS system (section 3.1) is discussed individually in the following section.

R.9   Download the reading material and metadata with the system.

The results in Section 6.6.3 indicated that adding items to the collection was more commonly used than other functions. Surprisingly, only P2 reported an issue arising while adding items. P2 claimed that the first time she used the software, she could not add a PDF file to Greenstone, although the file was downloaded in Zotero. However, when we met her in the third phase (the interview after intervention), she had added three items and all of them are added successfully. The reason for this is not clear, but it may have something to do with adding an item while the PDF has not yet downloaded (one of the limitations described in Chapter 5, Section 5.4). One point that annoyed P1, P2 and P3 was having to click the add item button each time they added an item.

R.10   Store the material itself as well as its metadata.

It was clear from the results that metadata stored without its PDF was frequently observed, and this was one of the issues mentioned by P2 and P4. In addition, P2

indicated that one of the PDFs did not have content and it contained lines. When a user cannot find or add reading material to a collection, this can lead to a lack of motivation to use the software. It can be observed that the participants were more interested in adding PDF files than metadata, so the material should become the first priority and the metadata the second.

R.11   Browse the material through the system.

The results shown in Section 6.6.3 indicate that PDF formatting was one of the issues that prevented P2 and P4 from reading through Greenstone. They indicated that the text was not convenient for reading, nor were the image sections. Moreover, the participants mentioned some tools that might help in reading, such as highlighting, taking notes and bookmarking.

R.12   Search the materials.

This feature was not experienced or mentioned by any participant. It is possible that these results are due to the size of the collection used or period of using the software.

R.13   Organise and manage materials.

The results of this study showed that the participants were not able to organise and manage their items. All the participants indicated that they could not group items or create collections and subcollections. Giving the control to the user for creating collections and moving items from one to another is one of the necessary improvements.

R.14   Indicate the reading status of materials.

This feature was used by P1 and P2. Although P2 did not mention this action in her diary, strong evidence was found in the data (see Section 6.6.3) that P2 had moved two items from the 'to read' phase to the 'currently reading' phase and had not written those actions in the diary. Both P1 and P2 indicated that this feature was helpful and they could see which items they had completed or not completed. However, the other participants did not mention this feature. A possible explanation for this is that the reading phase buttons are not displayed until the user logs in and edits the content. The participants would like to see the buttons

every time they open an item. In addition, building the library each time they move an item is inconvenient because they must wait for a few seconds.

In conclusion, using Greenstone as a part of TRMS was successful for helping people who need to track their material, however it did not match participants satisfaction in other requirements.

6.8   Summary

This chapter has described the results of the participants' evaluation of the TRMS for tracking reading materials. The study has evaluated the features developed for helping people to track material downloaded for reading. Those participants who were accustomed to organising their material into its reading status found that feature useful and easy to use. The process of downloading metadata and material needs to be automated to be more helpful. The organising and managing material feature was unsatisfactory, and did not help the participants to organise material into groups. More refinement of this feature is needed. The search function was not used by any of the participants. The browsing feature was not satisfactory. The formatting needs to be more comfortable for reading, and the metadata needs to be displayed alongside the text. Importing metadata worked but the material needs to be improved.

Chapter 7

7 Conclusion and Future Work

In this chapter, the whole thesis will be summarised in Section 7.1 and the positional possibilities will be described in Section 7.2.

7.1 Conclusion

This paper presents the results of a user study investigating methods for organising reading material. Researchers go through three reading phases with books and articles: have read, currently reading, and plan to read. The results of this user study revealed that users have well-established methods for keeping track of reading material in the first and last reading phases. However, most users did not have established approaches for the second phase. We believe that supporting this second phase could address some of the users' problems in keeping track of their reading.

From the study conducted, we derived four aspects for a proposed system: localisation, organisation and management, indication, and annotation. We looked at the available software used for organising and managing reading materials. All the systems were explained to identify which requirements are fulfilled by each one. This review demonstrated that none of the systems examined support all the requirements; thus, we needed to go a step further.

One of the best options for finding a workable system was to extend one or more of the available software systems that meet some of the proposed system's requirements because using existing systems saves on time and effort. Zotero and Greenstone proved to be the best options because they are freely available and use open source code. Each one was reviewed in detail to investigate how each one works, as well as to examine their functions that fulfil the requirements of our system. An analysis of both Zotero and Greenstone has shown that neither of the two systems is a clear favourite as a foundation software to be extended. However, another option was taken that combining the two systems and extend them to work together to reach the goal of this project.

The system was designed to use Greenstone for storing and organising the documents and Zotero for capturing document and metadata. Tracking Reading Material System was designed and implemented to transfer the data automatically from Zotero into Greenstone and to extend Greenstone to capture information about the reading status—have read, currently reading, and to be read. The processes of TRMS were performed by the three components, i.e., Zotero, TRMS Bridge and Greenstone. To identify the effectiveness of the process of TRMS and whether it is able to achieve the goal of this study, an evaluation study was conducted and it showed that TRMS was successful in tracking reading material feature. However, managing and organising, browsing and storing features were not fulfilled promised goals completely.

This project might assist Greenstone and Zotero Developers or any researcher who wants to develop any kind of tracking reading material system. In addition, it might help and encourage Greenstone users to add items from the internet easily because Greenstone does not provide this function as simply as TRMS does.

7.2 Future Work

Due to the limitations of evaluation study, the time and number of participants were not sufficient to give accurate results. Therefore, the system needs to be evaluated for a longer time with more participants. This might provide more beneficial results and additional features.

The potential areas for future research: (1) improving TRMS features; (2) extending Zotero; or (3) establishing new software.

Improving TRMS features: this possibility can be the easiest option because the information needed about the components (Greenstone and Zotero) is provided in this study. In this option, we need to focus on Greenstone because most of our features are based on Greenstone capabilities. In Greenstone, adding an item, editing metadata or modifying the library design cannot be done until the library is built. This point raises an important question to be considered before achieving the improvements recommended. Is it possible to use Greenstone without building the process?

The improvements recommended:

- Improve the reading phase function to make it more intelligent. For example, if any item is open, it should be automatically be moved to currently reading shelf unless the user marks it as a have read item. In addition, the user should be able to move more than one item at the same time from phase to another;

- Hide the TRMS Bridge and automate the process. For example, when the user adds an item, it should be added immediately to the collection without having a step in between;

- Managing and organising the items should be more flexible. The user should be able to create collections and sub-collections and move the items among them easily.

   The user could be motivated to use system by:

   1. browsing the material in an appropriate format and displaying the metadata beside the content,

   2. adding utilising tools such as add notes, tags, highlight the text; and

   3. storing a variety of file types such image, html and doc;

- Add items not only from the internet but also from the desktop;

- Delete items or collections easily; and

- Cite from the collection while writing and with citation information automatically inserted. .

Extending Zotero: We have reviewed in the related work that Zotero supports five requirements out of eight. The second possibility in the future research can be extending and incorporating the other three requirements: (R3) browsing material, (R6) indicating reading phase, (R7) indicating whether used for citation. To accomplish this work, we need to Investigate and understand how Zotero system works.

Establishing new software: Developing an independent new system is better than extending existing software but it may consume much effort and time.

References

About | LibraryThing. (2013). Retrieved December 6, 2012, from

    http://www.librarything.com/about

About Goodreads. (2013). Retrieved December 6, 2012, from

    http://www.goodreads.com/about/us

Boyce, P., King, D. W., Montgomery, C., & Tenopir, C. (2004). How electronic

    journals are changing patterns of use. *The Serials Librarian*, *46*(1-2), 121–

    141.

Bunkell, J., & Dyas-Correia, S. (2009). E-Books vs. Print: Which is the better

    value? *The Serials Librarian*, *56*(1-4), 215–219.

Butros, A., & Taylor, S. (2010). Managing information: evaluating and selecting

    citation management software, a look at EndNote, RefWorks, Mendeley

    and Zotero. In *IAMSLIC Conf Proc*. Retrieved from

    http://www.pitt.edu/~ctomer/lis2600/Butros-Taylor_iamslic2010.pdf

Barsky, E. (2010). Electronic resources reviews and reports. *Science and*

    Technology Librarianship. doi:10.5062/F4S46PVC. Retrieved from

    http://www.istl.org/10-summer/electronic.html

EndNote Online User Guide - Getting Started. (n.d.). Retrieved November 25,

    2012, from http://endnote.com/training/mats/enuserguide/

George Mason University. (n.d.). Zotero: The Next-Generation Research Tool.

    Center for History New Media. Retrieved December 9, 2012, from

    http://www.zotero.org

Hull, D., Pettifer, S. R., & Kell, D. B. (2008). Defrosting the Digital Library:

    Bibliographic Tools for the Next Generation Web. *PLoS Computational*

    Biology, *4*(10), e1000204. doi:10.1371/journal.pcbi.1000204

Jeon, S., Bang, J, Byun, K., & Lee, S. (2012) Recovery method of deleted record
for SQLite Database. *Personal Ubiquitous Computing,* 16(6),707–715.
Kessler, J. (2007). Zotero. *The Charleston Advisor*, *9*(2), 52–56.

Mathew Willmott, R. G. (n.d.). Research Guides. Overview of Citation Software

    at MIT: Managing Your References. Comparison of Citation Software at

    MIT. Retrieved January 13, 2013, from

    http://libguides.mit.edu/content.php?pid=55486&sid=427307

Overview | Mendeley. (2012). Retrieved November 19, 2012, from

    http://www.mendeley.com/features/

Puckett, J. (2011). *Zotero: A Guide for Librarians, Researchers, and Educators*.

    Assoc of Cllge & Rsrch Libr.

RefWorks: About Us. (2009). Retrieved November 15, 2012, from

    http://www.refworks.com/content/about_us.asp

rmzelle. (2012). searching [Zotero Documentation]. Retrieved January 10, 2013,

    from http://www.zotero.org/support/searching#fulltext_pdf_indexing

forums.Zotero.org(n.d)."moving metadata online" [admin responseto dstillman,

    Jun 22nd 2012].Retrieved From

    http://www.zotero.org/support/zotero_data

Trinoskey, J., Brahmi, F. A., & Gall, C. (2009). Zotero: A product review.

    Journal of Electronic Resources in Medical Libraries, *6*(3), 224–229.

Witten, I. H., Boddie, S. J., Bainbridge, D., & McNab, R. J. (2000). Greenstone: a

    comprehensive open-source digital library software system. In

    Proceedings of the fifth ACM conference on Digital libraries (pp. 113–

    121). Retrieved from http://dl.acm.org/citation.cfm?id=336650

Witten, Ian H., Bainbridge, D., & Boddie, S. J. (2001). Greenstone: open-source

    digital library software with end-user collection building. *Online*

    Information Review, *25*(5), 288–298. doi:10.1108/14684520110410490

Witten, Ian H., Bainbridge, D., & Nichols, D. M. (2009). *How to Build a Digital*

    Library. Morgan Kaufmann.

Appendix A: First user study documents.

A.1 Ethical approval

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

3 July 2012

Mohammed Alanazi
C/- Department of Computer Science
THE UNIVERSITY OF WAIKATO

Dear Mohammed

Request for approval to conduct an experiment study involving human participants for your Masters Degree

I have considered your request to conduct an experiment for your research project involving human participants, *Keeping track of reading material,* commencing 9th July this year. The purpose of the research is to develop a system that will help people track their reading material easily.

The procedures described in your request are acceptable.

I note your statement that no personal participant details will be recorded for the study itself, their questionnaires will be numbered sequentially. If they have given their names for contact for a summary of the results, these will not be disclosed in any resulting publications. All information/data and audio recordings will be stored in the FCMS data archive for three years and then destroyed.

The research participants' information sheet, consent form, and questionnaire meet the requirements of the University's human research ethics policies and procedures.

Yours sincerely,

La Hunt.

Lyn Hunt
Human Research Ethics Committee
Faculty of Computing and Mathematical Sciences

A.2 Questionnaire

Keeping Track of Reading Material

Questionnaire

Participant No:

Initial Questionnaire

I am Mohammed Alanazi. First of all, I would like to thank you for agreeing to participate in this research project. Your help is very much appreciated.

As previously communicated to you, the purpose of the interview is to investigate how people keep track their reading material. I'm doing this research as part of my master's degree in computer science. Please feel free to ask questions if there's anything you're not clear about.

I'd like to reiterate that the information you give will be treated confidentially and collectively, not individually. There are no right or wrong answers—I am interested in your own experiences and opinions. So please be as honest as you can in your responses.

Before starting the interview, I would appreciate if you can complete this questionnaire.

1) What is your gender?

☐ Male   ☐ Female

2) What is your age group?

☐ 18 - 28   ☐ 29 - 38   ☐ 39 - 48   ☐ > 48

3) Do you work?

☐ Yes. (If yes please, specify your occupation).   ☐ No.

Occupation: ....

....

4) Do you study?

☐ Yes. (If yes please, specify your qualification).   ☐ No.

Qualification: _____

_____

5) What is your highest education level?

☐ Nothing   ☐ High School   ☐ Diploma   ☐ Bachelor
☐ Postgraduate Diploma   ☐ Masters   ☐ PhD

6) Do you use a computer?

☐ Yes   ☐ No.

(If your answer is No, please disregard questions a & b)

a. How many hours do you spend on computer in a day?

☐ < 1   ☐ 1 - 3   ☐ 4 - 6   ☐ > 6

b. How many years have you used a computer?

☐ < 1   ☐ 1 - 4   ☐ 5 - 10   ☐ > 10

7) Do you have experience in writing documents and referencing? (If yes, please specify for how long).

☐ Yes ☐ No.

....

....

....

A.3 Interview

Keeping Track of Reading Material

Interview

Participant No: ....

The aim of this interview is to investigate people's experience with their track reading material in terms of the ways they use them and the problems they face.

1. Are you interested in reading?

_____
_____
_____

2. Do you read newspapers, magazines or novels?

_____
_____

2.1. How often: printed copy, electronic copy, kindle reader or anything else?

_____
_____
_____
_____
_____

3. Do you read books, articles, conference papers or any such material?

_____
_____
_____

3.1. How often: printed copy, electronic copy, kindle reader or anything else?

....
....
....
....
....

4. How many hours do you spend reading? (daily/weekly/monthly)

....
....
....

5. Do you keep a list of all (physical / online) materials you have read? (If yes Please, explain the methods you use).

....
....
....
....
....
....

5.1. Have you ever faced any problems with (physical / online) materials you have read? If yes please, explain.

....
....
....
....
....
....

6. Do you keep a list of (physical / online) materials you are currently reading? (If yes Please, explain the methods you use).

....
....
....

6.1. Have you ever faced any problems with (physical / online) materials you are currently reading? If yes please, explain.

....
....
....
....
....

7. Do you keep a list of (physical / online) materials you are planning to read? (If yes Please, explain the methods you use).

....
....
....
....
....

7.1. Have you ever faced any problems with (physical / online) materials you are planning to read? If yes please, explain.

....
....
....
....
....

8. Do you have a library / archives at home and office?

....
....

8.1. How many items do you have?

8.2. How do you organize them?

9. Have you ever written any material such as books, articles, conference papers? (If yes, Please list them if you can).

10. What kind of tools/ software do you use for writing documents (Word, Latex or open office)?

10.1. Which software system do you use for referencing?

....

....

....

....

....

Thank you very much for your time. It was kind of you to share with me your experiences of keeping track of reading material. Your help is greatly appreciated. If you think of anything else, or if you have any questions, please feel free to contact me.

Appendix B: second user study documents

B.1 Ethical approval

Computing and Mathematical Sciences
Rorohiko me ngā Pūtaiao Pāngarau
The University of Waikato
Private Bag 3105
Hamilton
New Zealand

Phone +64 7 838 4021
www.scms.waikato.ac.nz

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

20 August 2013

Mohammed Alanazi
C/- Department of Computer Science
THE UNIVERSITY OF WAIKATO

Dear Mohammed

Request for approval to conduct an experiment study involving human participants for your Masters Degree

I have considered your request to conduct an experiment for your research project involving human participants, *Keeping track of reading material*. The purpose of the research is to conduct a diary study with users using the system already developed to make tracking reading material easy.

The procedures described in your request are acceptable.

I note your statement that you will obtain written permission from the Head of Department to conduct this study and that your software will be tested by Technical Services before commencement.

No personal participant details will be recorded for the study itself, their questionnaires will be numbered sequentially. If they have given their names for contact for a summary of the results, these will not be disclosed in any resulting publications. All information/data and audio recordings will be stored in the FCMS data archive for three years and then destroyed.

The research participants' information sheet, consent form, and questionnaire meet the requirements of the University's human research ethics policies and procedures.

Yours sincerely,

Mike Mayo
Human Research Ethics Committee
Faculty of Computing and Mathematical Sciences

B.2 Interview before intervention

Keeping Track of Reading Material

Interview (before intervention)

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato
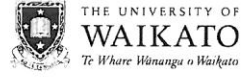
Participant No: ....

The aim of this interview is to investigate people's experience with tracking reading material and the problems they face before using TRMS.

1. Have you ever used Greenstone? If yes, for how long?

    1.1 Describe your experience? Things you like and dislike?

    1.2 Any issues you have ever encountered?

2. Have you ever used Zotero? If yes, For How long?

    2.1 Describe your experience? Things you like and dislike?

    2.2 Any issues you have ever encountered?

3   While you are doing your project or research you read digital papers related to you topic area, do you download those papers to your computer?

     3.1 If you do, how you do so?

     3.2 If you do, what kind of files do you usually download?

4   How many hours do you spend reading? (daily/weekly/monthly)

5   Do you keep track the materials you have read for your research / PhD (Never - rarely - sometimes - often - always)? If you do, please describe your experience?

     5.1. Have you ever encountered any problems? If yes please, explain.

7.1. Have you ever encountered any problems? If yes please, explain.

....

....

....

....

....

....

Thank you very much for your time. It was kind of you to share with me your experiences of keeping track of reading material. Your help is greatly appreciated. If you think of anything else, or if you have any questions, please feel free to contact me.

B.3 Diary study.

Date: _____

1- What you are trying to do?

2- Did you observe any issues?

3- How did you resolve the issue?

4- Further notes or questions :

B.4 interview after intervention

Keeping Track of Reading Material

Interview (after intervention)

THE UNIVERSITY OF
WAIKATO
Te Whare Wananga o Waikato

Participant No: ------------------------------

The aim of this interview is to investigate people's experience and the problems they faced with tracking reading material while using TRMS.

1. How long have you used TRMS?

....

....

....

2. What were your experiences using TRMS?

....

....

    2.1. What did you like particularly?

....

....

....

    2.2. What did you dislike particularly?

....

....

....

    2.3. Have you faced any problems? If yes please, explain.

....

....

....

....

....

2.4. Are there any features that were missing or that you would like to see in future?

....
....
....
....
....
....

3. Specific questions about the data the participant accumulated.

    3.1 How many papers have you added?

....
....

    3.2 How many papers that have PDF and non-PDF?

....
....

    3.3 How many papers in To Read / Currently Reading and Have Read shelves?

	To Read	Currently Reading	Have Read
PDF			
Non-PDF			

4 Did you notice any problems:

    4.1 For particular web sites?

....
....
....

    4.2 For papers with and without PDF?

....
....

4.3 For Greenstone showing paper content?

....

....

....

5   Any other comments or recommendations?

....

....

....

....

....

....

Thank you very much for your time. It was kind of you to share with me your experiences of keeping track of reading material. Your help is greatly appreciated. If you think of anything else, or if you have any questions, please feel free to contact me.