Personal digital libraries: Keeping track of academic reading material

Mohammed Al-Anazi, Annika Hinze, Nicholas Vanderschantz,
Claire Timpany, Sally Jo Cunningham

Dept. of Computer Science, University of Waikato,
Private Bag 3105, Hamilton New Zealand
msma1@students.waikato.ac.nz,
hinze,vtwoz,ctimapny,sallyjo@waikato.ac.nz

Abstract. This paper discusses options for tracking academic reading material and introduces a personal digital library solution. We combined and extended the open source projects Zotero and Greenstone such that material can be easily downloaded and ingested into the combined system. Our prototype system has been explored in a small user study.

1 Introduction

Researchers have to keep track of an increasing number of electronic publications. Not all articles and papers are read immediately, and keeping track of one's reading list is not straightforward [4]. Once a text has been evaluated, read or identified to be read later, many researcher instigate their own routine for managing the process of recording and remembering the context, location and identified need for that book or article. This paper is our first investigation into how to manage this problem for eReading environments. We report on a project that aimed to identify the issues academics encounter as they keep track of their reading material, and a software solution to support this process. We explored popular citation management software systems and rather than devising a completely new system, we instead re-use and extend available systems (the two open source projects Zotero and Greenstone) to form a personal digital library system for tracking digital reading material. The resulting prototype was evaluated in a small-scale user study.

2 User study

The study investigated the kinds of reading materials that students and academics keep track of, as well as the software or methods they use for tracking. Each participant answered questions in a questionnaire and an interview. Twenty people participated (14 male/6 female; 2 under 29, 10 under 39, 5 under 48 and three 48 and over; 14 students/6 academics; 13 from IT and 7 from other academic fields). All were highly experienced

adfa, p. 1, 2011.

with computers (>10 years) and 18 used computers more than 4 hours each day. In our evaluation here we focus on academic reading.

Reading material. 14 participants indicated that academic reading occupies about 66% to 85% of their reading materials, for two it is more and for the other four it is less. For 15 participants, more than 60% of their reading is articles and conference papers (>80% for 9 of these), and all 20 rea about 20% are books and theses. 18 of 20 read articles and papers wholly or predominantly in electronic form, and 14 read books predominantly as hardcopies. The remainder of academic reading is largely done electronically.

Tracking intention. None of the participants were interested in tracking which non-academic reading they had done or were currently engaged in, but only the reading they wished to do in future. For academic reading, all participants were interested in tracking what they *will* read and what they *had read*; 10 wish to track their current reading. All participants keep copies of electronic materials on their computers, and 10 print the materials in various circumstances. All participants noted that they tend to track their academic digital materials only when they do a project, research, or write a paper.

Tracking have-read material. Participants use three main methods to track materials they have read (some use several): folders and subfolders (19 of 20), a list (4), and software (7). No-one keeps only a list, and only one uses only software. Within folders and subfolders, documents may be sorted according to the structure of the document to be written; sometimes with a further structure according to importance. Participants indicated that not all the materials they place in their folders need be read or used – for this distinction they rely on their memory based on title or abstract. The software used is: Endnote (2), Mendeley (2), Zotero (1), NVivo (1) and Safari bookmarks (1). Each of the four participants who kept lists had their own method of what to record.

Tracking current reading. Only half of the participants keep track of current reading, of which four use software tools. Five participants leave the materials open on screen, two leave sticky notes in the office/on screen, one uses the Safari reading list, another Excel sheets and Endnote, and another one uses PDF reader software on a phone/tablet.

Tracking future-read material. Five different methods are used for keeping track of future reading material: software tools (5), download of documents (9), email with material/title or link (2), sticky notes (3), and print-outs (1). The software tools used were the online system for to-do lists Remember-the-Milk (named twice), a referencing system, reading lists in Safari, and bookmarks. Two participants also wished to track future reading but did not currently have any method for doing so. None of the participants who download documents had a specific location on their computers for storing these documents, but some reported maintaining a folder called ‘Want to read’ or similar.

Tracking Problems Encountered. A number of problems were reported about the methods used to track electronic reading materials. 10 participants mentioned that they spend quite some time looking for specific papers, and that often very little of the stored information is used. Four people reported reading a paper more than once because they do not know if it has been read or not. Papers that need to be used in two or three different projects are hard to place, replicating folders and applying different highlights (4 participants). Two participants mentioned saving a hyperlink, and the paper being gone when they returned to the link later. Two participants talked of the difficulty in identifying downloaded files by file names containing just numbers and letters. They also mentioned not finding files again, or downloading papers more than once.

Referencing material. 14 of 20 participants used software for referencing, e.g., Endnote (8), MS Word (2), and Bibtex, Jabref, Mendeley, and Zotero (1 each). 6 of 20 do their referencing manually. The participant using Bibtex was the only one writing papers in LaTex; all others write using MS Word.

We observe a discontinuity in the processing/tracking of reading materials as the participants use different methods for each stage. For previously read material most use folders and subfolders, which is not used at all for ‘currently reading’ and ‘planning to read’. It seems as if participants try different methods, but have no effective common strategy. As a consequence, several participants observed spending too much time looking for a specific paper, that often very little of the information is used, and that duplication of material and information is problematic. Further time is wasted by re-reading material unintentionally. Most participants use software tools to format references when writing, often those that integrate with the users chosen word processing software.

3 Requirements

We now define the requirements for a system for tracking academic reading, based on the user study results. The system needs to provide software features to

- R1) Download the reading material and its metadata
- R2) Store the material itself as well as its metadata
- R3) Browse the reading material
- R4) Search within the material and the metadata
- R5) Indicate the reading status of materials
- R6) Reference material when writing papers/articles
- R7) Annotate material

We will use these requirements to analyse the related work, and to design our own system (the prototype supports R1 to R5, leaving R6 and R7 for future work).

4 Related work

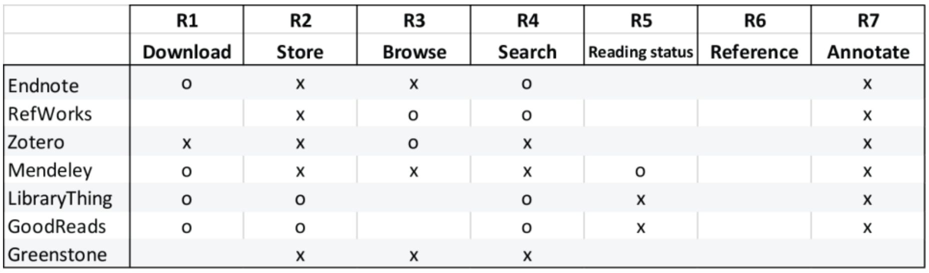
We reviewed four popular citation management systems (Endnote, RefWorks, Zotero and Mendeley), two social networking systems (LibraryThing and Goodreads), and one digital library (Greenstone).[1] Endnote is a commercial tool for managing references that can manage large libraries in a desktop system. Trinoskey et al. [5] concluded that it was particularly suitable for academic writing. Endnote fully supports R2, R3 and R7; it partially supports the R1 (download metadata only) and R4 (search in metadata only). RefWorks is a web-based system with features similar to those of Endnote. RefWorks supports R2 and R7; it partially supports R4 through search in metadata only. While RefWorks and Endnote were designed to manage citations, Zotero and Mendeley were developed to manage publications [3]. Zotero is a free open-source system that can be used as desktop software (for citations during writing) and as a Firefox browser extension (for import/download of documents). Zotero automatically recognises the records types (e.g., 'conference papers and books'), and its collections can be synchronized over multiple computers. Zotero fully supports R1, R2, R4 and R7. It partially supports R3, although one cannot browse the document through Zotero itself.

Mendeley is a free web-based system with a synchronising desktop component. Mendeley imports and organises PDFs and bibliographic citations via manual upload or metadata import from web-sources. It organises material into a folder structure. Mendeley fully supports R2, R3, R4, and R7. It partially supports R1 (metadata download) and R7 (indicating read/unread). Barsky (2010) states that Zotero and Mendeley are easier to use than Endnote and RefWorks [2]. LibraryThing is a social online service for cataloguing books. Users can add books to the collection by searching through the Library of Congress and over 695 world libraries. Any item in the collection can be tagged, reviewed, annotated, and rated, and can be shared with friends. Users can organise their collections in folders and sub folders, such as 'To read' and 'Currently reading'. LibraryThing supports R5 and R7; it partially supports R1, R2, R4 (only for metadata). Goodreads is a social online service for book recommendations and private library catalogues. Users can add books to their bookshelves by searching online or adding them manually. Users may organise the books into shelves such as 'Read', 'Currently reading', or 'Want to read'. Because Goodreads is a social network digital library like LibraryThing, it supports similar features: R5 and R7 fully and R1, R2, and R4 partially for metadata. Greenstone is an open source system 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 [6], and provides functions for accessing items (browse, search, and index). Greenstone fully supports R2, R3, and R4. The other requirements are not supported.

Summary. We identified seven requirements for software that tracks reading materials. Seven systems were reviewed according to these requirements. Our findings are summarised in Table 1, and it is apparent that none of the systems support all our requirements. The next section introduces our system design that combines and extends a combination of two of these systems.

[1] The systems are available at http://endnote.com, www.refworks.com, www.zotero.org, www.mendeley.com, www.librarything.com, www.loc.gov/index.html, www.goodreads.com

Table 1: Summary of requirements comparisons (x=fulfils requirement, o=partially fulfils requirement, empty = system does not fulfil the requirements)



	R1	R2	R3	R4	R5	R6	R7
	Download	Store	Browse	Search	Reading status	Reference	Annotate
Endnote	o	x	x	o			x
RefWorks		x	o	o			x
Zotero	x	x	o	x			x
Mendeley	o	x	x	x	o		x
LibraryThing	o	o		o	x		x
GoodReads	o	o		o	x		x
Greenstone		x	x	x			

5 System design & Prototype implementation

Here we introduce our system design from initial decisions to final prototype.

5.1 Conceptual Design

We decided to create a system that is a combination of Zotero and Greenstone, extended by additional functionalities. Both these systems are open source and freely available, and from our initial investigation it is clear that together they already fulfil a number of our requirements. In our design, we focus on requirements R1 to R5 (options for how to incorporate R6 and R7 will be discussed in Section 7). Although Greenstone and Zotero fulfil similar requirements, they follow very differently philosophies and provide different user experience:

- **R1:** users can download metadata and documents through the Zotero browser add-on. Greenstone can incorporate metadata and documents through the librarian interface, but does not support easy online incorporation.
- **R2:** In both Zotero and Greenstone, material and metadata can be stored long-term.
- **R3:** Users can browse the material full-texts as well as the metadata in greenstone. Zotero keeps the full-texts but does not support browsing through its software.
- **R4:** Both systems support search in metadata. Greenstone provides search in full-texts and indicates results within the full-texts.
- **R5:** Neither system directly supports indication of the reading status, but Greenstone can hold additional indexes and folders that can be extended for this purpose.

Zotero provides easy incorporation into web-search and automatic meta-data analysis, while Greenstone provides support for browsing, full-text search and easy extension for indexes. In our combined system, we use Greenstone as the document management system and Zotero as the metadata and document provider. We designed additional data transfer modules from Zotero into Greenstone, and extended Greenstone to capture information about the current reading status of documents. The PDF full-texts and its metadata usually reside on a website, and users capture these using Zotero. The system then imports this data into the Greenstone collection, and tracks the reading status.

5.2 Implementation

Our system has several modules as shown in Figure 2. Electronic articles and papers located on websites often have two components, i.e., PDFs and metadata (grey and yellow elements in Figure 2). Elements of Zotero are indicated in red, parts of Greenstone are shown in green and our system extensions are shown in blue.

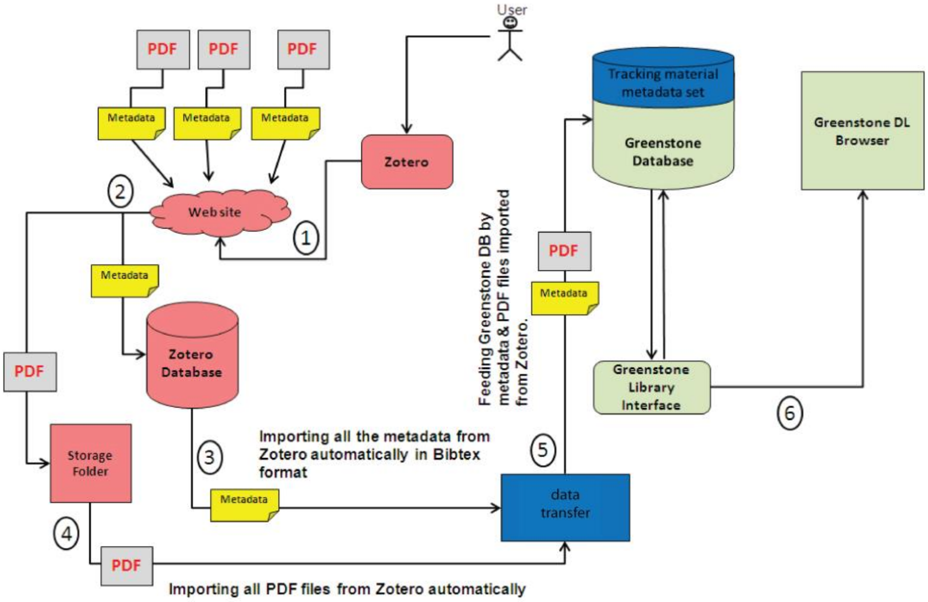
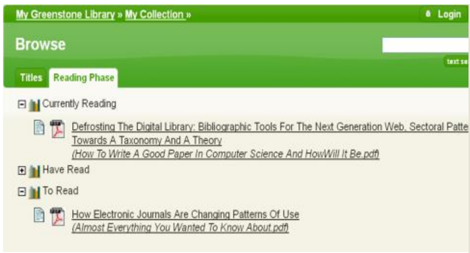

User
PDF
PDF
PDF
Metadata
Metadata
Metadata
Zotero
Web site
1
2
Metadata
Zotero Database
PDF
Storage Folder
3
4
Importing all the metadata from Zotero automatically in Bibtex format
Metadata
PDF
Importing all PDF files from Zotero automatically
data transfer
5
Feeding Greenstone DB by metadata & PDF files imported from Zotero.
PDF
Metadata
Tracking material metadata set
Greenstone Database
Greenstone Library Interface
Greenstone DL Browser
6

Figure 2: System architecture and walk-through

We now explain the working of the system architecture using a walk-through (numbers in steps are also shown in the architecture).

(1) Using the Zotero tool, the user adds records from websites.
(2) Zotero stores the metadata of the records added (type of source, title, author...), and available source files (e.g., PDFs or Web pages) in a ‘Storage Folder’.
(3) Our system imports the metadata automatically in Bibtex format from Zotero.
(4) Our system imports the metadata files automatically from Zotero.
(5) Our system 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.

The system is now ready to track reading information. Initially all reading material is classified as “to read”, later to be changed into “reading” or “have read”. Indexes hold the information for each of the three types. Figure 3 shows the interface for the reading phase. The example shows one document each in “to read” and in “have read”. Changing the reading phase of a document can be done for each document separately (see Figure 4). Further details about the system implementation can be found in [1].



⊟ Currently Reading

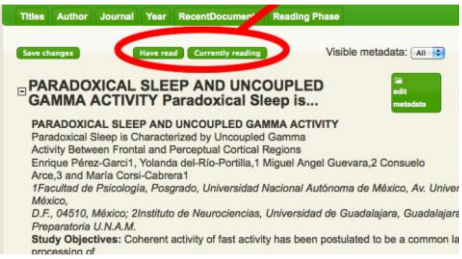
Defrosting The Digital Library: Bibliographic Tools For The Next Generation Web, Sectoral Patte
Towards A Taxonomy And A Theory
(How To Write A Good Paper In Computer Science And HowWill It Be.pdf)

⊞ Have Read

⊟ To Read

How Electronic Journals Are Changing Patterns Of Use
(Almost Everything You Wanted To Know About.pdf)

Figure 3: Reading phase overview



⊟ PARADOXICAL SLEEP AND UNCOUPLED GAMMA ACTIVITY Paradoxical Sleep is...

PARADOXICAL SLEEP AND UNCOUPLED GAMMA ACTIVITY
Paradoxical Sleep is Characterized by Uncoupled Gamma
Activity Between Frontal and Perceptual Cortical Regions
Enrique Pérez-Garci1, Yolanda del-Río-Portilla,1 Miguel Angel Guevara,2 Consuelo
Arce,3 and María Corsi-Cabrera1
1Facultad de Psicología, Posgrado, Universidad Nacional Autónoma de México, Av. Univer
México,
D.F., 04510, México; 2Instituto de Neurociencias, Universidad de Guadalajara, Guadalajara
Preparatoria U.N.A.M.
Study Objectives: Coherent activity of fast activity has been postulated to be a common la
processing of

Figure 4: changing the reading phase

6 Evaluation: Exploratory User study

Our exploratory user study consisted of three phases: an interview before using the software, a diary study while using the software, and an interview after using it. We recruited four participants and installed our software on their computers, encouraging them to use the software and to fill in a diary about their experiences. All participants were from Computer Science (three faculty members and one Master's student).

Participant background. Even though three participants had previously used the Greenstone interface, none of them had any experience with building collections in Greenstone (i.e., the librarian interface). Two participants had previously used Zotero for a long time (P1 and P2), while the other two had never used it. All participants were experienced in downloading papers for their research. For tracking reading material, P1 and P2 use a folder structure (to read/have read), while P3 accumulates electronic papers for writing articles but does not have any specific structure. P4 uses a folder structure that assigns topics once a paper has been read. None of the participants tracked their current reading. P3 and P4 do not track papers that they wish to read in the future.

Observations while using the system. ll four participants added material to the system: three participants encountered no problems but P2 felt that there were too many steps involved in successfully adding an item. P2 and P4 also added items that did not come with PDF full-texts. For other items, they observed that reading a PDF in Greenstone is not always convenient, and they instead opted to use an external PDF viewer. Three participants (P1, P2 and P3) wished to not need to click the "add item" button for each item. P1 suggested that the system should detect the reading of material automatically, and adjust the reading phase information. P3 wished to group items in the way they are structured in Zotero and asked for this feature to be added to the system. P1 wanted to be able to mark the reading place in the electronic document so they could come back to it later on and finish reading; similarly P4 wanted to bookmark places in the documents. Missing PDFs or inconvenient display of PDFs was commented on by all participants. Additionally, they wished to convert the webpages into PDF (P1), and reformat the way that Greenstone displays the PDF (P2). P1 also suggested connecting the system with Google scholar to download items more easily.

Discussion: P1 and P2, who were familiar with Zotero, organized their downloaded papers into folders and used tags, while P3 and P4 put all items into one folder. All four felt the system addressed a lack they had noticed in their own strategies. P1 and P2 particularly liked the reading phase feature (R5), as it was useful to know how many papers they had read or have started reading. P3 and P4 mentioned the automated adding of items and metadata feature (R1), and liked how a file is connected to its metadata. We believe that the differences (as much as anything can be concluded from such a small sample size) are due to the participants' different prior experiences in using a referencing system. This needs to be explored further.

7 Discussion & Conclusion

This paper explored ways keeping track of electronic reading materials. We analysed popular citation management systems as related work, and used a combination of Zotero and Greenstone to create a personal digital library system for tracking electronic reading material. The resulting prototype was evaluated in a small-scale user study. There are a number of necessary improvements that became apparent even from our very small exploratory study. For example, the downloading of metadata and material should support ingest of several documents at one time, instead of a one-by-one approach. Within the system, better organization should be supported by tags and folders. The formatting of articles needs to be more comfortable for reading, and the metadata needs to be displayed alongside the text. We conclude that our system fulfilled its basic requirements, but better user support and more elaborate methods for organizing content are needed. Future work in this project is manifold: once some of the elementary difficulties have been addressed, the system needs to be explored in a larger user study. More complex extensions may include an automated document ingest with transparent apparent internal workings, automatic detection of "current readings", and noting the last accessed position within documents. Finally, we wish to explore requirements R6 (citing from collections while writing), and R7 (full-text annotations of documents).

References

[1] Al-Anazi, M.S. (2014) *Keeping Track of Electronic Reading Material*, Master's Thesis, Computer Science Department, University of Waikato,

[2] Barsky, Eugene (2010) "Electronic Resources Reviews and Reports." Issues in Science and Technology Libraries 62

[3] Butros, A., and Taylor, S. (2010). "Managing information: evaluating and selecting citation management software" Proceedings of the 36th IAMSLIC Annual Conference. 1-27

[4] Hinze, A., McKay, D., et al. (2012). "Book selection behavior in the physical library: implications for ebook collections". *Proceedings of the JCDL,* 305-314

[5] Trinoskey J, Brahmi FA, Gall C. (2009) Zotero: A Product Review. Journal of Electronic Resources in Medical Libraries. 6(3):224-9.

[6] Witten, I., Bainbridge, D., & Boddie, S. J. (2001). Greenstone: open-source digital library software with end-user collection building. *Online Information Review, 25*(5), 288–298.