

BOOK DIGITIZATION: A PRACTICAL AND URGENT NECESSITY  
FOR THE WISCONSIN EVANGELICAL LUTHERAN SYNOD

by

Seth A. Georgson


A Senior Thesis and Project Submitted to

Wisconsin Lutheran Seminary

in Partial Fulfillment of the Requirements for  
the Master of Divinity degree

Professor John P. Hartwig, Advisor

Approved at Mequon, Wisconsin, on 3/29/2012



\_\_\_\_\_  
Advisor's Signature

BOOK DIGITIZATION: A PRACTICAL AND URGENT NECESSITY FOR THE  
WISCONSIN EVANGELICAL LUTHERAN SYNOD

BY  
SETH A. GEORGSON

A THESIS SUBMITTED TO THE FACULTY IN PARTIAL FULFILLMENT OF THE  
REQUIREMENTS FOR THE DEGREE OF MASTER OF DIVINITY

PROF. JOHN P. HARTWIG, ADVISOR  
WISCONSIN LUTHERAN SEMINARY  
MEQUON, WISCONSIN  
MARCH 28, 2012

**ADVISOR APPROVAL PAGE**

## **ABSTRACT**

More than a decade into the 21<sup>st</sup> century, life is becoming ever more digitized. A major part of this movement is the digitization of print literature. This paper examines digital libraries around the globe and discusses how current technology and methods might be applied to the Wisconsin Evangelical Lutheran Synod (WELS), especially concerning the rare books kept at Wisconsin Lutheran Seminary (WLS).

This study intends to be an overview to aid in the starting of a digitization initiative at WLS. Current digital libraries are examined for reference in their various purposes, formats, and Internet presence. Commercial digitization equipment is reviewed, with a look at various methods and price ranges. The nuts and bolts of digitization are taken apart with a brief overview of file formats and metadata. Finally, the question is asked, “How might this be accomplished for WELS?”

The conclusion of this research is that it is fully possible for WELS to create a digital library using limited resources. In order to best preserve the rich history of WELS and advance scholarship of theologians both at the seminary and far away, a digitization initiative should be considered an urgent necessity. This paper coincides with the building of a book scanner for the WLS library.

## CONTENTS

ADVISOR APPROVAL PAGE .....	i
ABSTRACT.....	ii
CONTENTS.....	iii
INTRODUCTION .....	1
CHAPTER 1 CURRENT DIGITIZATION PROJECTS .....	5
A. The Library of Congress .....	5
B. The Digital Library for Dutch Literature .....	7
C. Project Gutenberg.....	8
D. Google Books.....	10
E. The Internet Archive and the Open Library .....	13
F. Digital Libraries of Major Church Bodies.....	15
G. WLS Essay File.....	16
CHAPTER 2 DIGITAL SCANNING EQUIPMENT .....	18
A. A Brief History of Book Digitization.....	18
1. Keying.....	18
2. Flatbed Scanners .....	19
3. Book Scanners .....	19
B. The Atiz BookDrive .....	21
C. The Treventus ScanRobot .....	21
D. Kirtas Technologies, Inc. ....	22
E. The Ion Audio Book Saver.....	23
F. Summary of Commercial Scanning Equipment .....	23
CHAPTER 3 DIGITAL FORMATTING.....	25

A. NARA Guidelines for Digital Archives .....	26
B. File formats for Digital Books .....	26
1. XML.....	27
2. EPUB .....	27
3. DjVu.....	28
4. PDF .....	28
C. Image Quality.....	29
1. Understanding Pixels from Camera to Digital Image.....	29
2. Recommended Archival Image Resolutions.....	31
D. Metadata.....	32
CHAPTER 4 SCANNING THE WLS RARE BOOKS LIBRARY .....	34
A. The DIY Book Scanning Community.....	34
B. The Basic Parts of a DIY Book Scanner .....	34
C. The WLS Book Scanner.....	35
D. Considerations for a WELS Digitization Initiative.....	36
1. Recommendations.....	37
2. A Sample Digitization Workflow.....	38
CONCLUSIONS.....	40
BIBLIOGRAPHY.....	42
APPENDIX I: BOOK SCANNER IMAGES .....	48
APPENDIX II: SEMINARY BOOK SCANNER OPERATOR’S MANUAL.....	51
A. Overview .....	51
1. Purpose and Concept .....	51

2. Parts .....	51
B. Operation.....	53
1. Operation Summary.....	53
2. Cameras .....	53
3. Electrical System .....	54
4. Mechanical System .....	55
C. Maintenance .....	56
1. Cleaning.....	56
2. Moving.....	56

## INTRODUCTION

There are certain developments that stand out as landmarks in the history of communication. Written language, the movable-type printing press, the telephone, email—these inventions shifted the paradigms of how people transmit information. We are currently experiencing another paradigm shift. Books are being written and published using only two characters: 0 and 1. We are living in an age where digital storage space is out-pacing the number of things we have readily available to store.<sup>1</sup> With the invention of the e-reader and the tablet computer, literature has begun a rapid shift to a digital platform. This is the first major shift of the book medium since the codex replaced the scroll under the Roman Empire. Some people are even predicting that physical books will soon go the way of the telegraph and the hand-written letter. While this is debatable, there is no denying the many benefits of digital books. They are easily transported, easily distributed, easily cataloged, perfectly preserved, and the text can be quickly copied and searched.

As with any paradigm shift, people are wrestling with how to adapt to and utilize the available technology. Some authors and publishers, fearing that their works will be copied without permission, are refusing to change the way they do things. Others are experimenting with closed digital systems, such as that of Logos Bible Software. Libraries are looking for ways to convert millions of pages of literature into digital formats. People are collecting libraries for their e-readers. The Google Books project has digitized millions of books in only a few years' time.

The Wisconsin Evangelical Lutheran Synod (WELS) recognizes the value of this technology and is also making the move to digital literature. The seminary now uses laptops in the classrooms for Bible study and for notes. The seminary's website hosts thousands of essays in digital format. A large number of documents are published on WELS Connect.

There is one conspicuous hole, however, in the landscape of digital literature in WELS. Locked in the basement of Wisconsin Lutheran Seminary lie thousands of dollars' worth of rare

---

<sup>1</sup> While higher definition videos and audio still require substantial space, trends show that fewer people are maxing out the space available to them. Lower demand for space has resulted in slower growth of consumer hard drives. "Drives have failed to meet the 54% level [of growth] for the last 3 years and they are also not expected to meet it for the next 3 years." (Kryder's Law: A Rule of Thumb for Hard Drive Growth. available at <http://www.mattscomputertrends.com/Kryder%27s.html>. Internet. accessed Feb 16, 2012).



books. These books certainly are worth taking special care of. Many of them remain untranslated. Many more would be difficult—if not impossible—to replace. The same can be said of many church archives across the Wisconsin Evangelical Lutheran Synod, archives that are often kept in church basements where accessibility is limited and any number of accidents or the simple test of time could permanently destroy a valuable piece of history.

This concern is not a case of fear mongering. The rare books room at the seminary library is no high-tech vault. The lights are standard fluorescent bulbs, which, while low in damaging ultraviolet (UV) radiation, are not UV-free. The humidity is kept down with simple home dehumidifiers. These measures are not enough to protect these books, as evidenced by the occurrence of mold on the books. The mold can be removed, but the process of cleaning the books is also damaging to them.

All these problems cannot be averted without a significant budget, but there are steps that can be taken to preserve the historical literature of WELS at the seminary and across the synod. The most significant step that can be taken is an effort to digitize the rare books and archives of the Wisconsin Evangelical Lutheran Synod.

Librarians and archivists all around the world are undertaking digitization<sup>2</sup> projects to preserve and distribute literature. Much information can be learned from their successes and failures. What have they done well? What needs improvement? These are issues that need to be considered.

Unfortunately, despite rapidly falling technology costs, commercial book-scanning equipment is still out of the reach of most individuals and institutions. Scanners can easily cost \$10,000 or more. On top of that, these machines are not always gentle with the books they scan. Some press books flat and in doing so they damage the bindings. Robotic arms can tear pages. Standard commercial scanners don't accommodate thick bindings or oversized pages. Some commercial machines require the buyer to purchase cameras separately and they are often limited to certain kinds of cameras. Image quality and resolution is therefore limited.

---

<sup>2</sup> Strictly speaking, “digitization” refers to the creation of a digital work from one that is not digital, while a digital library is the place where these digital works are stored. Since a digital library is nothing without its books and a digitized book will only be lost or go unused without a place to store it and access it from, digitization and a digital library are considered to be two parts of one single project in this paper. In order for one to be done well, the other must also be done well.

These all appear to be difficult obstacles, especially considering that WELS has not historically budgeted much for the preservation of archives.<sup>3</sup> However, around the globe, people who call themselves "do-it-yourselfers" are finding ways to scan their books on a budget. They join together in online communities to share their ideas and advice.<sup>4</sup> The methods range from simple to complex, from cardboard to machined metal. They design scanners to accommodate newspapers, magazines, spiral-bound notebooks, and hardcover books of all sizes and types. They experiment with a wide variety of materials and methods and share their experiences for the betterment of the community.

This experience and this community can be used to bring affordable book digitization to WELS. Digitizing books is not something that only needs to be talked about. It is something that needs to be done. Digital literature is not a technology that is going to disappear. The quality of WELS rare books and archives is not improving. The sooner WELS can begin the work of preserving these books in digital format, the better.

Not only is digitization an excellent way to preserve books, it is also a way of making them accessible. Currently, access to the rare books at the seminary involves being on location and having an appointment with an authorized person. Few people are allowed to take books out with them and making copies on a photocopier can be extremely hard on a book.

Digital copies of books can be copied infinitely without any loss in quality. They can be printed and adapted to different formats. They can also be read directly in their digital formats. E-readers have been available for a few years and more recently the rise of affordable, quality tablet computers makes reading digital literature not only possible, but practical.

Finally, digital literature is able to be accessed over the Internet. This is an enormous advantage over physical books. Physical books must be located and either checked out of a library or bought. With newer books this is not a problem and in fact many people prefer a physical copy for various reasons. However, with old and rare books it is not so easy to obtain a

---

<sup>3</sup> For the past several years, there has actually been nothing in the budget for the archives.

<sup>4</sup> Description on one website: "We are a community of people crazy enough to build our own book scanners. We also write Free software for book scanning. We are the missing link between your bookshelf and your e-reader. Join us! Get involved by trying a simple scanner, or push the limits of scanning technology." (*DIY Book Scanner*. diybookscanner.org. Internet. accessed Feb 29, 2012.)

physical copy. Even if they are found, the owners often market them at a high price as antiques.<sup>5</sup>

The Wisconsin Evangelical Lutheran Synod can avoid many problems and gain many advantages by making an earnest effort to create a digital library. The goal of this project is to explore the viability of beginning this effort on a small budget. For this paper, I will look briefly into current digitization projects around the globe to see what might be gained from their experience. I will then discuss a number of popular commercial scanning solutions and my thoughts on them. I will give an overview of the accepted standards of a digital library and summarize the necessary attributes and capabilities a WELS digital library would have. Then I will document the process of actually building a budget book scanner and provide insights gained through that experience.

---

<sup>5</sup> Browsing eBay will yield many results for antique Christian books, most of which are Bibles. Few of the things listed are worth looking at, but the truly rare ones can usually be noticed by their price. On 2/8/2012 one seller had a 1592 Geneva Bible listed for \$4,245.75. Another had a 1629 first edition King James Bible in "museum quality" listed for \$29,000.00.

## CHAPTER 1

### CURRENT DIGITIZATION PROJECTS

Massive digitization projects are currently being undertaken in many places around the globe.<sup>6</sup> Libraries need to convert millions of books to digital formats. They are pioneering various technologies and working to standardize methods and formats. A survey of some of these projects will be helpful to starting a digitization project for WELS. Since a project like this will take many hours of labor to complete, it is imperative that the people involved know what they are doing so that they do not have to repeat work any more than is necessary. The successes and failures of others are an excellent place to gain this knowledge.

This research alone could fill a book but will be limited to a brief overview of the purposes and methods that some of these libraries are using and how they might affect the project at hand. Much of this information is publicly available and will serve as a sort of standard to which WELS can compare goals and methods.

#### **The Library of Congress**

An easy place to start looking is the Library of Congress. On the Library of Congress website ([www.loc.gov](http://www.loc.gov)) there is a link to "Digital Collections." The description on their page is:

The Library of Congress has made digitized versions of collection materials available online since 1994, concentrating on its most rare collections and those unavailable anywhere else. The following services are your gateway to a growing treasury of digitized photographs, manuscripts, maps, sound recordings, motion pictures, and books, as well as "born digital" materials such as Web sites. In addition, the Library maintains and promotes the use of digital library standards and provides online research and reference services.

The Library provides one of the largest bodies of noncommercial high-quality content on the Internet. By providing these materials online, those who may never come to Washington can gain access to the treasures of the nation's library. Such online access also helps preserve rare materials that may be too fragile to handle.<sup>7</sup>

The Library of Congress is a member of the Digital Library Federation (DLF) and has

---

<sup>6</sup> Karen Coyle. "Mass Digitization of Books." *Journal of Academic Librarianship*. Vol. 32, #6. available from <http://www.kcoyle.net/jal-32-6.html>. Internet. accessed Feb 11, 2012.

<sup>7</sup> "About Digital Collections." *The Library of Congress*. available from <http://www.loc.gov/library/about-digital.html>. Internet. accessed Feb 11, 2012.

made available a number of documents on standards for file formatting and for metadata. It was an early promoter of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). The OAI-PMH is intended to make gathering information simple across many different libraries and resources. It defines a framework for searching and accessing data over the Internet.

The Library of Congress does not have much information on the methods used for digitizing literature, but they have a wealth of information on formats, access, and preservation. They do have some contract information posted and according to one document from 1996, "Most collections are digitized by contractors who specialize in various types of originals: unbound paper, bound paper, searchable texts, moving images, still-pictorial images, microfilmed documents, sound recordings."<sup>8</sup>

The online interface for the Library of Congress could use some improvement. It appears to be categorized well but actually accessing various works is difficult. Different types of materials are jumbled together without any clear indicator to show if the item is a catalog entry, a digital book, an audio recording, or a photograph. Titles and descriptions of search results do not always seem connected to the keywords put in the search box. The site is strictly library-like in content, with no reviews or recommendations anywhere, no social aspects of any sort. This means that a person must base searches solely on keywords and titles and cannot expect recommendations from the Internet community. Some pages are very basic in appearance and the styles and general layout look rather dated. Text follows one left margin and titles in all capitals and bold blue underlined links abound. This may seem to be a small issue, but a poorly designed layout not only lacks aesthetics but can also make important information more difficult to locate quickly. Ultimately the Library of Congress has a wealth of information and resources, but the use of it is not terribly intuitive.

An important point to take away from the Library of Congress website is the necessity of keeping the interface for a digital library up-to-date. Their website may have been state-of-the-art at one point, but now it feels unnecessarily complex, especially when compared to other digital libraries. People will be more interested in using a website that is intuitive and visually appealing.

---

<sup>8</sup> Carl Fleischhauer. "Steps in the Digitization Process." January 1996. available from <http://lcweb2.loc.gov/ammem/award/docs/stepsdig.html>. Internet. accessed Feb 22, 2012.

## The Digital Library for Dutch Literature

Another national library system that is worth looking at is the Digital Library for Dutch Literature (*Digitale Bibliotheek voor de Nederlandse Letteren*, DBNL) because their initiative is somewhat newer, the work of digitization having been begun in 2003.

René van Stipriaan, the editor in chief, said in an interview done in the spring of 2009<sup>9</sup> that the planning had begun in 1998, but that interest was low for a couple years before funding was available. At the time of the interview however, they had over 1.5 million pages available online and logged between 100,000 and 140,000 pages downloaded daily. Van Stipriaan noted, however, that there was still much work to be done, having just contracted out another 10 million pages to be digitized.

The DBNL is not doing any of the work of digitizing themselves but is contracting with a company named SPi, located in the Philippines. SPi advertises itself as having 1,500 "digitization professionals" and a 120-person technology team.<sup>10</sup> They offer a wide range of services—including digitizing, editing, proofreading, and indexing—and formats, including high resolution, low resolution, text-only pages, and images. Originally the DBNL contracted with them for text-only digitization encoded in XML (*eXtensible Markup Language*), but more recently they began using searchable PDFs (*Portable Document Format*). This allows them to preserve the look of the original pages along with the text. Van Stipriaan admitted that when they began to work together quality was below an acceptable level, but they cooperated to design a better workflow that would catch more mistakes.

The DBNL has a very nice website ([www.dbnl.org](http://www.dbnl.org)) with the small problem for this writer of its being in Dutch. The interface is much more visually appealing than the Library of Congress and the home page has space for new or recommended works. The result of the carefully thought-out layout is that it is not too difficult to navigate, even in a foreign language. Users can tell where they are on the site by the images and colors of the website even without the descriptive text. It is unfortunate that many of the works are in text-only format because some

---

<sup>9</sup> René van Stipriaan, interview with. "Digitization helps future proof Dutch literature." *Research Information*, April/May 2009. available from [http://www.researchinformation.info/features/feature.php?feature\\_id=212](http://www.researchinformation.info/features/feature.php?feature_id=212). Internet. accessed Feb 29, 2012.

<sup>10</sup> *SPI Global: Digitization Services*. available from <http://www.spi-global.com/content-solutions/our-services/digitization-services>. Internet. accessed Feb 29, 2012.

people might enjoy seeing images of the original pages especially on the centuries-old books they have.

### **Project Gutenberg**

While national efforts to create digital libraries have access to many resources, private and community-driven online-only libraries might have more drive to do their job well since their existence depends on their success. They may also have an advantage over public libraries in that they often have a higher percentage of people who are qualified to handle the digital side of things.<sup>11</sup> The difference is in the starting point. Physical libraries are built around physical books and staff people who handle physical books. When they begin to work with digital books it is not as high of a priority for the whole system as caring for the physical books. They may also have to overcome hurdles with the people involved, whose expertise doesn't necessarily lie in the digital world. Managing physical bookshelves is quite different from organizing and maintaining files on computer hard drives and a transition will require training.<sup>12</sup>

On the contrary, online-only libraries are often started by people who have a passion for both computers and books. Many of them are volunteer-based and so only people who truly care about making information available to the greater community will use their time and effort to contribute. Others exist in some way to make a profit and so they must find ways to innovate and make their services attractive to gain users.

One of the oldest digitization initiatives is Project Gutenberg. Michael Hart (1947-2011) began this project in 1971 when he was given a wealth of access time<sup>13</sup> to the Xerox Sigma V

---

<sup>11</sup> One survey of digital librarians at libraries found that only 15% had a background in computer science. Youngok Choi & Edie Rasmussen. "What is Needed to Educate Future Digital Librarians." *D-Lib Magazine*. Vol 12, No. 9. 9/2006. available from <http://www.dlib.org/dlib/september06/choi/09choi.html>. Internet. accessed Feb 29, 2012.

<sup>12</sup> Margaret Hedstrom. "Are We Ready for New Skills Yet?" *New Skills for a Digital Era*. ed. By Richard Pearce-Moses and Susan E. Davis. (Society of American Archivists, Chicago, IL. 2008): 35.

<sup>13</sup> Time spent with a research mainframe computer was expensive in 1971. Michael Hart was given an operator's account in order to become proficient at using the computer, and the account came with "\$100,000,000 worth of access time." The biography does not record how many hours this equates to, but Hart clearly considered it a great gift and felt the need to use the time to do something worthwhile with it. He believed the development of the e-book was worth that \$100,000,000 because "a copy of the Declaration of Independence would eventually be an electronic fixture in the computer libraries of 100,000,000 of the computer users of the future." While it is hard to say how many people keep a copy of the Declaration of Independence on their hard drives, at that time Hart apparently did not foresee the day when over 2 billion Internet users could freely access it and other e-books.

mainframe at the Materials Research Lab at the University of Illinois. Project Gutenberg's website claims Michael Hart as the inventor of the e-book, as he was the first person to record a published document in a digital format. One of his first projects in 1971 was typing in the Declaration of Independence. He declared that "the greatest value created by computers would not be computing, but would be the storage, retrieval, and searching of what was stored in our libraries."<sup>14</sup>

Project Gutenberg has a somewhat different approach than many of the newer digital libraries. They purposely use only the most basic text format, which they call "plain vanilla ASCII." What they mean by this is that it is encoded in the simplest format possible in order to be compatible with the largest possible percentage of machines. There are a few images here and there, but the pages are all entered in plain text and formatting is even removed so that italics, bolds, and underlines are all converted to all capital letters. They claim that 99% of computers are able to read and search the books and that any additions to the format will reduce that number.

The Project Gutenberg library holds over 36,000 books, which is a smaller number than many digital libraries but is still a respectable number especially considering their strict formatting guidelines. Their compatibility requirements result in the website lacking in visual appeal since attractive graphics and complex web coding might not work well on all machines. However, in spite of the basic interface it is a simple matter to locate books or search for content across all the books in the library.

The best lesson that anyone can learn from Project Gutenberg is the importance of compatibility. Ideally, a digital library will contain multiple formats in order to be usable by the largest possible percentage of people. The people who create and manage the digital library ought to test their system on as many machines as possible to ensure functionality. And finally, while visual appeal is a plus for a web interface, sometimes compromises might have to be made

---

*The History and Philosophy of Project Gutenberg* by Michael Hart. August 1992. available from [http://www.gutenberg.org/wiki/Gutenberg:The\\_History\\_and\\_Philosophy\\_of\\_Project\\_Gutenberg\\_by\\_Michael\\_Hart](http://www.gutenberg.org/wiki/Gutenberg:The_History_and_Philosophy_of_Project_Gutenberg_by_Michael_Hart). Internet. accessed Feb 20, 2012.

*Internet World Stats: Usage and Population Statistics*. available from <http://www.internetworldstats.com/stats.htm>. Internet. accessed Feb 12, 2012.

<sup>14</sup> *The History and Philosophy of Project Gutenberg* by Michael Hart.



for the sake of older machines.

### Google Books

A newer but nevertheless massive digitization effort has been taken on by Internet superpower Google. Google Books has only become well-known in the last few years, but it actually has a history that goes back to the mid-1990s, before Google existed.<sup>15</sup> In fact the search engine that would eventually power Google was first conceived as a way of indexing and searching books.

Google Books recounts the history:

In 1996, Google co-founders Sergey Brin and Larry Page were graduate computer science students working on a research project supported by the Stanford Digital Library Technologies Project. Their goal was to make digital libraries work, and their big idea was as follows: in a future world in which vast collections of books are digitized, people would use a "web crawler" to index the books' content and analyze the connections between them, determining any given book's relevance and usefulness by tracking the number and quality of citations from other books.<sup>16</sup>

In 2002, Larry Page and a handful of other people officially began a secret project to digitally scan "every book in the world."<sup>17</sup> As part of his research, Page traveled to the University of Michigan, a pioneer in digitization with the JSTOR (Journal STORage) online journal archive, where he was told that the U of M estimated that it would take 1,000 years to scan their seven million volume library. Page told the university that he believed Google could do it in six.

Page and his team began doing research on non-destructive methods of scanning books and using software to automatically convert text in images into searchable text. In 2005 Google made a \$3 million donation to the digitization project at the Library of Congress and officially named their project Google Books.

---

<sup>15</sup> The founders of Google, Larry Page and Sergey Brin, began working together in 1995, but the name "Google" didn't come into play until 1997. *Google History*. available from <http://www.google.com/about/company/history.html>. Internet. accessed Feb 16, 2012.

<sup>16</sup> *About Google Books: History of Google Books*. available from <http://www.google.com/googlebooks/history.html>. Internet. accessed Feb 16, 2012.

<sup>17</sup> *Ibid.*

Today, Google Books has not disclosed the number of books in their library, but it is well into the millions. They didn't make their goal of scanning all of the University of Michigan's books in six years, but the university did announce that one million books had been scanned in February of 2008. Google Books is growing fast and has digitization contracts with a number of different universities, including Princeton, Stanford, Harvard, Oxford, the University of California, the University of Texas, and the University of Wisconsin.<sup>18</sup> An article from 2007 estimated that they are processing 10 million books per year.<sup>19</sup>

Google has not been eager to share much information on their methods of digitization, but in 2009 they received U.S. Patent #7,508,978, for "a system and method [to] locate a central groove in a document such as a book, magazine, or catalog."<sup>20</sup> The system is designed to automatically detect the "central groove" or binding of the book in order to separate the two pages. According to the patent, they use an infrared system to detect the shape of the book so that they do not have to use any sort of platen to hold the pages flat. Software then de-warps the images before they are saved. Due to the vast number of books being scanned, some revealing mistakes have been found which show that the pages of their books are turned by hand. For example, a book called "Rules for regulating the subscription library at Stamford; and a list of the committee, subscribers, &c. to which is added, a catalogue of the books in the library at its first opening in February 1787"<sup>21</sup> has numerous glaring errors throughout that show gloved hands holding the book open. A blog called "The Art of Google Books" frequently posts images of uncorrected mistakes.<sup>22</sup>

---

<sup>18</sup> Sophia Jih. "University and Google Books Move Forward with Digitization." *The Daily Princetonian*. May 8, 2010. available from <http://www.dailyprincetonian.com/2010/04/08/25772>. Internet. accessed Feb 16, 2012.

<sup>19</sup> Daniel Pudles. "The Future of Books: Not Bound by Anything." *The Economist*. March 22, 2007. available from [http://www.economist.com/node/8881446?story\\_id=8881446](http://www.economist.com/node/8881446?story_id=8881446). Internet. Feb 16, 2012.

<sup>20</sup> Francois-Marie Lefevre and Marin Saric. 2009. Detection of grooves in scanned images. US Patent 7,508,978, filed September 13, 2004, and issued March 24, 2009.

<sup>21</sup> "Rules for regulating the subscription library at Stamford; and a list of the committee, subscribers, &c. to which is added, a catalogue of the books in the library at its first opening in February 1787." Newcomb and Peat, 1787. available from [http://books.google.com/books?id=L3pbAAAAQAAJ&printsec=frontcover&source=gbs\\_ge\\_summary\\_r&cad=0#v=onepage&q&f=false](http://books.google.com/books?id=L3pbAAAAQAAJ&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false). Internet. accessed Feb 16, 2012.

<sup>22</sup> "The Art of Google Books." available from <http://theartofgooglebooks.tumblr.com>. Internet. accessed Feb 16, 2012.

Google is king of interfaces and is known in all their services for their clean, minimalistic, and tasteful design.<sup>23</sup> They have an enormous budget and since the majority of their income is advertisement revenue from people using their services, they have a strong incentive to make their web pages attractive.<sup>24</sup> Google Books is organized into "bookshelves" where a user can save books for different reasons. Default bookshelves include "Favorites," "Reading now," and "Have read." More bookshelves can be added or removed as needed. EBooks are viewable in plain text, in the original scanned images, or both.

Google Books has several extremely useful and intuitive tools. Citations are easy to make and hyperlinks can be created and shared online with only a few clicks. Text can be highlighted and searches show the image of the page with the searched-for text highlighted but also showing the surrounding context.

Google Books has both free, out-of-copyright books and copyrighted books that are available for a price. These are cleanly separated so that the searcher doesn't accidentally stumble into books that must be bought when looking for free resources. Searches can include all books or can be narrowed to free books only. The Google eBookstore has its own separate page, but is linked to the regular Google Books page for convenience and books bought there will appear on the user's digital bookshelves.

Possibly the most interesting part of Google Books is the way it is connected to Google's social network. Through Google+, the Google Books users can let their friends know what books they are reading and give recommendations. This is an extremely useful thing to have, especially with an enormous library like the one Google Books has. Between this feature and a nice review system, it is much easier to find books that are worth taking the time to read.

---

<sup>23</sup> In a study of the development of the web, Tim O'Reilly remarks, "If Netscape was the standard bearer for Web 1.0, Google is most certainly the standard bearer for Web 2.0" "What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software." *Communications & Strategies*. No. 1, p. 19, First Quarter 2007.

<sup>24</sup> In 2009, Google's total revenue was 23.6 billion, 22.9 billion of which came through advertising revenues (97%) and in 2010 28.2 billion out of their total 29.3 billion was through advertising (96%).

Google, Inc. (2009). Google 2009 Annual Report. available from [http://investor.google.com/pdf/2009\\_google\\_annual\\_report.pdf](http://investor.google.com/pdf/2009_google_annual_report.pdf). Internet. accessed Mar 25, 2012: 65.

Google, Inc. (2010). Google 2010 Annual Report. available from [http://investor.google.com/pdf/2010\\_google\\_annual\\_report.pdf](http://investor.google.com/pdf/2010_google_annual_report.pdf). Internet. accessed Mar 25, 2012: 29.

## The Internet Archive and the Open Library

Google Books is an excellent digital library and bookstore, but it is still owned by a company and is ultimately designed for profit. Probably the largest online collection that is completely public is the Internet Archive ([www.archive.org](http://www.archive.org)). The Internet Archive is more than a library for literature; it is a library for collections of many types, including audio, motion pictures, and it even has 150 billion web pages archived. It is non-profit and in many ways reminds one of a physical library with its patrons and volunteers.

The Internet Archive was founded in 1996 and its purpose is stated: "Libraries exist to preserve society's cultural artifacts and to provide access to them... The Internet Archive is working to prevent the Internet - a new medium with major historical significance - and other "born-digital" materials from disappearing into the past."<sup>25</sup>

There are currently over three million texts available, all free. Books are mostly uploaded by volunteers. The interface is not as clean as Google Books and it is missing some of the nice tools such as the bookshelves and quick citations, but is not difficult to use. Most books are available in a wide variety of formats such as PDF, ePub, plain text, and DjVu. Comprehensive metadata is also displayed with each book. Metadata is information which describes the content of a file. Common examples of metadata might be title, author, date, and publisher. Additional metadata might include scanning methods and what optical character recognition (OCR) software was used. Not everyone needs this information and so it probably doesn't need to be on the main page for each book and could be considered clutter, but it is useful information to have readily available for people who are doing research. Overall, the interface is dated because of its simple format, overuse of text, and under-use of graphics, but nevertheless it is usable.

The Internet Archive may have a mediocre interface now, but they are working to migrate their resources to a new system, called the Open Library ([openlibrary.org](http://openlibrary.org)). Already their main page says over one million titles are available there. The Open Library was created in 2008 with a different approach to hosting books online. Instead of listing each book individually, the Open Library has one page for each book. This doesn't sound all that different, but in practice it opens up many possibilities. The page for each book contains the entire history of that book, including

---

<sup>25</sup> *The Internet Archive: About the Internet Archive*. available from <http://www.archive.org/about/about.php>. Internet. accessed Feb 12, 2012.

every known printing.

The Open Library has a very nice interface with a wealth of information organized for quick deciphering. Some books have numerous editions digitally scanned, each edition with its own list of available formats for download. There is space for links to physical libraries and online library catalogs such as WorldCat to borrow the books, and there is also a spot where links can be placed to buy them.

Open Library takes a Wikipedia-style approach to digital libraries. Anyone is able to edit information, upload images, and add descriptions to the books. A history is kept of all the changes that have been made. This is a nice feature to have because it allows the community to contribute. It does not appear that they have implemented any way for contributors to modify the actual text of a book. This is unfortunate as there are frequent computer errors in character recognition. The platform is under constant development and so perhaps such functionality will be added in the future.

One other interesting section of the Open Library is their eBook lending service. They have partnered with a number of physical libraries to offer a lending system that works just like lending a physical book. Only one person is allowed to check a copy out at a time, and after a certain amount of time the book is "returned." This is a unique idea and gets around the issue of copyright, but a lending system is an awkward fit with a digital medium where infinite reproduction is possible. In the digital world there is no real reason that only one person can access information at a time like in the physical world. However, as long as book publishers are thinking with the mindset of books as physical properties such restrictions are necessary. For this to change, an entirely different profit model will have to be accepted where books are understood and treated as intellectual properties.

The Open Library does not have reviews or social features like Google Books. There is, however, the ability for users to create public lists of books categorized according to some theme. Active lists at the time of this writing included such titles as "Historical Cookbooks," "Civil War Accounts," "Parish Registers," and "Fun Books I Like."<sup>26</sup> Anyone can view the contents of the lists and the creators of the lists can edit them at will.

---

<sup>26</sup> These were all front-page active lists when the page was accessed on Dec 19, 2011. The lists on the front page change daily to reflect activity. *Open Library: Lists* available from <http://openlibrary.org/lists>. Interenet.

Overall, the Internet Archive is probably the best resource for people who are looking to start their own digital library. Their website has a wealth of information for people interested in digital libraries. They have information on every type of archive and links to myriad articles and archivist groups. These are divided under headings such as “Internet Libraries and Librarianship,” “Archiving Technology,” “Internet Mapping,” and “Copyright.” They also have some information on their storage system, a computer server array called a “Petabox.”

What may be the most useful part of the Internet Archive, and the Open Library in particular, is that the system is entirely open-source. This means that the software is entirely open, free, and available to any person who wants to use it. Because the source code for the software is freely available, a number of variants have been created by different groups for their own purposes.

### **Digital Libraries of Major Church Bodies**

To round out our look at digital libraries, it will be beneficial to take a glance at what other church bodies are doing. The Evangelical Lutheran Church in America (ELCA) keeps extensive archives, but little is available online. They have a nice website and a few small collections of historic photographs can be viewed there, but there is surprisingly little in the way of literature.<sup>27</sup>

The Lutheran Church—Missouri Synod (LCMS) has a large library in the Concordia Historical Institute ([www.lutheranhistory.org](http://www.lutheranhistory.org)). They describe themselves:

Concordia Historical Institute is one of the world's largest repositories of information on Lutheranism in North America. With over 1,400 individual collections and more than 2.6 million documents, the archives and manuscript holdings document American Lutheran history from the 19th century "Old Lutheran" immigration movements to the present.<sup>28</sup>

It seems, however, that all they have online is a catalog for their physical library. Searches bring up authors and descriptions of the works found. The interface is clearly dated and cluttered. In red letters the web page says, “Database updated on Monday, November 05,

---

<sup>27</sup> *ELCA Archives*. available from <http://www.elca.org/Who-We-Are/History/ELCA-Archives.aspx>. Internet. accessed Feb 18, 2012.

<sup>28</sup> *Concordia Historical Institute: Collections: Archives & Manuscripts Collections*. available from <http://www.lutheranhistory.org/collections/search.asp>. Internet. Accessed Feb 18, 2012.

2001.”<sup>29</sup> It was designed for function and not for aesthetics, with very basic formatting and no graphics.

A final place to look for a digital library was the Roman Catholic Church. One would think that they would have the resources to create a good digital library and they certainly have enough physical volumes to do so. However, it appears that they also do not have any central effort to create an online library. There are several smaller Catholic groups that have assembled some books online, but most of these are not very impressive.<sup>30</sup> Perhaps this is on purpose, since the Vatican doesn't typically grant access to all their collected writings to just anyone. According to one article, the Vatican archives have only 30 archivists employed to handle their millions of books stacked on 50 miles of bookshelves. Other scholars must obtain permission to enter with an escort and are only allowed to view the specific work they request, never to browse.<sup>31</sup>

### **WLS Essay File**

In all this, it is clear that the digital library WELS has in the essay file at the Wisconsin Lutheran Seminary Library is at least on par with—if not better than—the digital offerings of other large church bodies. The essay file is actually quite visually appealing and well-organized. It does not have any reviews or social features. This might be expected of a scholarly digital library of this sort, though such things still might prove useful. At the time of this writing, it contains 2153 essays, 48 audio essays, and 14 videos.<sup>32</sup> One nice thing that the essay file has which is very useful is a section with links sorted by Bible reference. This is an excellent feature to have in a religious library. The only odd part about this is that the links are sorted alphabetically rather than by book, so that a person who wants to browse to 2 Samuel must click the link for “2” and then scroll past 2 Corinthians, 2 John, and 2 Peter. I assume this is due to the

---

<sup>29</sup> *Ibid.*

<sup>30</sup> One such example is the Catholic Archive, which says, “The mission of The Catholic Archive is to create an all-encompassing digital library of Catholic documents, books, art, and prayers.” So far, the “all-encompassing” library includes one work, “The Imitation of Christ” by Thomas a Kempis. *The Catholic Archive*. available from <http://catholicarchive.org>. Internet. accessed Feb 18, 2012.

<sup>31</sup> John Preston. “The Vatican Archive: the Pope's private library.” *The Telegraph*. June 1, 2010. available from <http://www.telegraph.co.uk/culture/books/7772052/The-Vatican-Archive-the-Popes-private-library.html>. Internet. accessed Feb 18, 2012.

<sup>32</sup> “Essay File.” *Wisconsin Lutheran Seminary Library*. available from <http://www.wlssays.net>. Internet. accessed Feb 18, 2012.

limitations of the software.

The essay file might be the best place to host digitized books, at least until other platforms can be considered. It has a good system of organization, good visual appeal, and has the potential for a lot of expansion. It would be best for any digital library system at the seminary to host both digitized books and the essays as well. There is no good reason to have them separated into different catalogs. Were something like the Open Library system to be considered, migrating the essays would have to be part of the considerations. One other question would be whether or not the Open Library software could be integrated with the digital catalog at the WLS library. The current essay file is integrated so that results for digital essays and physical books show alongside each other. This is a valuable convenience and should not be lost.



## CHAPTER 2

### DIGITAL SCANNING EQUIPMENT

#### A Brief History of Book Digitization

The question that defies most people when contemplating digital libraries is, “How do I get the book from paper to the computer?”<sup>33</sup> In the digital age where most documents are created in word processors, this is simple. The document is created digitally and so it only needs to be correctly formatted and stored in an accessible place. Many millions of books, however, were created before the digital age, and making them into digital books is not always an easy process.<sup>34</sup>

#### **Keying**

One possible way of digitizing a book is to type it into the computer by hand, a process sometimes called “keying.”<sup>35</sup> This is how Michael Hart began Project Gutenberg.<sup>36</sup> Early computers had no other way of digitizing texts. Furthermore, storage limitations demanded that they not be copied in any way that would take up too much costly space, as with images. Manually typing out books does have the positive side of ensuring that the text is entered fairly accurately, depending on the skill and care of the operator. However, the obvious downside is the extreme time and concentration requirements of the person doing the work. Any pictures are lost, as well as any markings peculiar to the volume, which might have historical significance. A whole new problem is introduced when multiple languages are involved. The person doing the typing needs at least to be able to recognize the letters and know how to enter them in correctly.<sup>37</sup>

---

<sup>33</sup> “From book to e-book.” *Robotic Book Scanning at Stanford University*. available from <http://library.stanford.edu/depts/dlss/bookscanning/process.shtml>. Internet. accessed Feb 22, 2012.

<sup>34</sup> “Digitizing of rare books is a very manual, labor, and ultimately cost intense work.” Renate Evers. “Digitizing the Rare Book Collection of the Leo Baeck Institute.” *Digitization in the Real World*. ed. by Kwong Bor Ng and Jason Kucsma . Metropolitan New York Library Council, 2010. available from [http://metroblogs.typepad.com/files/ditrw\\_12.pdf](http://metroblogs.typepad.com/files/ditrw_12.pdf). Internet. accessed Feb 22, 2012: 194.

<sup>35</sup> *The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials*. National Initiative for a Networked Cultural Heritage, 2002. available from <http://www.nyu.edu/its/humanities/ninchguide>. Internet. accessed Feb 22, 2012: 85.

<sup>36</sup> *The History and Philosophy of Project Gutenberg* by Michael Hart.

<sup>37</sup> This becomes a hurdle more with languages that do not use the Latin alphabet. For WELS purposes, Greek, Hebrew, and Fraktur (used for German and Latin) occur frequently in the archives. A person who does not have any familiarity with these languages could easily confuse a Fraktur ŷ with an f, a Greek o with an σ, or a

### *Flatbed Scanners*

A better way of digitizing books came about with the advent of the flatbed scanner.<sup>38</sup> Flatbed scanners are able to make very accurate copies of pages and store them as digital images. Many modern scanners are capable of high resolutions in excess of 1,000 pixels-per-inch (ppi). They can store a page with every detail and color preserved. Flatbed scanners are ideal for loose pages, and scanners with paper feeders can do entire stacks of pages with little manual labor.

There are some problems though with flatbed scanners when it comes to digitally scanning bound books. Clearly the automation of a paper feeder is useless in this case, unless the pages are first removed from the binding. Another problem is that bound books do not flatten well on a single planar surface. The pages curve and the inside margin is easily lost. Opening a book so that the pages are at a 180 degree angle actually requires bending the spine past that point, which can be very hard on the binding. The older and more valuable the book is, the more costly accidental damage can be. Finally, there is the issue of speed. The vast majority of flatbed scanners do not scan quickly enough to be practical for entire books. The time it takes to turn a page, flip a book over onto the machine, line up the book, run the scan, and flip it back up to turn the next page makes scanning an entire book of possibly hundreds of pages a highly time-consuming task.<sup>39,40</sup>

### *Book Scanners*

Manual entry of texts and flatbed scanners have limitations that keep them from being ideal for digitizing books, but new types of machines have been developed to tackle the areas

---

Hebrew ה with a ה or a ת.

<sup>38</sup> The charge-coupled device (CCD) was invented in 1969 by Willard S. Boyle and George E. Smith. This is the main piece of technology behind the flatbed scanner. Boyle and Smith were awarded the 2009 Nobel Prize in Physics for their invention. "The Nobel Prize in Physics 2009." Charles K. Kao, Willard S. Boyle, George E. Smith. Press Release, Oct 6, 2009. Nobelprize.org: The Official Website of the Nobel Prize. available from [http://www.nobelprize.org/nobel\\_prizes/physics/laureates/2009/press.html](http://www.nobelprize.org/nobel_prizes/physics/laureates/2009/press.html). Internet. accessed Feb 28, 2012.

<sup>39</sup> "The fastest flatbed scan times are between ten and forty seconds per page, due to the mechanical movement of the imager across the scanning bed. This represents a hard limit to the number of pages that can be digitized in a day. **For** example, at thirty seconds per page, a 400-page book represents over three hours of nonstop scanning." Daniel Reetz. "The Why in DIY Book Scanning." *New York Law School Law Review*. Vol 55, No 1. 2010/11. available from <http://www.nylslawreview.com/201011-volume-55-number-1>. Internet. accessed Feb 22, 2012: 253.

<sup>40</sup> Evers, 189.

where these methods struggle. These more recently developed machines are commonly called “book scanners” and are designed specifically with the digital scanning of bound books in mind.<sup>41</sup> There are a number of companies that sell book scanners of different types and at widely varying price points.<sup>42</sup>

There are two different styles of book scanner available. The first can be called a “flat” scanner. Much like a standard flatbed scanner, it requires the book to be opened so that the pages can be pressed to a 180 degree angle. These have little advantage over flatbed scanners, and in fact are usually made with flatbed scanners. The only thing that makes most of them “book scanners” is the software that comes with them for processing books.<sup>43</sup> Another type of flat scanner has the book opened upwards. Instead of a moving scan head that travels across the page like in a flatbed scanner, they utilize a digital camera to record images of the pages. This does increase the speed of scanning, since the book does not need to be flipped. However, it still tends to have problems imaging two pages at once. The pages still need to be held as flat as possible to get a good image and the binding of the book can still be damaged.

The second type of book scanner is called a “planetary scanner” or an “orbital scanner.” Like the second type of flat scanner mentioned above, it uses digital cameras. However, it holds the camera at an angle and images each page separately. This way, the book can be held in a V-shaped cradle and the page can be snapped by the camera. The V-shaped cradle is much gentler on the binding than any of the flat methods. Planetary scanners often have a pair of cameras, one for each page, which can be triggered simultaneously. Most of these scanners still require manual page turning, but some companies sell robotic ones with automatic page turners.<sup>44</sup> The robotics raise the cost considerably<sup>45</sup> and their reliability is somewhat questionable.<sup>46</sup>

---

<sup>41</sup> Alexander Geschke and Eva Fischer. *Memorial Project – A Complex Approach to Digitisation of Personal Records*. available from [http://www.canfm.de/memorial/documents/Memorial\\_A\\_Geschke\\_EVA\\_03.pdf](http://www.canfm.de/memorial/documents/Memorial_A_Geschke_EVA_03.pdf). Internet. accessed Feb 23, 2012: 7.

<sup>42</sup> Reetz, 253.

<sup>43</sup> “Buying a Book Scanner.” *The Best Scanners – Top Scanners of 2011*. <http://www.thebestscanners.com/buying-a-book-scanner.html>. Accessed Feb 23, 2012.

<sup>44</sup> Examples include Treventus and Kirtas.

<sup>45</sup> Reetz, 253.

<sup>46</sup> “Though fully-automated page-turning scanners do exist, most organizations that employ such scanners

### **The Atiz BookDrive**

These book scanners come in a wide array of styles and prices. One of the more well-known companies is Atiz.<sup>47</sup> Their scanners are used by a number of universities and library associations.<sup>48</sup> There is a good reason for this. They offer a number of different models to fit different budgets and purposes. They offer all-inclusive kits with everything a person needs to get started. Their machines, while attractive looking, are fairly basic in capabilities. There are no automatic page turning systems available. There appears to be no way to adjust for books with a thick spine. Their machines are branded “BookDrive” and there are three models: BookDrive Pro, BookDrive Mini, and BookDrive DIY. The Pro has a base cost of \$13,895, the Mini is \$6,195, and the DIY is \$7,295. These costs do not include the cameras, which must be Canon brand DSLR-type cameras. A pair of these will add at least \$1000 for the lower-end model cameras and can easily go significantly higher.<sup>49</sup>

Were WELS to look for commercial book scanners, the Atiz BookDrive machines should be considered. However, the features are somewhat limited compared to other machines and the cost, although lower than many machines, would still be a significant investment for a synod with a tight budget.<sup>50</sup>

### **The Treventus ScanRobot**

Another intriguing company is Treventus and their ScanRobot.<sup>51</sup> The ScanRobot is a

---

also employ an individual to watch the mechanism. For this reason, most projects rely on humans to turn the pages, leaving the limit at two to fifteen seconds to lift the platen, turn the page, place the platen, and press the capture button.” Reetz, 254.

<sup>47</sup> *Atiz Innovation Co.* available from <http://www.atiz.com>. Internet. accessed Feb 29, 2012.

<sup>48</sup> Some of these include Stanford University, the Boston University, and the City of Toronto Archives Digitization Program. *Who are ATIZ customers and What are they saying about BookDrive?* available from <http://www.atiz.com/customers>. Internet. accessed Feb 23, 2012.

<sup>49</sup> Canon currently sells an EOS Rebel T3i Kit for \$549.99 and a EOS-1Ds Mark III for \$6,999.00, with six more models inside that price range. *EOS Digital SLR Cameras*. Canon U.S.A. available from [http://www.usa.canon.com/cusa/consumer/products/cameras/slr\\_cameras](http://www.usa.canon.com/cusa/consumer/products/cameras/slr_cameras). Internet. accessed Feb 23, 2012.

<sup>50</sup> “Book of Reports and Memorials.” Wisconsin Evangelical Lutheran Synod, May 2011. available from [www.wels.net/sites/wels/files/synodreports\\_2011boram.pdf](http://www.wels.net/sites/wels/files/synodreports_2011boram.pdf). Internet. accessed Feb 23, 2012: 77.

<sup>51</sup> *ScanRobot 2.0 MDS*. Treventus. available from [http://www.treventus.com/bookscanner\\_pageturner.html](http://www.treventus.com/bookscanner_pageturner.html). Internet. accessed Feb 23, 2012.

slightly different design that holds the book open at a slight 60 degree angle, compared to the Atiz BookDrive with its 100 degree angle. The lesser the angle, the gentler the machine is on the book. The ScanRobot has its own custom imaging system instead of cameras. The scan head appears to be similar to the scan head on a flatbed scanner, but instead of pressing the book onto a glass surface, the scan head is lowered into the open book.

The ScanRobot, as its name implies, is a robotic machine that automatically turns pages. The Treventus website<sup>52</sup> advertises it as capable of scanning up to 2500 pages per hour, but says that it may need to be slowed on fragile books. It can handle fairly large books, up to 15 centimeters (5.91 inches) thick and page sizes of up to 32 x 32 cm (12.6 x 12.6 in). Because of the method of scanning, the resolution is constant, with 300 ppi being standard but with an option to upgrade to 400 ppi. This is a decent resolution but nothing special. However, good image quality (sharpness, contrast, etc.) may compensate for average resolution.

The range of features of the Treventus ScanRobot comes at a price. Dealers must be contacted to negotiate exact pricing, but the Treventus website lists the MSRP for Europe at 65,000 to 75,000 Euro (\$85,000 to \$100,000). They offer three pieces of software that make up a suite, and each piece costs around \$2500. The price may seem extravagant, but it is not out-of-the-ordinary for a device like this.<sup>53</sup> Unfortunately, it does place the Treventus ScanRobot well out of the reach of most library budgets. While a tool like the ScanRobot would be a great asset for WELS archives, the cost would be difficult to justify. The only way it might be feasible would be through some sort of leasing deal or a large gift intended for this kind of use.

#### **Kirtas Technologies, Inc.**

A third large company that produces book scanners is Kirtas Technologies.<sup>54</sup> Kirtas has a number of choices to fit various needs. They sell an overhead flat scanner, a large format scanner for maps and newspapers, and they have a line of robotic scanners as well. Their robotic scanners are more of the standard style of planetary scanner and use Canon cameras like the Atiz scanners.

---

<sup>52</sup> *TREVENTUS Mechatronics GmbH*. [www.treventus.com](http://www.treventus.com). Internet. accessed Feb 23, 2012.

<sup>53</sup> David Rapp. "Product Watch: Library Scanners." *Library Journal*. Jul 15, 2011. available from [http://www.libraryjournal.com/lj/home/891007-264/product\\_watch\\_\\_library\\_scanners.html.csp](http://www.libraryjournal.com/lj/home/891007-264/product_watch__library_scanners.html.csp). Internet. accessed Feb 24, 2012.

<sup>54</sup> *Kirtas Technologies, Inc.* [www.kirtas.com](http://www.kirtas.com). Internet. accessed Feb 24, 2012.

They list their resolution as up to 400 ppi, the same as Treventus. Their vacuum robot arm can turn pages at a rate of 2,900 per hour and they advertise it as gentler than the human hand. This scanner ranges in price from \$69,000 all the way up to \$129,000,<sup>55</sup>

There are many other companies with varied but similar offerings as the ones mentioned here. The Crowley Company, Indus, Ristech, and ST Imaging are among the companies with scanners in use in libraries in the United States.<sup>56</sup> The most budget-friendly models are around \$5,000, but these often come with limitations, and nothing under \$50,000 comes with robotics.

### **The Ion Audio Book Saver**

One final system worth noting is one that is not for sale yet but is slated to be released early next year. Ion Audio is developing a device they call the Book Saver.<sup>57</sup> The great difference between the Book Saver and the previously mentioned devices is that the planned MSRP is between \$100 and \$200, though the exact price and release date have not been announced. Not many details have been publicized, but it will have two parts, a small book cradle and a unit that sits on top of the book, holding the pages flat and taking the picture. For each page the operator will have to pick up the top unit to turn the page.

While offering a budget scanner like this is nice for the hobbyist, it is hardly a tool that serious digitization projects would want to use. The camera specifications have not been released but at such a low price point the quality is most likely lower than is desirable for long-term archives. The pages will not be imaged consistently with a loose camera unit that is set on top without any fixed placement. Also, it will be limited in the size of book it can hold.

### **Summary of Commercial Scanning Equipment**

This survey of commercial book scanning machines reveals a few important points. First of all, the ideal type of scanner is not completely agreed upon. Book scanners aren't like copy machines, where a thousand different companies make them, but they basically look and function identically. Instead, nearly every company has their own take on how a book scanner should

---

<sup>55</sup> Rapp. "Product Watch: Library Scanners."

<sup>56</sup> *Ibid.*

<sup>57</sup> "Ion Announces Book Saver Book Scanner." *Ion Audio Press Releases*. Jan 6, 2011. available from <http://www.ionaudio.com/news/press-releases/ion-announces-book-saver-book-scanner>. Internet. accessed Feb 24, 2012.

work. The machines can be divided into categories, but within those categories there are many variants.

A second important point is that not only do the machines vary, but it seems that each company produces its own software suite to go with their machine. There is no gold standard in book scanning software. One piece of software that did get mentioned a few times was ABBYY Finereader.<sup>58</sup> This is an OCR program, for converting scanned images into searchable text. Other than that, most software to manage machines and workflow is created in-house.

A third point that can be learned from this overview is that these commercial book scanning machines are very costly. Even very simple machines with limited options generally cost between \$5,000 and \$15,000. The high-quality machines, especially the robotic ones, can run well over \$100,000. The extreme cost can be a prohibitive problem for many libraries. WELS might be able to budget for the cheaper scanners, but even \$5,000 is a significant amount when WELS has not had a line item in the budget for maintaining archives in the past several years.

---

<sup>58</sup> *ABBYY Software, Ltd.* [www.abbyy.com](http://www.abbyy.com). Internet. accessed Feb 24, 2012.

## CHAPTER 3

### DIGITAL FORMATTING

The most complicated part of building a digital library may have less to do with the “library” part and more with the “digital” part. When digital music started to become popular, many people already owned most of their music on compact discs. Their music already was digital; they just needed to transfer it from the CD to the computer. The process wasn't so easy for people who wanted to digitize their LP records. They had to have a record player with an audio output which could then play the record while a computer re-recorded it in a digital format. For both CDs and records, people had to decide how they wanted to store their music. Did they want to keep it at the highest quality possible, in a lossless format such as WAV? Or did they prefer to save space and use a lossy format that degraded the sound quality to do so? In the end, most people decided that the losses of quality in the MP3 format were negligible for the space saved. Of course, when the first portable MP3 players were released, memory was not cheap, and so the choice was a few tracks at high quality or a few hundred at a quality which sounded almost the same to most ears.

There are several similar challenges in the digitizing of books. What format is the best? Where is the trade-off between file size and image quality? High resolution color images take up a lot of space,<sup>59</sup> and when a book has several hundred pages, it can quickly eat up hard drive space.<sup>60</sup> An even bigger difficulty with large files might surface when people try to access the book remotely. What kind of network bandwidth is available? There are many different formats that can be used for digital books. Which is the best? What kind of metadata should be recorded with the book?

---

<sup>59</sup> Standard uncompressed (RAW or tagged image file format (TIFF)) color images are 24 bits per pixel (3 colors at 8 bits each). Grayscale images are generally 8 bits, and black and white lineart is 1 bit per pixel. A single byte is made up of 8 bits, and so an uncompressed, 18 MP color image will take up 54 million bytes (megabytes; MB) of space (18,000,000 pixels x 3 bytes per pixel). Compression reduces this number dramatically. For example, the Canon T2i advertises the file size for 18 MP RAW color images at 24.5 MB. Even with this reduction, a single book of uncompressed images at the largest file size with 400 pages will take up nearly 10 gigabytes of space. *EOS Rebel T2i EF-S 18-55IS II Kit: Specifications*. available from [http://www.usa.canon.com/cusa/consumer/products/cameras/slr\\_cameras/eos\\_rebel\\_t2i\\_ef\\_s\\_18\\_55is\\_ii\\_kit#Specifications](http://www.usa.canon.com/cusa/consumer/products/cameras/slr_cameras/eos_rebel_t2i_ef_s_18_55is_ii_kit#Specifications). Internet. accessed Feb 27, 2012.

<sup>60</sup> On space requirements for the British Library: “The digitised books with all their metadata and OCR are about 1Mb per page and the BL expects to have 25-30Tb of data by the end of the project. This data is replicated in three locations for preservation and disaster recovery.” Siân Harris. “Robotics Speed up Book Digitisation.” *Research Information*. August/September 2008. available from [http://www.researchinformation.info/features/feature.php?feature\\_id=184](http://www.researchinformation.info/features/feature.php?feature_id=184). Internet. accessed Feb 22, 2012.



## NARA Guidelines for Digital Archives

The U.S. National Archives and Records Administration (NARA) has put together an extensive guide to the digital formatting of literature. It is entitled “Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files – Raster Images.”<sup>61</sup> This document distinguishes between “production master” files which are intended for long-term archives and any sort of derivative files intended for access. Production masters are recorded in such a way as to preserve the details of the original and to facilitate batch software operations, such as file conversions and OCR processing. Access copies are often a lower resolution version or a text-only version that is compressed to allow for quick access over a network. The technical guidelines also describe “preservation masters” which require great care to capture page images at a pre-defined standard so that they can be reproduced in exactly the condition in which they were scanned. A good example of a preservation master would be the scanning of the Great Isaiah Scroll, which has been reproduced with such care that most people cannot tell the difference between the copies and the original.<sup>62</sup>

### File formats for Digital Books

There are many types of encodings to save digital books in. If a person were looking for the simplest method of storage possible, the best options would be either a plain text document or a directory of images of the pages.<sup>63</sup> These options have their downsides however. A plain text document does not fit the standards for production masters and preserves very little detail of the original. It also requires typing by hand if mistakes are to be minimized. A directory of images, on the other hand, preserves all the detail of the original pages, but does not function like a book.

---

<sup>61</sup> Steven Puglia, Jeffrey Reed, and Erin Rhodes . “Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files – Raster Images.” U.S. National Archives and Records Administration, June 2004. available from <http://www.archives.gov/preservation/technical/guidelines.pdf>. Internet. accessed Feb 27, 2012.

<sup>62</sup> The company which produced the copies goes into detail about how they created them using lasers to cut even the smallest cracks in the parchment and modern techniques to mirror even the smallest warps and buckles in the original. At the time the facsimiles were created in 2006, they were made using photographs taken in 1948. This means that the facsimiles are actually closer to the condition of the Isaiah scroll when it was discovered than the real Isaiah scroll after being handled for 60 years since its discovery. This kind of faithful reproduction does come at a cost. Facsimiles of the Isaiah scroll can be purchased for \$60,000. *Facsimile Editions: The Dead Sea Scrolls*. Facsimile Editions Limited, 2012. available from <http://www.facsimile-editions.com/en/ds>. Internet. accessed Feb 27, 2012.

<sup>63</sup> NARA Guidelines, 10.

The text is not understood as text by the computer and the pages are not linked together.

## ***XML***

There are several formats that are most commonly used by digital libraries. One is Extensible Markup Language (XML). This is a structured text-only format. It is most commonly found in books that are published in a digital format to begin with. XML works well for such works because of the good structure and small file size, but it is not a good format for digitally scanned books, since it does not preserve images of the original pages.<sup>64</sup>

## ***EPUB***

Another format similar to XML in its use is EPUB, short for “electronic publication.” EPUB is another format that is commonly used for digitally published works. According to the developers,

EPUB is the distribution and interchange format standard for digital publications and documents based on Web Standards. EPUB defines a means of representing, packaging and encoding structured and semantically enhanced Web content — including XHTML, CSS, SVG, images, and other resources — for distribution in a single-file format. EPUB allows publishers to produce and send a single digital publication file through distribution and offers consumers interoperability between software/hardware for unencrypted reflowable digital books and other publications.<sup>65</sup>

---

<sup>64</sup> XML books are encoded with a markup language such as the following example:

```
<?xml version="1.0"?>
<oldjoke>
<burns>Say <quote>goodnight</quote>,
Gracie.</burns>
<allen><quote>Goodnight,
Gracie.</quote></allen>
<applause/>
</oldjoke>
```

Norman Walsh. “What Do XML Documents Look Like?” *XML.com: A Technical Introduction to XML*. O’Reilly Media, Inc., 2010. available from <http://www.xml.com/pub/a/98/10/guide0.html?page=3>. Internet. accessed Feb 27, 2012.

<sup>65</sup> “EPUB.” *International Digital Publishing Forum*. 2012. available from <http://idpf.org/epub>. Internet. accessed Feb 27, 2012.

EPUB is an excellent format for publishing new digital books in, but it is not designed for archiving digitized books. For that reason, it is not recommended as the primary format for the WLS library.

### ***DjVu***

A digital library will want to use a format that keeps the images of the pages intact, especially for rare books.<sup>66</sup> In order to best preserve a book as much information as possible needs to be saved. This includes the original appearance—the typeface and layout. Also, older books tend to have a higher error rate in OCR conversions and an image of the original page allows the user to discern what the correct reading is.<sup>67</sup> One such format is DjVu (pronounced as *déjà vu*). DjVu was developed to preserve the images of book pages at a high compression rate. The files can be multi-layered, so that one layer might be the page image and another might be a text only layer. This way the text can be searched and copied without losing the appearance of the original.

### ***PDF***

DjVu is an excellent format, but it is not as widely accepted as a similar format, the Portable Document Format (PDF).<sup>68</sup> Since the creation of the word processor, people have had to deal with documents varying in appearance between computers.<sup>69</sup> Fonts, margins, and tabs often haven't translated as expected between different machines and different word processors. PDF was created by Adobe Systems in 1993 with the goal of being able to display documents identically regardless of the system the document is being displayed on. PDF, like DjVu, is capable of multi-layered documents. PDF does not have as high of a compression rate as DjVu,<sup>70</sup>

---

<sup>66</sup> Harris. “Robotics Speed up Book Digitisation.”

<sup>67</sup> Alison Babeu. “‘Rome Wasn’t Digitized in a Day’: Building a Cyberinfrastructure for Digital Classicists.” *Council on Library and Information Resources*, August 2011. available from <http://www.clir.org/pubs/abstract/pub150abst.html>. Internet. accessed Feb 27, 2012: 18.

<sup>68</sup> Nadir Weibel, Moira C. Norrie, and Beat Signer. “A Model for Mapping between Printed and Digital Document Instances.” *DocEng’07*, August 28–31, 2007. Winnipeg, Manitoba, Canada. ACM 2007.

<sup>69</sup> Terry Kuny. “A Digital Dark Ages? Challenges in the Preservation of Electronic Information.” *International Preservation News*, 1998. available from <http://ifla.queenslibrary.org/iv/ifla63/63kuny1.pdf>. Internet. accessed Feb 27, 2012: 5.

<sup>70</sup> Compression rates are a complex issue and much space could be dedicated to a full comparison of the compression rates and methods between DjVu and PDF. “A Model for Mapping between Printed and Digital

but its use is far more widespread.<sup>71</sup>

While there are many usable formats, PDF is to be recommended above the others at this time, mostly because it is the most commonly used. Although Adobe holds the patents to PDF, it is an open standard and can be accessed and modified by many different programs, including ones that are freely available for download, such as ePDFView (Linux)<sup>72</sup>, Foxit (Windows)<sup>73</sup>, and APV PDF Viewer (Android).<sup>74</sup>

### **Image Quality**

The format of the files that the books will be saved in is only one piece of the digital side of a digital library. The quality of the images is another aspect that needs to be considered.<sup>75</sup> There is a wide range of digital imaging devices available. Digital cameras range from webcams and cell-phone cameras up to DSLR cameras with interchangeable lenses that cost several thousand dollars. Most people understand that the more expensive the camera is, the better the pictures it will produce. It is not, however, always so easy to define what makes the images better. Nor does everyone know what kind of quality should be their goal.

#### ***Understanding Pixels from Camera to Digital Image***

The most advertised number on a digital camera is the megapixel rating (MP or Mpx). This corresponds to how many millions of dots (pixels) make up an image. The more megapixels, the more detail the camera can pick up. But, while cameras are rated by megapixels, the images they store are rated in pixels-per-inch (ppi) or dots-per-inch (dpi). Technically

---

Document Instances.” by Weibel, Norrie, and Signer dealt with this briefly. Compression sizes on their own are not enough for a full study, but quality of the compressed product makes a difference as well. This is why NARA Guidelines give different specifications for master files and compressed files for common use.

<sup>71</sup> K. Dennis, G. O. Michler, G. Schneider and M. Suzuki . “Automatic reference linking in distributed digital libraries.” *Conference on Computer Vision and Recognition Workshop*, 2003. Vol 9. Madison, WI. June 16-22, 2003. available from [http://nguyendangbinh.org/Proceedings/CVPR/2003/pdf/papers/DIAR\\_06.pdf](http://nguyendangbinh.org/Proceedings/CVPR/2003/pdf/papers/DIAR_06.pdf). Internet. accessed Feb 29, 2012: 26.

<sup>72</sup> *ePDFView* Trac Integrated SCM and Project Management. available from <http://trac.emma-soft.com/epdfview>. Internet. accessed Feb 27, 2012.

<sup>73</sup> *Foxit Reader*. available from [http://www.foxitsoftware.com/Secure\\_PDF\\_Reader](http://www.foxitsoftware.com/Secure_PDF_Reader). Internet. accessed Feb 27, 2012.

<sup>74</sup> *APV PDF Viewer*. available from <http://code.google.com/p/apv>. Internet. accessed Feb 27, 2012.

<sup>75</sup> NARA Guidelines, 25.

speaking, ppi more accurately describes a digital image and dpi describes a printed image or a printing device, but the terms are often used interchangeably.<sup>76</sup>

Megapixels are calculated by the total width of the image in pixels times the total height of the image in pixels. Pixels-per-inch are calculated by the pixels along the height or width of a single inch of the image, or to say it another way, the height in pixels divided by the height in inches or the width in pixels divided by the width in inches. These two numbers should be the same, as long as pixels are not being stretched.

The math for converting pixels-per-inch to megapixels is not too involved. Find the total pixels of the height and multiply them by the total pixels of the width. For example:

PPI: 300

Image size: 8.5 inches by 11 inches.

Equation:  $(PPI \times width) \times (PPI \times height) = total\ pixels$

$$\frac{total\ pixels}{1,000,000} = MP\ Rating$$

Solution:  $(300 \times 8.5) \times (300 \times 11) = 8,415,000$

$$\frac{8,415,000}{1,000,000} = approx.\ 8.4\ MP$$

This equation has limited use as written. It cannot be used to accurately calculate needed camera resolution, because it assumes that every pixel will be used and that the ratio is ideal to the book or picture that is being photographed. In reality, it is difficult to match the edges of a page exactly with the edges of a camera sensor. However, it is possible with this equation to provide a rough estimate of what is needed. At the very least, the MP rating cannot go below the calculated number.

Converting the other direction, from megapixels to pixels-per-inch is more complicated. Most lower-end cameras have an image ratio of 4:3 and most higher-end ones are 3:2. Here is an example:

Camera ratio: 4:3, where height is 3 and width is 4

Camera MP rating: 3 MP

Let *height* = *a*

---

<sup>76</sup> NARA Guidelines, 21.

Let  $width = b$

Let  $(MP \times 1,000,000) = Y$

Equations:  $Y = \frac{a}{b}$

$$\frac{a}{b} = \frac{3}{4}$$

Solution:  $3,000,000 = \frac{a}{b}$

$$a = \frac{3}{4}b$$

$$\frac{3}{4}b^2 = 3,000,000$$

$$3b^2 = 12,000,000$$

$$b^2 = 4,000,000$$

$$b = 2,000$$

$$\frac{a}{2000} = \frac{3}{4}$$

$$a = \frac{3}{4} \times 2000$$

$$a = 1500$$

Image height: 1500 pixels

Image width: 2000 pixels

Now the ppi can be calculated by the size of the page photographed. If the page is ten inches wide and the page fits exactly to the width of the image, then the image is 200 ppi.

### ***Recommended Archival Image Resolutions***

Now that the math is out of the way, what kind of ppi is considered ideal? According to the NARA guidelines, this will depend on the size of the smallest significant character. For example, if the smallest significant character is one millimeter, 400 ppi is the recommended

resolution, provided the image is in 8-bit grayscale. If the image is pure black and white, then the recommended resolution is 600 ppi.<sup>77</sup> If the optics of the camera are of a lower quality, then it might be necessary to get higher resolution images and downsample them to the needed resolution.<sup>78</sup> There is no listed recommendation for color images.

The NARA guidelines have many technical recommendations for the various aspects of digitization. These all connect in some way to accuracy and readability. They even have recommendations for viewing the files—things like the kind of lighting that the room should have when viewing digital books.<sup>79</sup>

### **Metadata**

Some parts of the guidelines may be overkill, since not everyone will be able to have a computer (or computers) dedicated for viewing digital books in a room that is painted the correct color and with the proper lighting. However, one important aspect of digital books that should not be overlooked is the metadata. Metadata is any information attached to the file that describes the book or the process of digitizing it. This is something that needs to be planned out before any major projects begin, because it is much easier to enter metadata right away than to go back and add it to an entire library later on. The metadata is useful in searching and categorizing digital books. It is also useful to anyone later on who might be researching how the books were digitized or who wants to know specific information about the file, including how it is structured. There can even be meta-metadata, which is data describing the metadata, such as the name of the person who entered the metadata.

Besides information about the file and how it was recorded, metadata should record whether or not optical character recognition (OCR) has been run on the file to create a text-only layer. If it has, the metadata should also record what software was used and any other pertinent OCR information.

The Dublin Core Metadata Initiative (DCMI) specifies a metadata standard for digital works.<sup>80</sup> Examples of standard metadata elements include title, subject and keywords,

---

<sup>77</sup> NARA Guidelines, 51.

<sup>78</sup> *Ibid*, 43.

<sup>79</sup> *Ibid*, 23,24.

<sup>80</sup> *Dublin Core Metadata Initiative*. available from <http://dublincore.org>. Internet. accessed Feb 28, 2012.

description, resource type, source, creator, publisher, date, language, and more.<sup>81</sup> DCMI should be kept as a reference by anyone who is creating a digital library. It is widely accepted as a standard and suggested by the NARA guidelines.<sup>82</sup>

---

<sup>81</sup> “Using Dublin Core - The Elements.” *Dublin Core Metadata Initiative*. DCMI, 2012. available from <http://dublincore.org/documents/usageguide/elements.shtml>. Internet. accessed Feb 28, 2012.

<sup>82</sup> NARA Guidelines, 7.



## CHAPTER 4

### SCANNING THE WLS RARE BOOKS LIBRARY

Wisconsin Lutheran Seminary has a wonderful collection of rare and historic books. Certainly having these books in a digital library would be a benefit for historians and theologians alike. These people would also benefit from the many historical records across the Wisconsin Evangelical Lutheran Synod being placed in a digital library.

#### **The DIY Book Scanning Community**

How can this be accomplished? Clearly the WELS does not have a spare \$10,000 for archives, much less a spare \$100,000.<sup>83</sup> Fortunately, it is not necessary to spend this kind of money to scan books, even to scan them well. There is an entire Internet community dedicated to building homemade book scanners. These can cost as little as a few hundred dollars or even less when made out of spare parts. In spite of the low cost, they can produce scans that rival much more expensive commercial machines.

The most active online community dedicated to do-it-yourself (DIY) book scanners is conveniently named [diybookscanner.org](http://diybookscanner.org). This community was begun by a man named Daniel Reetz. Reetz believed he could make digital copies of college textbooks for much less than he could buy them. With a little bit of experimenting, he figured out a cost-effective way to build a book scanning rig. He felt that his idea could be useful to many people and so he put the design online and the community was born.<sup>84</sup>

Since that time Reetz' book scanner has gone through a few major revisions and many other people have posted their own designs online. The forum contains a wealth of knowledge as woodworkers and engineers, experts and beginners alike, all share ideas and help each other with questions. This forum will be the base of knowledge used to design WELS' first book scanner.

#### **The Basic Parts of a DIY Book Scanner**

Every book scanner needs the same basic parts. It must have a v-shaped cradle to hold the

---

<sup>83</sup> Roughly the cost of a low-end and a high-end commercial scanner, respectively.

<sup>84</sup> Kim Lacey. "Interview with Daniel Reetz, founder of the DIY Bookscanning project." *HASTAC: Humanities, Arts, Science, and Technology Advanced Collaboratory*. May 31, 2011. available from <http://hastac.org/blogs/kimlacey/interview-daniel-reetz-founder-diy-bookscanning-project>. Internet. accessed Feb 28, 2012.

book open, ideally at about a 100 degree angle.<sup>85</sup> The majority of book scanners use a platen to hold the pages flat. The platen can be made out of glass or acrylic, and must be v-shaped at the same angle as the cradle. The book scanner needs a lighting system so that pages are brightly and evenly lit. Finally, there needs to be at least one camera, preferably two.

From that point there are many options as to how the scanner is designed. Most are set up so that the platen can be lifted to turn the book's page, but some instead raise and lower the book so that the cameras and lights don't have to be adjusted. Some people build their scanners with foot pedals. Some add LCD monitors to easily see what the cameras are seeing. Most are built primarily out of wood, but some are metal, some are purely acrylic, and some are even cardboard.

The designs are generally customized to a degree for the application for which they will primarily be used. If a person only has a handful of books to scan, a cardboard scanner is probably sufficient. A person who is scanning a library will want something that is comfortable to use for an extended period of time and gets work done as efficiently as possible.

### **The WLS Book Scanner**

The scanner for the WLS library is an oversized design to accommodate even the largest books in the rare books room. The design and operation is inspired by Daniel Reetz' most recent design which raises and lowers the book to meet the platen. This way the cameras and lights can be set once for each book and should not have to be further adjusted. However, the increased size of the scanner creates a number of issues that required design changes. Most notably, a simple levered system was not stable enough and had to be redesigned with crossbar legs (resembling a scissor-lift). Along with that change cables and counterweights were added in order to balance out the weight of the lift.

The function of the machine is similar to most non-robotic scanners. A book is loaded into the front. The cameras are adjusted to capture the page. One hand is then used to raise and lower the book and the other is used to turn pages. Users have recorded speeds of over 1,000 pages per hour using similar designs.<sup>86</sup>

---

<sup>85</sup> "Platen: Theory and Practice." *DIY Book Scanner*. Sept. 4, 2011. available from <http://www.diybookscanner.org/forum/viewtopic.php?f=1&t=1149&p=11165&hilit=platen+angle#p11165>. Internet. accessed Feb 29, 2012.

<sup>86</sup> "A Book Scanner in Every Hackerspace /DIY Kit." *DIY Book Scanner*. Jan 25, 2012. available from

WLS only has a budget for one single camera at present, but the machine is designed to easily add a second camera to capture two pages with a single lever-press. The original plan was to have capability for four cameras to capture extra-large pages at high resolution (two cameras per page), but time ran short for design and construction work and it was deemed unnecessary since only one camera is currently available. WLS has this year purchased a Canon T2i (also called the 550D) camera. This camera has an 18 MP resolution and outputs pictures up to 5184 pixels in height and 3456 pixels in width.<sup>87</sup> Recommended size for a page with this resolution is no more than 8.5" x 13" to achieve a resolution of 400 ppi for archival quality. For a passable resolution of 300 ppi page sizes can reach up to 11.5" x 17". With only a single camera, books will have to be run through twice, once to capture the left page and once to capture the right page.

The cost of this book scanner is far less than a commercial scanner. It was built for approximately \$350 in materials (not including cameras). Some money was saved by using recycled scrap for certain parts, but some extra expenditure was also incurred to experiment with different ideas. Despite the low cost, this machine is capable of everything a commercial scanner costing thousands of dollars can do. In fact, it is built to handle books even larger than most commercial scanners. Its custom construction also means it can be easily modified if a need arises.

### **Considerations for a WELS Digitization Initiative**

The creation of a book scanning machine is only a part of the work that needs to be done to digitize the WLS library. Some decisions will have to be made. First there are questions as to how to approach a digital library. Does Wisconsin Lutheran Seminary want to commit to digitizing the entire rare books room or will books simply be digitized as needed? Does WELS value the preservation of our heritage enough to fund it? Are there people who are committed to making this work?

---

<http://www.diybookscanner.org/forum/viewtopic.php?f=14&t=1192&p=12643&hilit=pages+per+hour&sid=513baae014919cb12cd5ed75e7ec27cb#p12643>. Internet. Accessed Feb 29, 2012.

<sup>87</sup> *EOS Rebel T2i EF-S 18-55IS II Kit: Specifications*. available from [http://www.usa.canon.com/cusa/consumer/products/cameras/slr\\_cameras/eos\\_rebel\\_t2i\\_ef\\_s\\_18\\_55is\\_ii\\_kit#Specifications](http://www.usa.canon.com/cusa/consumer/products/cameras/slr_cameras/eos_rebel_t2i_ef_s_18_55is_ii_kit#Specifications). Internet. accessed Feb 27, 2012.

Then there are the practical questions. Who will head up a digitization initiative? Who will perform the scanning? Where will funding come from? What software is needed? What hardware is needed? What steps should be implemented in getting books from print form to digital form?

### ***Recommendations***

After looking at the different aspects of several different digital library initiatives, these are my recommendations: A digital library is certainly worth some time and money. Because it will preserve priceless history and be an invaluable tool to present and future scholars, WLS should commit to creating a complete digital library. This would involve creating a plan to digitize  $x$  number of books in  $x$  years. Then a priority list would have to be created. This would depend on usage and value. The long-term plan might involve digitizing most of the out-of-copyright works throughout the library.

Financial constraints in WELS are real and so this digital library should be done with careful consideration of expenses. One book scanning machine is being created with this project. Others, if needed, can be built for a few hundred dollars. Cameras are the most expensive part. However, for archival quality photographs high quality cameras need to be purchased. Cameras should be carefully evaluated, but saving money here might mean more work in the future when books need to be re-scanned.

Of course, a digital library is far more than book scanning machines. Computers are needed to store the files. Ideally, there would be a dedicated library computer for processing image files into books. Software needs to be obtained for doing this. As mentioned before, ABBYY Finereader is a popular solution. Books should be created out of the images in PDF format with an OCR text layer. The original images should be archived as master copies and the derivative PDF files should be hosted on a web server. The seminary essay file can serve as the web server to begin with, but a more featured solution such as the Open Library should be implemented in the future. All computer systems should be carefully backed up on a regular basis, with a copy of the files kept off-sight.

Since the project will take place at the WLS library, it should be overseen by the management of the WLS library. Students employees of the library can work doing digitization. Consider: if employees take turns and scan one book per day, 5 days a week, then in 40 weeks 200 priceless rare books could be digitized. Few books exceed 1000 pages, and at 1000 pages per

hour it would require less than an hour of operation each day. Certainly at this rate it would take quite a bit of time to digitize the whole library, but it would be a huge step forward from what is currently being done. The amount of work being done can always be ramped up later on.

Note that this does not include OCR time. OCR accuracy and speed will be affected by computer processing power, the software being used, the clarity of the text being scanned, and the experience of the operator.<sup>88</sup> However, while OCR is an extremely useful thing to have, it is not essential to a digital library, meaning that if the books would be imaged at a high quality the possibility would exist to run OCR on them later. This is not to say that waiting to run OCR would be ideal; rather, it would be creating more work for the future. However, it does remain as a possibility.

### ***A Sample Digitization Workflow***

A digitization workflow might look like this: A worker images a book using a book scanner. The book is 500 pages and at an unhurried pace takes just under an hour to scan. The worker then takes the images off the two cameras and puts them onto a processing computer. On the computer, he loads the images into a new folder for the book and uses a batch file naming program to rename them all as pages in sequential order. He then loads up OCR software, setting it to adjust the pages as necessary (crop, skew, keystone, sharpen, etc) and runs text recognition to create the text-only layer, finally compressing everything into a PDF. Most rare books can have their final production copy saved in grayscale. Some very clear texts might be compressed down to black and white lineart, and the oldest texts and any with illuminations will need to be kept in color. This is a judgment on the individual text, but could be set in a policy.

Metadata must be entered by the person who is working with the file, according to any WLS guidelines that may be established and following the Dublin Core in general. After the book is processed, the original master images and a copy of the processed book will be stored in a permanent location and another copy of the processed book will be uploaded to the web server. Automatic backup systems ensure that copies are made to multiple computers for safe storage.

This workflow would be a simple and efficient way to convert a printed book into a digital record which is both preserved in great detail and available across the Internet to anyone

---

<sup>88</sup> Daniel Walker, William Lund, Eric Ringger. "Evaluating models of latent document semantics in the presence of OCR errors." *EMNLP '10 Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. (2010): 240

who wishes to use it. With the proper equipment and software, a single book can be done in a reasonable amount of time and the worker should not need to be extensively trained.

One issue is the proofing of the OCR output. As stated previously, the accuracy of the output from such software might vary widely depending on the document. Finding the best OCR software will be essential to getting good results. Further study is required to determine the best options. Unfortunately, some errors will occur no matter how good the software is. There are a couple possible ways to deal with this. Employees or volunteers could go through each book comparing the OCR results to the text, finding and manually correcting all errors. This would yield the best results but would be extremely time-consuming. Another possibility is that the OCR results could simply be left with any errors that occur. This would limit their usefulness in searches but would still give more results than no OCR at all. It seems that this is the approach that many digital libraries, such as Google Books and the Internet Archive take, presumably because their main concern is getting the books archived rather than devoting the resources to correcting every error. The best approach to take might be a combination of the two methods, doing a complete correcting by hand of certain texts and leaving the ones that are less commonly used.

The ultimate goal of a WELS digital library ought to be a comprehensive collection of WELS history and theology. This means having a complete library in digital format. Every major work and the majority of minor ones in the library should be preserved and available digitally. As copyright models for digital libraries are developed, a WELS digital library should stay up-to-date, offering as much as is legal. One final major part would be digitizing congregational histories. It is impractical for congregations to develop their own digital libraries, so the synod could begin a program where congregations can send in their records to be digitized and put in the library.

## CONCLUSIONS

Several things became apparent in this study. Digitization initiatives are numerous and widespread. They exist for various purposes and the quality of their offerings varies greatly. In a world of fast-moving technology, they may struggle to keep up with the available technology, the current standards, and the changing trends. Church bodies especially are falling behind in the transition to digital literature. It is clear that digitization is not a mere fad, but represents a major shift in the way people interact with literature. For this reason a serious scholarly institution should consider maintaining a quality digital library a high priority.

Many commercial solutions exist for digitizing books. They come in a variety of styles, the best of which are designed specifically for books, as opposed to all-purpose scanners. Commercial scanners are also priced within a large window. The cheapest scanners are priced under \$10,000, and the most expensive are well over \$100,000. These might be affordable for large institutions but for a small synod such as WELS the cost is significant. When coupled with the costs of the computers and software needed for a digital library, the total expense is a deterrent even to beginning a digitization effort.

To preserve books for the long-term, file formats need to be understood. Many different formats exist, each with their own advantages and disadvantages. PDF may be the most widespread and capable format for eBooks. Besides choosing a format, standards need to be set for image quality and metadata. Metadata needs to be carefully recorded right away so that it is available to anyone who needs it. Good documents to reference for these standards are the NARA Guidelines and the Dublin Core Metadata Initiative.

Despite the hurdles in going the commercial route for digitization, a digitization initiative is not beyond the capability of WELS, thanks largely to the online DIY community. People have spent a great deal of time and energy developing ways to accomplish the same things that commercial equipment can do, but at a significantly smaller cost. They offer their designs and advice to the public over the Internet for no cost. For a few hundred dollars a hobbyist with a few tools can build a highly capable book scanner. The biggest cost then might be the time it takes to develop and assemble such a machine. The “standard” scanner the DIY community is developing may be a great asset for WELS in the future.

The book scanner that accompanies this paper will serve as the WELS' first book scanner. It will ease the difficulty of entering into a digitization initiative. However, it will not do the

project on its own. A digital library requires leadership and hard work. It requires standards to be set and workflows to be defined. It requires an up-to-date interface and content management. In summary, it requires people who love books and want to see those books preserved and shared with hungry minds of today and of the future.

The question is not, “Can WELS afford a digital library?” The real question is whether or not WELS can afford *not* to begin a serious digitization effort. The WLS library contains many priceless works, both as pieces of history and for their faithful representation and exposition of God's word. This is a heritage from our fathers. What better way is there to honor them than to preserve and study what they have left us? What better way is there to pass this heritage on to our sons? Surely as a synod we are able to accomplish such a great task, *Deo volente*.

*S.D.G.*



## BIBLIOGRAPHY

- \_\_\_\_\_ *ABBYY Software, Ltd.* www.abbyy.com. Internet. accessed Feb 24, 2012.
- \_\_\_\_\_ “About Digital Collections.” *The Library of Congress*. available from <http://www.loc.gov/library/about-digital.html>. Internet. accessed Feb 11, 2012.
- \_\_\_\_\_ *About Google Books: History of Google Books*. available from <http://www.google.com/googlebooks/history.html>. Internet. accessed Feb 16, 2012.
- \_\_\_\_\_ *APV PDF Viewer*. available from <http://code.google.com/p/apv>. Internet. accessed Feb 27, 2012.
- \_\_\_\_\_ *The Art of Google Books*. available from <http://theartofgooglebooks.tumblr.com>. Internet. accessed Feb 16, 2012.
- \_\_\_\_\_ *Atiz Innovation Co.* available from <http://www.atiz.com>. Internet. accessed Feb 29, 2012.
- Babeu, Alison. “Rome Wasn’t Digitized in a Day”: Building a Cyberinfrastructure for Digital Classicists.’ Council on Library and Information Resources, August 2011. available from <http://www.clir.org/pubs/abstract/pub150abst.html>. Internet. accessed Feb 27, 2012.
- \_\_\_\_\_ “Book of Reports and Memorials.” Wisconsin Evangelical Lutheran Synod, May 2011. available from [www.wels.net/sites/wels/files/synodreports\\_2011boram.pdf](http://www.wels.net/sites/wels/files/synodreports_2011boram.pdf). Internet. accessed Feb 23, 2012.
- \_\_\_\_\_ “A Book Scanner in Every Hackerspace /DIY Kit.” *DIY Book Scanner*. Jan 25, 2012. available from <http://www.diybookscanner.org/forum/viewtopic.php?f=14&t=1192&p=12643&hilit=pages+per+hour&sid=513baae014919cb12cd5ed75e7ec27cb#p12643>. Internet. Accessed Feb 29, 2012.
- \_\_\_\_\_ “Buying a Book Scanner.” *The Best Scanners – Top Scanners of 2011*. <http://www.thebestscanners.com/buying-a-book-scanner.html>. Accessed Feb 23, 2012.
- \_\_\_\_\_ *The Catholic Archive*. available from <http://catholicarchive.org>. Internet. accessed Feb 18, 2012.
- Choi, Youngok & Edie Rasmussen. “What is Needed to Educate Future Digital Librarians.” *D-Lib Magazine*. Vol 12, No. 9. 9/2006. available from <http://www.dlib.org/dlib/september06/choi/09choi.html>. Internet. accessed Feb 29, 2012.
- \_\_\_\_\_ *Concordia Historical Institute: Collections: Archives & Manuscripts Collections*. available from <http://www.lutheranhistory.org/collections/search.asp>. Internet. accessed

Feb 18, 2012.

Coyle, Karen. "Mass Digitization of Books." *Journal of Academic Librarianship*. Vol. 32, #6. available from <http://www.kcoyle.net/jal-32-6.html>. Internet. accessed Feb 11, 2012.

Dennis, K., G. O. Michler, G. Schneider and M. Suzuki . "Automatic reference linking in distributed digital libraries." *Conference on Computer Vision and Recognition Workshop*, 2003. Vol 9. Madison, WI. June 16-22, 2003. available from [http://nguyendangbinh.org/Proceedings/CVPR/2003/pdffiles/papers/DIAR\\_06.pdf](http://nguyendangbinh.org/Proceedings/CVPR/2003/pdffiles/papers/DIAR_06.pdf). Internet. accessed Feb 29, 2012.

\_\_\_\_\_ *DIY Book Scanner*. available from [diybookscanner.org](http://diybookscanner.org). Internet. accessed Feb 29, 2012.

\_\_\_\_\_ *Dublin Core Metadata Initiative*. available from <http://dublincore.org>. Internet. accessed Feb 28, 2012.

\_\_\_\_\_ *ELCA Archives*. available from <http://www.elca.org/Who-We-Are/History/ELCA-Archives.aspx>. Internet. accessed Feb 18, 2012.

\_\_\_\_\_ *EOS Digital SLR Cameras*. Canon U.S.A. available from [http://www.usa.canon.com/cusa/consumer/products/cameras/slr\\_cameras](http://www.usa.canon.com/cusa/consumer/products/cameras/slr_cameras). Internet. accessed Feb 23, 2012.

\_\_\_\_\_ *EOS Rebel T2i EF-S 18-55IS II Kit: Specifications*. available from [http://www.usa.canon.com/cusa/consumer/products/cameras/slr\\_cameras/eos\\_rebel\\_t2i\\_ef\\_s\\_18\\_55is\\_ii\\_kit#Specifications](http://www.usa.canon.com/cusa/consumer/products/cameras/slr_cameras/eos_rebel_t2i_ef_s_18_55is_ii_kit#Specifications). Internet. accessed Feb 27, 2012.

\_\_\_\_\_ "ePDFView." *Trac Integrated SCM and Project Management*. available from <http://trac.emma-soft.com/epdfview>. Internet. accessed Feb 27, 2012.

\_\_\_\_\_ "EPUB." *International Digital Publishing Forum*. 2012. available from <http://idpf.org/epub>. Internet. accessed Feb 27, 2012.

\_\_\_\_\_ "Essay File." *Wisconsin Lutheran Seminary Library*. available from <http://www.wlsessays.net>. Internet. accessed Feb 18, 2012.

Evers, Renate. "Digitizing the Rare Book Collection of the Leo Baeck Institute ." *Digitization in the Real World*. ed. by Kwong Bor Ng and Jason Kucsma . Metropolitan New York Library Council, 2010. (p.185-194). available from [http://metroblogs.typepad.com/files/ditrw\\_12.pdf](http://metroblogs.typepad.com/files/ditrw_12.pdf). Internet. accessed Feb 22, 2012.

\_\_\_\_\_ *Facsimile Editions: The Dead Sea Scrolls*. Facsimile Editions Limited, 2012. available from <http://www.facsimile-editions.com/en/ds>. Internet. accessed Feb 27, 2012.

Fleischhauer, Carl. "Steps in the Digitization Process." January 1996. available from <http://lcweb2.loc.gov/ammem/award/docs/stepsdig.html>. Internet. accessed Feb 22, 2012.

\_\_\_\_\_*Foxit Reader*. available from [http://www.foxitsoftware.com/Secure\\_PDF\\_Reader](http://www.foxitsoftware.com/Secure_PDF_Reader). Internet. accessed Feb 27, 2012.

\_\_\_\_\_"From book to e-book." *Robotic Book Scanning at Stanford University*. available from <http://library.stanford.edu/depts/dlss/bookscanning/process.shtml>. Internet. accessed Feb 22, 2012.

Geschke, Alexander and Eva Fischer. *Memorial Project – A Complex Approach to Digitisation of Personal Records*. available from [http://www.canfm.de/memorial/documents/Memorial\\_A\\_Geschke\\_EVA\\_03.pdf](http://www.canfm.de/memorial/documents/Memorial_A_Geschke_EVA_03.pdf). Internet. accessed Feb 23, 2012.

\_\_\_\_\_*Google History*. available from <http://www.google.com/about/company/history.html>. Internet. accessed Feb 16, 2012.

\_\_\_\_\_*Google, Inc. (2009). Google 2009 Annual Report*. available from [http://investor.google.com/pdf/2009\\_google\\_annual\\_report.pdf](http://investor.google.com/pdf/2009_google_annual_report.pdf). Internet. accessed Mar 25, 2012.

\_\_\_\_\_*Google, Inc. (2010). Google 2010 Annual Report*. available from [http://investor.google.com/pdf/2010\\_google\\_annual\\_report.pdf](http://investor.google.com/pdf/2010_google_annual_report.pdf). Internet. accessed Mar 25, 2012.

Harris, Siân. "Robotics Speed up Book Digitisation." Research Information. August/September 2008. available from [http://www.researchinformation.info/features/feature.php?feature\\_id=184](http://www.researchinformation.info/features/feature.php?feature_id=184). Internet. accessed Feb 22, 2012.

Hedstrom, Margaret. "Are We Ready for New Skills Yet?" *New Skills for a Digital Era*. ed. By Richard Pearce-Moses and Susan E. Davis. Society of American Archivists, Chicago, IL. 2008.

\_\_\_\_\_*The History and Philosophy of Project Gutenberg by Michael Hart*. August 1992. available from [http://www.gutenberg.org/wiki/Gutenberg:The\\_History\\_and\\_Philosophy\\_of\\_Project\\_Gutenberg\\_by\\_Michael\\_Hart](http://www.gutenberg.org/wiki/Gutenberg:The_History_and_Philosophy_of_Project_Gutenberg_by_Michael_Hart). Internet. accessed Feb 20, 2012.

\_\_\_\_\_*The Internet Archive: About the Internet Archive*. available from <http://www.archive.org/about/about.php>. Internet. accessed Feb 12, 2012.

\_\_\_\_\_*Internet World Stats: Usage and Population Statistics*. available from <http://www.internetworldstats.com/stats.htm>. Internet. accessed Feb 12, 2012.

- \_\_\_\_\_  
“Ion Announces Book Saver Book Scanner.” *Ion Audio Press Releases*. Jan 6, 2011. available from <http://www.ionaudio.com/news/press-releases/ion-announces-book-saver-book-scanner>. Internet. accessed Feb 24, 2012.
- Jih, Sophia. “University and Google Books Move Forward with Digitization.” *The Daily Princetonian*. May 8, 2010. available from <http://www.dailyprincetonian.com/2010/04/08/25772>. Internet. accessed Feb 16, 2012.
- \_\_\_\_\_  
*Kirtas Technologies, Inc.* [www.kirtas.com](http://www.kirtas.com). Internet. accessed Feb 24, 2012.
- \_\_\_\_\_  
*Kryder's Law: A Rule of Thumb for Hard Drive Growth*. available at <http://www.mattscomputertrends.com/Kryder%27s.html>. Internet. accessed Feb 16, 2012
- Kuny, Terry. “A Digital Dark Ages? Challenges in the Preservation of Electronic Information.” *International Preservation News*, 1998. available from <http://ifla.queenslibrary.org/iv/ifla63/63kuny1.pdf>. Internet. accessed Feb 27, 2012.
- Lacey, Kim. “Interview with Daniel Reetz, founder of the DIY Bookscanning project.” *HASTAC: Humanities, Arts, Science, and Technology Advanced Collaboratory*. May 31, 2011. available from <http://hastac.org/blogs/kimlacey/interview-daniel-reetz-founder-diy-bookscanning-project>. Internet. accessed Feb 28, 2012.
- Lefevre, Francois-Marie and Marin Saric. 2009. Detection of grooves in scanned images. US Patent 7,508,978, filed September 13, 2004, and issued March 24, 2009.
- \_\_\_\_\_  
*The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials*. National Initiative for a Networked Cultural Heritage, 2002. available from <http://www.nyu.edu/its/humanities/ninchguide>. Internet. accessed Feb 22, 2012.
- \_\_\_\_\_  
“The Nobel Prize in Physics 2009.” Charles K. Kao, Willard S. Boyle, George E. Smith. Press Release, Oct 6, 2009. [nobelprize.org: The Official Website of the Nobel Prize](http://www.nobelprize.org/nobel_prizes/physics/laureates/2009/press.html). available from [http://www.nobelprize.org/nobel\\_prizes/physics/laureates/2009/press.html](http://www.nobelprize.org/nobel_prizes/physics/laureates/2009/press.html). Internet. accessed Feb 28, 2012.
- \_\_\_\_\_  
*Open Library: Lists* available from <http://openlibrary.org/lists>. Internet. accessed Dec 19, 2012.
- O’Reilly, Tim. “What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software.” *Communications & Strategies*. No. 1, p. 19, First Quarter 2007.
- \_\_\_\_\_  
“Platen: Theory and Practice.” *DIY Book Scanner*. Sept. 4, 2011. available from <http://www.diybookscanner.org/forum/viewtopic.php?f=1&t=1149&p=11165&hilit=plate>

n+angle#p11165. Internet. accessed Feb 29, 2012.

Preston, John. "The Vatican Archive: the Pope's private library." *The Telegraph*. June 1, 2010. available from <http://www.telegraph.co.uk/culture/books/7772052/The-Vatican-Archive-the-Popes-private-library.html>. Internet. accessed Feb 18, 2012.

Pudles, Daniel. "The Future of Books: Not Bound by Anything." *The Economist*. March 22, 2007. available from [http://www.economist.com/node/8881446?story\\_id=8881446](http://www.economist.com/node/8881446?story_id=8881446). Internet. Feb 16, 2012.

Puglia, Steven, Jeffrey Reed, and Erin Rhodes . "Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files – Raster Images." U.S. National Archives and Records Administration, June 2004. available from <http://www.archives.gov/preservation/technical/guidelines.pdf>. Internet. accessed Feb 27, 2012.

Rapp, David. "Product Watch: Library Scanners." *Library Journal*. Jul 15, 2011. available from [http://www.libraryjournal.com/lj/home/891007-264/product\\_watch\\_\\_library\\_scanners.html.csp](http://www.libraryjournal.com/lj/home/891007-264/product_watch__library_scanners.html.csp). Internet. accessed Feb 24, 2012.

Reetz, Daniel. "The Why in DIY Book Scanning." *New York Law School Law Review*. Vol 55, No 1. 2010/11. available from <http://www.nyslawreview.com/201011-volume-55-number-1>. Internet. accessed Feb 22, 2012.

\_\_\_\_\_ "Rules for regulating the subscription library at Stamford; and a list of the committee, subscribers, &c. to which is added, a catalogue of the books in the library at its first opening in February 1787." Newcomb and Peat, 1787. available from [http://books.google.com/books?id=L3pbAAAAQAAJ&printsec=frontcover&source=gbs\\_ge\\_summary\\_r&cad=0#v=onepage&q&f=false](http://books.google.com/books?id=L3pbAAAAQAAJ&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false). Internet. accessed Feb 16, 2012.

\_\_\_\_\_ *ScanRobot 2.0 MDS*. Treventus. available from [http://www.treventus.com/bookscanner\\_pageturner.html](http://www.treventus.com/bookscanner_pageturner.html). Internet. accessed Feb 23, 2012.

\_\_\_\_\_ *SPI Global: Digitization Services*. available from <http://www.spi-global.com/content-solutions/our-services/digitization-services>. Internet. accessed Feb 29, 2012.

van Stipriaan, René, interview with. "Digitization helps future proof Dutch literature." *Research Information*, April/May 2009. available from [http://www.researchinformation.info/features/feature.php?feature\\_id=212](http://www.researchinformation.info/features/feature.php?feature_id=212). Internet. Accessed Feb 29, 2012.

\_\_\_\_\_ *TREVENTUS Mechatronics GmbH*. [www.treventus.com](http://www.treventus.com). Internet. accessed Feb 23, 2012.

\_\_\_\_\_ "Using Dublin Core - The Elements." *Dublin Core Metadata Initiative*. DCMI, 2012.

available from <http://dublincore.org/documents/usageguide/elements.shtml>. Internet. accessed Feb 28, 2012.

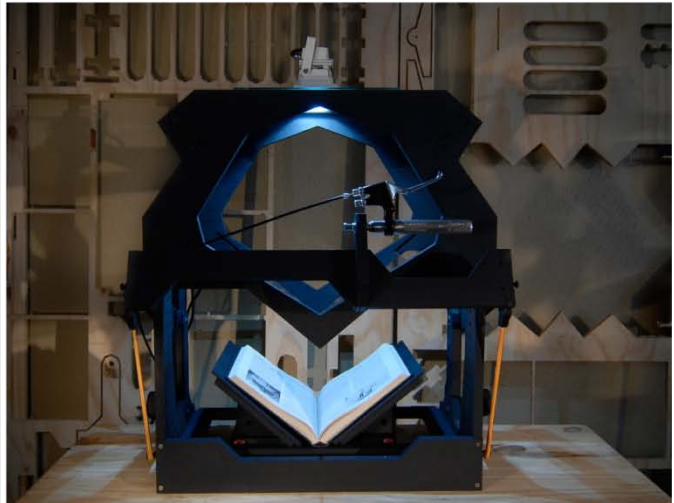
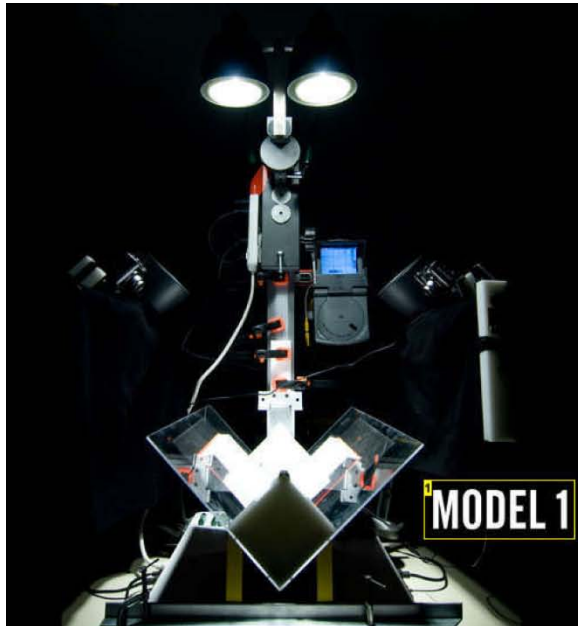
Walker, Daniel, William Lund, and Eric Ringger. "Evaluating models of latent document semantics in the presence of OCR errors." *EMNLP '10 Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. (2010): 240-250

Walsh, Norman. "What Do XML Documents Look Like?" *XML.com: A Technical Introduction to XML*. O'Reilly Media, Inc., 2010. available from <http://www.xml.com/pub/a/98/10/guide0.html?page=3>. Internet. accessed Feb 27, 2012.

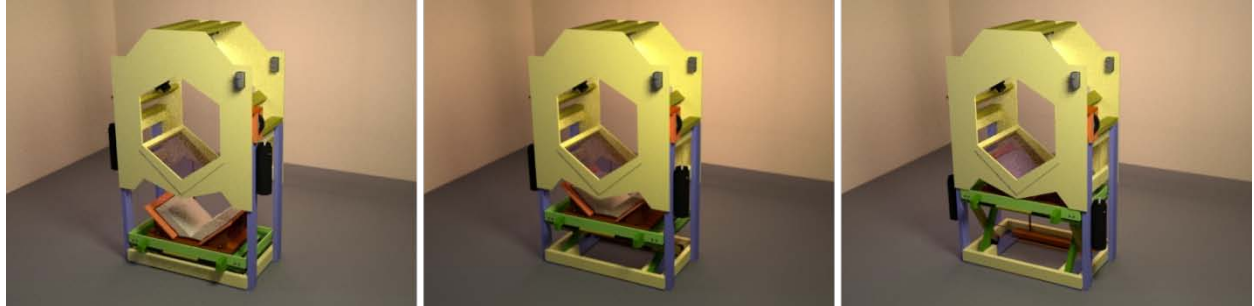
Weibel, Nadir, Moira C. Norrie, and Beat Signer. "A Model for Mapping between Printed and Digital Document Instances." *DocEng'07*, August 28–31, 2007. Winnipeg, Manitoba, Canada. ACM 2007.

\_\_\_\_\_. *Who are ATIZ customers and What are they saying about BookDrive?* available from <http://www.atiz.com/customers>. Internet. accessed Feb 23, 2012.

**APPENDIX I:  
BOOK SCANNER IMAGES**



Daniel Reetz' first DIY book scanner (left) and the most recent DIY community model (right). (Images available from [diybookscanner.org](http://diybookscanner.org))



3D renderings of the design for the seminary book scanner, with cradle in different positions.



Back of 3D rendering





Seth Georgson and father Steve Georgson during construction of the seminary book scanner.

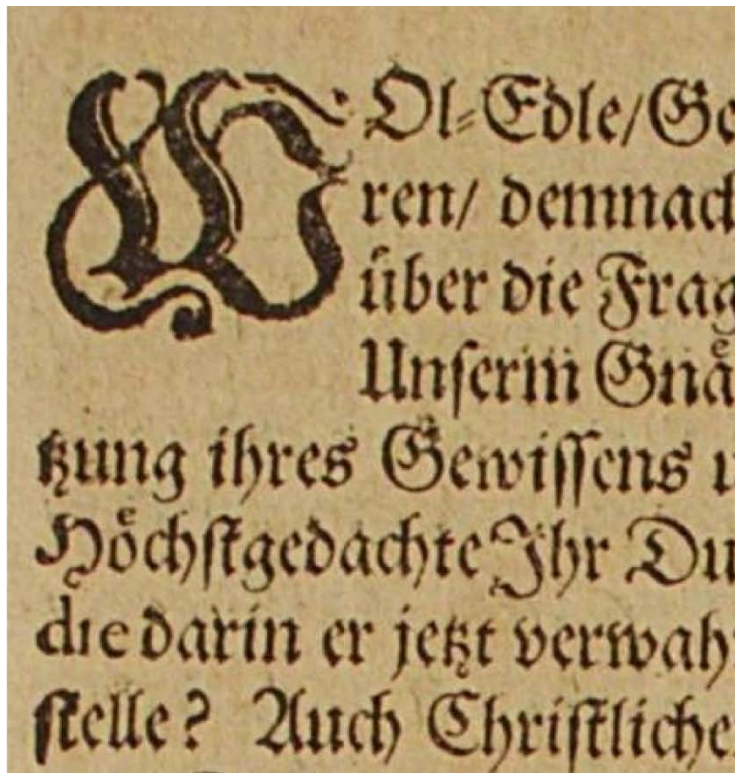
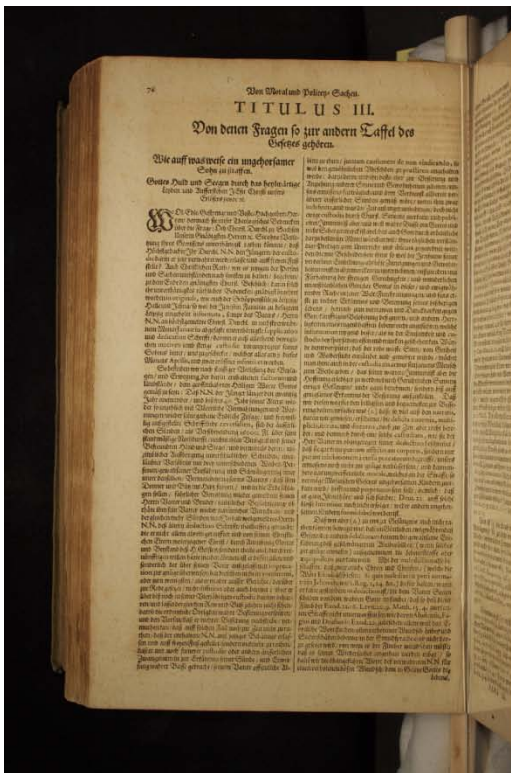


Completed book scanner set up in the seminary archives room.





Scanner with cradle in “up” position.



A scanned page (left) and a portion of the text expanded (right).

## **APPENDIX II: SEMINARY BOOK SCANNER OPERATOR'S MANUAL**

### **Overview**

#### ***Purpose and Concept***

Welcome to the world of book digitization. This is the manual for the seminary book scanner. It covers all aspects of usage and maintenance. It would be wise to skim through the entire manual before attempting to scan any books.

The seminary book scanner is a complex machine with a simple concept. The goal is to transfer books from a paper format to a digital one. While this could be done with a simple copy machine or flatbed scanner, the process is slow and tedious. A book scanner is designed specifically for bound books and increases the speed of digitization by a large margin.

Like most book scanners, the seminary book scanner must be operated manually. This means for each page the book must be lifted into position and a button must be pressed to photograph the page. Then the book is lowered and the page is turned by hand. Robotic scanners do exist, but their costs are extravagant (often over \$100,000) and even they must be closely monitored by workers to watch for sticking pages or the occasional bookmark that was left behind.

Despite manual operation, the seminary book scanner is capable of digitizing over 1000 pages per hour.

The seminary book scanner is best used for any bound book, large or small. It may be used for large projects, such as digitizing and preserving sections of the library. It can also be used for smaller projects, such as copying a few pages for a translation project. Expect to spend a few minutes calibrating the machine for the best images. Because of this, set-up is more complex than a copy machine. However, once the machine is calibrated for a book, the entire book can be scanned with few or no further adjustments.

This manual contains suggestions for further improvements on the design of the book scanner. Some may be added to the current design and others are considerations for future designs.

#### ***Parts***

A book scanner has several important parts:

Platen (1) – Glass panes used for pressing the book flat. The platen on this machine is in

a fixed location, which means the cameras and lights never have to be adjusted after they are calibrated for the book which is being scanned.

Cradle (2) – The cradle is the place where the book is held. The cradle is fully adjustable to accommodate books of different sizes. It is also removable to ease adjustment. Rollers allow the cradle to keep the book centered as pages are turned through it. On this machine, the cradle is raised in its track to meet the platen.

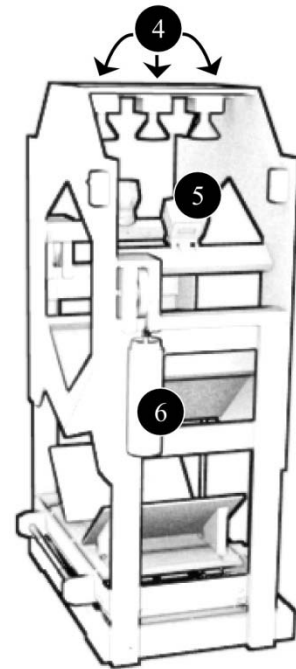
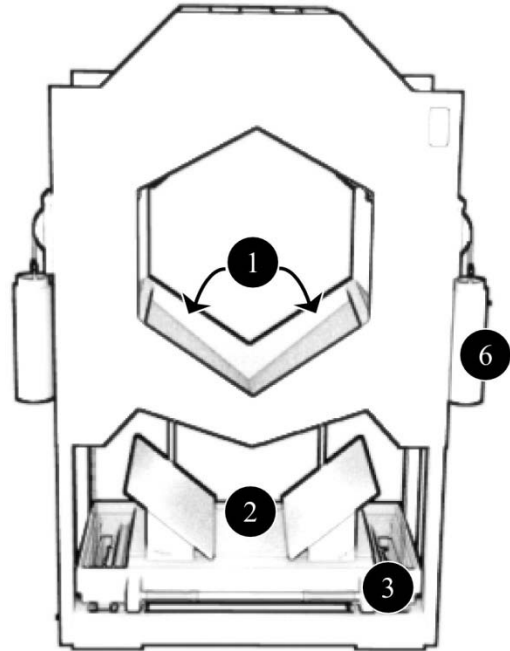
Cradle Track (3) – this is the frame in which the cradle rides. The track is lifted with the cradle in it to press the book against the platen. The track has crossed struts and dowel guides to keep it as level as possible, and is connected by steel cables to the counterweights to ease operation.

Lights (4) – Using the proper lights keeps the book evenly lit but with low glare for high quality photographs. This machine has sockets for three standard sized light bulbs, but not all three must be used.

Cameras (5) – Two cameras are triggered from one button to simultaneously photograph both pages. One camera can be used, but this means that each book must be run through twice; once for the right side and once for the left. The cameras and their mounting brackets are adjustable.

Counterweights (6) – Two separate weights are hung from the sides. Weight can be added and subtracted to balance the varying weights of books, to the user's preference. These weights are connected to the cradle track by steel cables.

Shades (not in diagram) – The shades are installed to keep stray reflections out off of the platen. The installed lights are carefully positioned for minimal glare, and any other light coming in might glare in an undesirable way. The front shade can be lifted over the head of the operator to watch how the book is meeting the platen.



## Operation

### *Operation Summary*

Before a book is scanned, the machine must be calibrated. There are three things that may need adjustment. First, the sides of the cradle should be moved so that the book is centered between them and sits flatly at the set angle. Second, the cameras need to be adjusted for focus and book size. Third, the counterweights may need to be added to or subtracted from in order to balance the book to the operator's preference.

Once the machine is calibrated, the book is scanned in four simple steps.

1. Raise book to meet platen.
2. Press button to photograph pages.
3. Lower book.
4. Turn page.

### *Cameras*

The seminary book scanner is designed to be used with any type of camera, though the camera triggering button may have to be converted for different types. The current button uses an electronic trigger with a 2.5mm TRS plug, compatible with a number of Canon DSLR cameras. There is a second button that operates the automatic focus, should it be enabled. One possible way to use this is to switch the cameras to automatic focus, press the automatic focus button, and then switch the cameras back to manual focus so that they preserve the focused setting.

The camera mounting bracket is held to the frame with two bolts. This mounting bracket can be adjusted up and down and tilted to align the camera properly. These should not need to be adjusted often and will require the use of wrenches.

There is a ¼-20 bolt to hold the camera down by its tripod mount. The camera should be angled to point straight at the glass and centered as much as possible. Software can perform minor corrections on the images taken.

For archive-quality images, 400 ppi is the recommended image resolution at color or grayscale. Since the pages will be cropped, it might be best to image a few pages and then process them on a computer to determine the resolution. This will mean measuring the pages in the book that is being scanned and then dividing the total resolution by the measurement.

For example, if a book is 7 inches tall and the final image is 2000 pixels, the resolution is

2000/7 = 285.7 ppi

This book would be less than acceptable archival quality. To remedy this, the cameras might need to be adjusted to more efficiently use the available resolution, or it may be that the current cameras simply cannot achieve the proper resolution. For the largest books, it may not be possible to achieve archival quality images until cameras with higher resolution are available.

In order to be most efficient and consistent it will be wise to set cameras to a manual mode for scanning a book. Each book may require its own settings. Scan a few images and then check them to make sure the settings are usable. If you don't check them, you might accidentally scan an entire book with blurred focus or with the wrong exposure!

Different cameras will require different settings but for reference these are some suggested settings for a Canon T2i DLSR Camera (This is the camera purchased by Wisconsin Lutheran Seminary this year):

ISO 200

Shutter 1/250

White Balance: Tungsten

Aperture: f/5.0

Flash Off

Manual Focus

These settings were determined by experimentation, and it may well be that with more research better settings are identified.

**Possible improvements:** Display screens can be added and connected to the cameras to show what they are seeing. Computers might be set up so that images are immediately copied, rather than saving the images to memory cards in the cameras and then transferring them. This option will depend on the type of cameras being used.

### *Electrical System*

The seminary book scanner has a simple electrical system consisting of a pair of electrical sockets, three standard light bulb sockets, and a switch which turns the entire system

on and off. The electrical sockets are intended for any peripheral use, such as power adapters for the cameras, task lighting, or display screens.

Not all of the light sockets must be used, but they are there for flexibility. The type of bulb used will make a difference for the quality of images. Halogen bulbs are recommended for their brightness and color. However, halogen bulbs emit heat and ultraviolet (UV) radiation which can be damaging to books.

Regular, tungsten filament light bulbs are similar to halogen, but not as bright. There is nothing to recommend them over halogen.

Compact fluorescent (CFL) bulbs are bright but generally have poor color reproduction. For this reason they are not the ideal bulb.

Light-Emitting Diode (LED) lights might be the best type, but at this time they are expensive. They are cool and have no UV emissions. They are capable of good color reproduction if they are built properly. As costs come down, they might become the best choice.

For now, halogen is the most recommended choice of lighting. It is the brightest and most cost-effective. The damaging heat and radiation are not exposed to any book long enough to cause damage unless the book is going to be scanned hundreds of times. The scanner is initially set up with two 75-watt halogen spotlights.

**Possible improvements:** The light switch on the front could be moved to a more convenient location. The halogen lights might be replaced with LED spotlights.

### ***Mechanical System***

The mechanical operation of the seminary book scanner is not complex. It involves holding a book open at a specific angle (100 degrees) and lifting it. The design is intended to expedite this process for the scanning of many pages quickly.

When the machine is not in use, the locks on the sides should be engaged in order to keep the cradle from springing up. The locks are simple bolts on each side, installed so that they can be slid to engage or disengage.

With the locks engaged, the cradle can be lifted off the cradle track. This simplifies adjusting it for a book. The book should be centered on the cradle base and the two sides should be lined up so that it can open flatly against the sides. Especially thick books may need extra support under the spine so dowels and foam pieces are provided to be used as needed.

The cradle can be put back in the track so that it moves freely side to side on its rollers.

With the book and cradle in place, the counterweights can be adjusted to the operator's preference. This is done by hanging weights off the bottom of the counterweight pipes on the sides. There are chains provided for doing this.

Once the weight is adjusted, scanning can commence. Lift using the handle on the front. Watch to make sure the book is meeting the platen properly. Press the button to snap the pictures. Make sure the cameras are properly calibrated before scanning a whole book, to eliminate having to re-scan any books.

**Possible improvements:** The counterweight system could be redesigned to be simpler and more aesthetic. A future design might rework the scissor-style struts to move more smoothly. Machined metal instead of wood would help for strength and weight. Greater precision would reduce areas of friction.

### **Maintenance**

The seminary book scanner is a fairly low-maintenance machine, only requiring occasional cleaning. It is possible that some moving parts will need adjustment from time to time.

There are skateboard bearings on most of the pivots of the machine. Note that the nuts holding these in place should not be over-tightened or they will cause the bearings to bind.

### ***Cleaning***

The only part of this machine that may need regular cleaning is the glass platen. A simple glass cleaner on the top and bottom of each pane will work fine. The glass can be removed but this is not terribly easy and will likely need cleaning again once it is re-installed.

Other parts of the machine should only require occasional inspection for dirt and dust build-up.

### ***Moving***

The seminary book scanner is a large and fairly heavy device and so it should not be moved more often than necessary. However, it is designed to facilitate disassembly.

The top half of the machine can be separated from the lower half of the machine with only a few bolts. First, remove extra parts that may get in the way. The shades should be removed. The counterweights should be disconnected from their cables and the cables that hold them should be removed from the machine. Depending on how far it is going, it may be advisable to remove the glass platens by sliding them out the front or back. The two wooden

dowels that serve as guides in the back should be removed by sliding them upwards.

Four bolts hold the top on each side, two on each leg. Once the bolts are removed, the top half may be lifted off. This should be done by two people. The legs may then be removed entirely by detaching the bolts that hold them at the bottom.

