Distributed Proofreading

Dr. Gregory B. Newby
Arctic Region Supercomputing Center
newby@arsc.edu

Charles Franks
Project Gutenberg Literary Archive Foundation
charlz@lvcablemodem.com

Abstract

Distributed proofreading allows many people working individually across the Internet to contribute to the proofreading of a new electronic book. This paper describes Project Gutenberg's Distributed Proofreading project, along with our general procedures for creating an electronic book from a physical book. Distributed proofreading has promise for the future of Project Gutenberg, and is likely to be a useful strategy for other digital library projects.

1. Introduction

According to the Internet Public Library (http://www.ipl.org), there are more than 20,000 publicly-accessible electronic books or eBooks on the Internet. About 30% of those originated with Project Gutenberg™ (http://www.gutenberg.net), one of the oldest all-electronic producers of content on the Internet.

Rather than emphasize particular authors, genres, languages or time periods, Project Gutenberg's content is very much determined by volunteer eBook producers' interests. All of the content comes from volunteers, rather than from a central collection development policy or work paid-for-hire. For a small number of eBooks, content was born digital – this occurs most frequently when a contemporary author donates his or her work for distribution by Project Gutenberg.

For well over 95% of the content, however, items were originally in print form and needed to be digitized for distribution. Creating and freely distributing eBooks is the mission of Project Gutenberg. General steps for digitization of Project Gutenberg eBooks are to:

1. Identify a printed copy of a book of interest
2. Assess whether this item is in the public domain in the United States, usually by seeking a copyright clearance from Project Gutenberg's copyright clearance coordinators
3. Scan the book
4. Perform optical character recognition (OCR) on the book
5. Proofread the OCR output
6. Confirm the formatting meets guidelines and submit the eBook for distribution

Over 2400 new Project Gutenberg eBooks were created in 2002 using this process. For 2003, we are striving to release our 10,000th eBook by completing an average of 75 new eBooks per week.

In order to accomplish this goal, we have embraced a new and innovative method for the most labor-intensive and time-consuming activity in eBook production: proofreading OCR output against actual page images. We call this method Distributed Proofreading.

In this paper, we will present a description of the Distributed Proofreading project, procedures, outcomes and plans. Future plans, as well as the possible role of DP for other digital libraries, follow.

2. Distributed Proofreading

Distributed Proofreaders (DP) is by far the easiest way for people to get involved with producing a Project Gutenberg eBook. We believe it can offer a useful training ground for many digital libraries. DP makes the process more efficient at every step in the eBook production phase. During 2002 DP contributed about 75% of Project Gutenberg's new eBooks.

The basis of DP is centralization. By centralizing copyright, scanning, proofreading and submission, economies of scale result, with enhanced qauality. For proofreading, the main bottleneck in eBook production, DP means that thousands of individuals with only a few minutes to donate can help to make a difference.

The general steps for DP are the same as for any other eBook, but modified to follow the centralized model. First is the identification of candidate books. At this phase, several volunteers are engaged in book-hunting activities. They seek book auctions, library sales, and auction opportunities, purchasing books by the box.

For items suitable for scanning, the title and verso pages are scanned and sent to the copyright team. The copyright team is a small number of experts who can confirm whether the book is in the public domain in the United States. They perform "due diligence" (as defined by copyright law) to insure no copyright violations are made. Scans are kept centrally in case of a copyright question in the future.

Once a book is copyright-cleared, it is sent to one of two locations where volunteers run page-fed scanners. These volunteers chop the spine from the book and are

then able to scan a complete book in only a few minutes. The scans are run through OCR, and the OCR'd text and images are uploaded to the DP server. DP uses ABBYY FineReader version 6.

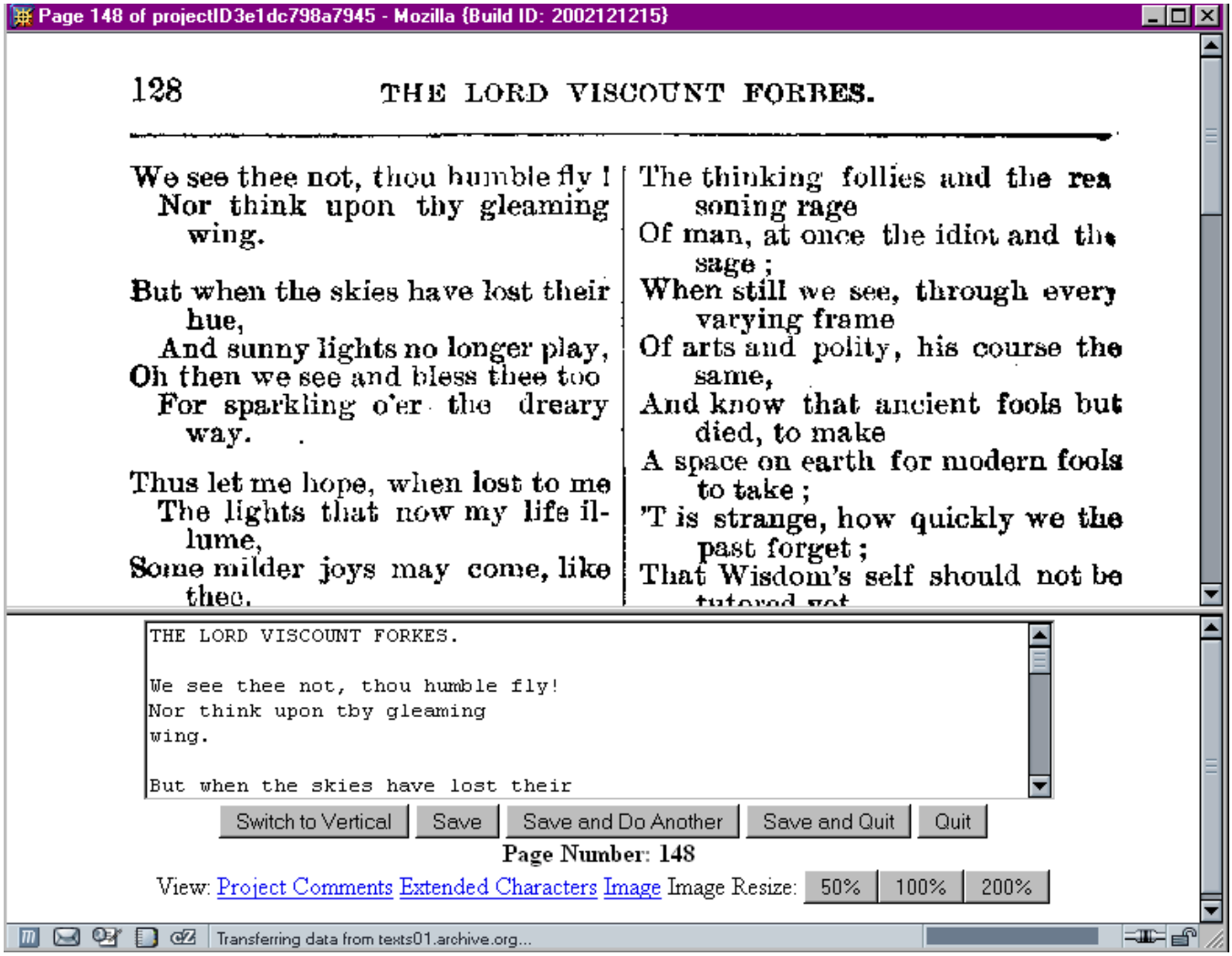
On the server, page scans and information about the book are input to the DP database system. When all is ready, the book is released for proofreading. The Project Manager (PM) for the book monitors its progress through the DP site.

The proofreading phase is where the greatest increase in performance is seen. Previously, one person would typically perform all of the steps, often taking months to complete the scanning, OCR and proofreading of an eBook. With DP, many individuals complete proofreading much more quickly.

Proofreading is accomplished using a Web interface (Figure 1). Proofreaders may take as long as they need to compare the OCR output to the page images, making necessary changes and updating formatting. Each page image is viewed by two different DP proofreaders. The first proofreader sees the original page scan and the raw OCR output. The second proofreader sees the original page scan and the output from the first proofreader. Proofreaders may complete as many pages as they would like in either phase of any of the available eBook projects. At any given time, there are usually forty or more eBook projects to choose from with different topics and in different languages. Some easier books (such as English fiction) are geared towards novices.

Once all pages have been proofread, the PM or a Post Processor (PP) performs final assembly of the eBook (making sure that page breaks, headings, and other structural items are properly formatted). Various off-the-shelf and in-house developed programs are used to assist in error checking and formatting.

Figure 1: Distributed Proofreading Web Interface



The PM/PP submits the eBook file(s) plus any additional information about the book to the posting team, who performs a final check, assigns a unique eBook number, generates metadata, uploads the files, and announces them.

The success of DP is evident in its numbers. In 2002, over 250,000 unique pages were proofread at DP, with each being examined twice. Over 6000 individuals performed proofreading ranging from a single page of work to many thousands. The page rate has gone up consistently, with a large spike thanks to Slashdot coverage in November (http://slashdot.org). DP is seeking to maintain or exceed December 2002's rate of over 110,000 pages per month throughout 2003.

3. The Future of Distributed Proofreading

Apart from an ongoing desire for growth, the biggest immediate goal for DP is to add facility for text markup. The first step towards this goal has been to enable tagging for italics and other text emphasis with and . DP volunteers are able to add such markup easily.

Other aspects of markup that are easily and automatically generated are paragraph and chapter breaks. For the future, though, we desire to add more complete XML markup (based on TEI LITE) to eBooks. This will enable conversion "on the fly" to different electronic formats, and will retain more of the information in printed texts. In the future, DP will markup and reformat all existing Project Gutenberg eBooks as XML documents.

DP is also seeking new volunteers and new opportunities. By coordinating with real-world organizations, we hope to have special projects – for example, for an entire class could spend an hour or so to complete a whole eBook.

4. Utility of Distributed Proofreading for Other Digital Libraries

Nearly all digital libraries are engaged in creating electronic content from physical artifacts such as books, catalogs, pamphlets, and so forth. While some have relatively deep pockets to pay for this digitization, and others might have rare or fragile materials that cannot be easily handled by volunteers, we believe that many digital library projects could benefit from the type of procedures that DP has developed.

For the casual volunteer, DP provides a very low barrier to getting involved with eBook creation. Using a standard Web browser, the process of signing up as a DP volunteer and viewing the first page for proofreading takes only a minute or two. Compared to the time, training and expense of setting up a scanning and OCR station to create an eBook from scratch, DP is extremely cost effective for volunteers.

For digital library projects engaged in work for hire or volunteers with more time or expertise, DP has created an infrastructure that scales well to make the creation of eBooks a routine task with high quality control. The author's experience with digitization and markup in a (non-Gutenberg) rare books project was that creating a single eBook cost about $1000. Similarly, the first eBooks produced with DP also took about $1000 worth of time and equipment. The difference is that the incremental cost with DP is very small – essentially, the cost of purchasing and shipping books. For professionally completed eBook generation the author was engaged in, *each* eBook cost $1000.

Project Gutenberg has a small number of key volunteers who contribute many hours per week, a larger number of volunteers who work somewhat less, and a very large number of volunteers who might only spend an hour or two per month working on eBook creation. This experience seems very typical for volunteer-driven organizations. Our goal, which is common to many digital libraries, is to make it easy for new volunteers to get involved – while at the same time having a scalable and efficient infrastructure to bring these new volunteers aboard. It is our contention that Distributed Proofreading represents the state-of-the-art in eBook creation for digital libraries in volunteer-driven organizations.

5. Conclusion

At the start of 2003, we expect approximately 1 million pages will be proofread during the year, significantly adding to the total number of eBooks online via Project Gutenberg. Visit the Distributed Proofreading site online at http://text01.archive.org/dp or http://pgdp.org.

5. Acknowledgement

Our thanks to the many Project Gutenberg volunteers for their work in making Distributed Proofreading effective. We are grateful to the Internet Archive for hosting DP and to iBiblio (http://ibiblio.org/gutenberg) as the main distribution point for Project Gutenberg eBooks.