GREENSTONE DIGITAL LIBRARY

FROM PAPER TO COLLECTION

Dr Michel Loots, Dan Camarzan and Ian H. Witten

Human Info NGO, Belgium
Simple Words, Romania
University of Waikato, New Zealand

We want to ensure that this software works well for you. Please report any problems to *greenstone@cs.waikato.ac.nz*

Greenstone gsdl-2.50 **March 2004**

About this manual

This document explains how to create CD-ROM collections from paper documents. It describes in full detail the procedures and economics involved in the scanning and optical character recognition (OCR) processes, so that you end up with text in the right format to apply the Greenstone software. It also describes how to create and edit the material associated with a collection.

We have tried to be as plain as possible in our explanation. Reference to any trade mark or company product is purely for illustrative purposes, and does not imply that we endorse or favor this product over any other.

Companion documents

The complete set of Greenstone documents include five volumes:

- Greenstone Digital Library Installer's Guide
- Greenstone Digital Library User's Guide
- Greenstone Digital Library Developer's Guide
- Greenstone Digital Library: From Paper to Collection *(this document)*
- Greenstone Digital Library: Using the Organizer

Acknowledgements

The scanning operation and other know-how relating to the creation of collaborative non-profit collections have been developed by Dr Michel Loots, MD, of Human Info NGO and HumanityCD, Dan Camarzan of Simple Words, and their team of collaborators in Brasov, Romania.

The Greenstone software is a collaborative effort between many people. Rodger McNab and Stefan Boddie are the principal architects and implementors. Contributions have been made by David Bainbridge, George Buchanan, Hong Chen, Michael Dewsnip, Katherine Don, Elke Duncker, Carl Gutwin, Geoff Holmes, Dana McKay, John McPherson, Craig Nevill-Manning, Dynal Patel, Gordon Paynter, Bernhard Pfahringer, Todd Reed, Bill Rogers, John Thompson, and Stuart Yeates. Other members of the New Zealand Digital Library project provided advice and inspiration in the design of the system: Mark Apperley, Sally Jo Cunningham, Matt Jones, Steve Jones, Te Taka Keegan, Michel Loots, Malika Mahoui, Gary Marsden, Dave Nichols and Lloyd Smith. We would also like to acknowledge all those who have contributed to the GNU-licensed packages included in this distribution: MG, GDBM, PDFTOHTML, PERL, WGET, WVWARE and XLHTML.

Contents

1 Introduction

One goal of the Greenstone Digital Library software is to empower organizations such as universities, United Nations agencies, non-governmental organizations, non-profit organizations and governments to create varied collections of information that can be delivered online or on CD-ROM.

Typical steps that have to be implemented are:

i. Selecting the documents to be included
ii. Securing copyrights permissions to use these documents in the digital library
iii. Scanning and OCR of the hard-copy documents which are not available in to digital form to have a perfect digital format
iv. Converting all documents to a format (integrating text and images) which can be imported into Greenstone (preferably HTML or Microsoft Word, but others are also covered at varying levels of precision by a "plugin" (see the *Greenstone User's Manual*)
v. Tagging the chapters, paragraphs and images of the digital documents
vi. Organising the collection into a optimally structured digital library
vii. Building the digital library using the Greenstone software
viii. Printing and distributing the collection on CD-ROM and/or distributing it over the Internet

In order to create a digital collection, the publications must be available in digital format. If books, newsletters or other documents are only available on paper, they will need to be scanned and processed into machine-readable form (step iii). Usually this is done using optical character recognition (OCR), but sometimes by manual retyping. This process is covered in Chapters 2-4 of this manual.

Step v. enables the different parts of a document to be independently

selected and displayed by readers in the final library, while step vi. involves assigning attributes to the documents such as subject categories, keywords and bibliographic data for ordering and searching the library. These steps are covered in Chapter 5 of this manual.

This manual introduces many issues that affect the editorial process of creating a collection from paper. Before reading on, you should consider these questions:

- What is the goal of your collection?
- What is your target group?
- How big is it—local, regional, or global?
- How many documents are you making available?
- How many pages?
- How much graphics content?
- Does the material split into parts that will be consulted by a limited audience and parts that need to be disseminated widely?
- Are the documents already available electronically?
- If so, in which formats? (Note incidentally that PDF files are not automatically equivalent to digital full-text form, as they often contain only page images.)
- What is the copyright status of the documents?
- Who owns the copyright?
- Are there other organizations with the same target audience?
- Are you willing to collaborate with other groups?
- What budget is available for the whole project?
- What human resources are available (in person-months) for co-ordination, editing, scanning and programming?
- How many computers are available for this project?
- How many CD-ROMs do you want to distribute?
- Will they be free, or for sale?

2
Scanners and scanning

The first step in converting paper documents into a digital library collection is to obtain images of all pages of all publications in digital format. The next stage is optical character recognition (OCR), and clean, high-quality images are essential for successful OCR. The digitization process requires a scanner capable of working at a resolution of 300 dpi (dots per inch). Most scanning can be done in black-and-white, but if color illustrations are included they must be scanned with a color scanner. In most cases the covers of the book contain colors and will have to be scanned as a color photographic image.

2.1 Scanners

Scanners are available in all price ranges, and all shapes and sizes. They range from $100 for flat-bed scanners to upwards of $50,000 for large industrial scanners from manufacturers such as Bell & Howell.[1] There are many websites that offer a wide range of scanners for sale. To locate them, just search for "scanners" in search engines like Google, Altavista, or Yahoo.

The output format of a scanned page is a computer file that is usually stored in TIFF or Bitmap format. Compressed TIFF IV is the best format to use. An average page scanned and converted to this format occupies only 50 Kb, compared to perhaps 2 Mb for the equivalent page in uncompressed Bitmap form.

Low-cost flat-bed scanner

Low-cost flat-bed units are the cheapest and most widely available type of scanner. There are many brands: HP, Agfa, Acer, etc. Prices range from $100 to $300. Both black-and-white and color images can be scanned. The low price allows each computer to have its own scanner.

[1] All sums of money mentioned in this document are in US dollars, and were current in 2001.

Disadvantages of these scanners include the medium quality of the result, the slow rate of scanning, unreliability in warm environments, and relatively frequent breakdown. Pages must be scanned manually, one by one. Each page must be positioned carefully on the scanning plate to ensure that it is aligned correctly. Productivity of these scanners is low. Despite manufacturers' claims that each page can be scanned in less than a minute, the fact is that rates exceeding twelve pages per hour are rarely achieved. The scanning process monopolizes the computer on which the work is being performed.

Consequently these scanners are useful only for small jobs with limited numbers of pages—no more than 200 to 400 pages a month on a regular basis, or one-time jobs of up to 1000 or 2000 pages.

Low-end scanner with sheet feeder

Low-end scanners with sheet feeders typically cost between $500 and $1200. Ten to fifty pages can be inserted, scanned and processed at once: thus the operator does not have to attend constantly to the machine. This increases capacity up to 150 to 200 pages per day. These scanners are more robust, and have a larger lifespan before repair—usually in the range 30,000 to 50,000 pages.

A disadvantage is that only one side of the page is scanned at a time—the stack of pages must be reversed and rescanned in order to obtain an image of both sides. This often creates problems because sheet feeders are never without problems and sometimes pages get blocked.

These scanners are useful for up to 1500 to 3000 pages a month.

Color scanners

Any scanning operation invariably involves some color images, so a color scanner will always be required. Generally speaking, less than 5% of any publication contains color images, plus the cover. Thus a low cost flat-bed scanner as described above suffices. It is advisable to select one capable of scanning up to 600 dpi resolution.

Professional duplex scanners

Professional scanners are reliable, heavy-duty machines capable of processing a large volume of pages—typically from 2000 pages to 10,000 pages per day. They have an automatic sheet-feeder tray system that processes batches of about 50 to 200 pages. The best and fastest are duplex machines that scan both sides of the page at once.

Professional duplex scanners require a powerful computer with a hard disk of at least 10 to 20 Gb. Prices range from $5000 to $50,000. For example, the Canon DR-6020 duplex scanner costs $5000 and works with double-sided documents. It has a capacity of about 2000 pages per day and a lifespan of 600,000 to 800,000 pages. Bell & Howell and Fujitsu scanners range from $10,000 to $50,000 and have a lifespan of many millions of pages.

Micro-fiche scanners cost from $15,000 for a semi-manual unit to $80,000 for one that operates fully automatically.

Scanning programs

Every scanner comes with its own software, which means that the program must be installed on the computer that manages the scanner. Some have a computer card that needs to be installed in your computer to speed up the scanning operation.

2.2 Preparing the documents

Before being scanned, documents must be properly prepared. Dusty documents must be cleaned, humid documents dried, clips removed, pages unfolded.

The spine of each book should be removed by cutting it off, straight and precisely. Books provided by libraries must often be rebound, and if so you should be particularly careful when removing spines in order to facilitate smooth rebinding.

If there are just a few documents, cutting can be done manually with a ruler and cutters. Be careful with your hands! For more documents, special manual cutting machines are available.

For high volumes—more than 20 documents—we recommend asking a printer or copy-shop if you can use their professional cutting machine. Do not forget to remove metal clips which could damage the cutting blades.

2.3 The scanning process

Using software provided with the scanner, a digital image of each paper page is scanned and transformed into a Bitmap or TIFF image. These images should be stored on hard disk with standard filenames. The OCR process starts once some or all of a batch of documents have been scanned. It can be undertaken by the person who operates the scanner, or by someone else.

Typically a scanning resolution of 300 dpi is needed, although sometimes 200 dpi is acceptable.

Quality control

The final goal of scanning is either to OCR the pages to obtain perfect word processor or HTML versions of the publications, or to produce enhanced image files such as PDF image files. In either case the quality of the image is very important. If quality is sub-standard, image files will not look good and will consume more memory. Image quality seriously affects the OCR process: with sub-standard quality, productivity deteriorates by up to 40%. OCR typically represents more than 90% of the total cost, so scanning quality can have a very substantial effect on the final cost.

The quality of the TIFF file can be enhanced by adjusting the scanning process to each type of paper, using settings provided by the scanner software. Relatively transparent kind of paper will require a lighter setting; the contrast must be adjusted depending on the quality of printing, and so on.

First divide the material into batches with similar paper and print qualities. Perform OCR tests on a sample from the first batch to determine the optimal settings. Then scan all material in this batch before proceeding to the next one.

Filename conventions

Give each book or document a job number or unique code, which will become the name of the folder that contains all TIFF images in the document. Depending on the computer system (DOS, Windows, UNIX, LINUX, etc) from 8 characters to 128 characters can be used in a filename. We recommend restricting this unique document identifier to 8 to 16 characters. The first five characters might identify the document, the following letter might contain a language code, and the remaining characters might identify the particular page. For example, the identifier *u7548e12.tif* might identify the TIFF image of page 12 of a book written in English with code *u7548e*.

Allocate one directory on the hard disk for scanning jobs, say *scanjobs*. Then make a subdirectory for each job. Within this make a subdirectory for each publication—say *u7548e* for the above document. Store all the TIFF images of the publication, including color images, in this folder.

2.4 Productivity and resources

You should not underestimate the magnitude of the scanning operation—and particularly the OCR process that follows. It is best to consider scanning and OCR as completely separate activities. The optimal choice from an economic and practical point of view should be made individually for each one.

Some points to consider are the investment in scanners and computers that is necessary; the availability of appropriate space and human resources; training the workforce; salary costs; the initial and total number of pages to be scanned; deadlines; and whether documents can be outsourced to third parties.

Scanning costs

An important decision is whether to invest in scanning equipment and perform all scanning oneself, or outsource it to a scanning company. The main considerations are:

- pressure of time for the scanning job;
- total number of pages;
- salary costs of those who perform the scanning.

The people who perform the scanning must be highly motivated, technically skilled, and quality-oriented.

The typical cost of scanning by a professional company is $0.06 per page. To this must be added the cost of shipment, which can be up to $0.03 per page for transport from developing countries to developed countries, and $0.015 per page for transport within countries.

Table 1 estimates the cost of doing it yourself, using various scanner types. Note that all figures are approximate. They are provided as rough guidelines based on the authors' experience. The first three columns concern labor costs. The first is the capacity in pages/month, assuming full-time work. The resources required in person-hours per page is obtained by dividing the number of working hours per month by the pages/month capacity in the second column. It is shown in the second column, which assumes 180 working hours per month.

Table 1. Scanning cost

	Capacity (pages/ month)	Hours/page (180-hour month)	Cost/page (assuming $4/hour)	Scanner acquisit-ion	Scanner lifespan (pages)	Outsourced pages for scanner cost (at $.06 each)
Flat bed scanner	2,500	0.072	$0.288	$300	7,000	5,000
Scanner with sheet-feeder	8,000	0.0225	$0.09	$800	30,000	13,000
Professional: low-end duplex	40,000	0.0045	$0.018	$6,000	600,000	100,000
Professional: high-end duplex	150,000	0.0012	$0.0048	$50,000	8,000,000	833,000

To determine the price per page, multiply the total hourly salary costs in your situation by the second column of Table 1. As an example, the third column gives the price of in-house scanning at a salary rate of $4/hour—not including investment costs.

These calculations assume that the scanner is used for a sufficient volume to justify the investment. The final three columns of Table 1 give more information about the cost of the scanner itself. The first of these shows the acquisition cost of the scanner, and the next gives its expected lifetime. The last shows the number of pages that could be scanned commercially, at a cost of $0.06/page, for the price of the scanner alone.

Of course, many other factors affect the choice of scanner: availability of funds, need to minimize dependence on others, desire to build local capacity, obligations to libraries to scan books locally and not transport them, and so on.

The above figures give some idea of the volume of pages needed to justify different levels of investment. Rarely will an institute or organization need to scan 800,000 pages. At such levels more complex issues arise—such as maintenance and the possibility of recouping costs by offering scanning services to others—that we will not discuss here.

It is tempting to regard the development of scanning capacity as a commercial venture, particularly in developing countries. But one should always bear in mind that scanning is not a repetitive business. Once documents have been scanned, clients never place new orders for the same documents—no matter how good the relationship with the scanning company. From a commercial point of view, intensive marketing efforts are needed. We do not advise NGOs or other non-profit organizations to venture into this realm without thorough initial trials and a carefully-considered business plan.

In conclusion, if 10,000 to 50,000 pages are to be scanned, one should consider outsourcing the job. A low-end professional scanner costing about $6000 can only be justified if more than 100,000 pages have to be scanned. You might consider banding together with a few other institutions—perhaps NGOs or libraries—to purchase such a scanner.

3
OCR: Optical Character Recognition

An optical character recognition or OCR system transforms a scanned image into text. The input is a digitized image in TIFF or Bitmap format—preferably a clean, high-quality image. The output is a word-processor or web file, typically in RTF, Word, or HTML format.

The following steps are involved in converting paper documents to computer form:

- scanning;
- page layout analysis;
- recognition;
- scanning images and tables.

Following these, you must perform quality checks on the resulting files, and save them in the appropriate format.

On the market are many good OCR programs, with prices ranging from $100 to $400.[2] For example, among many others are:

- *Read-Iris* (http://www.readiris.com/)
- *Omnipage* (http://www.omnipage.com/)
- *Fine-Reader* (http://www.finereader.com/)

All information, including lists of local distributors, can be found on the manufacturers' websites. Among these, in the authors' experience the most user-friendly are Fine-Reader and Omnipage. Fine-Reader is cheapest, costing about $100. It offers a great deal of flexibility, and the widest range of different language options.

[2] Recall that all sums of money are expressed in 2001 US dollars.

A choice must be made between undertaking the scanning and OCR in-house or outsourcing it to a commercial organization. To do it in-house requires a scanner, OCR software program, OCR skill development, and a quality-conscious, highly motivated workforce.

3.1 The OCR process

The OCR process differs from one OCR program to another, and each one requires a considerable amount of learning. The program's manual will explain this process in detail. Four points deserve particular attention: quality control, tables, images, and specialized material such as formulas, foreign characters etc.

Quality control

We cannot place enough emphasis on quality control. Quality checks are best performed by native speakers, or people with an excellent command of the language to check. The best people are at the university or high-school level. We should also note that young people tend to sustain higher concentration than older people for this kind of work.

Normally there are four quality checks.

The first is performed at the same time as OCR. Every OCR program has a built-in spell-checker that highlights every suspect letter. At the same time the image of the word appears too, making it easy to check and correct the error.

The second is a general check of the text once the OCR process is finished. Common errors are to miss a page, a paragraph, chapter titles, and so on. A general overview is necessary to check if pages are missing. It is essential to check titles, chapter headings, paragraphs, and tables.

The third is a spelling check using Microsoft Word. This program has a dictionary that is often more sophisticated than the one embedded in OCR programs. By importing the book into Word and performing a spelling check there, more errors can be found and corrected. Be sure to add to the spell-checker any particularly difficult or error-prone words, or scientific and technical terms common in that type of publication.

Finally, the completed document should be checked by an independent person who samples the complete book and checks for errors, problems with tables and images, tagging, and the general look of the resulting text. Only after this final check can a book be considered ready for digital

dissemination.

Tables

OCR programs do not cope well with tables. Moreover, tables are hard to check. They contain many digits, sometimes with points and commas, and entries are easily misplaced into the wrong row or column. They require concentrated effort, dedicated work, intensive proof-reading, careful checking, and good quality control. They can be handled in three basically different ways.

First, tables can be treated as images. This involves scanning them as black-and-white images and placing them in this form at the appropriate point in the document. This is the easiest solution. There are no errors, and the only time taken is that involved in creating the image. However, this solution consumes more memory than others. Also, the resolution is not always sufficient when large tables are displayed on a computer screen. If you make the complete table fit, the resolution is too small. If you make the table over-wide, the user must scroll to see all columns and rows, and cannot get an overview of the contents.

Second, tables can be recreated manually by making a table with the same number of rows and columns and filling the entries by typing them in, character by character.

Third, the table can be OCR'd. This saves time compared to the manual process, but has a potential for more errors. Columns sometimes get merged, and commas and points are not recognized.

Images

Publications contain three different general types of image:

- black and white line art;
- black and white photographs;
- color photographs.

Black and white line art should be scanned in line art mode and saved as GIF or PNG files. Black and white photographs should be scanned in greyscale mode and saved as GIF or JPEG files. Color photographs should be scanned in color mode and saved as JPEG files. Generally speaking, medium-quality JPEG provides adequate resolution.

For most collections, images consume the bulk of the space required on a hard-disk or CD-ROM. This makes it important to optimize each image for clarity and visibility, while minimizing its size. To save space you might drop some or all of the images if they are not relevant to the text.

Images should be scanned separately, one by one. We recommend giving the image files a name that consists of the first five or six characters used to denote the document followed by the number of the page on which the image was found. An alternative, assuming each document is in its own directory, is to simply use the letter *p* followed by the page containing the image. If there are several images on a single page, append an additional letter *a*, *b*, *c* … to the filename. For example, if a JPEG image appeared on page 36 of the publication *u7548e* discussed earlier, it would be placed in a file named *u7548e36.jpg* or *p36.jpg*.

Once the images have been scanned, you can put batch-processing programs to work to resize or enhance all the images at once.

Specialized material

Many documents contain specialized material such as special characters, formulas, and difficult pages. Special characters generally relate to different languages and diacritical marks. The language option for the OCR program should be set for the specific language being read. Formulas will have to be recreated manually. Sometimes this is not possible in the OCR program, but only in a word processor like MICROSOFT Word. Difficult pages that contain complex material or are damaged so that a clear image cannot be obtained might have to be retyped manually.

3.2 Productivity and resources

As mentioned earlier, you should not underestimate the difficulty of OCR. Although the economic and practical options for OCR should be considered separately from scanning, similar points arise: the necessary investment in computers; the availability of human resources and management skills; training the workforce; salary costs; the total number of pages to be processed; and whether documents can be outsourced to third parties.

In this section we share our experience of OCR operations in Belgium, Romania and India. All case studies, calculations and figures assume average situations, documents of standard difficulty (including tables and images) such as are found in most archives or libraries, very high-quality

results, and a medium- to long-term operation.

Intensive OCR

OCR is difficult. It demands great concentration and much skill. Before attaining peak productivity level and quality, a learning period of about six weeks is needed.

Typically, best results and productivity are achieved during the first hours of each day. After three hours of OCR work, productivity declines very rapidly, perhaps to 50% of the initial level. After six hours most people become very tired.

The same kind of evolution occurs over the initial weeks. In the first few weeks everyone achieves fairly high productivity, but after that up to two-thirds of people become bored and frustrated. These people either quit or perform poorly in terms of quality and productivity. Even those who pass the first three to five critical weeks and become part of the regular work team often leave in search of a better position after 6 to 12 months.

The remarks made in Section 3.1 about personnel apply particularly to intensive OCR. Quality checks are best undertaken by native speakers or people with a good command of the language being checked. Young people generally sustain higher concentration than older people for OCR work. As a rule-of-the-thumb, people aged between 18 and 23 years tend to be better suited than those over 25.

Finally, OCR can be a boring job, which makes motivation and sustained commitment to quality exceptionally important.

These facts about OCR lead to the following guidelines:

- Young people between 18 and 25 are best suited for this job.
- Because the first hours are always the most productive, the work should either be organized on a part-time basis or only the most motivated and concentrated people should be selected for full-time work.
- Two-thirds of people tend to quit or get bored after about three to five weeks. This translates into poorer quality and low productivity in the last weeks.
- A regular supply of work is needed to justify the necessary training, to maintain concentration, and to keep spirits high.

Achievable productivity

Table 2. OCR productivity

	Working hours/day	Pages/day	Pages/month
Initial training (6 weeks)	3	6	120
Optimal productivity level	3	9	150 to 200
	7	28	500 to 600

Table 2 gives typical OCR productivity figures. Documents come in all sizes and qualities, and these figures assume that the mix of documents contains an average number of images or tables—say one image and one table of five rows by five columns every 8 pages. They also assume that the page images are of medium to high quality—note that, as discussed above, this depends on the quality of scanning—and that the OCR workers have a good command of the language.

Table 2 gives separate figures for people undergoing training and for those who have reached their optimal productivity level. If a member of the administrative staff were to allocate three hours a day to OCR, they could achieve 180 to 200 pages OCR per month. For full-time staff with proper training, high concentration and dedication to quality, 500 to 600 pages a month can be achieved.

However, the rates that are achieved on difficult pages of low quality, with many columns or many tables, are far lower—perhaps 300 to 400 pages per month for full-time work.

Assume that the salary cost for dedicated and motivated full-time OCR workers is $400 per month, and the overhead—including management costs, computers, office space, utilities, etc.—comes to another $300 to $400 per person per month. Then the cost of OCR comes to about $1.2 to $1.6 per page. Taking into account the training period, total volume, time-span, and layoff costs should the operation close down for lack of work, these figures rise to $1.5 to $2.5 per page.

The cost of in-house OCR should be weighed against the cost of outsourcing the work to a professional OCR company. These typically charge from $1.5 to $4 per page, including images and tables. Human Info NGO/Simple Words has such a unit in Romania, and charges humanitarian non-profit organizations a special price that ranges from

$1.2 to $2 per page. Please contact us at scanning@humaninfo.org for further information and advice.

3.3 Alternatives to OCR

There are two alternatives to OCR that we discuss here.

Manual retyping

One, which eliminates most scanning as well, is to retype the documents manually, using a word processor. This still requires the images and front cover to be scanned, but the remaining pages need not be scanned—thus one can dispense with both powerful scanners and OCR software.

The people who do this work do not have to understand the text. They must be accurate typists and re-key exactly what they see. Retyping does introduce errors, and double-keying is often used to find and correct these. This method involves two people who independently re-key the same document, after which both digital versions are compared word for word using a special software program by an operator who has the original document in front of them. The assumption is that if the same word has been typed independently twice in the same way, it is correct. However, this is not always true, and for extremely high precision, triple-keying is performed.

The advantage of rekeying is that cost is saved because an OCR program is not needed and so the computers can be older, lower-range, or second-hand models—whereas powerful computers are needed for OCR. Also, the work can be performed by people with a lower level of skill. The disadvantages are that a training period of at least two months is needed. Single keying usually produces too many errors, and double or triple keying is needed.

The cost depends entirely on salary level. Typically, re-keyers in developing countries are paid on the order of $150/month. Their productivity could be twenty to thirty pages per day—corresponding to 400 pages per month, images included. With double-keying, this makes the total salary costs around $300 per month, plus overheads.

Image files

A very low cost alternative to OCR is simply to use a PDF image version of the document pages. The cost is only a fraction of OCR's—about $0.1 per page.

Once scanning has been completed and TIFF files are available, an automatic converter (usually Adobe Acrobat or Adobe Photoshop) converts all TIFF files of book pages into PDF files.

The downside is that these files are not searchable. Also, they are quite large—usually 50 Kb per page, plus or minus 20% depending on the quality of the original TIFF file.

PDF image files are slow—sometimes, in developing countries, impossible or prohibitively expensive—to download. They rarely fit on a floppy disk, and do not support text manipulation functions such as cut-and-paste.

The PDF image file method should only be used if no OCR budget is available, and for documents that are likely to be used by a small number of people who have high-speed low-cost Internet access.

3.4 Combining scanning and OCR

If a scanner is connected directly to the computer that runs the OCR software, most OCR programs can scan a page and perform OCR immediately. Page-by-page scanning and OCR is a reasonable strategy for low volumes, but will prove time-consuming for bigger and more continuous jobs.

For up to 100 to 150 pages per month, this solution may suffice. For higher volumes it is faster and more efficient to scan the document first, then perfom OCR on all the pages as a separate step.

4
Three examples: 1000 to 100,000 pages

4.1 Typical small collection: 500 to 1000 pages

Most NGOs have 500 to 1000 pages to scan. This volume can be OCRed in-house if motivated volunteers are available.

SCANNING

The first step is to scan the publications to generate a high-quality TIFF file of each page, and a separate line-art, grey-scale or color bitmap image for each illustration. Assuming that 1000 pages have to be scanned, this might represent a part-time job of about one month—just for scanning. The TIFF files would consume 60 to 80 Mb of hard-disk space, and a good policy is to create a CD-R containing these files. A low-cost flatbed scanner of $100 to $300 will be sufficient for the job. Scanning can be done after working hours or during the weekends by a volunteer in the office or at home.

OCR

The second step is OCR by another volunteer, or team of volunteers, skilled in language and correction. The TIFF files can either be shared between computers, or one computer can be used for the entire job. Typically, it will take five or six months of part-time labor (e.g. 20 hours a week) to convert 1000 pages into perfect Word or HTML documents.

OUTSOURCING

An alternative is to outsource the scanning and OCR process. It would probably cost $1500 to $2000 to convert everything into perfect Word and HTML files.

4.2 All publications from an organization: 5000 pages

Many larger organizations have archives of around 5000 pages of currrent or out-of print books, journals, newsletters, grey literature, etc.

SCANNING

This is too much for a flat-bed scanner. Scanning should either be outsourced (approximately $400 for 5000 pages) or a sheet-feeder scanner purchased (approximately $900). Alternatively, a more expensive scanner could be bought together with a few other institutions or NGOs ($6000 costs divided by the number of participants). All 5000 pages in TIFF format will take about 300 to 400 Mb of hard-disk space. Again, a good policy is to create a CD-R containing these files.

OCR

The second step is OCR by another volunteer, or team of volunteers, skilled in OCR and correction. Again, several computers might be used, or one computer for the whole job. It would take 25 to 30 months of half-time labor (assuming 20 hours a week) to convert 5000 pages into perfect Word or HTML. In practice this is too long and too computer-intensive to manage on a volunteer basis. One would have to pay volunteers, monitor them for performance and quality, provide adequate space, etc, in order to have the job finished within reasonable time at a high level of quality.

Alternatively one could create image PDF files, which would take 300 to 400 Mb of space and would be harder to download over the Internet.

OUTSOURCING

An alternative is to outsource the scanning and OCR processes. It would probably cost $7500 to $10,000 to convert everything into perfect Word and HTML files.

4.3 A small library: 100,000 pages

Larger organizations, universities, governments, and specialized libraries might have a whole library to digitize—say 100,000 pages. The first issue to consider is the copyright status of the publications. If they are not in the public domain, explicit permission to digitize them must be obtained from the copyright holders. You should also check whether the files are already available digitally.

20 THREE EXAMPLES: 1000 TO 100,000 PAGES

SCANNING

The volume is too high for a sheet-feed scanner. Scanning should either be outsourced ($8000 for 100,000 pages), or a more expensive scanner purchased together with a few other institutions or NGOs ($6000 shared between the participants). 100,000 pages in TIFF format will take 6 to 8 Gb of hard-disk space. The best plan is to create a set of CD-R copies containing these files.

OCR

The second step is OCR (or creation of PDF files for less widely used documents). It would take 500 to 700 months of half-time labor to convert 100,000 pages into perfect Word or HTML. This is impossible to realize with volunteers, and the job must be done on a professional basis.

To save cost, some of the less-frequently-used pages—say 80% or 80,000 pages—could be transformed into PDF, and the other 20,000 pages into Word and HTML. The PDFs would take 4 to 6 Gb space and be harder to download on the Internet, but would cost only $0.2 per page to create by a professional organization (total of $16,000). If 80,000 PDF files were created from TIFF files by volunteers using PDF conversion programs like Adobe Acrobat, 10 to 20 months of part-time work would be necessary on a powerful computer.

OUTSOURCING

An alternative is to outsource the work. If the 80% PDF and 20% HTML mix were maintained, the PDF would cost around $16,000 and the HTML $30,000 to $40,000—a total budget of around $50,000. If everything were OCRed, it would cost $150,000 to $200,000 to convert the entire collection into perfect Word and /HTML files.

5 Creating an electronic collection

Three important aspects should be kept in mind when deciding to create digital collections. First, the collection must be organized. The more content there is, the greater the need for indexes and powerful search systems. For collections of 3000 to 5000 pages or more, indexes and search systems are essential. Second, the needs of end-users must prevail. The target groups that will use the collection should be identified, and a process of regular consultation set up. Third, the available budget will determine how much can be done.

5.1 Methods of collection building

There are many examples of excellent CD-ROMs that are created on the web-page model. HTML, PDF or Word documents are added and linked using hyperlinks. Navigation is made simple and attractive by the use of hyperlinks, frames, keywords, indexes and so on. Such systems work well up to a few thousand pages, but from 3000 to 5000 pages onwards it is important to have a well-structured collection and a powerful search facility. This is where the Greenstone software can help.

The Greenstone Digital Library software creates a structured digital library including a very powerful search and retrieval engine. Up to 150,000 pages can be indexed on a single CD-ROM. Every CD-ROM can become an Internet server. Greenstone is open-source software, and is freely available under the GNU license.

The companion manuals describe how to build Greenstone collections. There are essentially three different ways of building collections:

- The librarian interface
- The Collector
- Building from the command line.

The first method is the "librarian" interface, described in the *Greenstone Digital Library User's Guide* (Chapter 3, "Making Greenstone Collections"). This is a comprehensive interactive facility for collection-building. With it, you can collect sets of documents, import or assign metadata, and build them into a Greenstone collection. The second method is the "Collector" subsystem, described in Chapter 4 of the *User's Guide*. This is an older facility that provides an alternative way of building collections of web pages or other documents. It guides you through a sequence of interactive web pages that request the information needed. However, it does not provide any way of adding metadata to the documents, and—because it is a web interface—it is not really suitable for collections that take more than a few minutes to build. The third method is to run the programs for collection-building directly from the command line; this is in the *Greenstone Digital Library Developer's Guide* (Chapter 1). This gives more flexibility in running programs individually and saving intermediate results, which may be desirable for collections that take many hours to build. You will also need to read Chapter 2 of the *Developer's Guide* in order to harness the full power of Greenstone to build advanced collections.

There is a fourth method for creating and editing the material associated with a collection, a program called the Collection Organizer. However, its functionality has been superseded by the librarian interface mentioned above. It is described in a legacy document entitled *Using the Organizer*.

5.2 Getting started in seven steps and 15 minutes

The best way of getting the look and feel of the librarian interface is to actually create a small test library. If you have 15 minutes please follow these steps and you will understand this program much better.

Before getting started, first install Greenstone (see the *Greenstone Installer's Guide*) which includes the Demo collection in DLS format and its source files. **Note, if you wish to be able to add to your collection any of the 140 documents in the DLS collection (instead of just the 11 of these documents in the Greenstone Demo collection), you should install DLS as one of the sample Greenstone libraries.** The Demo and DLS collections will be installed in *C:\Program Files\gsdl\collect*, in subdirectories *demo* and *dls* respectively. If you previously installed Greenstone without DLS and wish to install it, then you may re-insert your Greenstone CD-ROM and add this collection. It is not necessary to uninstall Greenstone first.

We suggest that you print the instructions below and follow them step by step:

1. Launch the librarian interface under Windows by selecting *Greenstone Digital Library* from the *Programs* section of the *Start* menu and choosing *Librarian Interface*. If you are using Unix, instead type

   ```
   cd ~/gsdl
   cd gli
   ./gli.sh
   ```

   where *~/gsdl* is the directory containing your Greenstone system.

2. Select *New* from the File menu in the horizontal menu bar at the top of the window. Give it a title, for example "My First Collection," and fill out your email address and a brief description of the collection. In the "Base this collection on" menu, choose "greenstone demo" or "Development Library Subset" (the effect is the same because these two collections have the same structure).

3. Add some documents from the Demo collection (or the DLS collection if it is installed) to your new collection. To do this, double-click the *Greenstone Collections* folder in the left-hand panel, then double-click the collection you desire. The documents in it are displayed underneath. Select one of these, drag it, and drop into the right-hand panel. This panel represents the collection you are building. Choose several documents and drag them into it one by one, or using multiple selection in the standard way.

4. Add some of your own documents that are not in the Demo or DLS collections. Close the *Greenstone Collections* folder in the left-hand panel and double-click the *Local Filespace* folder. Navigate to a directory that contains some documents (e.g. small Word or HTML files). Drag a few of these into the right-hand panel to include them in your collection.

5. Add metadata to the documents in your collection. So far you have been operating under the *Gather* panel, indicated by the *Gather* tab underneath the horizontal menu bar at the top of the window. Click the *Enrich* tab beside it. The documents in your collection now appear in the left-hand panel: click one and examine the metadata associated with it in the "*Element ... Value*" table at the top right. Use the panel underneath to change individual values by selecting the desired *Element* and either choosing an existing value from the list or typing a new value into the box near the bottom. Add *Title*, *Organization*, and *Keyword* metadata to each of your own documents that you put in the collection. After you type each value you need to click "*Append"* to add that value to the metadata.

6. Click the *Create* tab to leave the *Enrich* mode and create your new collection. Click the *Build Collection* button at the bottom. While the computer is building the collection you will receive some feedback on what it is doing.

7. When it has finished, click the *Preview* tab to view the collection from within the librarian interface. Check the *titles a-z*, *organisations* and *how to* lists to ensure that your documents have been included in the collection. You will also find when you visit your Greenstone home page that the collection has been installed as one of the regular collections.