Scan This Book!

By John Bruno Hare

(c) 2004, John Bruno Hare, All Rights Reserved. This text is copyrighted. It may not be reproduced or redistributed without the explicit permission of the author.

Introduction: Why there is no Bookster

This page explains some of the techniques which I've evolved which allow me to quickly and efficiently scan, proof and produce HTML versions of books. If it was as easy to do this as it is to rip an Anthrax CD, the lawyers would probably be trying to shut this movement down. As it stands, it takes a lot of talent, an obsession with detail, and a long attention span to convert a book into Internet content.

The technical term for this process is 'redaction'. REDACTION comes from roots meaning 'again' and 'fingers' (hence writing), essentially 'rewriting' or 'rewriter'. Perhaps I should call the act of transcribing paper books into electronic format 'e-daction' (but I won't!).

At this point in time producing etexts of public domain books is considered a harmless fringe hobby. However, I think most etext redactors are the successors of all those anonymous monks and archivists who saved civilization during the dark ages by devoting their lives to transcribing and preserving documents.

Why is this? In recent years corporate interests have essentially taken over the publishing industry, including publishing, distribution and bookstores. In this 'bottom-line' environment marginal books, even those of great historical or cultural importance are not being kept in print, even cheap paperback editions. Sometimes these books end up as expensive 'Print on Demand' versions which are out of the price range of most readers. Budget cuts and lack of interest have forced libraries to warehouse or even discard books and documents. Since there is little commercial or institutional demand to retain 'obsolete' books, a mass extinction is going on of much of the intellectual capital of the past few hundred years.

Libraries are not helping either. Old books are falling to pieces on the dusty shelves of libraries, unloved and unread. Truly rare items get locked up in 'special collections' which are inaccessible to anyone except for qualified scholars. Worse still, some libraries are dumping books which don't circulate; in some cases in landfills! So this is a rescue mission.

We need more workers in this field. The work is tedious, the rewards are non-existent, but if you think you have the 'right stuff' to get involved--a long attention span and a love of books--by all means, buy a flat-bed scanner and get started!

1. Selecting and Obtaining Public Domain Material

My first advice is to pick a topic which you enjoy and feel comfortable with, and then specialize in it. You will have to research the topic in great detail, and spend weeks or months proofing obscure books on that subject. You might as well enjoy it, or you will quickly get bored! For instance, if you like gardening, there are plenty of public domain books on gardening that could potentially be scanned. It will help when you are trying to decide whether obscure botanical or horticultural terms are spelled correctly. You can also sort out material that is of historical significance for your topic, so that you can prioritize which material to scan first.

It is easy to find public domain material to scan. Despite years of effort, and in spite of what you might have heard, there are hundreds of thousands of books in the public domain, possibly millions--on every topic--which have never appeared on the Internet. The libraries are full of volumes gathering dust which should be converted into ASCII or HTML as soon as possible. Often I find books which have immense historical value, which get hundreds or thousands of citations when you type their name into a search engine, WHICH HAVE NEVER BEEN POSTED ON THE INTERNET. It is easy pickings still, and will be for years to come. When I go to the library, the last time many of these books have been checked out is years ago, if ever.

Case in point: the Rig Veda: there are dozens of sites devoted to this Hindu text, which is considered by all authorities in comparative religion one of the most important scriptures in the world. Until I scanned a public domain English translation in 1999 and posted it when sacred-texts went live, no complete English translation of the Rig Veda had ever been posted on the Internet.

So where do I find books to scan? There are basically only two sources: Libraries and used book stores/sites.

Libraries

It helps if you have borrowing privileges at a large university library, or regular access to a large big-city used bookstore. If you aren't a student, but a college graduate, join the alumni program: this will get you basic borrowing access to the library. Another good alternative is to get a municipal library card with a system which allows interlibrary loans.

Why not just photocopy books from the library? Well, this is feasible, but scanning from a second generation copy creates massive character recognition problems; the best way to scan is from the original book. OCR (Optical Character Recognition) only works well if you have a 300 DPI (dots per inch) resolution source image; most photocopies are, practically speaking, 60 DPI. Anything less than 300 DPI and you get massive spelling errors. Also, it is costly and time consuming to photocopy the books at the library; the machines are seldom well-maintained and you'll be competing with students for copying time, particularly at finals. I highly recommend that you scan from an original printed copy. One exception is articles in bound serials, which often can only be checked out for a very limited time (48 hours is typical); this doesn't give you a lot of time to scan and proof. If the article is short (less than 100 pages) it's easier to put up with OCR errors. Or you can check it out and scan and OCR them using the book, return it promptly, and proof later at your leisure from a photocopy (or the page images).

All that said, if you have access to library stacks, you can just walk into the section of your choice and probably find dozens of candidates. Look for the dustiest books and worn bindings and start checking the copyright pages. Most libraries have shelves and shelves of books printed in the 19th and early 20th Century which are ripe for scanning.

If you do scan public domain books from a library, *DON'T TELL THEM WHAT YOU ARE DOING*. Librarians get nervous if you tell them you are scanning their books and posting them

on the Internet! I don't know why, but in this case, ignorance is bliss. Just pretend you are checking out that dusty 19th century chemical engineering text for a little light bedtime reading. Don't worry, though, you are doing them a favor which librarians and book lovers will be thankful for in the future. And as long as the texts are in the public domain, you are not breaking any laws.

Finding good material in the libraries can often require a sixth sense. If I know what I want, I will look it up and locate it. However, some of my best finds have come from walking through unfamiliar portions of the stacks, and opening books at random. Books are categorized by human beings; they are shelved by human beings, and library users have a tendency to migrate books and reshelve them in a remote portion of the stacks.

Used Bookstores

Even if you can't borrow a book you want from a library, you may be able to get information on various editions from the online catalogs or bibliographies in books on the shelves. Then you can try the used bookstores and used book sites

Don't bother with new bookstores for public domain material, unless a photographic reprint of an original is available. The corporate takeover of the book industry has eliminated any commercial motivation to reprint public domain texts; the last heyday of this was back in the 1960s. Even Dover Press, which once printed *extremely* high quality photographic reprints of numerous public domain books, has succumbed, and is now turning out a fraction of the good stuff it was printing in quantity back in the 1960s. Their best public domain stock is now out of print, and you'll have to find used copies of it.

In addition, corporate chain bookstore clerks are, for the most part, not savvy about public domain books. Books are just a commodity to these corporate puppets; they might as well be selling hamburgers. It's useless to even attempt to engage them about this topic. Do you want fries with your NY Times Bestseller? Independent bookstores are a better bet, particularly if they have a domain expert in your field. Then they can help you special order some of the material that you will be scanning, and can probably point you in the direction of books that you may have overlooked.

The used bookstore game is pretty simple: find a good one, with an active buying program, and check it *every single day*. Check the shelves which have books in your topic and other related topics. Use your 'book sense'; often inexperienced clerks mis-shelve books. For instance, if you're looking for books on travel in Africa, look at travel in Asia. Geography is not studied in much detail at the secondary level anymore in the US, so some under-educated clerk might think that Zimbabwe is an Asian country!

If they have an intake section, visit it once a day, and look for hardbound books, books in bad condition, and old paperbacks. Open them (even if they don't appear to be in your topic) and check the copyright date. You never know! If the book is an obvious photographic reprint, but bears a recent copyright, you may have a spurious copyright, which I'll discuss shortly.

Is the book printed prior to, or during 1922? YOU HAVE A HIT. Is it a photographic reprint of a book printed prior to, or during 1922? BINGO. Tuck that book under your arm and take it directly to the checkout.

Used book sites

The alternative to waiting for a book to show up in your local used bookstore is to go to the Internet and buy it used.

Of course, if the book is in print, or widely available used, you usually can get it at amazon.com, but if it isn't, you do have some alternatives. There are a lot of places to buy used books on the Internet, but there are only two that I patronize regularly:

ABEBooks (http://www.abebooks.com/)

Alibris (http://www.alibris.com/home.cfm)

ABEBooks has listings from hundreds of used bookstores around the world. Alibris is more of an integrated online used bookstore. Both sites are highly recommended. If you can't find a used copy at one of these sites, it probably isn't available anywhere used, or is extremely rare and only rarely appears used.

Print on Demand.

Sometimes your 'holy grail' book will only be available on 'Print on Demand' format (POD); this is a last resort. POD books usually have similar *or worse* quality than a photocopy. Kessinger Press, in particular, will be happy to ship you a print on demand copy of a rare document which will set you back 50-100 dollars, but which is basically *a crappy photocopy*. They are costly.

Many POD publishers list their books at ABEBooks or at Amazon.

2. Copyrights

Has the book already been put on the Internet? Check the Online Books Page (http://digital.library.upenn.edu/books/), by all means. It's likely that a well known text has already been put online. In this case, you may wish to examine the available copies. Are they complete? Do they include material such as footnotes, introductions, indices, illustrations? If not, it's probably a good idea (but not high priority) to do a more extensive version. However, many well-known books in the public domain remain to be scanned. If it's not listed at the Online Books Page, go to google.com and type in variations of the book name, author, etc. You may find many pages about the particular book, or bibliographies, Amazon links, etc., but no actual copy of the book. This strongly indicates that you should make this book your next project. Online books also lists various archives, some of which may not be fully indexed, so it's also worth checking specialized archives in your field to see if they have the book in progress or done already.

You need to ascertain whether it is possible to legally post the book on the Internet. Unless you don't plan to share the book with others (which is kind of useless), this is the key task which you will need to accomplish before you fire up the scanner.

The following advice applies if you are going to republish the book or article in the United States; and where I don't specify it, it is assumed that the book was published in the United States.

If you consult most of the material on the Internet about the topic of public domain books, you'll read some very scary stuff which may discourage you from embarking on your project. I think that a lot of this advice has been provided by lawyers who are just trying to be responsible and end up erring on the conservative side. Like many others who produce free ebooks, I feel that the Sonny Bono 'Digital Millenium' copyright act is nonsense, and should be revised, particularly the part which imposes draconian measures for the those who attempt to reverse engineer copy-protection measures. The reality is nobody in the free ebooks movement has been subjected to a Napster-like attack at this point in time, nor is one likely in the future.

If the book was published prior to, or during 1922, or it is a photographic reprint of a book published prior to, or during 1922, **it is in the public domain in the US**. In the latter case, if there is added material (normally a preface or appendix), this will usually bear an explicit copyright message; you can simply omit the copyrighted material from your scan.

How can you tell if it's a photographic reprint of a public domain book? There are many hallmarks of this kind of reprint. There will be archaic illustrations, obtrusive and superfluous graphic elements used to separate chapters, large initial capital letters in each chapter, and extensive footnotes. Illustrations which were in color in the original will often be reprinted as black and white halftones in the cheaper reprints. Look at the font. Irregular spacing between letters indicate handset type, which became obsolete long before 1922. The font will *look* archaic, with subtle drop-outs in the serifs, and 'blobby' letters. A non-serif font, on the other hand, is more likely 20th Century. A photographic reprint often will have marginalia (hand-written markings from the original), irregular light or bold areas due to warped or faded pages in the original, and visible artifacts such as specks or creases. Look for oddly spelled or hyphenated words; English (both American and British) spelling and hyphenation was not completely standardized until the early 20th Century. For instance, 'today' spelled as 'to-day'.

In some cases there may be a copyright message on a photographic reprint of a public domain book. This is, in a word, bogus. Ignore these messages without any guilt. Academic and reputable reprinters such as AMS and Dover will give the complete printing history of the original on the copyright page and not claim a copyright on actual public domain material. If one of these reprinters has published the book, and not claimed a copyright, you can safely assume that it's in the public domain and fair game. Conversely, if you see a Dover publication with a copyright, you can be certain that it should be left alone. If only everyone else was as straightforward. Note that Dover's science and math books, in particular, are mostly covered by copyright.

Reprints with bogus copyrights will often have additional text on the copyright page like "It is forbidden to reproduce this work by mimeograph, Morse code, Whale Song, Semaphore flags, store it on punch cards or any data retrieval device which currently exists, existed at some point, or in a parallel universe, or which may be devised in the future. Violators will be pummeled by flying monkeys" *Riiiiiight*. Ignore this silly crap; it is not legally binding, nor is there any criminal or civil basis to pursue anyone who might dare to violate such a warning. It is just there to scare you. And anyway, most of the public domain reprinters from the 1960s who employed this kind of boilerplate are out of business anyway, or have been swallowed by some conglomerate, which probably doesn't care either.

Although the bogus copyright books often go to some lengths to conceal the publication dates of

the original source material, sometimes the date can be found at the end of the preface. If not, you can usually find the actual publication date by looking it up in any decent online library catalog, such as MELVYL, or the Library of Congress online catalog. This may not be feasible if you are in the bowels of a used bookstore (unless you have a laptop with wifi!) so when in doubt, buy the book and check its status when you get home. It may not be there when you get back.

If the book was published after 1922, but before or during 1964, there are some loopholes which may put it in the public domain. These loopholes fall into some easy to check categories.

If the book does not have an explicit copyright, and was published between 1922 and 1964 (inclusive), it is in the public domain. Also, if the book was banned for some reason, the publisher may not have wanted to identify themselves. There is a lot of erotica (or books which were once considered sexually oriented, such as birth control manuals, or cutting edge fiction) which was printed without such a copyright page. However, this must be a first edition; sometimes (often, actually) pirate editions were printed without copyright; this is a counterindication.

GATT Restorals

There is one thing to note, however. If you find a reprint by Dover or some other publisher of a book originally published outside the US after 1922, and republished as a public domain work prior to the late 90s, the 'GATT Restoral' may apply to it. In this case, certain works published outside the US which *had been* in the public domain had a full US copyright term restored by international treaty and US legislation. There is one loophole to this loophole. If it had been published prior to the restoral there is a grandfather clause (the 'reliance parties' provision) which allows *that edition* of the book to stay in print and continue to be sold. However, this does not apply to *you*. For instance, Dover has long republished the works of the Egyptologist, E.A. Budge, many of which were published in Britain after 1922. I don't know whether this is the case, but apparently Dover can continue to publish these books (which had a GATT restoral) because they had published them prior to the passage of GATT.

GATT has lead to some strange glitches, which are typically not in to the advantage of the public domain. For instance, there are many books which are public domain in their country of original publication, but are technically under copyright in the US. For instance, Pakistan has life+50 rules for copyright. Assume that a book published in Pakistan by an author in 1945; if the author died in 1950, the book is now in the public domain in Pakistan. However it will be copyrighted IN THE UNITED STATES until 2035! (date of publication+90, the current US term of copyright).

There is a loophole to the loophole to the loophole. If the author was not a citizen *or* resident of the country in which the book was published, GATT does not apply. This can take a bit of sleuthing, but if you can determine that the author was a US citizen and not living in the UK at the time their book was originally published in London, GATT does not apply, and if they didn't renew their book's copyright, it is in the public domain.

So this is why I generally avoid anything printed after 1922 outside the US. This includes books published simultaneously (technically, within one month) in the US and abroad. So if you find a work that appears to be in the public domain in the US because of a defective copyright, it's

important to verify (using library catalog searches) that there were no foreign editions. It is usually impossible to get the exact date of publication for a book without some additional footwork, so if a book published in the US lacks a renewal and there was another edition printed in, say, London the same year, it's generally a good idea to avoid it.

If the book was not registered or renewed in a timely fashion at the US copyright office, and there is no sign of a publication abroad in the same year, it is probably in the public domain. There are two ways to do this.

Searching Physical Records For Renewals

The first option is to check the physical records, which is sort of difficult if you are not located in Washington, DC, or near certain libraries which have the bound volumes of the copyright renewals.

Your best option is to use a copyright search firm such as Thompson and Thompson, (1-800-356-8630) they can get you the information in a few days for about \$100. Ask for a 'renewal search'; a full report from them can be a lot more costly. If T&T signs off (they can't find a registration and/or renewal), you are good to go. The alternative is to use the copyright office, which bills at \$60 an hour but which can take months.

Before you call T&T, you should have already checked off the following facts:

- 1) The book was first published in the US prior to 1964 and after 1922 (inclusive).
- 2) There were no simultaneous or prior foreign editions.

T&T will simply need the name of the book and the author to conduct the search.

Searching For Renewals Online

The second option is to check the online records, starting with the Catalog of Copyright Renewals. This can be found at the following URL:

http://www.kingkong.demon.co.uk/ccer/ccer.htm

Using the CCR is a bit difficult but it is the only way to check records prior to the mid-1940s. The CCR has links to page images of the US Copyright renewals for all years from 1950 to 1978. This corresponds to books published from 1923 to 1950.

Go to the URL above, and click on the year the book was published. For each year, from 1923 on, you need to check two years worth of records (the book had to be renewed 23 years after publication, with six months leeway on either side, e.g. books published in 1923 could be renewed in 1951 or 1952, and occasionally 1953). Certain early years have been scanned and OCRed, which makes consulting them easy. For later years, there are two sets of page images. This is because the records were published twice a year. Therefore, minimally, you will need to download four page images. The page images are in TIF and sometimes PNG format; TIF can't normally be viewed in a browser so you will need to download them to your hard drive and open

them up in an image processing program.

There is one other problem with the CCR. Up to 1973 (renewals for 1945-1946), the renewals were sorted by author; after that they are sorted by an index number. This means that the records from 1946 on are useless unless you download the entire set for that year. So the CCR is ideal for checking whether books originally published between 1923 and 1945 was renewed in a timely fashion between. From 1945 the amount of work required goes up exponentially.

If the book was published after 1950 but before 1964 (inclusive) they are much easier to check. If there was a renewal (or registration) it will be listed in a searchable database at the US copyright office website. This can be found at the following URL:

```
http://www.copyright.gov/records/
```

If it falls into the 1950-1964 range and was published in the US, it can be easy to determine quickly whether it was registered or renewed using this database.

Other Exemptions And Caveats

If the book or article was published by an agency of the US Government, it is in the public domain, no matter which year it was published; the taxpayers own the rights; and that's us! There are a few exceptions to this (notably some Smithsonian publications), but these will bear explicit copyright messages, so it will be obvious which ones not to mess with.

If the copyright is after 1964, **FORGET IT**, unless you can track down the copyright holder and talk them into it.

If the book or article is based on an unpublished manuscript which was not published until 1964, it is not in the public domain. For instance, some of the rites of the Golden Dawn ritual magic movement were written down during the 19th Century, but not published until recently. The book in which they were published is *not* public domain.

After 1964, the copyright laws gradually eased requirements to file a notice with the Copyright Office, or even display a notice in the document in a specific format (although it is conceptually easier to prosecute copyright violators if you have done so). Today, anything which has been given fixed expression as text is automatically copyrighted, for instance *this* document.

If you are tempted to post a text or book which falls outside of these categories, you may wish to consider that a copyright infringement legal action (which is very unpleasant) will only get triggered if you step on someone's financial toes. For instance, bootlegging a high profile book like the Lord of the Rings or Harry Potter would be asking for trouble. Scanning and publishing a book which is out of print, with an out of business publisher, of very limited interest, whose author who died long ago, presents a minimal legal exposure, particularly if you aren't doing it for financial gain. However, there is still an outside possibility that it might lead to legal action. *I certainly don't and can't recommend it.*

As a last resort, if your website is a 'nonprofit archive', there is an exemption if the work can't be obtained easily and it is nearly out of copyright. This normally means that there may be a few copies on the used market at high prices, and you aren't charging anything to view the etext.

However, this is rarely applicable.

3. Scanning and OCRing the Book

The basic requirements to produce a new etext are a flat bed scanner, a commercial quality OCR (Optical Character Recognition) program, and a decent word processing program (such as Microsoft Word). You can omit the word processor program, but you need a spell checker to proofread the text.

While it is possible to type in a document by hand, it is not recommended unless it is so poorly printed it can't be easily recognized by OCR, or it is in a non-Latin alphabet which must be transliterated. And this should *not* be your first etext project.

An OCR program is like a non-English speaking typist with the IQ of a goldfish; the good news is that they can type 5,000 words per minute.

It is very labor intensive to retype a book. For instance, Moby Dick is about one megabyte of text. Assuming an average of 5 characters per word, at 60 words per minute, that is 55 hours of continuous typing. Who can type at 60 words per minute without typos? In practice, you're probably looking at 80 hours of typing, spread over a month. On the other hand, I'd estimate Moby Dick would be about 3 to 5 hours of scanning, and about 20-40 hours of proofreading, depending on how good the OCR is. Proofreading is infinitely less taxing than typing; the spell checker will do most of your work, anyway.

Before scanners, all etexts were created by typing. The Gutenburg Project produced only one actual text during the 1980s: the King James Bible, and even still, this had some errors which are still being corrected. Once flat bed scanners got so cheap that they were basically giving them away with new computers, and OCR software such as OmniPage became available on a retail basis, the pace picked up. If you bought a retail system in the recent past, you may already have a flatbed scanner gathering dust, or still in the original packing box.

Most people use scanners to get hard copy photos into their computer; we're going to use the scanner to launch public domain text onto the web. This still involves a lot of trial and error; there is a learning curve to overcome, some of which involves simply putting aside your assumptions as to what is possible and having reasonable expectations.

Start small. Find a public domain book which has less than 200 pages, printed between 1900 and 1922, with no illustrations, and few footnotes, italics, accents, or strange formatting. I made the mistake of starting with Max Müllers' Upanishads...not recommended as a first project.

There are three phases.

First you need to scan the pages into your system.

Secondly, you need to proofread the text.

Thirdly, you need to format the text.

The second and third phases usually are performed at the same time, although there are formatting tasks which can take place after proofing is finished, such as breaking it down into

separate files, and proofreading tasks which occur before formatting, such as global search and replace.

Here are some issues which you will need to deal with if you are OCRing a book.

A hardbound book is bound in 'signatures'; these are bundles of pages which have been printed on a single sheet, folded cut and then bound together. In some cases (normally books from the 19th and early 20th century) you will find adjacent pages in books that have not yet been cut apart; in this case I recommend that you carefully take an very sharp knife and separate the pages. Place the knife parallel to the uncut pages, cutting away from yourself, and slowly draw the blade through the pages. Do not 'saw': if you find that you have to move the blade back and forth, use a sharper blade. In a pinch you can separate the pages by hand, but this is not recommended, as you may tear the page.

You will be scanning two pages at a time, unless the pages are too big to fit on your flatbed scanner. A book which has been sitting upright on a shelf for decades tends to sag in the center. Also, facing pages which are poorly bound (which is the vast majority of books) will be misaligned. Facing pages on different signatures will also be misaligned.

This presents some problems, particularly with used books and library books. One page of the two will be misaligned by 2 to 15 degrees. This throws off the OCR on that particular page; normally the first couple of lines on the page will be jumbled, or superfluous line breaks inserted in the first paragraph. While this isn't a disaster, it will add proofreading time. When multiplied by many hundred pages, this can be a real drain. However, in most cases this is unavoidable.

It is possible to ensure that both page images will be aligned correctly. There is a technique to this which takes a bit of practice. Most scanners have a slight indentation surrounding the glass surface. Align the book against the edge of the indentation; then press *gently but firmly* on the spine so that both pages are aligned properly. This requires a slight lateral pressure so that both pages are flush with the indentation. This is of course fairly stressful on the binding, so it may cause permanent damage to the spine. Worse case, the spine can fracture on a poorly glued signature. This is not necessarily a problem if it's your book but if it's the libraries', you will have some explaining to do. If properly done, however, this can greatly improve the accuracy of the OCR.

The 'sag' can differ depending on how far away you are from a signature boundary in the book. In the middle of a signature, it can be easy sailing. Close to the signature, there was less pressure on the pages; so the most misaligned pages will be close to the edges of the signature. This will be apparent while you monitor the page images as they are scanned. If this is the case, you can adjust where you place the pressure on the book spine. If you press near the top or bottom of the spine, this may help align the off-kilter pages better.

In a pinch, don't hesitate simply to scan one page at a time for a few pages until they straighten out.

Most OCR programs have ways to adjust the tilt of a page image manually ('*de-skew*'). In the case of OmniPage, there is no way to specify an exact angle. You just press a button and it attempts to find the optimal tilt. This is potluck; in most cases it produces better results. Occasionally it may produce better results on one page and worse results on the other. You can't

undo OmniPage's automatic de-skew. In most cases it's not a good idea to use OmniPage's automatic de-skew feature, unless both pages are parallel and you scanned it at an angle accidentally. The worst case to use auto-de-skew is when both pages have significantly different alignments.

If you find that your OCR yield isn't up to par, it may mean that you need to adjust the page scanning contrast. In order to reliably recognize letters, it helps if the page image is as dark as possible. However, the more you darken it, there will appear artifacts (particularly close to the inside edge of each page). These artifacts may obscure text. So the trick is to adjust the contrast to the point where all the text is dark and legible but not covered up by artifacts.

One thing that can throw off the contrast adjustment is illustrations, particularly if they have a lot of dark areas. This will lead to the text scanning too light. To solve this I put a white post-it over the illustration, or if it covers a whole page, just slip a sheet into the book at that point.

Commercial scanning houses take the book, cut off the binding with an industrial cutting tool and feed the sheets into a high speed scanner with a page feeder which scans the book in a few minutes. You probably don't have this luxury; certainly not if you are scanning a library book or one that you want to keep intact after the scanning process.

Scanner software has its limitations. I scan about 50 images (100 pages) at a time. After about 50 pages, OmniPage tends to crash unpredictably. Fortunately, when OmniPage is relaunched, it is able to restore the page images. Consequently, I scan about 100 book pages at a time; I attempt to end each file on a 100 page boundary.

Naming the files you create systematically is important. I select a three or four letter 'signature' for each text. Then each successive file I generate I append a zero-prefixed two digit number. For instance, if the book is named 'Ham and Eggs' I might give it the signature 'heg'. The first file I scan will be heg00.htm, the second heg01.htm, and so on. The zero prefix is important, because the collating order when you list the files will be sequential. If you name the files heg1.htm, and so on, then heg10.htm will follow this file in the directory listing; heg2.htm will follow heg19.htm. Of course, if you anticipate more than 99 files, use three digit zero prefixed numbers. You can also name them by appending letters, e.g. hega.htm, hegb.htm, etc.

After scanning, you must 'zone' the pages; this helps the OCR software to determine which areas of the page to recognize (convert to text), and in which order to recognize them. This is because the OCR program can't perform this task; a human being must accomplish it. Zoning consists of drawing rectangular or irregular areas on the page in the correct order. For instance, if there are two columns on a page, you will need to construct a zone for each column, with the left column sequenced before the right column.

It is important to capture the page numbers (either at the top or bottom of the page). This will insert the correct page numbers into the output file. This is essential for proofing. In most cases you will be throwing away most of the text in the page header other than the page number.

Tips and Tricks

Save the OCR output as HTML or plain text, not Word, or RTF. Then edit it as an HTML or text file in your word processor. Because Word is a proprietary format, third party programs don't

always generate valid files in it.

Do not let the OCR program apply any formatting to the document. Today's OCR programs are incapable of recognizing styles correctly (despite the marketing literature). If you can't figure out how to get the OCR program to stop applying formatting, the first thing you should do to the document is to 'select all' and then apply 'normal' to the entire document. Then if you are going to apply styles like italics and type sizes, do so manually.

You *are* going to be using your spell checker on the OCR results. Become very familiar with how your word processing program's spell checker works. Detailed instructions on MS Word's spell checker will be found elsewhere in this document.

Again, when you scan the file, *do not* discard the page numbers, even if you will be leaving them out of the final product. You need the page numbers to orient yourself while proofing. So scan and OCR the entire page including the headers and footers. It's always easier to delete material than put it in later.

OCR programs normally are able to recognize blocks of text (i.e., paragraphs) correctly. In some cases, especially when the page images are marginal, it may not be possible for the OCR program to do this correctly, particularly with poetry. For instance, the OCR program may collapse blocks of poetry into a paragraph, or leave the line separators in a paragraph. Specifically OmniPage has two 'Save' options: one which preserves line breaks and one which doesn't. I usually save two copies of the OCR output, one with line breaks and one without, the reason for which will become apparent below.

When you construct your 'proof' file (the one you will be actively editing), if the book is mostly poetry, use the OCR without line breaks. Otherwise use the normally formatted one.

4. Scanning Illustrations

Illustrations need to be scanned separately. This is because the optimal resolution for OCR is monochrome, while you will need either gray scale or color for the illustrations. Preparing illustrations for posting on the web is a fine art which I will discuss in a following chapter; this section just discusses what you will need to do during the 'capture' phase of the project.

The key to this part of the project is getting as high-resolution images as you can. I typically create a directory called 'hrimg' under my project, and systematically scan all of the illustrations right into separate files in that directory after I'm done with OCR phase. This way, the structure of the book is fresh in my mind. I save the files in JPG format to save some disk space; however, if you are doing an archival version of the book and disk space and bandwidth is not an issue, you might want to use a non-compressed format such as BMP.

I name the illustration files systematically. Typically I name them using a five digit file name. The first three digits are the zero-padded page number. The second two digits are a zero-padded sequence number which starts over on each page at zero; that is, the first illustration gets sequence number 00, the second 01, and so on. The sequence number goes from top to bottom, left to right. Thus the first image on page 14 is 01400.jpg. If the illustration is a plate, I number it according to the facing page. If the page numbers are Roman numerals, I normally append an underscore and translate the numbers into Arabic: for instance the second illustration on page

xxvii will be named _02701.jpg.

Scan black and white images as gray scale, and color images as 24 bit color. Set your scanner for its highest possible resolution; 300 dpi is ideal.

If a illustration is line art without a lot of detail, sometimes you can get away by saving the OCR program's monochrome page images and then selecting, copying and pasting the image into a fresh file for each one. However, this does not work well for woodcuts or photographs, and color images will be an unrecognizable hash.

When you view the high resolution images at various magnifications, keep in mind that certain magnifications (particularly those with odd denominators such as 1/3 size) will produce on-screen 'moiré' effects; there are artifacts and will not appear in the final version.

I don't do any other modifications to the images at this point other than crop them. If, however, you are scanning images printed on non-opaque paper or low-quality paper with a visible grain, particularly gray scale images, you may wish to adjust the contrast at this point. This is indicated if you get any 'bleed-through' from the back of the page--you can see faint backwards type--or if your images have a porridge-like speckling in the background. While paper grain might be an nice effect in some cases, mostly it is a distraction. You can erase this nicely without losing any detail by applying a contrast of about 20% to 40% in your image processing program. It can also help if you apply brightening of about 20% in some cases. If you do this while doing the scanning, it will save time later.

5. Proofing

Now the hard part starts. It takes a few hours of labor to scan a book and OCR it. It will take at least *ten times* that amount of time to proofread it.

One of the first questions that comes up is how closely your etext must resemble the original document. Current theory is that you must not change any of the spelling or punctuation in the original book in your etext. In practice, we must make some compromises.

You might think that correcting typographical errors, repeated words, punctuation errors or extrapolating words with broken type will mean that you have produced a 'new version' of the document. However, there is no need to become overly obsessive. In theory, there is actually some leeway allowable. Courts have ruled that simply editing a public domain book in a way that *does not involve any creative effort* does *not* produce a new copyright. Specifically, an edition of a public domain text which had all of its spelling Americanized and punctuation modernized, which another publisher subsequently pirated, was not actionable. However, for documentation purposes, most etexts attempt to preserve as much of the original text and formatting as possible.

Personally, I don't think there is any harm in correcting a few typographical errors per document; however, egregious changes like converting all of the English or archaic spellings in a document to American spellings is the sign of an etext novice. And adding or worse yet, altering material in the text, such as extensive comments or footnotes to the etext is definitely frowned upon, even if you note that they are inserted by the redactor.

My practice is to scrupulously document all spelling errata that I encounter using the 'sic' STML tag. This is in the format: {sic "bad" "good"} where "bad" is the misspelling and "good" is the corrected spelling. I don't alter British, archaic, or idiosyncratic spelling in the document (i.e. if a word is spelled strangely twice, I don't 'sic' it.) My STML parser converts this into a link from the corrected spelling to an 'errata' page. This allows me to correct spelling but document my changes. In addition, if you wanted a document with the incorrect spellings left intact, it could be generated from the STML source as well.

One issue you'll have to decide how to handle is transcribing character sets with accents, Greek, Egyptian, or so on. The easiest way to deal with this is to just transcribe texts that are in English. However, that may not be an option, particularly if you are dealing with scholarly literature (or non-English literature, naturally). This is such a big topic I will deal with it below in a separate chapter.

When spell-checking I edit the files as HTML. This is because you can use HTML styles in Word to mark up the document, which is pretty essential if you are going to post the result on the web. In fact, I don't use native Word format much any more. I save the files from my OCR program as HTML. You can convert a Word document to HTML through the File menu: select "File|Save As HTML". You can open any HTML document from the desktop in Word by right clicking on it and selecting 'Open with Word'.

Punctuation

Punctuation is a touchy subject, particularly if you are preparing etexts for Project Gutenberg. In this case it will be run through a program which will critique the punctuation, particularly missing quotes. It doesn't help that many older books vary widely on how they treat extended quotes. They may quote only the first line in a paragraph in a long quote. Or in some cases there may be a quote at the start of each line. These are insidious to remove because the OCRed version will wrap differently than the printed one and these quotes will be peppered throughout the paragraph (hence the term 'salt and pepper quotations'. (I do modernize this style of quotes). Another problem is how embedded quotes are treated. In some cases, double quotes surround single quotes; the problem being when there is a third-level of nesting.

My practice is to leave the punctuation as intact as possible, within the limitations of the HTML standard character set. I represent 'em' dashes by two ASCII dash characters, because not all browsers seem to do well with the HTML em dash. I eliminate superfluous space in quotations. For instance:

" Here is a quote, ' she said .' "

I change into:

"Here is a quote, 'she said.""

Lastly, I fix some obvious punctuation errors, for instance, when there is an obvious missing close quote, and so on. Very occasionally I will add a comma or other punctuation mark if it is missing and there is obvious white space in the printed version--punctuation marks sometimes didn't show up if the book was typeset by hand because they tended to wear out faster than the rest of the type set. I don't make any attempt to document punctuation changes except in extreme

cases.

Fixing Common OCR problems

A common OCR problem is that it will often insert a space before certain punctuation marks, particularly semicolon, colon, comma, question mark and exclamation mark. These are easy to fix using a global search and replace. I do this before anything else when I start editing a new file. Note that you don't want to do this with the period, as space before period is a valid sequence occasionally; you will have to fix the periods manually as you go.

One common glitch in OCR (particularly OmniPage) is that zero is confused with capital O (and sometimes lower case O as well when adjacent to another digit); and one, lower case L, and occasionally single quotes can all be confused. This is sometimes difficult to spot on the fly. However, in the case of zero and O is it fairly easy to handle. When I edit the document for the first time, I replace every instance of zero surrounded on both sides by a space by capital O. Then I go through each combination of single digit and lower case 'o' and do a global search and replace. For instance 60 I globally replace with 60 (for some reason this seems to be the worse offender), and likewise for 10, 20 and so on.

Fixing Quotation Marks

The worst problem with OCRed punctuation is quotation marks, both single and double. These can disappear completely (typically at the beginning of paragraphs or blocks of poetry), turn up as the digit 1 or 11, or some strange thing like 'cc' or two adjacent commas.

If you see the 'eleven' for double quotes, it is typically for an open double quote. You can then change < 11 > for < "> pretty reliably and possibly save yourself hundreds of manual replacements.

I use normal quotation marks normally. If you want to insert curved quotation marks, Word uses ctrl-backquote backquote to insert an open curved single quote, and ctrl-singlequote singlequote for closed curved single quote. You can insert a curved open double quote using ctrl-backquote doublequote and a closed curved doublequote using ctrl-singlequote doublequote. These are a bit hard to do at first but once you get used to them they are pretty natural. It's just more typing, and can substantially increase the time to proof a text. You may wish to only use the curved quotes for words like 'til and O'Grady. The curved quotes are also handy for transcriptions of non-English languages such as Arabic or Native American, where the curved quote indicates a 'glottal stop' and the straight quote accent.

Wrangling Word's Spell Check.

One of the distinctive features of the texts which I prepare for sacred-texts is that I actually spellcheck them. You would think that this would be the case, but not all electronic text redactors seem to do this.

Proofing a document without using the spellchecker is ignoring one of the key benefits of using word processing software. Without using the spellchecker, it is easy to ignore dozens of typos a page.

So why is this the case? Some may not be using a word processor, but a simple text editor. Other

redactors may simply be intimidated by the interface of their spellchecker. Word's spellchecker has a substantial learning curve and some huge fatal (but avoidable) bugs which can really hit you hard--you can lose a lot of work if you haven't been careful with backups. It is also poorly documented.

I use Microsoft Word 97 to proofread my texts, and use a custom dictionary and leave on the spell checkers' highlighting. The problem with Word (as I will call it hereafter) is that its spellchecker eventually crashes on long documents; it appears to be suitable for checking short memos, but it will definitely crash on a typical 500Kb raw scan. The symptom is that it will pop up a message box stating that 'There are too many spelling errors in this document'. Or it may do a hard crash and lose all your changes since the last save!

On the face of it, Word's spellchecker is easy to get started with. Turn on interactive spellcheck by selecting 'Tools | Options' from the main menu and clicking on the 'Grammar & Spelling' tab. Turn on 'Check Spelling as you type.' (You can leave off grammar checking if you are redacting, as the goal is to reproduce the original text as exactly as possible, not nit-pick their clause structure). Close the dialog box.

Then to 'add' a word, you will right click on it and select 'add' from the popup menu. Word will also suggest possible corrections, which can occasionally be useful to fix OCR errors (or at least a bit of comic relief). You probably will not want to select 'ignore': Word will *no*t remember these words the next time you open the document (if you are editing an HTML file).

So where do these words you 'add' go? Into a custom spellcheck dictionary. There are a standard set of words which Word keeps in an inaccessible file which you can't add to or subtract from. Then there is a custom spellcheck dictionary (by default CUSTOM.DIC) which you can edit. You can also set up other custom dictionaries.

It *is* possible to use the spellchecker on a long document without crashing it by babysitting it a bit. Word will not complain if you correct the words one screen at a time. Do not let too many errors (redlined words) show up at one time. If you have an extensive block of text in, say, French, you will want to scroll down until only one or two lines show up at the bottom of the screen. Proof one line at a time, then scroll down a line. If you are doing a walkthrough search and replace on a file with a lot of spell-check red marks, you will want to turn off the interactive spellchecker for the duration--this can also trigger a soft spellchecker crash.

You will need to become familiar with Words' spell checker custom dictionary interface. *Don't* just click on the little spellcheck icon on the toolbar while viewing a raw OCR file, unless you want to get the 'Too many spelling errors' box. The ultimate goal should be to get 'No spelling errors' when you click on the spellcheck icon. You can do this using a custom spellcheck dictionary.

To create a custom spell-check dictionary, select Tools | Options from the main menu, then click on the 'Spelling & Grammar' tab. There is a button labelled 'Dictionaries' on this tab. Click on it.

A dialog box is presented labelled 'Custom Dictionaries'. This has a listbox of dictionaries with a checkbox next to each dictionary. When you first install Word there will be only one blank custom dictionary, named 'CUSTOM.DIC'. This is where words go if you 'Add' them when right clicking on a redlined word in the document window.

However, you want to start a custom dictionary for a specific text. This will hold all of the idiosyncratic words which you find in the text, after you verify their spelling from the printed copy of the book.

To start a new dictionary, click on New and fill in a name (normally my document prefix). This new dictionary will appear in the list box with a check next to it. Close the Custom Dictionary dialog box. On the Spelling & Grammar tab, look for the pull-down list next to 'Custom Dictionary:'. This will have 'CUSTOM.DIC' still listed. Select your custom dictionary from this list instead.

You can keep multiple custom dictionaries active (used to suppress red-lined words), but only the one selected in the pull-down list will get added to; in practice I only keep the text-specific custom dictionary and CUSTOM.DIC active. You *can't* deactivate 'CUSTOM.DIC' for some reason.

It is easy to edit and backup the custom dictionaries. These are just text files with one entry per line and you can open them in Word, Notepad or any text editor. The files have a '.dic' extension and reside, by default, in "\Program Files\Common Files\Microsoft Shared\PROOF". I keep a shortcut to this folder on my desktop for ready access.

A common mistake, for instance, is to add a word that you didn't want to, put text-specific words in 'CUSTOM.DIC', and so on. You can simply open up the .dic file in question and edit as much as you want. Word will put spell check on hold if you open a spellcheck file which is active.

To edit an active or inactive spellcheck file inside Word you can leave your document open; select Tools | Options from the main menu and click on the Spelling and Grammar tab as before. Click on the 'Dictionaries' button. With the 'Custom Dictionaries' dialog displayed, select your custom dictionary from the listbox and click on the 'edit' button. Word will warn you with a nag dialog that it is turning off spellcheck while you are editing the custom dictionary file.

Word automatically saves your currently active custom dictionary when you exit Word. You can also save it manually at any time by using the above procedure to edit it, and then just File | Save.

Word will more that often crash in the middle of an edit session. When this happens your carefully compiled spell check dictionary may be completely wiped out! For this reason, back up the spell check dictionary often. I back it up as a txt file. Edit the custom dictionary file using the steps in the previous paragraph. Then 'Save as' a txt file. Close all the dialogs. Don't forget to re-initialize spellchecking (with the appropriate custom dictionary) when you get back to the Tools | Options | Spelling and Dictionaries tab. Then if Word crashes, you can just copy the txt backup file over the empty dic file (from the desktop) and everything will be fine.

Sometimes there may be words you want to insert into the custom dictionary en masse, such as all of the Roman numerals between ten and a hundred (Word doesn't have these in the standard spellcheck dictionary). You can just open the dic file and insert these manually, one per line.

One other quirk of Word's spellchecker: you can turn off all of the grammar checks except for repeated words. This can be the result of an actual repeated word in the book, normally printing the same word at the end of one line and the start of the next. More often, it is exactly what the

author intended, particularly if they leave out some punctuation, for instance:

The knight pressed on on the dismal moorlands.

which should be punctuated:

The knight pressed on, on the dismal moorlands.

However, we are preserving the original book, warts and all, so you want to leave the on on in place.

To turn off Words' repeated word nag, simply substitute a non-breaking space for the space between the words.

Reduplication of words is also common in poetry and some languages. You can insert a nonbreaking space to tally the fact that this is correct.

Tips and tricks

Capitalizing entire phrases

Often you will need to change the capitalization of the OCR output, for instance it will occasionally put a lowercase letter in the middle of a long all-caps sequence. The OCR can also stick in an all capitalized word for no reason at all (particularly if the shape of the lower case letter is the same as the upper case, e.g. OF and of). Select the block of capitalized letters and press Shift-F3. This will rotate between all-caps, lowercase and initial caps. Note that if you want to have a standard all initial caps sentence (as in a chapter heading) you will need to manually down-case the shorter words such as 'of', 'the', 'and' etc.

Paragraphs and Linebreaks

OmniPage outputs linefeeds as paragraph separators. If your book is mostly or entirely prose, you will have to globally replace all of the linefeeds with normal paragraph separators as an initial step in processing your text.

Here's how to do this in MS Word. Either select all (or with no selection active) bring up Words' Find and Replace dialog (shortcut key Ctrl-H). Then enter ^l (that's uparrow and lowercase L) in the 'Find what' field and ^p in the 'Replace with' field. Press the 'Replace All' button (or Alt-A).

OCR programs will sometimes format a paragraph into individual lines. Or you may have a block of poetry with linebreaks instead of paragraphs at the end of each line.

To fix up the linebreaks in MS Word there are two procedures.

To merge a number of lines into one paragraph, select all of the text from the first line down to just before the final line break. Then do a Find and Replace with ^p (a paragraph mark) in the 'Find What' field and a single space in the 'Replace With' field. After you press 'Replace All' (and the merged paragraph is created), Word will prompt you if you want to perform this replacement on the rest of the file. Press ESC to cancel the global replacement, (or your document will be turned into one big paragraph!).

Fixing up a block of poetry which has been collapsed by the OCR program into a paragraph is easily fixed if you saved a version of the file without line breaks. Simply open up the file without line breaks and copy and paste the properly formatted poetry into the proof file.

If the line breaks in a block of poetry were recognized correctly, but have normal paragraph markers, you can also simply replace all of the paragraph makers in the block with line feeds instead. Select all of the text from the first line down to just before the end of the final line of poetry. Do a Find and Replace with ^p (a paragraph mark) in the 'Find What' field and ^l (up-arrow and lowercase L) in the 'Replace With' field. After you press 'Replace All' (and the merged paragraph is created), Word will prompt you if you want to perform this replacement on the rest of the file. Press ESC to cancel the global replacement (or all paragraphs in your document will be turned into linefeeds!).

Font Leakage

There is a problem with Word which I call 'Font leakage'. This mostly comes up with italics, but it can also be a problem with font sizes. The problem is as follows: when Word outputs HTML, it closes out any font markups after the closing </P>, on the next line next to the next opening <P>. This isn't a problem in Word, but it gets displayed incorrectly in some browsers.

For instance:

<P>Refer to <I>The King in Yellow.</P></I></P>.....

As you can see, the nested $\langle P \rangle$ and $\langle P \rangle$ overlap with the $\langle I \rangle$ and $\langle I \rangle$. The result is in some browsers that the italics start in this paragraph at the $\langle I \rangle$ and go all the way to the end of file! The document looks fine in Word, but when you load it into a browser half the file is italicized. It's a real difficult problem to fix if you don't know what's going on.

The solution is what I call *font firewalling*. You basically need to fool Word into terminating the italics before the end of paragraph, by ensuring that the last character in the paragraph is *not* italicized. So if you have a punctuation mark at the end of a paragraph, *don't* italicize it. This will produce:

<P>Refer to <I>The King in Yellow</I>.</P>

instead, and the font leakage stops before the period (the firewall).

If there is no punctuation mark at the end of paragraph, insert a non-breaking space (ctrl-alt-space) and terminate the italics there (that is, italicize everything up to the non-breaking space).

I've gotten in the habit of doing this throughout my files, not just at the end of paragraphs. This is in contravention to the 'book' style where the italics include the terminating punctuation mark, but I'd rather be consistent to avoid subliminally irritating the reader.

Note that 'Smart Selection' in Word is one of the sources of this problem. If you can't live without Smart Selection, get in the habit of pressing left-arrow after doing a smart selection to back up to the actual end of the word. For instance, Smart Selection selects the following space

after a group of words. If you convert this selection to italics, the concluding space is italicized in addition, so you will have some lurking italics markup which is invisible. If you're just selecting *one* word to italicize, click anywhere in the body of the word, and then press ctrl-i to italicize; this does the correct thing and won't select following punctuation or whitespace.

This applies to other HTML formatting in MS Word, for instance font size changes and block indents. Inserting a firewall character will fix this problem.

Word also has a quirk of applying centering and block indentation to the successive paragraph, particularly if it is in the same style and empty. This is related to the font leakage bug but it can be even more maddening. Again, this does not show up until you save the file and load it again (or view it in a browser). This bug can also be prevented by inserting a firewall character. For instance, if you center or indent a paragraph or table, insert a left-indented paragraph consisting of one non-breaking space afterwards.

Submarining

One of the most irritating Word bugs occurs while editing documents. More that occasionally an entire line of text will disappear onscreen, just pop out of visibility. The text is still there, it's just not being displayed. I call this 'submarining'. To force it to display, simply scroll it off top or bottom of the window and then scroll back. There doesn't seem to be any reason for this and there isn't any way to prevent it.

Beyond 'Z'

When I was growing up one of my favorite books was Dr. Seuss' "On Beyond Z." Little did I know that eventually the alphabet would go quite a ways beyond Z.

There are three different issues here: accented characters, characters with Unicode representations, and those that can't be represented easily using the other two methods.

I use the ISO-8859-1 character set for common accented characters. This has the umlaut, acute, grave and circumflex vowel sets, as well as a few useful accented consonants such as c cedilla and 'thorn' and 'eth'. This covers all of the accents you'll need for Western European languages, such as French and German.

These are represented in standard HTML as 'character entities', and will show up in all standard browsers. Also, they can be converted to '8 bit' characters in a text file and for the most part will show up correctly. Sometimes Project Gutenberg releases two versions of text files, 7 and 8 bit.

If all you have is a couple of names with umlauts or accents, it may be sufficient to leave them out, or use 'ue' for ü and 'oe' for ö.

I have standard transliteration systems for Greek and Hebrew, and when I encounter text in these scripts I make an attempt to transliterate them; however, this is only recommended if you have actually studied that language and feel comfortable transliterating it. Of course, you could scan images of all of the non-Roman text. This applies to other text which is hard to represent in standard HTML character sets such as mathematical equations, chemical formulae, and so on. The brute-force solution is to scan images of these and insert them into your text. This is hard to get right, however, and requires some image processing black magic. I will discuss image

processing for etext production below.

Appendix: THROW THE BUMS OUT!

The easy errors to correct are those that the spellchecker flags right off the bat. After that you have words that are flagged but are 'correct', that you have to add to the spellcheck dictionary. What's left after that? 'Bums'! What are 'bums'? It helps to understand how an OCR program works.

First an OCR algorithm tries to figure out what pixels translate into glyphs (characters). It has to put together connected pixels into lumps, and work out which lumps might be connected (for instance, the dot of the 'i'). Once it has the glyphs on the page worked out, it lumps those glyphs together into words. It then attempts to take the characters that it has pieced together and find the best match for it that is an actual English word. 98% of the time this works fairly well.

However, we are left with a number of words in our scanned text which are valid English words, but not the word in the original text. The difficulty is that if you use a spell checker it will not flag these words. Even if you are reading the OCR output, your brain will not recognize all of these, since after awhile it is simply looking for misspelled words, no matter how careful you are.

One example is 'bum'. The word 'burn' is sometimes mis-recognized as 'bum', because the rn combination gets misrecognized as 'm'. For this reason, I call this type of OCR error, 'bums'. Another example is 'he' and 'be', 'bad' and 'had'. Obviously, 'h' and 'b' get confused, and often. Other problem areas are h being mis-recognized as li, confusion of c and e, and so on.

You need to train yourself to look for certain words that fall into this category and correct them by hand while doing the main proof pass. Once you know what to look for, they will pop out at you, even while reading texts proofed by others.

Certain combinations of font, OCR software settings and so on will tend to produce more or less instances of a particular bum. When you have an idea of which ones are showing up in a particular file, you can also fix them globally, (but one at a time); this can be done as a post-proof pass or while you are proofing.

Obviously, what you *don't* want to do is simply globally replace 'be' with 'he'! What you need to do is walk through the file doing a find on each instance of the bum word and correcting only the ones that are wrong. The way to do this in MSWord is to do a 'find' (Ctrl-F) on one of the bums, say 'bad', once. Then use Ctrl-PgDn to search for additional instances of 'bad'. If it needs to be changed to had, type 'had'. Then the next time you have to change an instance, use Ctrl-Y to repeat the previous typing. This can make a fixing a text with a lot of bums that are common words such as be/he easy (if not particularly enjoyable) to fix quickly.

A comprehensive list is impossible to construct. Again, the *bums* in a particular document will depend on the font of the source document and the OCR software you are using. Here are some of the most common bums to watch for. The first word in each pair is the 'bum' and the second is the correct word (although in some cases, the second word can also be a bum of the first. In this case, they are separated by a slash). The most common bums are marked with an asterisk. These are the ones which show up in nearly every OmniPage OCR, and are worth doing a quick scan

for at the end of each job.

b/h

This is the most common substitution. It probably depends on how wide the serif at the bottom of the 'h' is in a particular font. Some scans, are rife with this problem others are completely free. It reminds one of the Monty Python sketch where the Nazis try refurbishing their image by substituting 'b' for 'h' and other letters: "I don't like Bitler, and I don't like those boncentration bamps, either!".

bad, had * bag, hag bail, hail (and bailing, hailing) ball, hall band / hand (and bands / hands) bang, hang (and bangs, hangs; banging, hanging) bard. hard bare, hare bat, hat bath, hath be / he * bead, head (and bead, heads; beading, heading) beard, heard beat, heat (and beating, heating) bell, hell bid, hid bide / hide (and biding / hiding, bidden / hidden) bight, hight bill, hill (and bills, hills) his, bis (yes, bis is a word) bit, hit bobble, hobble (and bobbles, hobbles)

bold, hold

bow / how

bowl, howl (and bowled, howled; bowling, howling)

bunch, hunch

bung, hung

bunt, hunt (and bunting, hunting)

but, hut

li/h

lie, he *

tile, the *

lien, hen

ri/n

arid, and *

riot, not *

rn/m

burn, bum

comers, corners

com, corn

modem, modern

tom, torn

stem, stern

c/e

This one seems to turn up when the horizontal stroke in the 'e' is a little narrower than usual.

arc / are * ease, case cast / east * ear / car *

```
eave, cave (and eaves, caves)
scat, seat *
tic, tie
sec, see
```

il/n

This one is usually a product of a skewed page or type.

mail, man *

oil, on *

li/u

This one is likewise a product of a skewed page or type.

lip, up *

ti/n

batik, bank (and batiks, banks)

ti/u

This one is likewise a product of a skewed page or type.

tip, up *

tinder, under

ll/n

ill, in *

x/r

axe, are

cl/d

clog, dog

clown, down

ni/m

Fortunately both of these bums don't appear outside of certain contexts, so they are usually easy to fix.

sonic, some

conic, come

(1

This one is usually because the curve in the 'd' in the font is a bit taller than normal. It doesn't happen a lot, but when it does, it is infuriating.

(lay, day

(laughter, daughter

(lone, done

Al, M

Alan, Man

Some of the bums can be fixed up using a simple global search and replace, if you are sure that they don't occur your file. For instance, the bum *modem* (for modern) is unlikely to show up in a 19th century book. If you're not sure, use the scan and replace method above. Some of the most likely culprits in this category are:

conic, come
sonic, some
modem, modern
clown, down
com, corn
bath, hath
bare, hare
bard, hard
batik, bank (and batiks, banks)
as appropriate that in addition to

It seems appropriate that in addition to tossing out the *bums*, we must also get rid of the *clowns*.

And of course, lest we forget:

Trish, Irish

Farce, Faroe (Islands)

Who wouldn't love to book a ticket to the Farce Islands, possibly with Trish? Thankfully, these are probably just confined to one specific book (*America Discovered by Norsemen*) with a combination of these particular specialized terms and a particular font. However, it just goes to show that you have to be on the lookout constantly for new and inventive OCR bums. There were three instances of 'Trish/Irish' in this book and it wasn't until I was almost done that I

noticed one of them. The human brain is very good at ignoring typos like this and sometimes they won't pop out at you until you look very hard at a particular passage.

Note that many of the 'bums' may be OmniPage specific. Other OCR packages seem to do a better job of eliminating bums.

This is one of the hardest problems to spot while proofing, yet one of the most embarrassing. It is also one of the easiest problems to fix. You can probably look at any published Project Gutenberg text and find at least a couple of bums from the list above in a few minutes.