Title: The Project Gutenberg FAQ 2002

Author: Jim Tinsley

Release Date: October, 2005 [EBook #9109]
[Yes, we are more than one year ahead of schedule]
[This file was first posted on September 7, 2003]

Edition: 10

Language: English

Character set encoding: ASCII

*** START OF THE PROJECT GUTENBERG EBOOK THE PROJECT GUTENBERG FAQ 2002 ***

The Project Gutenberg FAQ 2002

by Jim Tinsley

Important: This file is posted to the Project Gutenberg archives
not as a current guide, more as a historical reference. I hope
that future FAQs will be posted, as the project evolves, but
this one is of its time.

If you want the most up-to-date information from PG, please see the current version of the FAQ, from the Project Gutenberg site, or, at the time of posting, at:

http://ibiblio.org/gutenberg/faq/gutfaq.txt
or
http://ibiblio.org/gutenberg/faq/gutfaq.htm

Acknowledgements

Writing a FAQ for an organization of fanatical proofreaders has its ups and downs! I'd like to thank all those who corrected my facts and my typos, and especially the people who pointed out the lack of clarity in certain answers. The remaining errors and opacity are all mine.

Preface to the archive edition

Ironically, Project Gutenberg, which preserves the writings of others, doesn't have much written history itself. There are scraps of e-mails and guidelines, but many newsletters and other internal writings before 1996 have gone to the great bit-bucket in the sky.

The later half of the '90s marked a graceful blooming of Project Gutenberg's growth. Three related technical factors contributed: the explosion in home PCs brought standardization, which made it easy for non-techies to install scanners, which, in response to the new demand, became plentiful and cheap. And, of course, these years saw the rise in popularity of the Internet, which has always been PG's main channel of communication and distribution.

However, while PG's production expanded geometrically, at Moore's Law rates, there were barriers to participation. Most volunteers had to find an eligible book, scan or type it, and proof the resulting text all by themselves. This was and is a fairly significant amount of work: 40 painstaking hours would be a typical commitment for one book.

Beyond that, simply learning the mechanics of producing e-texts could be a serious challenge for newcomers. Nearly all internal PG communication, except for the Newsletter, was by private e-mail, and instructions had to be repeated many times to individual new volunteers, all of whom showed up with great good will, but most of whom vanished after a week or two.

Michael Hart was unstinting in his editing of incoming texts and

handling questions by e-mail, but any one person has only so many hours.

The Directors of Production at the time -- Sue Asscher, Dianne Bean, John Bickers and David Price -- served as contact points for advice and help, made enormous efforts of production themselves, and tried to share the scanned texts among new volunteers for proofing. They made a huge contribution to building community in PG.

Pietro Di Miceli set up a web site for the project in 1996, and with the popularization of the Web (as opposed to the Internet), this became a beacon for readers and new volunteers.

All of these people reached out to willing volunteers, drew them in, helped them, encouraged them. The Project and all of the readers of the books, now and in the future, owe these people a great debt. Without them, Project Gutenberg could not have achieved what it has. But still, for the most part, each volunteer worked alone.

In 1999, I wrote, in response to an offer to volunteer:

> I think I can best answer your offer, and many others like it, by giving an extended description of what actually happens in the making of PG texts, and why it's often not easy to get started.
>
> There is no agenda, no master list of tasks ready to be given to volunteers. This is often the hardest thing to get across to new volunteers. I know I waited quite a while after volunteering for someone to give me a job to do before I realized it.
>
> Exactly five steps are normally performed in the publishing of an e-text.
>
> 1. Someone, somewhere gets a public-domain copy of a text they want to contribute.
>
> 2. That volunteer confirms its PD status by sending TP&V to Michael, and getting copyright clearance.
>
> 3. Someone, usually the same volunteer, scans and corrects the text, or, if skilled in typing, types the book into an e-text.
>
> 4. Someone, often a different volunteer, second-proofs the e-text, removing the smaller errors.
>
> 5. The e-text is sent to Michael for posting.
>
> There are three barriers which make it difficult for most people to contribute:

1. Getting a PD book.

2. People without scanners and typing skills have no way of turning a book into an e-text.

3. Even with a scanner, turning a book into an e-text is not easy or quick.

Since, generally, people who have a PD book don't just want to send it off to a stranger for scanning, the people who produce e-texts have to get over all three of these barriers. This is the bottleneck in production. It's relatively easy to get an e-text second-proofed; making it in the first place is the hardest part. You need to have a book, the means to turn it into an e-text and the time and will to do it.

After that comes second proofing. There are two problems here. One is that there may not be enough texts for all the people who want to second-proof; the other is that a lot of beginners just abandon texts given to them for second-proofing, which holds up the process and is discouraging for others. So a lot of volunteers do their own second-proofing or send their texts to established contacts with a track record of finishing the job, rather than making them available to newbies. The Directors of Production do serve as contact points, and at any given moment may have some texts for proofing, but they can only distribute the texts that have already been made.

With that explanation out of the way, I can better address your question of what you can do.

Second-proofing is an easy way to start, but material isn't just waiting for you. If you want to look for some, post your offer here and wait a week or so. If no takers by then, e-mail Michael and ask if there are any texts available; he may be able to refer you to a Director of Production who has something current. You may not get an e-text immediately, but you will get one. Of course, you can also look here for offers of e-texts ready to proof.

Your other option is to take on a book yourself. In your case, you already have a scanner, so you are equipped to become a producer. You need to find a PD book.

Getting PD books means finding and borrowing or buying them. You can do this through used bookshops, libraries or book sites on the Internet. I mention a few net sites in the FAQ in the link below. I get all my books through them, since they make it easy for me to find the books I want. Prices range from $5 up to (in my case) about $30.

> The best advice I can offer here is: pick a book that you _want_ to contribute, and a book you'll enjoy working with--you'll be living with it up close and personal for quite a while.

In March and April of 1999, Pietro created the PG Volunteers' WWWBoard and Greg Newby set up the mailing list gutvol-d, and, for the first time, volunteers who hadn't been introduced to each other by Michael or the Directors could meet online and communicate directly. A few FAQs and HOWTOs were written, covering the basics, the nitty-gritty of producing books. All of this activity made it much easier for people to get involved, and the Project experienced a new influx of interested volunteers. Improved OCR software was also a factor at this time: in response to the commoditization of scanners, there was rapid improvement in the quality of OCR, and better OCR made for easier production of e-texts. More work was shared out in co-operative proofing experiments.

It was in this new, expansive atmosphere, with ideas flooding in from enthusiasts newly energized by the project, that Charles Franks (Charlz) came up with the idea of a web site that would serve to distribute the work of proofing a book among many volunteers. But not only did he think of the concept; he went ahead and did it!

In April 2000, Charlz first requested comments on his idea in a post on the Volunteers' WWWBoard, and by the end of September, the first e-texts were queueing up on the production line.

On October 9th, Charlz wrote:

> Number of pages proofed by date:
>
> 2nd 6
> 3rd 6
> 4th 20 <-- Newsletter
> 5th 27
> 6th 25
> 7th 29
> 8th 30
> 9th 45!! (and the day ain't over yet)

(The "Newsletter" is a reference to the site being mentioned in the PG Newsletter on October 4th, 2000).

Distributed Proofreaders, or DP, simply kept growing from there, as Charlz kept scanning and adding more books and features and proofers, and its simple organic growth produced 600 e-texts in two years, but when Charlz asked for more help on Slashdot, a popular technical news site, on November 8th, 2002, the response blew the roof off! The pages per day figure jumped from 1,000 to about 10,000 for a while, then settled down at its current 4,000. 4,000 pages,

even given that each page is proofed twice, is a lot of pages. 2,000 produced pages per day is about five full books per day. DP has formed the backbone of PG's production ever since. Whatever the future of DP's production, its effect on shared knowledge and resources, and the communication and community it has built, ensures that Project Gutenberg will never be the same again.

I began writing this FAQ in March 2002, and was essentially finished around December 2002. It sat around, with a few tweaks here and there in response to comments, until the start of September 2003.

Today, it is a useful guide to Project Gutenberg norms and practices. By the time you read it, it may be ancient history ("Hey, Grandad, did you REALLY scan things from paper? Why didn't you use your brain implant?" :-) But it is one record of How Things Were in Project Gutenberg during this time of change.

jim
September 7th, 2003.

Project Gutenberg FAQ 2002

I have a question not answered in this FAQ. How do I ask it?

If it's about how to produce a text, the Volunteers' Board at <http://www.gutenberg.net/vol/wwwboard/> is generally the best place to ask.

If it's a question of active interest to the general body of volunteers, you can ask it on the gutvol-d mailing list. See <http://www.gutenberg.net/subs.html> for joining it.

For other questions, you should check our Contact Information page at <http://www.gutenberg.net/contactinfo.html> and e-mail the appropriate person.

About Project Gutenberg:

G.1. What is Project Gutenberg?
G.2. Where did Project Gutenberg come from?
G.3. What has Project Gutenberg achieved?
G.4. Who runs Project Gutenberg?
G.5. How many people are in Project Gutenberg?
G.6. How can I contact Project Gutenberg?

G.7. How can I help Project Gutenberg?
G.8. How can I keep in touch with what Project Gutenberg is doing?
G.9. What is the relationship between Project Gutenberg, Projekt Gutenberg-DE, Project Gutenberg of Australia, and Project Runeberg?

About Project Gutenberg publications:

G.10. Does Project Gutenberg publish only books?
G.11. What books does Project Gutenberg publish?
G.12. What other things does Project Gutenberg publish?
G.13. How does Project Gutenberg choose books to publish?
G.14. What languages does Project Gutenberg publish in?
G.15. Why don't you have any / many books about history, geography, science,
G.16. Why don't you have any books by Steven King, Tom Clancy, Tolkien, etc.?
G.17. Why is Project Gutenberg so set on using Plain Vanilla ASCII?

Readers' FAQ

About Finding eBooks:

R.1. How can I find an eBook I'm looking for?
R.2. Can I get a complete list of Project Gutenberg eBooks?
R.3. How can I download a PG text that hasn't been cataloged yet?
R.4. You don't have the eBook I'm looking for. Can you help me find it?
R.5. Where else can I go to get eBooks?
R.6. I see some eBooks in several places on the Net. Do different people really re-create the same eBooks?

About Using the Web Site:

R.7. Why couldn't I reach your site? (or: Why is your site slow?)
R.8. I get an error when I try to download a book.
R.9. I searched for a book I know is in Project Gutenberg, but got no results.
R.10. Can I copy your website, or your website materials?
R.11. Your site doesn't look right in my browser. I clicked on a button, and nothing happened.
R.12. What does that thing about "Select FTP Site" mean?
R.13. What exactly is an FTP site anyway?
R.14. Can I become an FTP mirror?
R.15. Can I make a private FTP mirror for my school, library or organization?
R.16. When I clicked on the file I want, nothing happened.
R.17. How many texts are downloaded through the web site?
R.18. What are the most popular books?

About Downloading and Using Project Gutenberg eBooks:

R.19. Should I download a ZIP or a TXT file?
R.20. I've got a ZIP file. What do I do with it?
R.21. I tried to unzip my file, but it said the file was corrupt, or damaged.
R.22. I see gibberish onscreen when I click on a book.
R.23. Can I download and read your books?
R.24. What am I allowed to do with the books I download?
R.25. Does Project Gutenberg know who downloads their books?
R.26. I've found some obvious typos in a Project Gutenberg text. How should I report them?
R.27. I've found some obvious typos in a Project Gutenberg text. Who should I report them to?
R.28. I've reported some typos. What will happen next?
R.29. I've got the text file, and I can read it, but it seems to be double-spaced or it has control characters like ^J or ^M at the end of every line.
R.30. When I print out the text file, each line runs over the edge of the page and looks bad.
R.31. I can read the text file, but a few characters appear as black squares, or gibberish.
R.32. Can I get a handheld device for reading PG texts? Which device should I get?
R.33. How can I read a PG eBook on my PDA (Palm, iPaq, Rocket . . .)

About the Files:

R.34. What types of files are there, and how do I read them?
R.35. What do the filenames of the texts mean?
R.36. What is the difference within PG between an "edition" and a "version"?
R.37. What is the difference between an "etext" and an "eBook"?
R.38. What are the "Etext/Ebook numbers" on the texts?
R.39. What do the month and year on the text mean?

Copyright FAQ

C.1. What is copyright?
C.2. Does copyright differ from country to country? From state to state?
C.3. What are the copyright laws outside the U.S.?
C.4. Why does Project Gutenberg advise only on U.S. copyright issues?
C.5. I don't live in the U.S. Do these rules apply to me?
C.6. What is the public domain?
C.7. What can I do with a text that is in the public domain?
C.8. How does a book enter the public domain?
C.9. How does a copyright lapse?
C.10. What books are in the public domain?
C.11. My book says that it's "Copyright 1894". Is it in the public domain?
C.12. How can a copyright owner release a work into the public domain?
C.13. When is an author not the owner of a copyright on his or her works?
C.14. What does Project Gutenberg mean by "eligible"?

C.15. I have a manuscript from 1900. Is it eligible?
C.16. How come my paper book of Shakespeare says it's "Copyright 1988"?
C.17. What makes a "new copyright"?
C.18. I have a 1990 book that I know was originally written in 1840, but the publisher is claiming a new copyright. What should I do?
C.19. I have a 1990 reprint of an 1831 original. Is it eligible?
C.20. I have a text that I know was based on a pre-1923 book, but I don't have the title page. Can I submit it to PG?
C.21. How does Project Gutenberg "clear" books for copyright?
C.22. I want to produce a particular book. Will it be copyright cleared?
C.23. I have some extra material (images, introduction, preface, missing chapter) that should go into an existing PG text. Do I have to copyright-clear my edition before submitting it?
C.24. I see some Project Gutenberg eBooks that are copyrighted. What's up with that?
C.25. What are "non-renewed" books?
C.26. How can I get Project Gutenberg to clear a non-renewed book?

Volunteers' FAQ

About the Basics:

V.1. How do I get started as a Project Gutenberg volunteer?
V.2. What experience do I need to produce or proof a text?
V.3. How do I produce a text?
V.4. Do I need any special equipment?
V.5. Do I need to be able to program?
V.6. I am a programmer, and I would like to help by programming.
V.7. What does a Gutenberg volunteer actually do?
V.8. Can I produce a book in my own language?
V.9. Does it have to be a book? Can I produce pieces from a magazine or other periodical?
V.10. Do I _have_ to produce in plain ASCII text?
V.11. Where do I sign up as a volunteer?
V.12. How do PG volunteers communicate, keep in touch, or co-ordinate work?
V.13. Where can I find a list of books that need proofing?
V.14. Is there a list of books that Project Gutenberg wants?
V.15. I have one book I'd like to contribute. Can I do just that without signing up?

About production:

V.16. How does a text get produced?
V.17. How long must a text be to qualify for PG?
V.18. What books are eligible?
V.19. Are reprints or facsimiles eligible?
V.20. What is the difference between a reprint and a facsimile?
V.21. What is the difference between a reprint and a "new edition"?
V.22. What book should I work on?
V.23. I have a book in mind, but I don't have an eligible copy.

V.24. Where can I find an eligible book?
V.25. What is "TP&V"?
V.26. What is "Posting"?
V.27. I think I've found an eligible book that I'd like to work on. What do I do next?
V.28. What books are currently being worked on?
V.29. How do I find out if my book is already on-line somewhere?
V.30. My book is not on the In-Progress list, and I can't find it on-line.
V.31. My book is on-line, but not in Project Gutenberg. What should I do?
V.32. My book is already on-line in Project Gutenberg, but my printed book is different from the version already archived. Can I add my version?
V.33. I see a book that was being worked on three years ago. Is anyone still working on it?
V.34. I've decided which book to produce. How do I tell PG I'm working on it?
V.35. I have a two- or three-volume set. Should I submit them as one text, or one text for each volume?
V.36. I have one physical book, with multiple works in it (like a collection of plays). Should I submit each text separately?
V.37. How do I get copyright clearance?
V.38. I have a two- or three-volume set. Do I have to get a separate clearance on each physical book?
V.39. I have one physical book, with multiple works in it (like a collection of plays). Do I have to get a separate clearance for each work?
V.40. Who will check up on my progress? When?
V.41. How long should it take me to complete a book?
V.42. I want/don't want my name published on my e-text
V.43. I'd like to put a copy of my finished e-text, or another Gutenberg text, on my own web page.
V.44. I've scanned, edited and proofed my text. How do I find someone to second-proof it?
V.45. I've gone over and over my text. I can't find any more errors, and I'm sick of looking at it. What should I do now?
V.46. Where and how can I send my text for posting?
V.47. What is the "Credits Line"?
V.48. How soon after I send it will my text be posted?
V.49. I found a problem with my posted text. What do I do?
V.50. Someone has e-mailed me about my posted text, pointing out errors.
V.51. Someone has e-mailed me about my posted text, thanking me.

About Proofing:

V.52. What role does proofing play in Project Gutenberg?
V.53. What is Distributed Proofing?
V.54. What do I need to proof an e-text?
V.55. Do I need to have a paper copy of the book I'm proofing?
V.56. What's the difference between "first proof" and "second proof"?
V.57. What do I do with an e-text sent to me for proofing?
V.58. What kinds of errors will I have to correct?
V.59. How long does it take to proof an e-text?
V.60. Are there any special techniques for proofing?

V.61. What actually happens during a proof?

About Net searching:

V.62. I've found an eligible text elsewhere on the Net, but it's not in the PG archives. Can I just submit it to PG?
V.63. I've found an eligible text elsewhere on the Net, but it's not in the PG archives. Why should I submit it to PG?
V.64. I have already scanned or typed a book; it's on my web site. How can I get it included in the Gutenberg archives?
V.65. I have already scanned or typed a book; it's on my web site. The world can already access it. Why should I add it to the Gutenberg archives?
V.66. I have already scanned or typed a book, but it's not in plain text format. Can I submit it to PG?

About author-submitted eBooks:

V.67. I've written a book. Will PG publish it?
V.68. I have translated a classic book from one language to another. Will PG publish my translation?
V.69. OK, this is one of the cases where PG will publish it. What do I do next?
V.70. I hold the copyright on a book. Can I release it to the public domain?
V.71. I hold the copyright on a book. Do I have to release the book into the public domain for Project Gutenberg to publish it?
V.72. I hold the copyright on a book, and would like Project Gutenberg to publish it. Can I choose what rights to assign?

About what goes into the texts:

V.73. Why does PG format texts the way it does?

About the characters you use:

V.74. What characters can I use?
V.75. What is ASCII?
V.76. So what is ISO-8859? What is Codepage 437? What is Codepage 1252? What is MacRoman?
V.77. What is Unicode?
V.78. What is Big-5?
V.79. What are "8-bit" and "7-bit" texts?
V.80. I have an English text with some quotations from a language that needs accents--what should I do about the accents?
V.81. I have some Greek quotations in my book. How can I handle them?
V.82. I want to produce a book in a language like Spanish or French with accented characters. What should I do?

About the formatting of a text file:

V.83. How long should I make my lines of text?
V.84. Why should I break lines at all? Why not make the text as one line per paragraph, and let the reader wrap it?
V.85. Why use a CR/LF at end of line?
V.86. One space or two at the end of a sentence?
V.87. How do I indicate paragraphs?
V.88. Should I indent the start of every paragraph?
V.89. Are there any places where I should indent text?
V.90. Can I use tabs (the TAB key) to indent?
V.91. How should I treat dashes (hyphens) between words?
V.92. How should I treat dashes replacing letters?
V.93. What about hyphens at end of line?
V.94. What should I do with italics?
V.95. Yes, but I have a long passage of my book in italics! I can't really CAPITALIZE or _otherwise_ /mark/ all that text, can I?
V.96. Should I capitalize the first word in each chapter?
V.97. What is a Transcriber's Note? When should I add one?
V.98. Should I keep page numbers in the e-text?
V.99. In the exceptional cases where I keep page numbers, how should I format them?
V.100. Should I keep Tables of Contents?
V.101. Should I keep Indexes and Glossaries?
V.102. How do I handle a break from one scene to another, where the book uses blank lines, or a row of asterisks?
V.103. How should I treat footnotes?
V.104. My book leaves a space before punctuation like semicolons, question marks, exclamation marks and quotes. Should I do the same?
V.105. My book leaves a space in the middle of contracted words like "do n't", "we 'll" and "he 's". Should I do the same?
V.106. How should I handle tables?
V.107. How should I format letters or journal entries?
V.108. What can I do with the British pound sign?
V.109. What can I do with the degree symbol?
V.110. How should I handle . . . ellipses?
V.111. How should I handle chapter and section headings?
V.112. My book has advertisements at the end. Should I keep them?
V.113. Can I keep Lists of Illustrations, even when producing a plain text file?
V.114. Can I include the captions of Illustrations, even when producing a plain text file?
V.115. Can I include images with my text file?

About formatting poetry:

V.116. I'm producing a book of poetry. How should I format it?
V.117. I'm producing a novel with some short quotations from poems.

About formatting plays:

V.118. How should I format Act and Scene headings?
V.119. How should I format stage directions?
V.120. How should I format blank verse?

About some typical formatting issues:

V.121. Sample 1: Typical formatting issues of a novel.
V.122. Sample 2: Typical formatting issues of non-fiction
V.123. Sample 3: Typical formatting issues of poetry
V.124. Sample 4: Typical formatting issues of plays

About problems with the printed books:

V.125. I found some distasteful or offensive passages in a book I'm producing. Should I omit them?
V.126. Some paragraphs in my book, where a character is speaking, have quotes at the start, but not at the end. Should I close those quotes?
V.127. The spelling in my book is British English (colour, centre). Should I change these to American spellings?
V.128. I'm nearly sure that some words in my printed book are typos. Should I change them?
V.129. Having investigated what looks like a typo, I find it isn't. Do I need to do anything?
V.130. Aarrgh! Some pages are missing! Do I have to abandon the book?
V.131. Some words are spelled inconsistently in my book (e.g. sometimes "surprise", sometimes "surprize"). Should I make them consistent?

Word Processing FAQ

W.1. What's the difference between an editor and a word processor?
W.2. Should I use an editor or a word processor?
W.3. Which editor or word processor should I use?
W.4. How can I make my word processor easier to work with for plain text?
W.5. What is the difference between proportional and non-proportional fonts?
W.6. I can't get words in a table or poem to line up under each other.

About using MS-Word:

W.7. I've edited my book in Word - how do I save it as plain text?
W.8. Quotes look wrong when I save a Word document as plain text.
W.9. Dashes look wrong when I save a Word document as plain text.
W.10. I saved my Word document as HTML, but the HTML looks terrible.

Scanning FAQ

S.1. What is a scanner?
S.2. What types of scanners are there?
S.3. Which scanner should I get?
S.4. What is ADF?
S.5. Should I get ADF?
S.6. What's a "TWAIN driver" and why do I need one?
S.7. How do I scan a book?
S.8. My book won't open flat enough for a good scan, and I don't want to cut the pages.
S.9. How long does it take to scan a book?
S.10. What scanner settings are best?
S.11. Can I use a digital camera in place of a scanner?
S.12. What is OCR?
S.13. What differences are there between OCR packages?
S.14. How accurate should OCR be?
S.15. Which OCR package should I get?
S.16. What types of mistakes do OCR packages typically make?
S.17. Why am I getting a lot of mistakes in my OCRed text?
S.18. I got an OCR package bundled with my scanner. Is it good enough to use?
S.19. I want to include some images with a HTML version. How should I scan them?
S.20. I want to include some images with a HTML version. What type of image should I use?
S.21. Will PG store scanned page images of my book?

HTML FAQ

H.1. Can I submit a HTML version of my text?
H.2. Why should I make a HTML version?
H.3. Can I submit a HTML version without a plain ASCII version?
H.4. What are the PG rules for HTML texts?
H.5. Can I use Javascript or other scripting languages in my HTML?
H.6. Should I make my HTML edition all on one page, or split it into multiple linked pages?
H.7. How can I check that I haven't made mistakes in coding my HTML?
H.8. Can I submit a HTML or other format of somebody else's text?
H.9. How big can the images be in a HTML file?
H.10. The images I've scanned are too big for inclusion in HTML. What can I do about it?
H.11. Can I include decorative images I've made or found?
H.12. How can I make a plain text version from a HTML file?
H.13. How can I make a HTML version from my plain text file?

Programs and Programming FAQ

P.1. What useful programs are available for Project Gutenberg work?
P.2. What programs could I write to help with PG work?

Formats FAQ

F.1. What formats does Project Gutenberg publish?
F.2. What is, and how do I make or use various formats?

Volunteers' Voices - Volunteers talk about PG

Amy Zelmer
Ben Crowder
Col Choat
Dagny
Gardner Buchanan
Jim Tinsley
John Mamoun
Ken Reeder
Lynn Hill
Sandra Laythorpe
Tony Adam
Tonya Allen
Walter Debeuf

Bookmarks - web pages commonly referred to in the FAQ

B.1. Project Gutenberg
B.2. Distributed Proofing Sites
B.3. Other On-Line eBook Pages
B.4. Lists of Suggested Books to Transcribe
B.5. Finding Paper Books On-Line

About Project Gutenberg:

G.1. What is Project Gutenberg?

Project Gutenberg is a volunteer effort to digitize, archive, and distribute cultural works.

G.2. Where did Project Gutenberg come from?

In 1971, Michael Hart was given $100,000,000 worth of computer time on a mainframe of the era. Trying to figure out how to put these very expensive hours to good use, he envisaged a time when there would be millions of connected computers, and typed in the Declaration of

Independence (all in upper case--there was no lower case available!). His idea was that everybody who had access to a computer could have a copy of the text. Now, 31 years later, his copy of the Declaration of Independence (with lower-case added!) is still available to everyone on the Internet.

During the 70s, he added some more classic American texts, and through the 80s worked on the Bible and the collected works of Shakespeare. That edition of Shakespeare was never released, due to copyright law changes, but others followed.

Starting in 1991, Project Gutenberg began to take its current form, with many different texts and defined targets. The target for 1991 was one book a month. 1992's target was two books a month. This target doubled every year through 1996, when it hit 32 books a month.

Today, we have a target of 200 books a month.

G.3. What has Project Gutenberg achieved?

Project Gutenberg is the original, and oldest, etext project on the Internet, founded in 1971.

In mid-2002, we are not only still going, we have made over 5,000 eBooks available, with a current production target of 200 more each month.

We have many mirrors (copies) of our archives on all five continents.

G.4. Who runs Project Gutenberg?

The Project Gutenberg Literary Archive Foundation is a 501(c)(3) organization. Dr. Gregory B. Newby <gbnewby@ils.unc.edu> is our volunteer CEO. Professor Michael Hart <hart@pobox.com> is our Founder and Executive Director.

In terms of the day-to-day production of eBooks, our volunteers run themselves. :-) They produce books, and submit them when completed. Our Production Directors help with general volunteer issues. The Posting Team check submitted texts and shepherd them onto our servers. You can find current contact information for these people on the Contact Information page at <http://www.gutenberg.net/contactinfo.html>.

G.5. How many people are in Project Gutenberg?

As of mid-2002, there are about 100 active producers, and 200 regular, active helpers doing tasks like proofing. Something like 1500 people

receive our Newsletter.

G.6. How can I contact Project Gutenberg?

There are lots of ways to contact us, depending on what you want to talk about. The Contact Info page <http://www.gutenberg.net/contactinfo.html> on the main web site lists them.

G.7. How can I help Project Gutenberg?

Donate money! We're an all-volunteer project, and we don't have much to spend, so even a little goes a long way. Our Donation page <http://www.gutenberg.net/donation.html> tells you how.

Produce a text! Turn an old book into an immortal etext.
The Volunteers' FAQ [V.1] tells you how.

G.8. How can I keep in touch with what Project Gutenberg is doing?

Subscribe to one of the Newsletters--weekly or monthly!

The page <http://www.gutenberg.net/subs.html> gives details of how to subscribe, unsubscribe and access the archives.

G.9. What is the relationship between Project Gutenberg, Projekt Gutenberg-DE, Project Gutenberg of Australia, and Project Runeberg?

These are all entirely separate organizations. Projekt Gutenberg-DE and Project Gutenberg of Australia use the "Project Gutenberg" trademark with permission, and they operate within the copyright rules of their respective countries. Project Runeberg has no specific connection with Project Gutenberg.

About Project Gutenberg publications:

G.10. Does Project Gutenberg publish only books?

No.

Project Gutenberg also publishes other cultural works like movies and music, but the bulk of our collection is books.

G.11. What books does Project Gutenberg publish?

Any books that we legally can, and that our volunteers want to work on.

We cannot publish any texts still in copyright without permission. This generally means that our texts are taken from books published pre-1923. (It's more complicated than that, as our Copyright FAQ explains, but 1923 is a good first rule-of-thumb for the U.S.A.)

So you won't find the latest bestsellers or modern computer books here. You _will_ find the classic books from the start of this century and previous centuries, from authors like Shakespeare, Poe, Dante, as well as well-loved favorites like the Sherlock Holmes stories by Sir Arthur Conan Doyle, the Tarzan and Mars books of Edgar Rice Burroughs, Alice's adventures in Wonderland as told by Lewis Carroll, and thousands of others.

These books are chosen by our volunteers. Simply, a volunteer decides that a certain book should be in the archives, obtains the book and does the work necessary to turn it into an e-text. If you're interested in volunteering, see the Volunteers' FAQ at [V.1] below.

G.12. What other things does Project Gutenberg publish?

We have published some music files, in MIDI and MUS formats. We have published the Human Genome. We have published pictures of the prehistoric cave paintings from the south of France. We have published some video files and some audio files, including a Janis Ian track and readings from public domain books.

G.13. How does Project Gutenberg choose books to publish?

Project Gutenberg, as such, does not choose books to publish. There is no central list of works that volunteers are asked to work on. Individual volunteers choose and produce books according to their own tastes and values, and the availability (or price!) of the book.

G.14. What languages does Project Gutenberg publish in?

Whatever languages we can! As above, this is decided by what languages our volunteers choose to work with.

G.15. Why don't you have any / many books about history, geography, science, biography, etc.?
Why aren't there any / more PG books available in French, Spanish, German, etc.?

If we can legally publish a book, and it isn't in the archives, it's because no volunteer has produced it yet. At the moment, we have a predominance of English language novels because that is what most people have chosen to work on.

We're always looking for new languages and topics, and always delighted to see people producing them. If we don't have enough of the types of books you would like to see, why don't you help us out by contributing one? If the people interested in a particular area don't contribute, we'll always be short in that area.

G.16. Why don't you have any books by Steven King, Tom Clancy, Tolkien, etc.?

Project Gutenberg can publish only books that are in the public domain [C.10] unless we have the permission of the copyright holder. Current bestsellers have not yet entered the public domain, and we're not likely to get permission from the authors to publish them.

G.17. Why is Project Gutenberg so set on using Plain Vanilla ASCII?

Don't misrepresent us--we support and publish many open formats, but, yes, we do want to have a plain text version of everything possible.

We're looking at our history, and we're planning for the long term--the _very_ long term.

Today, Plain Vanilla ASCII can be read, written, copied and printed by just about every simple text editor on every computer in the world. This has been so for over thirty years, and is likely to be so for the foreseeable future. We've seen formats and extended character sets come and go; plain text stays with us. We can still read Shakespeare's First Folios, the original Gutenberg Bible, the Domesday Book, and even the Dead Sea Scrolls and the Rosetta Stone (though we may have trouble with the language!), but we can't read many files made in various formats on computer media just 20 years ago.

We're trying to build an archive that will last not only decades, but _centuries_.

The point of putting works in the PG archive is that they are copied

to many, many public sites and individual computers all over the world. No single disaster can destroy them; no single government can suppress them. Long after we're all dead and gone, when the very concept of an ISP is as quaint as gas streetlamps, when HTML reads like Middle English, those texts will still be safe, copied, and available to our descendants.

The PG archive is so valuable, yet free and easily portable, that even if every current PG volunteer vanished overnight, people around the world would copy and preserve it.

If the ZIP format loses popularity, and is replaced by better compression, it will be easy to convert the zip formats automatically (and we post all plain-text files in unzipped format as well). If hard drives are replaced by optical memory, it will be easy to copy the files onto that. If even ASCII is superseded by Unicode or one of its descendants, it will be possible for our grandchildren to convert it automatically (and ASCII is included in Unicode anyway).

By contrast, many of us have files saved in proprietary formats from word-processors only 5 or 10 years old that are already impractical for us to read. Some of our files produced just a few years ago using non-ASCII character sets like Codepage 850 are already giving problems for some readers. Some eBook reader formats launched within the last few years are already obsolete. We have learned from that experience.

We also encourage other open formats based on plain text, like HTML and XML, and even occasionally not-so-open ones when simple formatting isn't enough, but plain text and ASCII is the only format and character set we're sure of in a rapidly-changing technological landscape.

Please see also the FAQ [F.1] "What formats does Project Gutenberg publish?" for more detailed discussion of formats.

Readers' FAQ

About Finding eBooks:

R.1. How can I find an eBook I'm looking for?

For PG books, the simplest way is to go to the home page at <http://www.gutenberg.net>, type the Author or Title into the search form, press the "Search" button, and follow the choices.

As of late 2002, there is a full-text search available at

<http://public.ibiblio.org/gsdl/cgi-bin/library.cgi>
where you can search not only for titles and authors, but any words or phrases you want to look up. For example, entering "Ample make this bed" and running an "entire books" search for all words leads you to Poems Of Emily Dickinson, Series Two.

R.2. Can I get a complete list of Project Gutenberg eBooks?

Yes. There are two main options:

GUTINDEX.ALL is the raw list of files posted. You will find it at:
<ftp://ibiblio.org/pub/docs/books/gutenberg/GUTINDEX.ALL>

PGWHOLE.TXT is the list of files cataloged. A Zipped version is:
<http://promo.net/gg/pgwhole.zip>

When we post a book, the posting information contains title and author, eBook number, base filename and schedule year and month. This raw information goes into GUTINDEX.ALL.

After posting, our catalogers get to work and add more information --things like full title, subtitle, author birth and death dates, Library of Congress Classification, full filenames and sizes. When a book has been cataloged, it is entered onto the website database so that you can search for it. PGWHOLE.TXT is a summary of the books in the website database.

People who want to bypass the search on the website and find books themselves will probably want to use GUTINDEX.ALL, since it doesn't wait for the cataloging.

R.3. How can I download a PG text that hasn't been cataloged yet?

In short, just browse to:

<http://www.ibiblio.org/pub/docs/books/gutenberg/>

choose the schedule year of the text (newly-posted texts will usually be in the latest year) and look down the list to find the filename you're looking for.

In general, you need to know:

a) the address of an FTP site
b) the schedule year of the text you want
c) the basename of the text you want.

The fastest and safest FTP site to use for this is ftp.ibiblio.org, which is the first of our two primary posting sites (the other being ftp.archive.org). We post to these two sites, and then other sites

copy from them at intervals, so with any FTP sites other than these two, the file may not be available immediately.

You can get the schedule year and basename of the text from its line in GUTINDEX.ALL. Let's take an example. The file

Mar 2004 The Herd Boy and His Hermit, by C. M. Yonge [#32][hrdbhxxx.xxx]5313

has been posted just a few hours ago as I write this. From the GUTINDEX entry, the schedule year is 2004, and the basename of the text is hrdbh.

We divide our texts into directories (folders) based on the schedule year, so this eBook will be in the directory for 2004, which will be named something ending in /etext04. All the directories are named etext plus the last two digits of the year. (Somebody's going to have to change that convention in about 87 years from now! :-) We currently have directories starting at 90, running through the 90s and then 00, 01, 02, 03, 04. All eBooks produced before 1991 are in the /etext90 directory, so if you're looking for

Dec 1971 Declaration of Independence [whenxxxx.xxx] 1
or
Aug 1989 The Bible, Both Testaments, King James Version [kjv10xxx.xxx] 10

you should look in /etext90.

As it happens, ibiblio supports both HTTP (web) and FTP access to the text, so we can just browse to

<http://www.ibiblio.org/pub/docs/books/gutenberg/>

and choose the 2004 directory from there.

If you want to automate this, you could also use the more direct address

<ftp://www.ibiblio.org/pub/docs/books/gutenberg/etext04/>

The equivalent address for ftp.archive.org is

<ftp://ftp.archive.org/pub/etext/etext04/>

Either way, we see a long page of files, in alphabetical order. Scroll down to the "H"s and look for hrdbh. We see four files with this basename:

hrdbh10.txt
hrdbh10.zip
hrdbh10h.htm
hrdbh10h.zip

This means that both plain text and HTML formats are available,

and you can choose to download them either zipped or uncompressed. For more detail about conventions for filenames, see the FAQ "What do the filenames of the texts mean?" [R.35]. The main thing you need to know is that any file beginning with hrdbh is some format or edition of this book.

Finally, all you have to do is click on the format you want to download.

R.4. You don't have the eBook I'm looking for. Can you help me find it?

Sorry, no. We can suggest (see below) some other places to look for publicly accessible books on the Net, but we can't do the search for you.

R.5. Where else can I go to get eBooks?

The On-Line Books Page <http://onlinebooks.library.upenn.edu/> and the Internet Public Library at <http://www.ipl.org/> are two sites that specialize in creating a list of all books on-line from any source. Searching them is a good place to start.

If you're looking for commercial books, like current textbooks or bestsellers, you're not likely to find them here, since recent books are not in the public domain. For these, you should look for commercial booksellers on the Net--any search engine will direct you to some if you enter search terms like "shop ebook".

R.6. I see some eBooks in several places on the Net. Do different people really re-create the same eBooks?

It does happen, but mostly by accident. Anyone experienced in eBook creation will first search the usual places to see whether anyone else has already transcribed the book they're interested in. If it has been transcribed, they will not duplicate the effort.

Etexts that are in the public domain very often float around the Net for years--stored in a gopher server here, posted to Usenet there, held on someone's local computer for a year or two and then reformatted as HTML and uploaded to a web site somewhere else. And this is good, because we want texts to be copied as widely as possible.

Public domain eBooks are fair game for anyone to copy, correct, mark up, package and post: that's what being in the public domain means.

Project Gutenberg eBooks are often quickly copied and reformatted, and

posted on other sites like Blackmask at <http://www.blackmask.com>.

If you find an eBook in many different places, the odds are good that it came from one original source, and was copied around.

It does sometimes happen that people duplicate the transcription of books already made into text. Sometimes it's because they didn't find the version already made. Sometimes they have a different edition, and want to transcribe that. Mostly, though, we all try not to do more work than we have to.

About Using the Web Site:

R.7. Why couldn't I reach your site? (or: Why is your site slow?)

This isn't common, but it happens. Project Gutenberg is a very busy site, probably one of the busiest non-commercial sites on the Web, and sometimes the amount of traffic causes a slowdown.

There may also be a bottleneck somewhere else between you and the site. If at first you don't succeed, _don't tell us_, just try, try again. The correct address is either:

http://promo.net/pg/
or
http://www.gutenberg.net/

R.8. I get an error when I try to download a book.

We do not keep e-text files on this site. Instead, many FTP sites throughout the world hold the whole Project Gutenberg archive of texts. An FTP site is just a computer on the Internet that specializes in holding files for download and sending them to people on request. You can find a list of FTP sites that hold Gutenberg texts at <http://www.gutenberg.net/list.html>.

When you're searching or browsing for titles and authors, you're on this Project Gutenberg site, but when you click on the book to download it, you are connected to an FTP site. At the time you click on the filename, your browser contacts an FTP site and tries to download the file from there. If you get an error, it could be because the FTP site is busy, or because there's a network traffic bottleneck between you and that FTP site, or because the text you're looking for is missing from that FTP site.

Usually, the easiest solution is to choose another FTP site to

download your text from. Go to the Search page, choose a different FTP site, and search again for your text.

Tip: You should always try to choose the FTP site closest to you. Not only are you helping to minimize Net traffic by choosing a nearby site, but your file will download faster!

If all else fails, note the year and the filename of the book you want, choose an FTP site from this list and click on one of them. Then browse your way through the listings to the file you want.

For example, if you find "Lady Susan" by Jane Austen, you will see that it was published by Gutenberg in 1997, and its filename is lsusn10.txt, so browse to one of the FTP sites, choose the directory called etext97 and click (or right-click and Save, depending on your browser) on the file lsusn10.txt.

R.9. I searched for a book I know is in Project Gutenberg, but got no results.

First go to the Advanced Search page. Sometimes you may miss in searching because of alternative spellings, so try searching separately using just one word in Author or Title. Read the Search Tips.

If that fails, you can Browse through the site catalog. Let's say you're looking for "The Wandering Jew" by Eugene Sue.

Go to the PG Home page: <http://www.gutenberg.net/>

Once on this page, click on: "Browse" in "Browse by Author or Title"

You are then brought to a new page, asking you to select an "FTP site". Further details on how and why to choose an "FTP Site" are available on this page.

Select an FTP Site from the Selection List available at the bottom of the page, then click on "Select".

You get a new page, Click on "S", initial for "Sue, Eugene"

You should now see a list of all of the Authors whose Last name starts with "S". Scroll down till you find the direct links to the Sue, Eugene works.

Click on the work you are interested to, then click on the file link found on the page you were brought to, Etext Card ID -3987- when selecting the work, as immediately above.

On this page, above the teaser, there are two working links:

DOWNLOAD:

- es12v10.txt - 2.95 MB
- es12v10.zip - 1.10 MB

Click on the link of your choice in order to get the book.

If you can’t find your text either way, the book has not been cataloged. The site catalog always lags behind the postings, since we need to collect extra information about the book and the author before it goes into the full catalog. If you know that the book has been posted recently, and maybe hasn’t made it into the catalog yet, read the FAQ "How can I download a PG text that hasn’t been cataloged yet?"

If even this doesn’t help, don’t despair! We don’t have it, but it may be elsewhere on the Web. Go to the major search engines and try there. You can also try looking in the Book Search section of The On-Line Books Page <http://onlinebooks.library.upenn.edu/> or the Internet Public Library <http://www.ipl.org>, and if you have no luck with that, you might be able to find it listed as being In Progress somewhere on their Books In Progress and Requested page at <http://onlinebooks.library.upenn.edu/in-progress.html>.

R.10. Can I copy your website, or your website materials?

No.

Keeping the PG site updated with the latest e-text releases is an ongoing job, and our experience is that people, however well-intentioned, do not keep copies up to date. We want there to be one clear source for people seeking the latest Project Gutenberg information, and we think that having a lot of out-of-date copies and partial copies scattered around the net would be a bad thing.

We welcome mirrors and copies of our e-texts, in new FTP sites [R.14], but the main web site itself is copyrighted and may not be copied.

R.11. Your site doesn’t look right in my browser.
I clicked on a button, and nothing happened.

We take a lot of trouble to ensure that our website uses only valid, standard HTML, and we’re not even slightly tempted to use glitzy features that look good in one browser but don’t work in another, so we can promise you that our site is not the problem.

The site uses Cascading Style Sheets (CSS), a W3C standard since 1996. Some older browsers have a buggy implementation of CSS, and this can cause some things to appear off-kilter. If your browser is even older, or doesn’t know about CSS at all (as in the case of Lynx, for example) it should have no problem.

If you actually clicked on a button, like the Search button or the Post button on the Volunteers' Web Board page, and nothing happened, you might be behind a proxy or web filter that doesn't like you making POST requests. If you have a web filter switched on, turn it off, reload the page and try again.

R.12. What does that thing about "Select FTP Site" mean?

Our texts are not actually held on the website. The website just holds an index; the files themselves are held on many sites throughout the world, called FTP sites. When you have found the book you're looking for, and you make that final click to get it, you're not actually talking to our website any more--you are transferred to the FTP site you selected. Some FTP sites are near you; some are far away. Some may be faster than others, even if they are about the same distance; some may have temporary technical problems.

You should usually select the FTP site nearest you. If you find you're having problems with that one, you can select another.

R.13. What exactly is an FTP site anyway?

FTP stands for File Transfer Protocol, one of the oldest and most reliable protocols of the internet. This is the method by which a file can be copied from one computer to another.

An FTP site, or FTP server, is a computer that holds files that people can upload and download. In the case of PG, the Posting Team upload our texts when they're ready to two main FTP servers, <ftp://ftp.ibiblio.org> and <ftp://ftp.archive.org>, which serve as our master copies.

Other FTP sites around the world automatically download the files from these master sites, so they have a full set of PG publications for you to download. Because they only check for updates and new files at intervals, some FTP sites may be a day or two behind. Some FTP sites don't have space available for everything, so they may hold only the zipped versions of the files. But most FTP sites will have the entire PG collection. These are called FTP "mirrors", since they are a copy of the original.

Many FTP sites exist that offer a full PG mirror but are not on our FTP sites list. Commonly, these are in schools, where they serve the local students, but don't have enough bandwidth to offer downloads to worldwide users.

R.14. Can I become an FTP mirror?

Yes! We're always looking for more FTP mirrors.

If you manage an FTP site with a few GB of space, please check our Contact Information page <http://www.gutenberg.net/contactinfo.html> and contact the appropriate person, who will make the arrangements for you. If space is a problem, you can consider holding only zipped copies of the texts. We can move you up or down the FTP site list as you want more or less traffic.

R.15. Can I make a private FTP mirror for my school, library or organization?

Yes.

We like all FTP mirrors to be open to as many people as possible, but we know that not all schools have the resources to be a public mirror, so we welcome all mirrors.

And anyway, you don't even have to ask, because we don't control what happens to our texts once we post them!

R.16. When I clicked on the file I want, nothing happened.

When you select a file for download, your request goes to the FTP site you selected, not to our website. If the FTP site you selected is having problems, or if there is the Net version of a traffic jam between you and it, you may have problems downloading.

Select a different FTP site [R.12] and try again.

R.17. How many texts are downloaded through the web site?

We don't really do statistics, but in one particular month for which we did, we had a figure of about 800,000 searches completed. Since the final request for download goes to the FTP site selected and not to our website, we can't confirm that all of these were actually downloaded, but we expect that most people who have gone all the way through the search will finish the job.

In another month, we had about 1,000,000 downloads of files from ftp.ibiblio.org, our main FTP site. This does not count downloads from other FTP sites, of course. Why are there more downloads than searches? Because people who are already familiar with getting PG texts can skip the website search and download straight from the FTP sites.

R.18. What are the most popular books?

We very rarely do statistics, but on one occasion in late 1999 when we did, we found the top author searches to be:

1 shakespeare
2 poe
3 doyle
4 melville
5 dante
6 joyce
7 shaw
8 christie
9 conrad
10 porter
11 verne
12 hemingway
13 darwin
14 miller
15 woolf
16 zola
17 king
18 eliot
19 churchill
20 smith
21 twain

and the top individual books searched for to the point of downloading were:

1. Lady Susan, by Jane Austen
2. 1st PG Collection of Edgar Allan Poe
3. The Adventures of Sherlock Holmes, by Arthur Conan Doyle
4. Moby Dick, by Herman Melville
5. A Christmas Carol, by Dickens
6. The King James Bible
7. Twelve Stories and a Dream, by H.G. Wells
8. Stories by Modern American Authors
9. Lock and Key Library, Magic & Real Detectives
10. [Hans Christian] Andersen's Fairy Tales
11. The Legend of Sleepy Hollow, Washington Irving

These numbers vary a lot. When a movie based on a classic is released, downloads of that eBook go through the roof!

About Downloading and Using Project Gutenberg eBooks:

R.19. Should I download a ZIP or a TXT file?

If you know how to unzip a file, then downloading the zip is faster. For some non-text eBooks that contain multiple files, like HTML with included images, only a zip file may be available. For some other formats, like MP3 or MPEG, there may not be a zipped version available because the native format of the file is already compressed enough that zipping it doesn't save much.

R.20. I've got a ZIP file. What do I do with it?

Unzip it.

If you want a free program, you could try the open source Info-Zip software available at <http://www.ctan.org/tex-archive/tools/zip/info-zip/> for Mac, MS-DOS, Unix, Windows and just about everything else you might have.

If you want a commercial program, PKZIP from <http://www.pkware.com> and WinZip from <http://www.winzip.com> are among many popular shareware utilities that allow you to unzip files.

Mac-users using Stuffit Expander may like to set a preference (File / Preferences / Cross Platform) to "Convert text files to Macintosh format . . . When a file is known to contain text". This gets rid of strange characters (linefeeds), which are not wanted on a Mac, at the beginnings of lines. MacZip is another free program for Macs. Mac users can also try ZipIt or other shareware programs available from the Info-Mac archives, e.g. from <ftp://mirrors.aol.com/pub/info-mac/_Compress_&_Translate/>.

R.21. I tried to unzip my file, but it said the file was corrupt, or damaged.

The chances are that it didn't download correctly. Try downloading it again. If you don't succeed the second time, try downloading the unzipped version.

R.22. I see gibberish onscreen when I click on a book.

To save download time, our etexts are stored in zipped form as well as text form. Zipped files are smaller, and take less time to transfer to your computer, but you need a program to unzip them. If you try to view a zipped file directly, it looks like gibberish.

You can recognize zipped files easily because their filenames end in .zip.

If this happens, either make sure you're asking your browser to Save the file rather than display it (often, you right-click the file and choose Save) or else click on the version of the file that ends in .txt instead of .zip. You don't need a zip program to view .txt files.

Looking at a zip rather than a text file is by far the most common reason for this problem, but there are some others. If you're quite sure that you're not looking at a zip file, then it could be that the file you downloaded is in a character set that your viewer doesn't recognize, like Big-5 [V.78] for Chinese texts, or Unicode [V.77]. If this is the case, you will have to find a viewer that works on your computer for the specified character set. We may also have an ASCII version of the same text available for you--we do try to have ASCII versions for everything [G.17], but some languages, like Chinese, just cannot be sensibly expressed in ASCII.

If you can see _most_ of the characters, enough to be able to make out the text, but there are regular gibberish characters, black squares, empty boxes or obviously missing characters scattered about through words, then you are probably looking at an "8-bit" text [V.79], with accented characters, and your viewer doesn't handle the character set. See the FAQ "I can read the text file, but a few characters appear as black squares, or gibberish" [R.31].

If there are a very few gibberish characters, black squares or obviously missing characters in the text, then it's likely that this was intended to be a 7-bit text, but a few 8-bit characters like the British pound symbol or accented letters slipped through.

R.23. Can I download and read your books?

Yes. That's what Project Gutenberg is all about--making texts available free to everyone!

R.24. What am I allowed to do with the books I download?

Most Project Gutenberg e-texts are in the public domain. You can do anything you like with these--you can re-post them on your site, print them, distribute them, translate them to other languages, convert them to other formats, or redistribute them in unchanged form. However, if you distribute versions under the Project Gutenberg trademark, we do impose some conditions, which are explained in the header and/or footer in each text.

Some Project Gutenberg e-texts have copyright restrictions. You can

still download and read these, but you may not be allowed to reproduce, modify or distribute them. When browsing or searching on the site, you will see these copyright-restricted texts indicated in the listings. For fuller information about them, download the e-text and read the header or footer of the file, which will spell out the conditions in detail.

R.25. Does Project Gutenberg know who downloads their books?

No, and we don't want to!

Like any Internet transfer, our sites have to know the IP addresses that contact them; without that, no communication is possible. But we do not trace, hold or examine them beyond what is necessary to deal with any problems or maintain logs or statistics. We never identify IP addresses with people.

Further, we encourage people, sites, schools around the world to mirror, or copy, our texts to their sites. Once that happens, we have no control over them, and we never have any idea who or even how many people access them after that.

Even further, we encourage people to distribute the texts on disks, CDs, paper, and any other storage format they can find. We encourage them to convert the texts to other formats, and share them.

For most people reading this, anonymity is probably not an issue, but you may live in a place or time where reading Paine, or Voltaire, or the Bible, or the Koran, is considered suspicious or even subversive. We don't know who you are, and what we don't know, we can't tell.

Currently (mid-2002), by means of DRM (Digital Rights/Restrictions Management) many commercial publishers can make a list of exactly who is reading which of their eBooks. We _don't_ know, and we don't _want_ to know.

R.26. I've found some obvious typos in a Project Gutenberg text.
How should I report them?

The first thing to remember is that the people who actually make the corrections you suggest are very experienced, and are used to seeing lots of different types of errata reports. So the exact format of your report isn't really very important--just get the report to us in any clear form that we can understand.

Beyond that, here are some tips to avoid misunderstandings.

It's always helpful if you report the full title, etext number, year and filename of the text you are correcting. We have multiple editions

and versions of some texts, like Homer's "Odyssey", and unless you tell us exactly what text you mean, we may have to spend some time searching and guessing.

Especially, _please_ check and report the exact filename of the text. It is amazingly common for people to report problems with abcde10.txt, when abcde11.txt is already posted, and has these and other errors already fixed.

When there are only a few errors, it's usually easiest to cut and paste the line or lines where the error is into your e-mail, with your comment.

It can also be useful to give the line number of the place where the error is, and some people who check texts regularly do this. If this seems natural to you, do it; if it doesn't, don't.

An ideal report for a typical errata list might look like:

Title: The Odyssey, by Homer
Translated by Butcher & Lang
April, 1999 [Etext #1728]
File: dyssy08.txt

Line 884:
back Telemachus, who bas now resided there for a month.
"bas" should be "has"

Line 1491:
Ithaca yet stands. But I wouldask thee, friend, concerning
"would" and "ask" are run together here

Line 1563:
in his father's seat and the elders gave place to him
This is the end of a paragraph, and needs a period at end.

Line 15346-7:
'Hearken to me now, ye men of Ithaca, to the
will say. Through your own cowardice, my friends, have
I think there is something missing between "the" and "will"

But the following would get the job done as well:

In Homer's Odyssey, translated by Butcher and Lang, from /etext99, file dyssy08.txt, I found the following errors:

Telemachus, who bas now resided
change "bas" to "has"

But I wouldask thee,
"would ask" run together

and the elders gave place to him
needs period

ye men of Ithaca, to the
will say.
line missing between "the" and "will"?

Where there are more than a few changes, it may be easiest all round just to submit a corrected version of the file. However, if you do this, please do not re-wrap the paragraphs unless it is really necessary; we need to check your suggestions before reposting, and if the file is very different, it is difficult and time-consuming for us to find your real changes among all of the changes in the lines.

R.27. I've found some obvious typos in a Project Gutenberg text. Who should I report them to?

The Posting Team, who post the books, also make the corrections, and ultimately, the corrections need to go to them.

Many producers put their e-mail addresses in their texts, specifically so that readers can contact them when errors are found. If you see that in your text, you should try to contact the producer first. This is especially true if the corrections aren't obvious, as in the case of missing words. The producer is likely to have the original book, and will probably be able to confirm your corrections without visiting a library. If the book needs the corrections, the producer can then notify the Posting Team.

If you get no response from the producer, or if there is no e-mail address listed, or if the corrections are small and obvious, you can send them to any or all of the Posting Team directly.

R.28. I've reported some typos. What will happen next?

This varies wildly. Sometimes, you may just get a response e-mail in a day or three saying thanks, and that we've fixed the typo. This is normal when you've just reported one or a few obvious typos.

Where there is some text missing, or the changes you suggest are otherwise not obvious, we may have to find someone with an eligible copy of the book to confirm the changes, and that might take time. Normally, you will get an e-mail explaining that within a week.

Sometimes, even though you've noticed only one or two small typos, one of the Posting Team who was looking at it may find many more, and decide that the whole text needs to be re-proofed. This may also take time.

If the text needs a lot of changes, we may post a new EDITION [R.35] of it, with a new filename: e.g. abcde10.txt may become abcde11.txt. In this case, you will receive a copy of the e-mail sent to the posted list announcing the new file. Our current rule of thumb is that we create a new edition when we make twelve significant changes, but we judge each on a case-by-case basis, and especially will usually not make a new edition if the original was posted recently.

R.29. I've got the text file, and I can read it, but it seems to be double-spaced or it has control characters like ^J or ^M at the end of every line.

This is most often seen on Mac or Linux. If you want to dig into why this effect happens, see the FAQ "Why use a CR/LF at end of line?" [V.85].

Perhaps viewing it in a different editor or viewer will help, but it's usually easiest just to globally replace all of the control characters (if you see them) with nothing, or to replace all double line-ends with single line-ends.

R.30. When I print out the text file, each line runs over the edge of the page and looks bad.

If you have a file ending in .txt from Project Gutenberg, it is usually formatted with about 70 characters per line, and with a Carriage Return/Line Feed pair (also known as a "Hard Return" or a "Paragraph Mark") at the end of every line.

This is the most widely accepted format for text files, but it's not ideal on all computers and all programs. 70 characters per line means that if you are using an unusually large or small font to print it, lines may wrap around or not reach across the page. The hard return means that on some systems, the lines may appear double-spaced.

Unfortunately, we can't advise you how best to format texts on all systems, mostly because we don't know every system! Here are a couple of tips you might try:

If your font is too big or too small, try setting the font to Courier size 10 or Times size 12. It may not be ideal, but it mostly works.

In a word processor, you may be able to remove the Hard Returns, but beware! if you remove too many, the whole text will become one paragraph. One common formula for removing the HRs goes like this:

1. First, all paragraphs and separate lines should be separated by two HRs, so that you can see one blank line between them. Where they aren't, as in the case of a table of contents or lines of verse, add the extra HRs to make them so.
2. Replace All occurrences of two HRs with some nonsense character

or string that doesn't exist in the text, like ~$~.
3. Replace All remaining HRs with a space.
4. Replace your inserted string ~$~ with one HR.

R.31. I can read the text file, but a few characters appear as black squares, or gibberish.

The text is using some character set that your editor or viewer isn't. For example, the text is using ISO-8859-1, and your viewer is using Codepage 850--or vice versa. You can see the plain ASCII characters, but non-ASCII characters like accented letters display as nonsense.

Look at the top of the file for a clue to the character set encoding: if it's there, it may help you to find which editor, or font, or viewer you should be using.

R.32. Can I get a handheld device for reading PG texts? Which device should I get?

To read eBooks on a handheld, you need three things: the eBook content itself (which you can get from PG and other sites), a device (which I will sometimes call a PDA, even though technically, the RocketBook isn't a PDA) and the reader software that runs on the PDA.

In mid-2002, there are three main families of handheld devices people use for reading eBooks: Palms, Pocket PCs and RocketBooks (or their successor, REB1100s). In general, it is possible to use any of these in combination with any common type of personal computer.

Palms are very common, especially when you count not just the Palm <http://www.palm.com> itself, but PalmOS-based devices from other manufacturers, like:

the Franklin eBookman <http://www.franklin.com/ebookman/>,
the Handspring Visor <http://www.handspring.com>.
the Sony Clie <http://www.sony.com> and

Because of the number of makers of PalmOS-based devices, you can buy them with lots of combinations of features--color screen, audio, different memory sizes. Of course, Palms have other applications besides eBook reading. Palms are the smallest and most portable of the three classes, and tend to have the best battery life for travelling, but they also have the smallest screen. Just about all reader software will run on Palms, except the Microsoft Reader, which runs only on Pocket PCs, but you don't need the Microsoft Reader for Project Gutenberg eBooks.

In Pocket PCs, the Compaq iPaq is by far the most common in mid-2002. More expensive and bulkier than a Palm, it does have a bigger screen.

Like the Palms, it can perform many functions besides reading eBooks. Only Pocket PCs can support the Microsoft Reader, but this is not necessary for reading Project Gutenberg eBooks. <http://www.compaq.com>

The RocketBook, and its successor the Gemstar REB1100, <http://www.gemstartvguide.com> are quite different from the others. These were built specifically for reading eBooks, and do not have additional functions. They are not, technically, PDAs. Their screens are bigger, and excellent for reading, but do not offer color. They also don't offer a choice of readers--the dedicated reader is built-in to the device. Both of them require the eBooks you load to be formatted for their reader, and files made for them usually have the extension .rb for RocketBook. The REB1100 does not come with the RocketLibrarian, which is the program you run on your PC to turn an etext into a RocketBook file, but people are still making .rb files, and the RocketLibrarian is still available and popular among an enthusiastic group of Rocket users. (The REB1200 is entirely different from the REB1100, and, as far as we know, PG etexts cannot easily be transferred to it.)

In summary, the Rocket/REB1100 is a dedicated reader, with a good screen, but limited to what it does.

Palms are relatively cheap and common, with a wide range of options, and the capacity to function as PDAs as well. They can run all common readers except the Microsoft one.

The iPaq <http://www.compaq.com> has a good color screen, but is bulkier than a Palm, and can run lots of readers, including the Microsoft one, but not all Palm readers are available for Pocket PC. Like Palms, the iPaq can do other jobs besides displaying eBooks.

Different people make different choices among these for reading their eBooks, and they all work well; it's a matter of personal taste.

R.33. How can I read a PG eBook on my PDA (Palm, iPaq, Rocket . . .)

To read a book on your PDA, you need to get the file into a format that your reader software understands. Each PDA reader program will work only with a specific format of file. Some will read several formats, but, in general, it's a jungle of competing options.

Unless you use a Rocket or REB1100, you will need to install at least one reader program, and many veteran readers install two or three to deal with different formats. There are many of them available. In a recent internal poll of Gutenberg volunteers who use PDAs,

C Spot Run <http://www.32768.com/bill/palmos/cspotrun/index.html>,
Mobipocket <http://www.mobipocket.com>,
PalmReader <http://www.peanutpress.com/>
Plucker <http://www.plkr.org>

were our favored choices for reader programs.

Further, the process may be different depending on which reader software you're using. Each format that a reader understands has one or more converter programs that run on your PC, and turn the plain text file into that format. So in general, you have to:

1. Download the PG text
2. Edit the text for the layout the converter wants (often HTML).
3. Use the converter to create a file of the format the reader wants.
4. Transfer the converted file to your PDA.

If all this sounds too complicated, remember that many people take and convert PG texts into many formats, and offer them for download from their sites. Of course, there is no guarantee that someone will have converted the particular eBook you want, but there are lots of options. Try Blackmask <http://www.blackmask.com>, which lists thousands of texts already converted for Mobipocket, iSilo, RocketBook and the Microsoft Reader.

There are many other sites that serve pre-converted PG texts.

MemoWare <http://www.memoware.com> is also a useful resource for converted eBooks, and has lots of information, including an excellent map of the readers and formats jungle at <http://www.memoware.com/mw.cgi/?screen=help_format>

Tecriture <http://www.tecriture.net> hosts a service that downloads and converts PG texts on the fly, and delivers them straight to you.

If you're "rolling your own", you'll probably need to convert our plain texts to HTML at some point, because a lot of converters require HTML as input, and this is a common theme in readers' explanations of how they get texts onto their PDAs. Don't panic! You don't have to be a HTML wizard to do this--in fact, you don't need to know anything about HTML at all! Usually, it's just a matter of removing some line ends and Saving As HTML. You won't get a lot of fancy markup, or images out of thin air, but you will get the book.

One of the main things you usually have to do in making HTML is unwrap the lines. If you're making your HTML manually, this is usually done by replacing two paragraph marks with some nonsense marker like @@Z@@, replacing all single paragraph marks with a space, and replacing the nonsense marker with a paragraph mark. After unwrapping, the text can just be Saved As HTML.

There are some applications that specifically assist with auto-converting text into HTML:

GutenMark <http://www.sandroid.com/GutenMark> was specifically written for the purpose, and knows enough about PG conventions to do a very good job.

InterParse <http://www.interparse.com> is a Windows-based generic text parser that is very easy and intuitive to use.

The World Wide Web Consortium lists some other options at <http://www.w3.org/Tools/Misc_filters.html>

If you're using a RocketBook or REB1100, you don't have either the choices or the confusion to deal with. One of our volunteers who uses a RocketBook offered this recipe for getting a PG text onto a RocketBook:

On converting to Rocket:

1. Download text file.
2. Using your utility for showing formatting, enter your word processing program's edit mode.
3. Replace all double paragraph marks with some nonsense sequence that can't possibly actually be there, such as @@Z@@.
4. Replace all single paragraph marks with one single space (enter).
5. Replace your nonsense sequence with one paragraph mark.
6. Convert all your double spaces to single spaces. Repeat this until you get "0" for how many replacements were made.
7. Save in HTML.
8. Go into your Rocket Librarian. Use "import file using Rocket Librarian." Go and pick up the file, which will be automatically converted to .rb in this process.

This sounds long, but it usually takes me under three minutes except for a very long text. I've never taken longer than five minutes. You can just go in and pick up the text file with Rocket Librarian, but what you get onscreen doing this looks very odd. Steps 2-7 are not essential, and if I'm in a hurry to read something once I might skip them, but if it's something I know I want to keep I use them.

This formula is not ideal for poetry or blank verse--if you want to keep the lines unwrapped, you should avoid removing the paragraph marks.

Another volunteer, who reads on Mobipocket <http://www.mobipocket.com> offered this suggestion:

I use the MobiPocket Publisher, available free from www.mobipocket.com. It wants to take a HTML file as input, so the first thing I have to do is convert my PG text to HTML.

I usually do this by running GutenMark, available at <http://www.sandroid.org/GutenMark>. I can also do it in Microsoft Word using the following sequence:

Edit / Replace / Special and choose Paragraph Mark twice (or, from replace, you can type in ^p^p to get two Paragraph Marks) and replace

with @@@@. Replace All. This saves off real paragraph ends by marking them with a nonsense sequence.

Now Replace _one_ Paragraph Mark (^p) with a space. Replace All. This removes the line-ends.

Finally, replace @@@@ with _one_ Paragraph Mark. Replace All. This brings back the Paragraph Ends.

Now I can Save As HTML.

GutenMark does a better job of converting to HTML than my simple Word formula, since it recognizes standard PG features, and sometimes Mobipocket doesn't like the HTML produced from Word--it complains of a missing file, or doesn't recognize quotation marks.

Having got my HTML file, I open Mobipocket Publisher, choose "Project Gutenberg", Add the File I created, and just Publish it to MobiPocket .PRC format. Then I pick it up on my iPaq the next time I sync. The whole process takes two or three minutes, and the results, since I discovered GutenMark, are good.

I recently came across InterParse 4 at <http://www.interparse.com>. It doesn't have the built-in knowledge of GutenMark, so the results aren't as good, but it's really easy to use, and you can see the effect of your changes onscreen as you do it. For most PG books, all you have to do is just Open the text file and choose Options / Remove all CRLFs (Except at Paragraph End), then Convert / Text to HTML and Save As the HTML filename you want. Quick and painless.

About the Files:

R.34. What types of files are there, and how do I read them?

The vast majority of our files are plain text. You can read these with any editor or text viewer or browser. Some are HTML. You can read these with any browser.

For a full listing of other file types as of mid-2002, and how to read them, please see the Formats FAQ [F.2].

R.35. What do the filenames of the texts mean?

PG files are named for the text, the edition, and the format type.

As of February, 2002, all PG files are named in "8.3" format--that is, up to eight characters, a dot, and three more characters.

The first five characters in the filename are simply a unique name for that text, for example, "Ulysses" by Joyce begins with "ulyss".

If the text has been posted as both a 7-bit and 8-bit text, then the first character of the filename will be a 7 or an 8, to indicate that. For example, we have both 7crmp10 and 8crmp10 for Dostoevsky's Crime and Punishment.

The 6th and 7th characters of the name are the edition number--01 through 99. We normally start at edition 10 (1.0); numbers lower than that indicate that we think the text needs some more work; numbers higher than that mean that someone has corrected the original edition 10.

The 8th character of the filename, if it exists, indicates either the version or the format of the file. When we get a different version of the text based on a different source, we give it an a, b, c, as for example if the text is from a different translation. Where we have posted a text in a different format, we also add an eighth character--"h" for HTML, "x" for XML, "r" for RTF, "t" for TeX, "u" for Unicode are established formats. There have been some experimental postings with "l" for LIT, and "p" for either PRC or PDB.

So, for example:

7crmp10 is our first edition of Crime and Punishment in plain ASCII
8sidd10 is our first edition of Siddhartha, as an 8-bit text
dyssy10b is our first edition of our third translation of Homer's Odyssey, in plain ASCII
jsbys11 is our second edition of Jo's Boys, in plain ASCII
vbgle10h is our HTML format of our first edition of Darwin's Voyage of the Beagle
7ldv110 is our 7-bit ASCII version of the first volume of the Notebooks of Leonardo da Vinci

To make it worse, we don't always stick to these rules, for example:

1ddc810 is our first edition of the first book of Dante's Divina Commedia in Italian, as an 8-bit text
80day10 is our first edition of Verne's Around the World in 80 days, in plain 7-bit ASCII in English.
emma10 is our first edition of Jane Austen's "Emma"--with a 4-character basename instead of 5.

Some series have special, non-standard names. Shakespeare is named with a digit representing the overall source (First Folio, etc), then "ws", then a series number, so for example 0ws2610, 1ws2610 and 2ws2610 are all versions of "Hamlet". The Tom Swift series is named with a two-digit prefix denoting the series number, then "tom", so for

example 01tom10 is "Tom Swift and his Motor-Cycle".

And what should we do with a text from a different source that is formatted as HTML? For example, if dyssy10b is the name of the third translation, what should the HTML version be named? dyssy10bh is obvious, but it uses 9 characters.

The problem, of course, is that we are trying to fit a lot of information into an 8-character filename, and as the collection grows, and the number of formats and versions increases, we come across more pressure on filenames, so while the filename is a good guide to the contents, it's not definitive.

R.36. What is the difference within PG between an "edition" and a "version"?

We give the name "edition" to a corrected file made from an existing PG text. For example, if someone points out some typos in our file of "War and Peace", we will fix them, and, if enough are found to warrant a "new edition", then instead of just replacing the file wrnpc10.txt, we may make a new file wrnpc11.txt, and leave the original alone. A new edition is always filed under the same year and etext number as the original--it's just an update.

We give the name "version" to a completely independent e-text made from the same original book, but a different source. For example, Homer's Odyssey was translated by many different people, but they all worked from the same book. The translations by Lang, Butler, Pope and Chapman are very different, but they all come from the same root.

Thus, these are all "versions" of Homer's Odyssey. We give them all the same basename--dyssy--and each gets a new number, but we keep the original basename, and add a letter to the filename to indicate that they are "versions" of the same original book:

dyssy10.txt	Butler's Translation
dyssy10a.txt	Butcher & Lang's Translation
dyssy10b.txt	Pope's Translation

The differences don't have to be as extreme as this for us to create a new version. "Clotelle"/"Clotel", for example, was a book published multiple times in English by William Wells Brown, and each time, he changed the text. We preserve three different texts of the same book as different versions: clotl10 clotl10a and clotl10b.

R.37. What is the difference between an "etext" and an "eBook"?

If there is any, it seems to be in the eye of the Marketing Department! Michael Hart started the whole thing, and coined the word "Etext". The term "eBook" is gaining in popularity, even for texts

that are not full books, so we've started using that more now.

R.38. What are the "Etext/Ebook numbers" on the texts?

These are simply a series of numbers. We give one to each etext as it is posted, so the earliest etexts have low numbers and later etexts have higher numbers. Etext number 1 is the Declaration of Independence, the first text that Michael Hart typed in to the mainframe that he was using in 1971.

A few numbers are reserved for books that we hope to have in the PG archive someday; for example, 1984 is reserved for Orwell's classic.

When we improve an text by making some corrections, we call it a new EDITION, and it keeps the same etext number, but when we post a different VERSION of the same text, from a different paper book--like different translations of Homer's Odyssey--each new version gets a new etext number.

R.39. What do the month and year on the text mean?

Project Gutenberg sets a production target for itself. The idea is that we try to produce X texts in a month, and we date the texts according to what month of our schedule they appear in. For example, if our target for September 2000 was 50 texts, and we actually produced 55, then the last five would be dated October 2000, and we'd get a head-start on the month. At the time of writing, in July 2002, that target is the publication of 200 books per month. However, our actual production has far outpaced our targets, with the result that the "head-start" has accumulated so much that we are currently releasing books scheduled for March, 2004!

The fact that we're so far ahead of schedule makes this quite confusing for newcomers. If it bothers you, just don't think about it! But at least it's better than being _behind_ schedule. We didn't always produce so many books. In the September 1994 newsletter, Michael Hart wrote:

> As always, I am terrified of the prospect of
> doubling our output to 16 Etexts per month for
> next year, we really need your help!!!

That was when the Project's target was 8 Etexts per month. Today, our target is heading towards 8 eBooks per _day_!

C.1. What is copyright?

Copyright is a limited monopoly granted to the author of a work. It gives the author the exclusive right, among other things, to make copies of the work, hence the name.

C.2. Does copyright differ from country to country? From state to state?

Copyright laws are constantly changing all over the world. Each country has its own copyright laws, some within the framework of international treaties, some not. Within the U.S., copyright laws are federal, and do not vary from state to state.

C.3. What are the copyright laws outside the U.S.?

Sorry, we can't advise on copyright law outside the U.S. We can point you to resources like <http://onlinebooks.library.upenn.edu/okbooks.html> which tries to summarize the various copyright regimes, but we can't guarantee that these are accurate. Even when they are accurate, it is very hard to express some of the subtleties of copyright law in a summary--for example, the question of what constitutes "publication" for copyright purposes is sometimes unclear.

C.4. Why does Project Gutenberg advise only on U.S. copyright issues?

The Project Gutenberg Literary Archive Foundation is registered in the U.S. as a 501(c)(3) organization, and our two posting servers are situated in the U.S., so we are subject to U.S. copyright law, and only to U.S. copyright law.

Because copyright laws are so tangled and different between countries, not only in the broad sweep but also in the detail, and because Project Gutenberg is subject only to U.S. copyright law, we just don't have the expertise, time or resources to research and advise on the law in other countries.

C.5. I don't live in the U.S. Do these rules apply to me?

Your country's copyright laws are different from those in the U.S., and understanding and dealing with them is up to you. If you have a book that is in the public domain in your country, but not in the U.S., it is perfectly legal for you to publish it personally there, but we can't.

Similarly, it may be legal for us to publish it here, but not for you to publish it, or perhaps even copy it, where you are.

There are organizations in other countries operating in more liberal copyright regimes that may be able to publish texts that we cannot. For example, Project Gutenberg of Australia at <http://www.gutenberg.org.au> can accept many works not eligible in the U.S.

C.6. What is the public domain?

The public domain is the set of cultural works that are free of copyright, and belong to everyone equally.

C.7. What can I do with a text that is in the public domain?

Anything you want! You can copy it, publish it, change its format, distribute it for free or for money. You can translate it to other languages (and claim a copyright on your translation), write a play based on it (if it's a novel), or a novelization (if it's a play). You can take one of the characters from the novel and write a comic strip about him or her, or write a screenplay and sell that to make a movie.

You don't need to ask permission from anyone to do any of this. When a text is in the public domain, it belongs as much to you as to anyone.

(However, when some character or part of the work is also trademarked, as in the case of Tarzan, it may not be possible to release new works with that trademark, since trademark does not expire in the same way as copyright. If you propose to base new works on public domain material, you should investigate possible trademark issues first.)

C.8. How does a book enter the public domain?

A book, or other copyrightable work, enters the public domain when its copyright lapses or when the copyright owner releases it to the public domain.

U.S. Government documents can never be copyrighted in the first place; they are "born" into the public domain.

There are certain other exceptional cases: for example, if a substantial number of copies were printed and distributed in the U.S. before March, 1989 without a copyright notice, and the work is of entirely American authorship, or was first published in the United States, the work is in the public domain in the U.S.

C.9. How does a copyright lapse?

Copyrights are issued for limited periods. When that period is up, the book enters the public domain.

Copyrights can lapse in other ways. Some books published without a copyright notice, for example, have fallen into the public domain.

C.10. What books are in the public domain?

Any book published anywhere before 1923 is in the public domain in the U.S. This is the rule we use most.

U.S. Government publications are in the public domain. This is the rule under which we have published, for example, presidential inauguration speeches.

Books can be released into the public domain by the owners of their copyrights.

Some books published without a copyright notice in the U.S. prior to March 1st, 1989 are in the public domain.

Some books published before 1964, and whose copyright was not renewed, are in the public domain.

If you want to rely on anything except the 1923 rule, things can get complicated, and the rules do change with time. Please refer to our Public Domain and Copyright How-To at <http://www.gutenberg.net/vol/pd.html> for more detailed information.

C.11. My book says that it's "Copyright 1894". Is it in the public domain?

Yes.

Its copyright date is 1894, which is before 1923, so its copyright has lapsed.

C.12. How can a copyright owner release a work into the public domain?

A simple written statement, which may be placed into the work as released, is sufficient. When a copyright holder places a book into the public domain and wants PG to publish it, all we need is a letter [V.70] saying that they are or were the holder of the copyright, and that they have released it into the public domain.

C.13. When is an author not the owner of a copyright on his or her works?

An author may sell, assign, license, bequeath or otherwise transfer his or her copyright to another party, such as a publisher or heir.

C.14. What does Project Gutenberg mean by "eligible"?

A book is eligible for inclusion in the archives if we can legally publish it.

We can legally publish any material that is in the public domain in the U.S. [C.10], or for which we have the permission of the copyright holder.

C.15. I have a manuscript from 1900. Is it eligible?

Maybe not.

Works that were created but not "published" before 1978 will not enter the public domain before the end of 2002. This gets complicated, and it’s not too common. If you have such a case, ask about it.

A borderline example is the classic "Seven Pillars of Wisdom" by T. E. Lawrence, which was actually printed and privately distributed, but not "published", in 1922. We haven’t been able to confirm any pre-1923 "publication" for this.

C.16. How come my paper book of Shakespeare says it’s "Copyright 1988"?

Shakespeare was published long enough ago to be indisputably in the public domain everywhere, so how can a Shakespeare text be copyrighted?

There are two possibilities:

1. The author or publisher has changed or edited the text enough to qualify as a "new edition", which gets a "new copyright".

2. The publisher has added extra material, such as an introduction, critical essays, footnotes, or an index. This extra material is new, and the publisher owns the copyright on it.

The problem with these practices is that a publisher, having added this copyrighted material, or edited the text even in a minor way, may

simply put a copyright notice on the whole book, even though the main part of it--the text itself--is in the public domain! And as time goes on, the number of original surviving books that can be proved to be in the public domain grows smaller and smaller; and meanwhile publishers are cranking out more and more editions that have copyright notices. Eventually it becomes harder and harder to prove that a particular book _is_ in the public domain, since there are few pre-1923 copies available as evidence.

Among the most important things PG does is preventing this creeping perpetuation of copyright by proving, once and for all, that a particular edition of a particular book _is_ in the public domain, so that it can never be locked up again as the private property of some publisher. We do this by filing a copy of the TP&V, the title page where the copyright notice must be placed, so that if anyone ever challenges the work's public domain status, we can point to a proven public domain copy.

C.17. What makes a "new copyright"?

1. New edition

When a text is in the public domain, anyone--from you to the world's biggest publisher--can edit it and republish the edited version. When the edits are substantial enough, the edited work is deemed a "new edition", and gets a new copyright, dating from the time the new edition was created.

How substantial must the edits be to qualify as a "new edition"? That is for a court to decide in any particular case. Changing some punctuation or Americanizing British spelling would not qualify a work for a new edition. Theorizing something about Shakespeare and rewriting lots of lines in "Hamlet" to emphasize your point _would_ make a new edition. In between those extremes is a grey area, where each new edition would have to be considered on a case-by-case basis.

A special case, that isn't quite a new edition, is when someone "marks up" a public domain text in, for example, HTML. Where this happens, the text is in the public domain, but the markup is copyrighted. We've already seen that when an editor adds footnotes to a public domain text, he owns copyright on the footnotes but not on the text: similarly, when he adds markup to the text, he owns copyright on the markup.

2. Translation

Translation is a common and justified special case of a new edition. When someone translates a public domain work from one language to another, they get a new copyright on the translation (but not on the original, of course, which stays in the public domain so that lots more people can use it.)

C.18. I have a 1990 book that I know was originally written in 1840, but the publisher is claiming a new copyright. What should I do?

From a practical point of view, there's not much you can do about it. It's a Catch-22 situation: in order to prove that the new printing should be in the public domain, you need a provably public domain copy to compare against the allegedly copyrighted edition, and if you have that, you don't need the modern edition anyway.

C.19. I have a 1990 reprint of an 1831 original. Is it eligible?

Yes, as long as we can _show_ that it is a reprint, which usually means that it has to _say_ that it's a reprint somewhere on the TP&V.

However, we need to be very careful in a case like this. Commonly, the book itself is eligible, but introductions, indexes, footnotes, glossaries, commentaries and other such extras may have been added by the modern publisher, so you should not include them except where you can prove that they are part of the reprinted material.

C.20. I have a text that I know was based on a pre-1923 book, but I don't have the title page. Can I submit it to PG?

Unfortunately, no.

What you "know" isn't proof that we could take into court if we were challenged about it in 20 years, and the whole problem of "new copyright" [C.17] makes it effectively impossible to tell for sure what is and isn't copyrighted anyway, without reliable evidence like the title page.

You need to find a matching paper edition for proof. See the FAQ "I've found an eligible text elsewhere on the Net, but it's not in the PG archives. Can I just submit it to PG?" [V.62]

C.21. How does Project Gutenberg "clear" books for copyright?

Usually, we just look at the TP&V. If it was published before 1923, or says it is a reprint of a pre-1923 edition, that's all we have to do.

In other cases, we may look up library publication data to prove, say, that a book published in the U.S. without a copyright notice was indeed published in the years when a copyright notice was required. Or we may simply see that a particular text was published by the U.S.

Government.

The bottom line is the question: if someone comes to us claiming to hold the copyright on a text, do we have proof to show that they're wrong?

Whatever proof or search we have to do, we then file it, either on paper or electronically, so that the proof will be available in 20 or 50 years' time, or whenever the challenge is made.

C.22. I want to produce a particular book. Will it be copyright cleared?

If it was published before 1923, you will have no problem with its clearance. If you're relying on one of the other rules, it may just be too much work to try and prove its public domain status.

C.23. I have some extra material (images, introduction, preface, missing chapter) that should go into an existing PG text. Do I have to copyright-clear my edition before submitting it?

Yes.

Otherwise we would have no proof that the extra material you're adding isn't copyrighted by someone. It's quite common for modern publishers to add introductions or illustrations to a public-domain novel, and we need the same standard of proof for these additions that we do for the main text.

This doesn't apply to an occasional word or two that was omitted by mistake when the text was first typed. For example, you don't need to clear another edition just to restore the words "thus perfected the" and "eliminating all" to the sentence:

> And while we Country, we were also sorts of tediums, disputable possibilities, and deadlocks from the game.

while fixing typos.

C.24. I see some Project Gutenberg eBooks that are copyrighted. What's up with that?

Authors or publishers may grant Project Gutenberg an unlimited license to republish their works. In this kind of case, the copyright holders still retain their rights, but grant permission for us to share these eBooks with the world.

These copyrighted PG publications can still be copied, but the

permissions granted are spelled out in their headers, and usually forbid anyone to republish them commercially.

C.25. What are "non-renewed" books?

Works published before 1964 needed to have their copyrights renewed in their 28th year, or they'd enter into the public domain. Some books originally published outside of the US by non-Americans are exempt from this requirement, under GATT. Some works from before 1964 were automatically renewed.

C.26. How can I get Project Gutenberg to clear a non-renewed book?

As of mid-2002, you probably can't. Because of all of the checks we need to do to ensure that the book wasn't renewed, or wasn't one of the exceptions that was automatically renewed, we just don't have the time to do it. But we're working on it. Right now, we're processing copyright renewal records with the aim of making them searchable.

Volunteers' FAQ

About the Basics:

V.1. How do I get started as a Project Gutenberg volunteer?

What you actually need to do to produce a PG text can be stated very simply:

1. Borrow or buy an eligible book.
2. Send us a copy of the front and back of the title page.
3. Turn the book into electronic text.
4. Send it to us.

That's it! All the rest of the producing parts of the FAQ are about the details of how different people approach these steps.

Different people find their own ways into PG work, and once in, find their own niches. If you have your own ideas, don't let anything here stop you from pursuing them.

Some people just read the FAQs, go up to their attic, pull an eligible book off the shelf, send TP&V [V.25] in, and start typing or scanning. Next time we hear from them is when they send in [V.46] the completed

eBook for posting. It can be as simple as that.

Some people just download existing PG texts, re-proof them very carefully and send in corrections.

Some people find regular collaborators through gutvol-d or the Volunteers' Board or the distributed proofing sites, earn a reputation as reliable proofers, and continue working as proofers.

Most people start small, and after a little experience of distributed proofreading or other proofing, begin their PG career as producers.

If you're a typist, cheer now, because you can ignore all the complicated paraphernalia of computer interfaces, and scanners, and the quality of OCR software and the mistakes it makes. You can just sit down at the keyboard with your eligible [V.18] book.

If you're not a typist, start thinking about scanners. It may be a while before you're ready to start scanning for yourself, but it's never too early to find out about them.

As soon as you have a solid grasp of how to turn a book into an etext, please start thinking about how you're going to become a producer. While proofing work is valuable, PG can only add books when someone makes the effort to actually make etexts from them, and the people who run distributed and co-operative proofing projects have to do a lot of work before and after the proofing step; we want to spread that around as widely as possible. Project Gutenberg needs more producers!

Whatever you do, _don't_ just hang around expecting someone to offer you a task to undertake. There is no "head office" where overworked staff occasionally need interns to do filing and odd-jobs. There are maybe 200 fairly regular contributors to PG, producers and significant proofers. We almost never meet each other in person. We have jobs, and families, and other interests. We work for PG when we can, and when we want to. In many ways, you could look at us as 200 unrelated people, each doing our own etext project, using Project Gutenberg as an umbrella group that sets loose standards, files copyright proofs and provides secure placement for the finished texts. Since we each have our own self-assigned single-person tasks, there isn't too much room to delegate some of that work to a beginner. By all means, volunteer for some tasks--on the Volunteers' Board, or in gutvol-d--but you should think in terms of defining your own tasks, and making your own contribution.

Orientation.

Absolutely everyone--scanners, typists, proofers--should first spend some time working on a distributed or co-operative proofing project. This will allow you to get a feel for what happens in making an etext from paper pages without committing you to more than a few hours'

work.

This is not in any way an institutional requirement, since we don't have any institutional requirements, but it is very good advice. Many volunteers start eagerly, wanting to do lots of PG work, and then drop out because they took on too much, too fast, without understanding the nature of the work. Don't let that happen to you. Take it in small chunks.

Check out these distributed proofing sites:

Charles Franks: <http://www.pgdp.net/>
JC Byers: <http://www.wollamshram.ca/1001/index.htm>
Dewayne Cushman: <http://www.metalbox.net/dcushman/pgroot.htm>

and spend a few hours over a couple of weeks just processing some pages for real.

While you're doing that, you should also join a couple of PG mailing lists [V.12]--gutvol-d and either the weekly or monthly Newsletter list. Reading these will start to get you connected to what's going on. Browse the Volunteers' Board--there may be some offers going, and there's a lot of experience captured in some of those "back-issues", so don't confine yourself to the front page.

Inform yourself on e-text issues generally, not just within Project Gutenberg. Explore The On-Line Books Page and the IPL [R.5] and from them find other eBooks available on-line.

Have a look at our In-Progress List and some lists of suggestions from others [B.4].

Look at sites like Blackmask <http://www.blackmask.com> and Pluckerbooks <http://www.pluckerbooks.com/> and Memoware <http://www.memoware.com> and Bookshare <http://www.bookshare.org> to learn how our work is being used as a basis and copied and converted and amplified in many other projects.

Above all, READ a few Project Gutenberg eBooks! You don't have to read them in full; you don't need to spend weeks poring over Dostoyevsky or studying Shakespeare. Just download a few and skim them--you'll absorb what a PG text should be quite painlessly, and maybe you'll get caught up in the story! If you're looking for light reading, and can't think of something that you specifically want, how about these all-time favorites:

The Gift of the Magi, by O. Henry.
The Lady, or the Tiger?, by Frank R. Stockton
A Christmas Carol, by Charles Dickens
Alice in Wonderland, Lewis Carroll
Anne of Green Gables, by Lucy Maud Montgomery
The Marvelous Land of Oz, by L. Frank Baum
A Princess of Mars, by Edgar Rice Burroughs

Heidi, by Johanna Spyri
A Connecticut Yankee in King Arthur’s Court, by Mark Twain
Black Beauty, by Anna Sewell
Tarzan of the Apes, by Edgar Rice Burroughs
Tom Swift and his Motor-Cycle, by Victor Appleton
Rebecca Of Sunnybrook Farm, by Kate Douglas Wiggin
Little Lord Fauntleroy, by Frances Hodgson Burnett
Aesop’s Fables
Grimms’ Fairy Tales
The Art of War, by Sun Tzu
Dracula, by Bram Stoker
Swiss Family Robinson, by Johann David Wyss
The War of the Worlds, by H.G. Wells

If you have a taste for detectives and mysteries, there’s

The Adventures of Sherlock Holmes, by Arthur Conan Doyle
Monsieur Lecoq, by Emile Gaboriau
The Mysterious Affair at Styles, by Agatha Christie
Arsene Lupin, by Edgar Jepson & Maurice Leblanc
Edgar Allen Poe’s "The Gold-Bug" and
"The Murders in the Rue Morgue" in The Works of Edgar Allan Poe V. 1

For the excessive buckling of various swashes, see:

The Prisoner of Zenda, by Anthony Hope
The Man in the Iron Mask, by Dumas, Pere
The Three Musketeers, by Alexandre Dumas
Treasure Island, by Robert Louis Stevenson
The Scarlet Pimpernel, by Baroness Orczy

Effen youse got a hankerin’ for a Western, there’s:

Riders of the Purple Sage, by Zane Grey
The Virginian, Horseman Of The Plains, by Owen Wister
Back to God’s Country, By James Oliver Curwood
Selected Stories by Bret Harte
Jean of the Lazy A, by B. M. Bower

Or if you prefer your fiction more domesticated, there’s:

Little Women, by Louisa May Alcott
Pride and Prejudice, by Jane Austen
The Warden, by Anthony Trollope
The Heir of Redclyffe, by Charlotte M Yonge
Mother, by Kathleen Norris

For something to raise a smile, you can rely on:

The Devil’s Dictionary, by Ambrose Bierce
The Wallet of Kai Lung, by Ernest Bramah
The Importance of Being Earnest, by Oscar Wilde
Three Men in a Boat, by Jerome K. Jerome
Piccadilly Jim, by P. G. Wodehouse

If poetry is your thing, you have lots to choose from:

Shakespeare’s Sonnets
Project Gutenberg’s Book of English Verse
The Home Book of Verse, edited by Burton Stevenson
The Complete Poems of Henry Wadsworth Longfellow
Leaves of Grass, by Walt Whitman

Now, that’s just a handful from our over 5,000 eBooks, so don’t tell me you can’t find anything to read! If you do have ideas of your own, download GUTINDEX.ALL or PGWHOLE.TXT and browse through the whole list, or Browse by Author on the website at <http://promo.net/cgi-promo/pg/cat.cgi>.

Download a few. Read them on your PC, or reformat them and print them out, or convert them for your PDA. Get used to working with and formatting text. Look at the formatting decisions that earlier volunteers have made--they’re not entirely consistent; different people make different choices, different books require different methods, and PG conventions have shifted slightly over the last 10 years--but they’re all perfectly readable and convertible today.

If you find typos [R.26] in any of them, tell us! That’s also a part of being a Gutenberg volunteer. Our eBooks _improve_ with time!

If you’re thinking of making the best use of your time looking for errors in posted texts, a good start would be to download 40 or 50 texts, and run a spelling checker and gutcheck [P.1] on them all, spending only 5 or 10 minutes on each. Having had a quick look at all of them, concentrate on the ones that seem to have most problems--where automated checkers see 10 problems, a careful human will usually be able to pick up 20.

Getting Productive

OK, so you’ve seen what etexts should look like, you know what we do, and proofing hasn’t scared you off. It’s time to step up and become a producer. If you’re not a typist and you don’t have a scanner, take a detour down to the Scanning FAQ [S.1] now, and come back when your scanner is set up. If you’re a typist or you’ve already got a scanner, read on . . .

Get a book. Just do it, OK?

Ya gotta start somewhere, right? And finding an eligible book is definitely somewhere.

Finding an eligible book is a threshold for many beginning volunteers--it's the first major step on the way to producing. For a lot of people, it's also the toughest barrier they have to cross. Fortunately, the barrier is only psychological, and can be crossed in a few minutes.

It's an unfamiliar process, and one that a lot of beginners feel some anxiety about. Don't. It's quite straightforward: it's just buying a book--you've done that, haven't you? Don't over-think it, don't worry about whether you're making the "right" choice, don't spend months comparing lists and choosing. Just do it. Once you've got your first, you'll wonder what all the fuss was about. Thanks to the wonders of the internet, your book can be on its way to you in an hour if you have $20 to spend.

Typists blessed with a good local library don't even have to buy their books--they can just borrow one and type it up! (You may be able to scan a library book, but get some experience with scanning first, and avoid damage!)

Let's deal with the decisions and other issues of picking one.

Copyright

For your first book, don't try getting fancy with copyright issues. Choose one that was published before 1923, and you're in the clear for U.S. and PG copyright purposes. You can read the dates just as well as we can--with books printed before 1923, there are no hidden catches: "Pre-'23 is free". Just read the TP&V [V.25] of the book, and see that it was printed before 1923, and you have no problems. Of course, reprints [V.19] of books copyrighted pre-1923 (and various other cases) are also clear, but if you have any concerns, just stick to pre-'23 editions.

Which book?

The answer to this question is different for everyone, but see how much you agree with the following statements:

"I have a favorite book, and I'd really like to produce that."

Well, hey, this is no problem! You already know what you want. Go check out whether the book is already on-line [V.29].

"I'd like to work on an important book, but I don't know which."

Well, everybody's definition of "important" is different, but some people have put their various ideas forward already; you can see whether you agree with them! The InProg List contains some, with the notation "Suggested book to transcribe" beside them. Steve Harris keeps a list of unproduced possibles at Steveharris.net. John Mark Ockerbloom's "Books Requested" page lists titles that people have asked for. [B.4] Your problem if you fall into this category is that other people probably wanted to produce "important" books too, and lots are already done.

"I just want an easy, trouble-free book to start with."

Your first book doesn't have to be War and Peace (we've already got that anyway!). Here's a tip: try looking for children's or what we would nowadays call "Young Adult" books. These are typically short, and may have large print, which makes life much easier if you're scanning. They age well: children's stories from a century or more ago are still readable and interesting to children today. We have many children's and YA eBooks: not just the classics like Grimm and Andersen and Heidi and Oz and Peter Pan and William Tell, but lesser-known but still enchanting stories like The Counterpane Fairy, or Lang's Fairy books. There are series, like the Motor Girls, or the (Country) Twins series, or the Bobbsey Twins. There is lots and lots of material here for you to start with, and these books are relatively plentiful, since they were made to take the kind of treatment children dish out, and many of them have been in school libraries or attics for years.

Whatever your choice, pick a book that you'll like; you'll be living with it up close and personal for a while. Light reading, adventure fiction, and books aimed at younger readers are safe first choices for most people. If you admire 19th Century scientists or scholars, and want to immortalize their work, great! But don't feel that you have to dive in at the deep end just because someone else wants you to.

Getting your book: a practical exercise

The Search

At this point, you've got a list of books--maybe just one, maybe several by an author or two, maybe just a genre like "Children's Books" with some specific ideas. Maybe your mind is still wide-open.

Before used booksellers had the Net, finding a particular old book was a daunting job. Booksellers had informal networks among themselves and exchanged catalogs so that each would know something about what was available elsewhere, but, for a buyer, finding a particular book was still hit-and-miss. Now, however, a number of large sites provide a service to booksellers, where they can list their inventories for

people to search from anywhere.

So now we go hunt for them on the Net. No, you don't have to buy them on the Net--you can rummage in booksales and garage sales and used bookstores, and that's its own kind of fun, though on a physical hunt, what you need is to bring a long list of "already done" books with you. But even if you never buy over the Net, it's a vast source of information about what books are available, which are plentiful, and which are cheap. It gives you some experience of what to expect when you do your in-person browsing.

Here's a story of a typical Net-hunt. And you can follow along with it at home. :-) Your results, and the sites you end up at, will be different from mine, but even if you don't end up buying a book on this hunt, you'll get some experience of what's involved. C'mon, do it with me--see if you can find a better bargain!

I'm starting with two lists, and I'll follow up whatever seems promising. I'd like to spend about $20--might go to $30. Definitely not interested in $50 and up. I'm keeping in mind that I'll have to add a bit for delivery--usually up to $10 within the U.S., but can get expensive if you're in Perth, and ordering from a bookstore in Munich.

I'm also avoiding anything that might be tricky to clear on this search, and confining myself to books printed before 1923.

Of course, by the time you read this, some of these books may already have been produced, so if you're actually thinking of buying any, check carefully first!

My first shortlist consists of books that caught my eye from David Price's In-Progress List, Steve Harris's site, and The On-Line Books Requested page [B.4], and it reads:

Louisa May Alcott: The Inheritance
E. W. Hornung: Irralie's Bushranger
E. W. Hornung: Stingaree
A. A. Milne: The Dover Road
A. A. Milne: Once on a Time
Samuel Richardson: Pamela
Oscar Wilde: The Critic as Artist

As well as following along with my list, you should try finding two or three books of your own, from those sites or from your own preferences, and search for them in the same ways that I do.

Everyone has their own searching technique and their own favorite sites to search. For this session, I'm opening up three copies of my browser--one for Alibris <http://www.alibris.com>, one for Abebooks <http://www.abebooks.com>, and one for the Catalog of the Library of Congress <http://catalog.loc.gov>. I'll do my initial searches on Alibris and Abebooks, and keep the LoC site handy for reference.

In Alibris, I head straight for the Advanced Search page, since they allow searching by date, and I immediately put "before 1923" into every search, which avoids having to scan through modern reprints. In Abebooks, I choose "Hardcover" in their advanced search, which is not quite as good a filter, but does at least screen out recent paperback editions.

In each of the sites, I just enter the author's surname and one word from the title of each book, and look at the search results.

Louisa May Alcott's "Inheritance" looks like it's going to be tough. I don't find it in either of my two bookstores. On doing a little checking with modern bookstores, I find it was her first novel, written when she was 17, and as far as I can see, not published during her life: apparently only recently published--the LoC site has nothing prior to 1997. A disappointing start to my search. I understand why it's very desirable to get it online, but this one's going to be very tough to clear, and I'm staying away from it.

E. W. Horning's "Irralee's Bushranger" is also elusive: it doesn't show up at either of my sites, so I check out the LoC to confirm I have the title right, and yes, there it is: "Irralee's Bushranger, a story of Australian adventure, 1896." So I widen my search by visiting <http://www.trussel.com/f_books.htm> and searching many of the sites there. Still no luck. If I were particularly eager to get this book, there are several things I might do at this point: I might register a "want" with one of the sites, asking to be notified when a copy is listed, I might use the OCLC WorldCat search (which Abebooks calls "Find it at a local library") where I can locate libraries that have copies, or I might even contact some individual booksellers and make a request that they look for it. Some booksellers actually specialize in looking for hard-to-find books; but of course I expect I'd have to pay a bit more for it when they do find it, and given my success with the rest of my list, and my price bracket, there seems no need to go that far today.

Horning's "Stingaree", by contrast, seems to be everywhere, in several editions, and cheap. It must have been a bestseller in its day--not surprising, from the author of "Raffles". 1902, 1905, 1909 editions abound. The cheapest are 1910 and 1907 editions for $4.95 and $5.00 from booksellers listed at Abebooks.

Milne's "Dover Road" is available from both sites. There seems to have been a Putnam's printing in 1922 of "Three Plays: The Dover Road. The Truth About Blayds. The Great Broxopp." of which lots of copies survive. There also seem to be later printings which would qualify as reprints if I were desperate, but the 1922 edition is priced from $12.00 to $50.00, so I'll take the 1922 $12.00 copy from Abebooks. As a bonus, I don't see the other two plays listed as being online anywhere, so I'll get three texts (and short ones, too!--279 pages for all three) for the price and effort of one.

Milne's "Once on a Time" is a bit less common, but once again a

Putnam's printing of 1922 keeps it in the race. There are a couple of booksellers in England selling for 15 pounds (which just about makes my $20 threshold) and 20 pounds, and an ex-library copy going for $25.

There are lots of eligible copies of "Pamela" available, ranging from a fourth edition at a mere $4,999 (no, thanks!) to a 1921 printing at $6.60 at Alibris. I'll take that one, please.

Wilde's "Critic as Artist" is fairly widely available. A 1905 edition of "Intentions: the Decay of Lying; Pen Pencil and Poison; the Critic as Artist; the Truth of Masks" is available at Alibris for $8.80, (and other copies of the same edition there and on Abebooks in the $20-$30 range) and Abebooks lists a London 1919 edition at $12.50. There are several copies listed in both places as "undated" and "reprints"--I'm avoiding these, since while it's quite likely that they might be clearable, I'm not taking risks on this search.

My second list isn't a list--just a vague category: children's books that are easy to do.

I go to Alibris' Advanced Search, and enter "Child's" in the title, and pre-1923 in the date, and, excluding titles already on-line, immediately get:

A Child's History of France $13.20
A Child's Story of the Bible $5.50
First Lessons in Botany or The Child's Book of Flowers $13.20
The Child's Book of American Biography $11.00
The Child's First Bible $8.80
The Child's Music World $8.80

and so on through quite a list.

OK. That's a good start. But my choice so far is unimaginative. I need better search terms. So I go to main search engines with the terms "children's antiquarian books" and find a half-dozen or so sites that specialize in them. I can browse around there, though it's slower going without searches to focus my results. I find <http://www.bookrescue.com>, specializing in children's books. Wading through the miles and miles of Alcotts and Barries and Burnetts, which are mostly already online, I think, I find a couple of authors from them who must have been popular, because they seem to have published lots of books before 1923: Angela Brazil and Dorothy Canfield. (I only got as far as the "C"s!)

I could of course stop here and buy some, but today I want to see what else is out there.

Back at Alibris and Abebooks, armed with my authors to search by, I turn up 4 pre-1923 books under $20 for Angela Brazil:

A Terrible Tomboy

The Youngest Girl in the Fifth
A Fourth Form Friendship
A Pair of Schoolgirls

and several between $20 and $30.

Dorothy Canfield immediately yields multiple copies of:

The Brimming Cup
Home Fires in France
Hillsboro People
Understood Betsy
Rough Hewn
The Real Motive

and others, and I haven't even got to $20 yet, nor to the letter "D".

A browse through the Ebay Collectible and Antiquarian Books section also throws up a respectable list of eligibles. I won't even bother counting that.

In 20 minutes, I have found five of the seven on my search list. In less than hour after that, I found over 16 eligible children's books, all under or around $20 and all available online.

Before committing to one, though, I would double-check that the book hasn't been transcribed online, and isn't In Progress.

Double-checking your selection

If you're concerned that the book you have chosen duplicates another that might be in progress, and want to double-check, you can e-mail the Posting Team asking them to check whether any recent clearances have come in for that title.

Duplications do happen--there's no way of avoiding them when different people are making independent decisions--but they are rare.

Dealing with used booksellers

As a class, used booksellers are very pleasant people--remarkably friendly, knowledgeable and helpful, even to people buying on a typical Gutenberger's budget.

Some of them are not, however, models of ideal data organization when it comes to Internet listings. There are lots of one- or two-person operations dealing with an inventory of many thousands of books, and having located your book online, you should check that it's still available.

You can place an order through the site and wait for the confirmation, or you can simply call the bookseller. Not all booksellers' contact details are listed, so it's not always an option, but when you do phone you're likely to be speaking immediately to someone who can tell you for sure whether the book is still there, can pull the book off the shelf and answer questions about it, and can take your credit card details on the spot and dispatch the book immediately.

Copyright Clearance

As soon as your book arrives, send us the information needed for Copyright Clearance first. Even if your book is a true-blue, no-questions-asked pre-1923 edition, we should know about it as soon as possible so that it can go onto the In-Progress list for others to see that someone has started on it.

Wait for the confirmation e-mail before starting any serious work. Some people have thought that "Copyright 1923" plus some wishful thinking would be good enough, and, unfortunately, it isn't. Some people have gone ahead and produced the whole book before sending in the clearance, only to be disappointed, all their work wasted.

Books published in 1922 or earlier are clearable, but some people, ever optimists, overlook that little "1927" in small print on the verso. Sometimes there is no copyright date on the front, and other optimists assume that these books are OK. They may be; they may not be. Don't get caught in the copyright trap.

As soon as you have what you think might be an eligible book, do not start on it. Do not ask another volunteer's opinion. Just send in the TP&V and wait for the confirmation e-mail to find out for sure.

Even when your TP&V clearly says "Copyright 1901", send it in. We need to get it into the clearance files so that we can register it as being In-Progress.

Producing

If you're a typist, there's not much more you need to know from this point: you can just get on with the job, with maybe a few tips from the FAQ. In fact, if you're a typist, you might wonder why the rest of us make such a fuss about scanners, and settings, and OCR. Take pity on us! we just can't produce the way you can. Smile indulgently, ignore all the scanner jargon, and submit your completed text while we're still saying bad words about the guttering on a greyscale image of page 372. :-)

If you are using a scanner to copy a book for the first time, be

patient with yourself. Some people start off with too high expectations of what they can achieve. Believe it or not, scanning does work effectively; it just doesn't work perfectly. And often, you need a little practice before your scans work right with your OCR. The Scanning FAQ [S.1] has lots of specific tips you can try. Start by scanning a double-page about a third of the way through the book. Scan in Black and White and in Greyscale, at 300dpi and 400dpi. Try 600 dpi if it seems like a good idea. Put it through your OCR and see what comes out. Move your scanner so that you can be comfortable while placing the book and turning pages. Allow yourself an hour to experiment with different settings, and different pages. Put the sample images included with the Scanning FAQ through your OCR and see how the output compares to the text produced by other packages. That first hour finding out about how your setup works will be the most valuable hour of scanning you will ever do.

Having figured out what settings you want to use for this book, make sure you implement the best speed you can. Usually this means telling the scanner to scan _only as much area as the book covers_. This is quite important, since the scanner will by default scan its whole area, and you don't need all that; it just wastes time and makes your images bigger.

You may also be able to set your OCR or scanner software to auto-scan pages with some preset delay, like 5 seconds. This also speeds things up, because the scanner isn't waiting for you to hit the keyboard, and you have both hands free at all times to turn the page and replace the book. It takes a few pages to get into the rhythm; if you miss a page-turn, don't worry--you can get it on the next scan.

Using a reasonably modern but quite ordinary home/office type flatbed scanner, you should be able to scan 200 pages an hour [S.9] of a typical book, at good quality. 400 pages an hour is not unheard-of. Now, it may fairly be said that scanning offers all the fun of ironing, without the sense of adventure :-), but if you have got your settings right, you will probably be able to do the whole job in less than two hours. And now you're really on the road!

V.2. What experience do I need to produce or proof a text?

None.

For producing, you will have to be able to type pretty well, or have a scanner.

For proofing someone else's text, when you don't have a copy of the book in front of you, you should be reasonably familiar with the language used in the book, and the styles of the time--Chaucer's English was quite different from ours, and even 19th Century novelists write some phrases unfamiliar to us today.

That's it. You don't need experience in publishing, editing, or computers.

V.3. How do I produce a text?

There are acres of words in this FAQ about that, but it all boils down to 4 simple steps:

1. Get an eligible book--pre-1923, or one of the exceptions. Pull it from your attic, borrow it from a library or a friend, buy it in your local bookstore, in a flea-market or on-line. We don't care which.
2. Send us a copy or the front and back of the title page so we can file proof of copyright clearance.
3. Copy the text from the book into a computer text file. We don't care whether you type it, scan it, voice-dictate it, or think of some totally new way to do it. Just get it into a file.
4. Send us the computer text file.

That's all there is to it!

V.4. Do I need any special equipment?

You need the use of a computer of some kind, and Internet access is usual, though we have had some volunteers contribute texts on floppy disks.

If you intend to scan books, you will need a scanner, but if you're just typing or proofing you won't.

V.5. Do I need to be able to program?

Absolutely not! Very little of Project Gutenberg's work involves programming, and it is never necessary to any part of volunteering.

V.6. I am a programmer, and I would like to help by programming. What can I do?

At the risk of sounding facetious, the very best thing you can do is figure out ways that more programming can help Project Gutenberg!

A lot of programmers work on PG books, and anything easy has probably already been done. The challenge for programmers who want to write something that will help to produce etexts is not in writing the code; it's in identifying ways that programs can help.

Please see the FAQ "What programs could I write to help with PG work?" [P.2] for some ideas in this direction. Whatever you do, don't just hang around waiting for someone to ask you to write something, because that's not going to happen. Think up a project, ask volunteers if they would use it, and dig in! Better still, produce a few etexts yourself, using the existing tools, and get a feel for the kinds of problems that new software could help with.

Apart from text production, we do develop some programs to help with posting work, but as of mid-2002, we have nothing like an ongoing programming project which people can join.

V.7. What does a Gutenberg volunteer actually do?

We buy or borrow eligible books, scan, type, and proofread. There are a few other activities, but they consume only a very small fraction of volunteer time.

V.8. Can I produce a book in my own language?

Yes! We want to encourage people to produce books in all languages, and we cheer when we can add a new language to the list.

V.9. Does it have to be a book? Can I produce pieces from a magazine or other periodical?

Magazines, newspapers, and other publications are just fine. For copyright clearance, they work just the same way as a book.

You do need to check the length of your piece [V.17]; we don't want a zillion separate one- or two-page files. If the piece you have in mind isn't long enough, you can add other pieces to it, or even most or all of the magazine. If the work was serialized over multiple issues, you can join them together for your PG text, but you do have to copyright clear every issue of the magazine from which you copy material.

If you have lots of old periodicals, you could even take one piece from several, and make a new text which is a "theme" anthology of those pieces. You can give it an appropriate title: "Civil War Commentaries from X magazine 1892-1898."

V.10. Do I _have_ to produce in plain ASCII text?

Certainly not if it doesn't make sense. To take an extreme example, if

you're working in Japanese or Arabic, or creating audio files, there is no point in trying to reproduce that in ASCII!

Where the text can largely be expressed in ASCII, we do want to post an ASCII version, even if it is somewhat degraded compared to the original. However, we will post your file in as many open formats as you want to create, so that your original work is available for those who have the software to read it.

V.11. Where do I sign up as a volunteer?

You don't. We have no formal sign-up process, no list of volunteers, no roll-call. If you produce a PG eBook, or help to produce one, you are a volunteer.

V.12. How do PG volunteers communicate, keep in touch, or co-ordinate work?

We are very scattered geographically: U.S., Australia, Brazil, Taiwan, Germany, South Africa, Italy, India, England, and all over the world, so we can't really meet for coffee on Thursdays. :-)

Most co-operation and co-ordination goes on by private e-mail. This is efficient for volunteers who have worked with each other before, since they know each other's interests and skills, but not so easy for beginners to break in on, since they don't.

The Volunteers' Web Board at <http://promo.net/pg/vol/wwwboard/> is a publicly accessible forum for volunteers or potential volunteers to post any question or information about how to create a PG eBook.

There are a few Project Gutenberg mailing lists. Information about joining them is available on the main site, at <http://promo.net/pg/subs.html>.

The Project Gutenberg Weekly and Monthly Newsletters, gweekly and gmonthly, are one-way announcements, which allow PG to communicate with non-volunteers who are interested in the eBooks we produce, but they also contain notes and requests for assistance from volunteers.

The Volunteers' Discussion Mailing list, gutvol-d, is a an e-mail discussion forum for subscribers about any Gutenberg topic.

The Volunteers' List, gutvol-l, is for private announcements for active volunteers.

The Programmers' List, gutvol-p, is for discussion of programming topics.

There are some other, specialized, closed lists for people who

do specific work within PG:

The "Posted" List, posted, is for people who perform indexing on our texts. An e-mail is sent to this list every time we post a text (see the FAQ "How does a text get produced?" [V.16] section 5: Notification) and the members of the list use it to update their catalogs.

The Whitewashers' List, pgww, is for Posting Team internal messages.

The Heroic Helpers List, hhelpers, is for people who can devote some fairly regular time to doing odd jobs.

V.13. Where can I find a list of books that need proofing?

There is no central list of this kind. There are distributed proofing projects, currently at

Charles Franks: <http://www.pgdp.net/>
JC Byers: <http://www.wollamshram.ca/1001/index.htm>
Dewayne Cushman: <http://www.metalbox.net/dcushman/pgroot.htm>

where you can proof parts of a book. This is advisable when you're just starting out because it gives you some feel for what the work is like.

You can also look up existing, posted texts from the archives and proof them. Just as there always seems to be one more bug in any given program, there always seems to be one more typo in any given text! Download a few, and scan quickly for problems by doing a spellcheck or other automated check; if you can find any problems quickly, then there are likely others to be discovered by a careful proofing.

V.14. Is there a list of books that Project Gutenberg wants?

No. Project Gutenberg, as such, does not "want" any specific books. Individual volunteers choose what books to produce. Nobody gives orders to volunteers about what they should work on. Nobody has an official "hit-list" of books to add to the archives.

Of course, individual volunteers and non-volunteers have their preferences, and may suggest books to transcribe, and such suggested lists pop up every so often, and are often useful to people looking for ideas.

There are usually some suggestions in David Price's InProgress list. The On-Line Books Page has a section where people can list requests,

and Steve Harris has a site devoted to lists of books not yet in Gutenberg or elsewhere. Treat all of these lists with some caution, since someone may have started or even finished one of their suggestions since they were last updated.

PG Books In Progress <http://www.dprice48.freeserve.co.uk/GutIP.html>
On-Line Requested List <http://onlinebooks.library.upenn.edu/in-progress.html#requests>
Steve Harris' "To-do"s <http://www.steveharris.net/PGList.htm>

V.15. I have one book I'd like to contribute. Can I do just that without signing up?

Well, since there is no formal sign-up, of course you can! A lot of texts have been contributed by people who just wanted to immortalize one favorite book. Many of them had already created the eBook before they even heard of Project Gutenberg, and we're always delighted to add these to the archive!

About production:

V.16. How does a text get produced?

As stated back in the Basics section, all you need to do is:

Borrow or buy an eligible book.
Send us a copy of the front and back of the title page.
Turn the book into electronic text.
Send it to us.

That's all you actually need to know in order to be a producer. But if you're interested in the details of how other people actually do this, and want to know what else happens behind the scenes, here's a full, blow-by-blow account.

1. Finding an eligible book

Volunteers find eligible books [V.18] in all sorts of ways. Some lucky people have them in their bookshelves, or their attic. A lot of people have a good library nearby, where they can find books, or request them on interlibrary loan. Some people are big eBay fans; others like to hunt for bargains on specialist booksites. And of course lots of volunteers enjoy rummaging through actual used bookstores, or local markets, or yard sales.

Even if you're not going to take on a book yourself right now, search for some on the Net and find out about how to get a copy. Next time you pass an antiquarian bookstore, or a book market, drop in and browse. Ask your local library about interlibrary loans. Eligible books aren't hard to find once you know where to look.

2. Copyright Clearance

New volunteers sometimes find it hard to understand why this is so important, and why, in particular, Project Gutenberg is so careful about it. At base, it's simple: by keeping a filed copy of the TP&V [V.25] of every book we produce, we can at any time protect our publications against claims from publishers that they "own" the work, and thus we can keep them available to the public.

The copyright laws can be difficult to understand, and sometimes it may take serious research to prove that a particular edition is actually in the public domain. If you're not legally-inclined, just keep repeating "Pre-'23 is free" if you're in the U.S.A. and stick to books published before 1923. If you do want to delve deeper, read our Copyright Rules page at <http://www.gutenberg.net/vol/pd.html> and then go on to reading the Library of Congress Copyright Office official papers at <http://www.copyright.gov/>. If you're in another country, find out about your own copyright laws.

Volunteers send in the TP&V from the book for us to inspect. This not only gives us the proof to file, it also lets us know that someone is really working on the text so that we can list it as being In Progress for the information of others who might be interested.

3. Scanning, typing, proofing and editing

This makes up the bulk of PG's effort, and is discussed at great length elsewhere in this FAQ. There are many, many ways to create an etext from a paper book, and different people use different methods, but it all boils down to making a text file. For a typical book, it will probably take 40 hours of a volunteer's time. All that happens here is that somebody makes the effort to transcribe one paper book into a file that can be shared around the world and for all time.

4. Posting

[Note: this information is quite specific to the process we go through now. It is quite likely to change as we improve the automation of the tasks.]

Posting is done by the Posting Team. The basic job is to receive the text from the producer, check that it has been copyright cleared, check that it conforms to Project Gutenberg standards, check it for correctness (which can be anything from XML validity to simple spelling), add the Project Gutenberg header and copy the text to the two PG servers.

In a simple case, where everything goes right, this can take as little as fifteen minutes. In a complicated case, where we have to convert formats, or there are a lot of errors in the text, or there are problems with the copyright clearance, it can take hours or even days while we wait for responses, or do a lot of editing, or find conversion tools.

Michael Hart used to do this work entirely alone, but in September 2001, he created the Posting Team to handle the load. (The Posting Team are nicknamed the "Whitewashers" in honor of Tom Sawyer's victims. :-)

Transferring the file

You send the text to us [V.46] either by Web, by FTP with a username and password that any of the Posting Team can give you privately), or by e-mail.

If you're FTPing, you should e-mail one or more of us as well, to let us know what you've uploaded.

One problem is files that don't transfer correctly. Especially by e-mail, some files get damaged on the way. It's better to ZIP the file before sending, if possible, to prevent some common problems with text files. The use of compression formats other than Zip can also create problems. Members of the Posting Team work on multiple platforms--DOS, Windows, Linux, Solaris--and zipping and unzipping programs are commonly available for all of these. Other compression methods, like Stuffit or bzip2, are not so readily available, and may give us trouble.

We login via ssh to beryl, which is the Unix system on which we work when posting, the same one that you FTPed the file to, unzip the file and glance at the top of it.

Checking Clearance.

We then check it for copyright clearance. The one and only absolute rule that we NEVER bend, no matter what, is that we WILL NOT post a file that doesn't have a clearance. If it ain't in the clearance files, it don't get posted.

Most regulars know that they should include their clearance line in the e-mail submitting the text, but not everybody does, and not

everybody remembers every time. This can be frustrating, when clearance is not included and not obvious.

When Michael gives you your clearance on a book, he sends you back an e-mail that has just one line, something like this:

The Works Of Homer [Iliad/Odyssey] Tr. George Chapman Jim Tinsley 06/14/01 ok

He saves these lines in files that we posters can access. We regard this information as private, so we don't publish the details of who has cleared what.

When we get the text, we check whether the submitter has cleared it. If there is a clearance line in the e-mail notifying us about the text, there's no problem. If we can find the title of the text under the submitter's name in the clearance files, there's no problem. Unfortunately, sometimes we can't find it. There are two usual reasons: either the text submitted is _part_ of the work cleared (for example, submitting one play from a collection), or the text hasn't been cleared yet. If the clearance isn't straightforward, we can go back and forth and round and round in e-mails for a while.

This is why it's a good idea to paste the clearance line into your e-mail.

If the title of the text you're sending isn't the same as the title of the text cleared, BE SURE to paste in the clearance line AND explain that the text you're sending is PART of the cleared book. Please also list the titles of the other parts; it really does cause confusion and delay when this is not clear.

Checking and Editing

Sometimes, people send in a book in a non-text format like Word Perfect or Microsoft Word, or send a text with unwrapped lines. In that case, we try to get the submitter to fix them, but if they can't, we have to convert the file to straight text before starting.

Some producers, particularly inexperienced ones, want to add non-standard annotations and mark-up and symbols to the text. This can get ticklish; we don't want to discourage them, but we need to keep texts reasonably standard. Usually, we can work something out. Maybe the book should be added in _both_ text and HTML, for example.

Assuming that it's a plain text file, we next run gutcheck and a quick spellcheck on the file. This will tell immediately if it adheres to PG standards and if there is any serious problem with it.

If the file looks clean, we may skim it, looking for potential problems or formatting issues. For clean texts, the only things we usually need to change are unindented quotations or inconsistent chapter headings (a lot of people seem to mix "CHAPTER III" with

"Chapter 14" and have irregular numbers of blank lines) or spacing and a few 8-bit characters. Occasionally, we have to rewrap a text. We also look out for included publishers' trademarks, which we normally prefer to remove (trademarks are NOT subject to copyright expiration: Macmillan(TM), the publishing house, is still around and trading), unnecessary or downright odd indentation or centering, stray page numbers, and prefaces or introductions or appendices that may not be in the public domain. If the file has lots of 8-bit characters, we probably need to make a separate 7-bit version, and post both.

If the gutcheck and spellcheck don't look clean, or if conversion is required, we may spend a lot more than 15 minutes on it. In a bad case, we may have to get the file re-proofed.

If you are conscious that you're doing something non-standard, and really mean it to stay, say so in your e-mail. (For example, I recently posted a text containing a family-tree representation that had lines over 80 characters. Now, I would have left that one alone anyway, but it helped that the submitter drew my attention to it in the e-mail.) If it's too non-standard, the poster may not allow it to stay, but at least you can discuss it. When a text needs a lot of non-standard formatting or markup, you really need to ask yourself whether you shouldn't be submitting it in HTML, with all the bells and whistles, and settle for something more normal in the text variant.

Mostly, errors are obvious, and there are at least some obvious errors in most texts. When errors are completely obvious, we just fix them without feedback to the producer unless you have specifically asked for feedback in your e-mail.

We're getting more HTML formats now, which is great, but incoming HTML often needs a lot of work, because people who are not experienced with HTML often make mistakes. The W3C <http://validator.w3.org> is the official standard for valid HTML, but, for the average volunteer, it's awkward to use. However, if you're submitting a HTML format, please use Tidy, which you can get from <http://tidy.sourceforge.net>, to check your text before sending it.

Header and Footer

We add the PG header and footer. If there is a header and footer already there, we strip them off first, since recent changes in the header mean that a lot of people send files with headers that are out of date. We have written programs to help with this.

We get the number for the text from a program on beryl called "ticket" that Brett Fishburne wrote, that dispenses the next number. That way, if two or three of us are posting at the same time, we won't all grab the same number. We create a 5-letter base filename, checking that it hasn't been used before, and finally zip up the file.

Posting

We now transfer the .ZIP and .TXT files to two servers:
ftp.ibiblio.org and ftp.archive.org. (This is usually the point at
which we realize that we forgot to make a change we noticed while
checking. Aaaargh!)

5. Notification

At this point, the book is posted, but nobody knows about it! We need
to do something about that. . . .

We compose an e-mail to the "posted" e-mail list, cc: the producer,
with the line that is to go into GUTINDEX.ALL, the master list of PG
files.

The "posted" list has only a few subscribers. These are the people who
index and create links to PG texts, and include both PG volunteers and
the maintainers of other sites that link to PG texts.

They also commonly download the texts to get more information for
their indexes, and tell us if there is anything wrong with the files.

This e-mail is simply the official notification to all these people
and the producer that the file has been posted. Here's a sample of
such an e-mail:

To: "Posted Etexts for Project Gutenberg" <posted@listserv.unc.edu>
Subject: [posted] Posted (#5301, Duncan) !
From: "Jim Tinsley" <jtinsley@pobox.com>
Date: Tue, 25 Jun 2002 06:21:27 -0400 (EDT)
Cc: you@example.com

Mar 2004 The Imperialist, by Sara Jeannette Duncan [SJD#4][mprlsxxx.xxx]5301

There may also be some remarks, if the text is in any way
non-standard, or if files other than plain text were posted with it.

From this e-mail, you can, if you want to see any corrections made,
immediately download the posted file and compare it to your version.
Since the notification is made _after_ the file has been copied to the
servers, it should be there waiting for you.

To find out how to download a book that has just been posted, see the
FAQ "How can I download a PG text that hasn't been cataloged yet?" [R.3]

6. Indexing

From the "posted" list, the posting line is added to GUTINDEX.ALL

and our indexers begin the cataloging process, which is much more thorough, for the website. This includes work like finding author's dates of birth & death, getting the Library of Congress classification, and the other information that makes up the website searchable index. That process takes extra time, which is why the website searchable catalog must always lag behind the actual titles posted.

7. Corrections

It's remarkable how many people who went over and over the text to the point of hating it suddenly see problems with it when they download it a couple of days after it's posted! Something psychological there, I expect. Anyhow, if you do download your text and see problems with it, don't worry, just e-mail whoever posted it, or any other member of the Posting Team. No, you're not stupid, or if you are, you're in good company, because we've all done it! There's no big deal about replacing the posted file with a corrected copy immediately.

Over time, other readers may submit corrections. If you find an error in a PG etext, see the FAQ "I've found some obvious typos in a Project Gutenberg text. How should I report them?" [R.26]

When the corrections are small, as most are, we will just make the change to the existing text. If there are a lot of changes, we may post a new edition [R.35] with a new edition number; e.g. if the file abcde10 was corrected, we may post abcde11. We never make a new edition when we get corrections immediately after posting.

V.17. How long must a text be to qualify for PG?

The rule of thumb is that we try not to post texts shorter than 25K, or about 350 lines of 70 characters. This rules out, for example, a lot of individual short poems. If you are interested in contributing this type of material, consider making a collection of similar texts--poems by the same author, or magazine articles on the same subject. We have made a few exceptions, like Martin Luther King's "I have a dream" speech, but very few.

V.18. What books are eligible?

A book is "eligible" for posting if we can legally publish it. This is the case if:

1. it is in the public domain in the U.S.A.,
   OR,
2. the copyright holder has granted unlimited

non-exclusive distribution rights to PG.

V.19. Are reprints or facsimiles eligible?

A reprint or facsimile of a book that would be eligible is itself eligible.

For example, if a book published in 1995 is a reprint of a book published in 1900, then it is eligible. However, the onus is on us to prove that it _is_ a reprint, and if it doesn't _say_ on the TP&V that it is a reprint, confirming its eligibility may be impractical.

V.20. What is the difference between a reprint and a facsimile?

A facsimile retains the page layout and formatting of the original. A reprint keeps the same words, but may lay the pages out differently. For our copyright purposes, there is no difference--we can use either.

V.21. What is the difference between a reprint and a "new edition"?

A reprint contains only the words and pictures that were printed in the original. A new edition is in some way changed; it has different text, or pictures. It may be abridged, or expanded. It may have material added or changed, using other versions of the book.

A new edition gets a new copyright, and has to be cleared based on its own copyright date and status, not the date of the original printing of the title. See also the FAQ "How come my paper book of Shakespeare says it's 'Copyright 1988'?" [C.16] for an example.

Please note that we are talking here about a new edition of the printed book, not a new (corrected) edition number for Project Gutenberg naming purposes.

V.22. What book should I work on?

Nobody in Gutenberg is going to set assignments for you. You decide what book to process. Just pick one that no-one else has already done, or is working on. It's also sensible to pick one that you'll like--you'll be living with it for a while. On a practical note, it's probably better to start with a short book or even a short story, since a long book can take quite a while to produce.

Start by thinking of books written before 1923. Pick a book you like, and check it out. If it's already done or still in copyright, try

other books by the same author.

Visit the Project Gutenberg site and download a full list of Gutenberg books in GUTINDEX.ALL. Have a look at the List of Books In Progress and Complete [B.1]. Look for authors you like, and see what books by them aren't yet available.

Check out your old books. Maybe you have an eligible edition that would be of great help to the project.

Try your library. They may have some eligible editions--books we can prove to be in the public domain--and you will certainly come away with ideas. Ask your librarian. Librarians are keen to help on projects like this.

Browse second-hand bookshops in your area. There are lots of treasures to be picked up very cheaply.

Search for literature pages and bookshops on the Internet.

If all else fails, you can always ask on the Volunteers' Board or try the gutvol-d mailing [V.12] list for ideas. Others may know of books that people are especially looking for, or projects already started where you could help out.

V.23. I have a book in mind, but I don't have an eligible copy.

First, determine whether there are any eligible copies of the book, by finding out the date it was published, possibly from the Catalog of the Library of Congress [B.5] and checking the Public Domain and Copyright Rules [B.1]. If there is a public domain edition, the next problem is to find one to work with.

V.24. Where can I find an eligible book?

The most commonly used outlets are used bookstores, garage sales, library sales, charity shops and any other place that sells old books.

The Internet is a wonderful medium for finding used and antiquarian books--used bookstores all over the world have found ways of co-operating and listing their inventories on the Net, so that whether you live in Los Angeles, Moscow or Perth, you can still find that book you're looking for in a shop in a laneway of Amsterdam. Most on-line listings will quote the publication year of the book, so you can check that it's pre-1923.

Two such sites that allow second-hand booksellers to list their inventory are:

Advanced Book Exchange <http://www.abebooks.com>

Alibris <http://www.alibris.com>

The book search page at trussel.com [B.5] has a list of many such Net bookshops, or you can simply visit any search engine and search for Used or Antiquarian Bookshops. You can often buy eligible books through these sites very cheaply.

If you still can't find the book you need, post a message on the Volunteers' Board or to the gutvol-d mailing list; maybe someone else can find it for you.

Sometimes, it may be possible for you to work from a later edition, so long as somebody who has an eligible edition can check it to make sure that no changes have been made. Sometimes, you may be able to find a modern reprint; reprints may be eligible, as long as they say they are reprints of an edition that would be eligible.

If you can type, or can scan without damaging the book, you can borrow books long enough to produce them. Even if your local library doesn't have the books you want, they may well be able to get them for you on inter-library loan. Ask your librarian about it.

V.25. What is "TP&V"?

This is an abbreviation for "Title Page and Verso", and means a paper or image copy of the front and back of the title page.

Even if the back is blank, we need to have an image of it for the files, to show that it _is_ blank, so that if, in ten years' time, somebody queries our right to publish, we can show that we haven't just lost it.

Publishers print copyright information, like title, author, copyright year and owner, and whether the book was a reprint, on the TP&V, and by filing this, we can prove that the book we produced was in the public domain.

Sending us the TP&V is the One True Way to getting PG copyright clearance [V.37].

V.26. What is "Posting"?

Posting is the final stage in the production process, where the file is given a number and official PG header, and copied onto our FTP servers for distribution. See section 4 of the FAQ "How does a text get produced?" [V.16] for a blow-by-blow account.

V.27. I think I've found an eligible book that I'd like to work on.
What do I do next?

Make sure nobody else is working on it, and that it's not already online somewhere.

V.28. What books are currently being worked on?

Check out David Price's In Progress List (a.k.a. "the InProg List") online at <http://www.dprice48.freeserve.co.uk/GutIP.html>. David gets the information from Copyright Clearances that have been done, and organizes it into a list. It can never be 100% up to date, since clearances come in all the time, but it's the best online facility we have, and it's much more clearly presented than the original clearance files.

V.29. How do I find out if my book is already on-line somewhere?

There's no foolproof method; some student somewhere could have scanned it and put it on her college web page without announcing it anywhere. However, there are some regular places to check.

It may sound obvious, but you should always look in the PG archives first. Download GUTINDEX.ALL and keep it handy. Search the InProg List [B.1].

The two other main places to search for your book are the Internet Public Library <http://www.ipl.org> and the On-Line Books Page <http://onlinebooks.library.upenn.edu/>. These projects specialize in indexing books that people make available on-line.

If you still don't see your book on-line anywhere, hit your favorite search engine, and give it the title, author's last name, and preferably a few uncommon words from the first page of the book. Sometimes one of those solo efforts shows up in a general search.

V.30. My book is not on the In-Progress list, and I can't find it on-line.
Is it safe to go ahead and buy it?

Probably. It could have been cleared, but not included in the InProg list yet. If the amount of money to buy it is a consideration, you can e-mail any of the members of the Posting Team, and ask them to check the latest clearances for you. Even this isn't foolproof; another volunteer could be placing their order at the same time you're placing yours. Such duplications do happen, but they are very rare.

V.31. My book is on-line, but not in Project Gutenberg. What should I do?

If the on-line file is from the same edition as the one you have (e.g. not a different translation) then you may be able to submit that file, perhaps slightly edited, to Gutenberg using the clearance from your paper copy. See "I've found an eligible text elsewhere on the Net, but it's not in the PG archives. Can I just submit it to PG?" [V.62] for how to do that.

And of course, you can always still make your own version for PG. It's surprising how often even very similar paper editions have small differences that can be interesting or significant.

V.32. My book is already on-line in Project Gutenberg, but my printed book is different from the version already archived. Can I add my version?

Yes! In fact, assuming that the version already there is in the public domain, you can piggyback on the work already done by what is called "comparative retyping". For example, let's say that you have a later edition than the existing file; you can just take the existing file, edit it to match your paper version, and submit it as a new file. Of course, you must have Copyright Cleared [V.37] your paper version as well.

V.33. I see a book that was being worked on three years ago. Is anyone still working on it?

Maybe, maybe not. Some people abandon books, some people who are regular producers clear them and put them at the bottom of the pile, perhaps for years (though they will get round to them sometime), and some people just simply take two or three years to produce a book.

Once, we put names and contact details on the public InProg list, but for privacy and spam-prevention reasons, we've taken them off. However, the Posting Team have access to the master list of cleared files, and will send a message on your behalf to the person who originally cleared the book, asking if the project is still active, or if the producer wants help.

So if you really want to check this situation out, e-mail one of the Posting Team.

V.34. I've decided which book to produce. How do I tell PG I'm working on it?

As soon as you get Copyright Clearance [V.37], your book is entered in the "cleared" files. David Price will take these, and add your entry in his next release of the In Progress List.

V.35. I have a two- or three-volume set. Should I submit them as one text, or one text for each volume?

Both.

Quite a lot of 18th and 19th Century books, even straightforward novels, were published as multipart sets. When you have such a set, you should usually submit one text for each volume, and a "complete" text with the contents of all volumes together.

People who do this often complete and submit one volume at a time, until they've finished, and then contribute the "complete" file.

V.36. I have one physical book, with multiple works in it (like a collection of plays). Should I submit each text separately?

If the works are clearly separate, stand-alone texts, and are long enough [V.17] to warrant inclusion on their own in the archives, then yes, you should, and you _may_ also submit a "complete" version as well, if it seems appropriate. This most commonly happens in a collection of plays, though essays and other works may also fit the criteria. Collections of poetry rarely do, since most poems are too short to submit as stand-alone texts.

Sometimes the book includes a preface or introduction or glossary covering all the works in it. In this case, you can decide whether to include these with each of the parts, or save them for the "complete" version.

V.37. How do I get copyright clearance?

Basically we need to see images of the front and back of the title page of the book, which is where copyright information is usually shown. This is called "TP&V", for "Title Page and Verso" [V.25].

To Submit Online:

As of late 2002, we have a new automated upload procedure using a web page. This is by far the fastest and easiest way to get clearance. You need scanned images (PNG, JPEG, TIFF, GIF), of the two pages, of good enough resolution that the text can be read clearly, though the files don't need to be huge.

Just go to <http://beryl.ils.unc.edu/copy.html> and follow the instructions.

There are two other, older ways to submit a text for clearance.

To submit by paper mail, photocopy the front and back of the title page, even if the back is blank, write your e-mail address on it, and send the photocopies to:

MICHAEL STERN HART
405 WEST ELM STREET
URBANA, IL 61801-3231 USA

This is called Title Page & Verso, or TP&V for short, and is needed for copyright research. A colored envelope is best, to make sure your letter is easily recognized as TP&V.

E-mail Michael hart@pobox.com when you send them, so he knows they're on the way. It's a good idea to check back with him by e-mail after a week or so if you haven't heard from him.

About this, Michael says: "Please include always your e-mail name and address, and mark the envelope with some distinctive mark and or color. Colored envelopes fine. Just something so I can find it easily, the mail here is slow and deep, like snow. Please send a note to: <hart@pobox.com> for more info."

To submit by e-mail, scan the front and back of the title page, even if the back is blank, and e-mail the images to Greg Newby <gbnewby@ils.unc.edu> as TIFF, JPEG or GIF in medium resolution. Make sure that the print is legible before you send.

Whichever method you use, you should expect to get an e-mail back after about a week, with one line containing the Author, Title, your name and date with the word "OK" at the end. This means that your text has been cleared.

A Clearance Line looks something like:

The Works Of Homer [Iliad/Odyssey] Tr. George Chapman Jim Tinsley 06/14/01 ok

If you don't get any response, e-mail to check that your TP&V was received OK. If the word at the end of the line is not "OK", then your text is not eligible, and a comment will probably be appended explaining why it is not eligible.

Don't start work on your book until you get that OK! It's very sickening to do all that work, and then find out that your text can't legally be put on-line!

V.38. I have a two- or three-volume set. Do I have to get a separate clearance on each physical book?

Yes.

Some multi-volume works, notably reference books and translations, were published in a series, and it may be that the first volume is 1922, but the others are 1923 or later, so we have to clear each individually.

V.39. I have one physical book, with multiple works in it (like a collection of plays). Do I have to get a separate clearance for each work?

No. Since they were all printed together, one TP&V will suffice for all, but . . .

You should list each separate title included, if you intend to submit each title separately (see the FAQ "I have one physical book, with multiple works in it like a collection of plays. Should I submit each work separately?" [V.36]). If, say, you clear a "Collected Plays of Sheridan", and later submit an eBook of "The School for Scandal", we will have trouble finding your clearance unless we have made a note that "School for Scandal" is part of the contents of "Collected Plays".

In a case like this, you should include, on your paper or e-mail, something like:

George Bernard Shaw. Plays Unpleasant. 1905.
Contents:
- Preface to Unpleasant Plays
- Widower's Houses
- The Philanderer
- Mrs. Warren's Profession

You only need to do this when you are going to submit each part separately, which is commonly the case with plays, and sometimes essays, stories and novellas. Taking a different example, the "Collected Poems of Emily Dickinson", we would not need to list the contents, since we wouldn't publish each poem separately.

There is one exceptional case: if your book was printed after 1923, but contains stories or plays some of which are stated to be reprints of pre-1923 editions, you should give as much detail as possible about what you intend to submit.

V.40. Who will check up on my progress? When?

Nobody. There are no schedules or timetables. You're welcome to contact other volunteers [V.12] with comments or questions, though.

V.41. How long should it take me to complete a book?

Most books get done in between one and three months, but this varies wildly. It depends on the amount of time you can afford to give it, the length of the book and, if you're not typing, the quality of the scan--if the book scans badly, you need to put more time into proofing.

Some very productive volunteers manage to turn out an e-text a week; some books can take a year or more.

Scanning itself doesn't take too long. Even if it takes you as much as two minutes per page to scan, you will still complete a 300 page book in 10 hours, and you will probably be scanning much faster than that [S.9]. The problem is that the text generated by the scanner and your OCR package is usually faulty. There are many cute scanner errors, mistaking b for h, or e for c, so that "heard" is scanned as "beard" or "ear" as "car". Makes the story more interesting sometimes!

So now you need to do a first proof of the e-text. Read it carefully, correct scanning mistakes, and make sure that you haven't left out pages or got them in the wrong order. Unless your scan was exceptionally good, this is the time-burner in the process.

When you've done the first proof, you can either do a second proof yourself, or send it to another volunteer for second proofing.

If you're a typist, of course, you can skip right over the messy scanning and scan-correction process. Yay typists!!

V.42. I want/don't want my name published on my e-text

No problem. When you send the e-text for posting, mention exactly what, if anything, you want the Credits Line [V.47] to say.

V.43. I'd like to put a copy of my finished e-text, or another Gutenberg text, on my own web page.

Great! PG encourages the widest possible distribution of e-texts. We like to publish everything in plain text, which is the most accessible format, since everybody can read plain text. But once it's available

in plain text, it’s open to you or anyone else to convert it to other formats like HTML for further distribution.

If you are reposting a text, though, please be careful to check that your posting complies with the conditions spelled out in the header, especially for copyrighted works.

V.44. I’ve scanned, edited and proofed my text. How do I find someone to second-proof it?

You can post a request on the Volunteers’ Board, or on the gutvol-d Mailing List. You will probably get some offers there. In a difficult case, you might ask Michael Hart to add it to the "Requests for Assistance" section of the next Newsletter.

In general, the best way to handle it is to make a co-operative proofing project out of it. This is like a miniature version of the distributed proofreading sites, without the page images.

There are always people looking for proofing work, but many beginners take on more than they can handle, and don’t finish the job, and this can be very disappointing if you give the whole thing to one volunteer who then vanishes without trace. You can minimize the risk of this by splitting the book into chunks of about 20-30 pages, or one chapter if that’s around the right size, each. Write explicit instructions about what you want them to do when they spot a suspected error, like fix it or mark it with an asterisk. (Marking is probably safer with beginners who don’t have the book or an image of the page to refer to.) Give the first chapter to the first person who responds, the second to the second, and so on. As you hand out the chapters, let the proofers know that if they’re not returned within three or five days, you’ll assume they’ve quit. Three days is more than plenty of time for 20 pages. If someone returns a chapter, you can give them another. If someone doesn’t get back to you within the time set, assume they’re not going to, and recycle that chapter to someone else. No hard feelings, no problem. This process of "co-operative proofing" ensures that beginning proofers don’t choke on the work, and that one vanishing volunteer doesn’t hold up the whole project.

V.45. I’ve gone over and over my text. I can’t find any more errors, and I’m sick of looking at it. What should I do now?

We all know that feeling! Particularly with your first book, you’ve probably gone through a patch when you thought you’d never finish--and when you do, you can’t stand the idea of looking at it again. Heh. Cheer up--the first twenty texts are the worst! :-) And you’ll feel a lot better when you see your text available for everyone to read.

You have three choices:

You can send it for posting as it is. [V.46]

You can put it aside for week or so, and come back to it with fresh eyes.

You can ask in any of the standard ways [V.12] for someone else to second-proof it for you. This has a lot to recommend it; it gets other sets of eyes looking at the text, it relieves the pressure that you may feel, it may rekindle your enthusiasm for the text, it allows you to "meet" other volunteers, and possibly form partnerships for future PG collaboration. Above all, it gives new proofers a chance to get their feet wet, and this is good for them, and good for PG. You are not only contributing a text, you're helping to train and encourage the next generation of producers.

V.46. Where and how can I send my text for posting?

As of late 2002, we have a new automated upload procedure using a web page. This has a lot of good things going for it, because we keep a record of what's uploaded, you get an e-mailed copy of the notification, you don't have to fiddle with FTP, and we can make up the header automatically from the information you enter, which saves time and prevents keying errors.

As always, it's better to ZIP your file first, because it'll take less time to transfer.

Just go to <http://beryl.ils.unc.edu/cgi-bin/upload>, fill in the form, specify the file to upload, and hit "Send" at the bottom.

And you're done!

If, for some reason, you can't use this page, there are two backup options: you can e-mail it, or you can upload it by FTP. Whichever you use, it is always best to ZIP the file first if you can.

If you are comfortable with sending files by FTP, this is better than e-mail, First, you will need a username and password, which you can get by e-mailing any of the Posting Team.

If you already know how to use command-line FTP, here's how to do it:

Log in to beryl.ils.unc.edu using the username and password supplied and change to the work directory by typing "cd work". Change to binary mode with the "bin" command and "put" your file.

Summary instructions:
ftp beryl.ils.unc.edu
login: yourlogin

password: yourpassword
cd work
bin
put yourfile.ext
quit

Here is a sample session:

>ftp beryl.ils.unc.edu
Connected to beryl.ils.unc.edu.
220-Access from unknown@127.0.0.1 logged.
220 FTP Server
User (beryl.ils.unc.edu:(none)): xxxxxxxx
331 Password required for xxxxxxxx.
Password: xxxxxxxx
230 User xxxxxxxx logged in.
ftp> cd work
250 CWD command successful.
ftp> bin
200 Type set to I.
ftp> put MYFILE.ZIP
200 PORT command successful.
150 Opening BINARY mode data connection for MYFILE.ZIP.
226 Transfer complete.
ftp: 172313 bytes sent in 17.34Seconds 9.94Kbytes/sec.
ftp> quit

When you are in the work directory, you will not be able to list
files, but they _do_ exist and they _are_ there.

When you have uploaded your file, e-mail a note to any or all of the
Posting Team, including your
1. filename
2. credits line as you want it on your text
3. clearance line you received [V.37]

An ideal note might be:

Subject: Beryl upload for posting: Hamlet

I have uploaded to beryl:
Hamlet, by William Shakespeare

File is: hamlet.zip

Credits line is:
Produced by John Doe <jdoe@example.com>

Clearance was given as:
Hamlet William Shakespeare John Doe 05/03/02 ok

If you'd rather send it by e-mail, send the e-mail, including the

Credits Line and Clearance Line as in the sample above, to any or all of the Posting Team, with your text as an attachment. Again, ZIPped is better, since it avoids certain damage that can happen to a plain text e-mail along the way.

Do not add the Project Gutenberg header or footer to your file, unless we specifically asked you to. If you do add it, we'll just have to strip it off again, since we add headers automatically when posting. There are times, perhaps when you're working in an unusual non-editable format, when we may give you a header and ask you to add it, but this is rare.

Please read section "4: Posting" of the FAQ "How does a text get produced?" [V.16] for more detail about what happens in posting. Especially, if you want to draw some peculiarities of this text to the Posting Team's attention, or want feedback on any minor edits done during posting, you should say so in the e-mail you send.

Don't assume that we know anything when you send the e-mail. We don't know what you want us to put on the Credits Line. We don't know that this is an unusual text, and needs some kind of special reformatting. We don't know that the text should be split into two volumes before posting. We don't know that you would really like us to check it closely before posting. You have to tell us, exactly and precisely, what you want on the Credits Line. If the text needs some specific work, you have to tell us exactly what that is. And please do that in your e-mail, not in the text itself. Remember that we could be dealing with five or ten other texts at the same time, and even if the poster you discussed it with two weeks ago is the same one who posts the book, he may not remember.

V.47. What is the "Credits Line"?

The Credits line is a line that the Posting Team can insert into each PG text naming the producer or producers of a particular text.

You should decide what you want on the credits line of your text; it's really not up to us.

Most credits lines are something like:

> Produced by John Doe <jdoe@example.com>.

If you don't want to be mentioned by name at all, just say, in your e-mail:

> Please omit the Credits Line for this text. I want to contribute it anonymously.

If you do want to be mentioned, please give the exact wording you want us to use. Some people want their name only; they don't want us to

include their e-mail addresses. Others want to make their e-mail addresses public so that readers can contact them with comments. That is entirely up to you, but you do need to tell us. If you do want to include your e-mail, remember that having it permanently on the net is a spam-magnet, and we can't effectively remove or change it later.

Occasionally, a Credits Line can spill onto more than one line, for example:

> This text was converted to HTML by Jane Roe <jroe@example.com>
> from an original ASCII text scanned by Jack Went
> and proofed by Jill Hill

V.48. How soon after I send it will my text be posted?

First read the "Posting" section of the FAQ "How does a book get produced?" [V.16] to understand the process.

You should expect some response within three or four days. We try to get to all submissions within that time. In most cases, that response will be simply the official notification that it has been posted. If there is a query on your text, for example if we can't find the copyright clearance or if we have trouble converting or correcting your text, we will probably e-mail you back directly with questions.

If you don't hear from us within four days, send a follow-up e-mail; it could be that your original note never got to us, or just fell through the cracks.

If your file happens to arrive while one of us is logged in and working, it could get posted within the hour. Some frequent contributors who know our habits know just how to time their uploads!

V.49. I found a problem with my posted text. What do I do?

Most postings go smoothly, but problems can happen. Sometimes, one of the servers is down. Sometimes a file gets corrupted for some unknown reason. Sometimes, let's face it, we screw up.

Usually, one of the indexers will tell us about it, but if you catch it first, e-mail whoever sent out your notification e-mail and explain the problem. Don't worry; your original file will be quite safe, since we keep these long after posting them.

V.50. Someone has e-mailed me about my posted text, pointing out errors.

Great!

Since you're the original producer, you're in the best position to decide whether these are real errors. If they're right about it, tell the Posting Team and we'll correct the text.

V.51. Someone has e-mailed me about my posted text, thanking me.

Nice feeling, isn't it? :-)

About Proofing

V.52. What role does proofing play in Project Gutenberg?

A very big one!

Typists' work doesn't usually need many corrections, but unfortunately, scanners and OCR packages are far from perfect, and scanned text varies from "almost-right" down to "maybe I should consider typing instead of scanning". Proofing is the process that turns a scan into a readable e-text.

Proofing a typist's work is straightforward; you just read it, and keep an eye out for mistakes. Typists typically have few mistakes in their texts, but the errors that they do make tend to be hard to spot. Proofing OCRed text has its quirks, and you can expect many, many errors to correct.

The only thing that all proofers agree on is to differ in their methods. Some people scan and almost complete the proofing process within their OCR package, others do no editing at all within their OCR. Some spell-check first, others spell-check last. Some work through in one pass, doggedly line by line, others make several light passes. Some start at the end and work backwards! Some proofers mark all queries with special characters like asterisks (*) in the text, most just make all the obvious changes and mark only the dubious ones. Some people always send their texts out for proofing, others prefer to do it all themselves.

So this guide is not prescriptive; this is not the "only way" to do it. The only rule is that, at the end of the process, your e-text should be as error-free as you can make it, and should conform to Gutenberg's editing standards, which are mostly just common sense guidelines to make readable text.

The aim of this FAQ is to give you an understanding of what text looks

like when it comes fresh off the scanner, and an overview of the whole process by which it becomes a publishable e-text.

V.53. What is Distributed Proofing?

It has always been common for volunteers to share proofing work among themselves--you take the first five chapters, I'll take the next, and so on.

When you're just starting as a PG volunteer, you should go to one of the Distributed Proofing sites [B.4] and do some work there to get a grounding in the basics and a feel for whether you would like to continue working in PG. In distributed proofing, you get a very short section, as little as a page of text at a time, and usually an image file of the page as it scanned. You then make the text match the image. This is a great start, since all you have to do is read, compare and correct. However, other work also needs to be done, and will normally be done by the project managers of these sites. The samples below give you an idea of the whole process, and also some ideas of what proofing a whole book from start to finish is like.

V.54. What do I need to proof an e-text?

You actually need only two things: the e-text itself and a text editor or word-processor that can handle book-sized files and save them as text.

Nearly all word processors and text editors in current use will work. Volunteers use many common programs, including WordPerfect, Microsoft Word, WordPad, DOS EDIT, vi, Brief, Crisp, EditPad, MetaPad, emacs, AbiWord, and the word processors from Open Office abd AppleWorks. And all of these are in actual use by volunteers today. Since all of them contain the necessary basic functions, the best program is the one you're most comfortable with.

Be cautious with recent, powerful word-processors that "auto-correct" text, or use "smart quotes" or any other such automatic retyping or formatting feature, since they can Do Bad Things to your e-text without your consent! When using any such package, it is best to switch off any feature that makes changes without asking you.

Two utilities which may come in useful are a spell-checker and a version difference checker. These may be built into your word processor, or you may have them as separate packages.

A spell-checker is like a chain-saw: a powerful tool, but one to be used very carefully. It is very easy to say "Yes" to the wrong change, and make a really bad mess of the text. Spell-checkers have problems

with proper names, foreign words, archaic usages, and dialects. Incautious use can leave you with a text such as that immortalized in the

> Owed two a Spell in Chequer.
>
> Eye half a spell in chequer,
> It cane with my Pea Sea.
> It plane lee marques four my revue
> Miss steaks eye can knot sea.

Every e-text should pass through a spell-checker at some point, but the human half of the partnership needs a very light hand on the confirmations of change!

A difference checker, such as FC or COMP for MS-DOS, diff for Unix or ExamDiff <http://www.prestosoft.com/examdiff/examdiff.htm> for Windows, may also come in handy. A difference checker compares two versions of the text, and points out the changes. This is important when you've sent a text out for proofing, and you get it back with changes. Rather than re-reading the whole text, you can use a difference checker to highlight the changes so that you can verify them against the printed text. As a proofer, you can use it to compare the original text with what you're sending back to ensure that you've only changed what you meant to change.

V.55. Do I need to have a paper copy of the book I'm proofing?

No.

Your job as proofer is to ensure that the e-text you're working on is readable in itself, and contains no obvious errors. Where you think there might be an error, but you're not sure, you mark the spot in the e-text, and let the volunteer who has the paper book look it up.

V.56. What's the difference between "first proof" and "second proof"?

These are fuzzy terms used to indicate how accurate the e-text is, and what type of work is needed to improve it. Quite commonly, the same volunteer who scans the book proofs the whole thing in one or two passes. Sometimes, given a good scan, the text can be sent out for "first proof" with little or no preparatory fixing-up. Often, the scanner makes quite a lot of corrections, then sends the text out for "second proof".

A text is ready for first proofing when it's obvious that there are plenty of errors, but it's possible to figure out, in almost every case, what the correct text should be without needing to refer to the book.

The objective of first proofing is to eliminate all the obvious errors, so that if you speed-read quickly through the text, you probably won't notice any.

Second proofing involves taking a text that has been first-proofed and correcting all the remaining, more subtle errors. Often, some simple errors such as incorrect spacing and quotes may be left for second proofing. Texts that have been typed instead of scanned will always be of at least second-proof quality.

V.57. What do I do with an e-text sent to me for proofing?

First, establish reasonable expectations. A typical book takes 10-15 hours of concentrated effort, and when you first start, you're climbing a learning curve. For your first session, decide to mark out a chapter or two--something like 500 to 1,000 lines--and work only on that. If you get through 1,000 lines in your first sitting, you have done extremely well! It's a good idea to send this first 1,000 lines or so back immediately. The volunteer who sent you the e-text will comment on it, and let you know about any style guidelines you may have breached or common errors you may have missed. Most beginning proofers do make mistakes, so don't worry about it--it's easier to correct these in 1,000 lines than to go back over them in 15,000 lines!

You will usually receive the e-text as an attachment to your e-mail. It's better to send e-texts as attachments than to paste them as text into the body of the e-mail to make sure that the text isn't changed by different e-mail clients. It's better to send e-mailed attachments as ZIP files [R.20], since e-mails sent as text can be damaged along the way. But whether you receive a TXT file or a ZIP file that you have to open, you should save the .TXT file to your hard disk and open it with your editor.

It may be that the text you see appears double-spaced--every second line is blank--or that all the text is on one incredibly long line. This is a familiar effect when moving between a DOS/Windows computer and a Mac or Unix system, but it can happen between any two editors. It is caused by the use of different characters to mark the end of a line. If you have this problem, ask whoever sent you the text to re-send it, telling them what kind of computer and editor you have.

Now you make any changes that obviously need to be made, and mark any places where the text looks wrong, but you're not sure what the right text should be. You can usually use asterisks (*) to mark these dubious spots, but you might use other characters if the text already contains asterisks. When in doubt, mark them all, and let the volunteer with the text sort them out!

It is usually best not to make global changes to line lengths by

reformatting lots of paragraphs, since the person who sent you the e-text may want to use a difference checker when you return it, and changed line-lengths throughout mean that every line will be different.

When working on a long text, or when making a lot of changes, it may be wise to save several versions of the text with different filenames at different stages so that if something goes badly wrong, you can revert to the last good version. This applies especially to saving the text just before performing a spell-check.

When you're finished with the e-text, make sure you save it as a plain text file (.TXT) and send it back by zipping it if you can, and attaching it to an e-mail.

V.58. What kinds of errors will I have to correct?

Each text has its own peculiarities, but there are a number of well-known scanning errors you will be dealing with all the time.

Punctuation is always a problem. Periods, commas and semi-colons are often confused, as are colons and semi-colons. There are also usually a number of extra or missing spaces in the e-text.

The problem of quotes can assume nightmarish proportions in a text which contains a lot of dialog, particularly when single and double quotes are nested.

The numeral 1, the lower-case letter l, the exclamation mark ! and the capital I are routinely confused, and often, single or double quotes may be mistaken for one of these.

Lower-case m is often mistaken for rn or ni.

The letters h and b and e and c are commonly mis-read, and these are probably the hardest of all to catch, since ear/car, eat/cat, he/be, hear/bear, heard/beard are all common words which no spell-checker will flag as problems.

For example:

" Hello1' caIled jirnmy breczily. 11Anyone home ? "

There seemed to he no-oneabout. Only tbe eat beard him."

should read:

"Hello!" called Jimmy breezily, "Anyone home?"

There seemed to be no-one about. Only the cat heard him.

As well as scanner errors, which affect one letter at a time, you have to keep an eye out for editing mistakes by the volunteer who scanned the text or by previous proofers. These are typically cases where a whole line, paragraph or page has been omitted or misplaced. They show up as sentences that don't make sense, or paragraphs that don't follow from the previous one.

This means that you have to keep reading the flow of the text, so that you can spot context errors as well as typos.

V.59. How long does it take to proof an e-text?

This depends on how long the e-text is, how clean the text is when you start, and how thorough you're being, as well as how much time per day you can give it and how fast you can proof.

On a first proof, it can take a very long time to get the e-text to a readable condition if it scanned badly. As a beginner, you would be unlikely to be given such a difficult text to work with. First proofs are usually done by the same person who did the scanning, and are only given out in the context of established scanning/proofing teams.

You might expect to proof anywhere between 500 and 2,000 lines per hour during a second proof. A short novel or novella might have as few as 6,000 or 7,000 lines; War and Peace weighs in at about 54,000 lines. Most novels run to 10,000 to 15,000 lines. So you might spend anything between 5 and 30 hours second-proofing a standard book, with 10 to 15 hours being typical.

For an average novel, a week or two for second proofing is good going. A month is reasonable.

Proofing an e-text is a significant amount of work, and you may find it psychologically more comfortable to take on a chunk at a time--say 1,000 lines per session--and send that proofed section back, rather than wait until the whole job is done before sending anything back. This helps to avoid the fairly common case where you keep falling behind where you expect to be until you dread the thought of getting back to the text, and finally just abandon it.

If you find after a while that you just don't want to continue, please tell the person who sent you the text that you're not going ahead with it. It's very frustrating for the volunteer who scanned the book, and who wants to get it posted, to wait for two or three months, only to have to start all over again with another proofer.

V.60. Are there any special techniques for proofing?

The classic way to proof is to open the text in your editor or word

processor, and just start reading carefully.

This method has received a major boost since editors and word processors have added a feature of showing squiggly red underlines under words not in their dictionary. While this is very useful, you still need to read carefully, since not all errors produce misspelled words. The classic, and very common, example of this is scanning "he" for "be". These visual spellchecks also commonly do not check words beginning with capitals. Capitalized words are commonly names not in the dictionary, and when checking of capitalized words is switched off, they will not query "Tbe". Other errors that a spellchecker doesn't look for include missing spaces, mismatched quotes and misplaced punctuation. For these, you can try gutcheck [P.1]. And of course, no automatic check will find omitted lines or words. Worse, spellcheckers will query words not in their dictionary that might be quite correct, and this can be quite troublesome when dealing with older texts or dialect.

Still, if your concentration is up to the job, scrolling through a text with non-dictionary words underlined in red is a fast and effective way of giving a text the final once-over.

Volunteers have also used other techniques for proofing. Some people can't sit at their screen and read for hours; many people don't want to.

Some people just use the good old-fashioned method of printing out the text to be proofed, and blue-pencilling the mistakes.

It is becoming fairly common now for people to load the text onto their PDA, and read it from that. Mistakes found can be bookmarked or jotted down and fixed when they go back to their PC.

Getting your computer to read the text aloud is a very effective way of achieving high accuracy. Modern PCs have audio capabilities built in, and it is possible to find free or cheap shareware "read-aloud" text-to-speech packages for just about everything. Some PDAs are also capable of doing text-to-speech.

The first time you try text-to-speech, it will probably sound and feel a little strange, but you will quickly learn to _hear_ errors in words. This can be very effective, but you should have given the text at least a light proofing before you begin; it is hard to deal with a high number of errors using a text-to-speech method.

When proofing by a speech program, you either set your text-to-speech program to pronounce all punctuation, or, if that is not possible, you make a special version of your text to feed it, first doing a global replace of "," with " comma ", ";" with " semi-colon ", and so on. Mark a block of 500 to 1,000 lines for reading aloud, and set the reading speed to whatever is comfortable for you. Then you sit down with the original book in front of you, and listen. When you hear an error, mark the place in the text with a light pencil. Stopping the

reading at every error, editing the text and restarting is possible, but it breaks the flow, and ends up taking longer. When the reading is done, go to your keyboard and correct the errors found.

V.61. What actually happens during a proof?

Stage One--The original Scan

We start with a scanned e-text, in this case a paragraph from The Odyssey. The paragraph used as an example here has been "enhanced" with more errors than in the real scanned text, so that you can see samples of many problems all in one place.

We begin by looking at the original OCRed text, of which our sample section reads:

```
 1There Periniedes and Eurylochus held the victims, but l
    drew my sharp sword from my thigh, and dug a pit, as it were
    a cubit in length and breadth, and about it poured a drink-
    offering to all the dead, first with mead and thereafter with
    sweet wine, and for the third time with water, And 1 sprink-
 BOOK XL
ODYSSEY X,  24-56.
        173

                  ODYSS.EY XI, %4-56.                  173
 lef white incal thereon, and entreated with many prayers
  strengthless beads of the dead, and prornised that on my
 return to Ithaea 1 would offer in my halls a barren heifer,
 the best 1 had, and fil  the pyre with treasure, and apart unto
 Teiresias alone sacrifice a black rarn without spot, the fairest
 of my flock. But when 1 bad hesought  the tribes of the
    d with vows and prayers, 1 took the sheep and cut their
       s over the trench. and the dark blood flowed forth,
          he spirits of the dead that he departed gathered
      from out of Erebus.
```

It’s clear that we should tidy up the page headings and numbers that have been scanned in with the main text, and that we should separate the paragraphs and remove the spaces inserted by the scan at the start of some lines. We also need to restore some of the text that got lost in the scan. Since there isn’t much of it, we just type it in. Having done this, we get to . . .

Stage Two--First pass through the scanned text

At this point, we have a complete text. All of the words are actually there, and we have eliminated page breaks and other extraneous artifacts of proofing. Again, mileage varies: some people like to preserve page breaks and numbering until much later, to make it easy

to refer back from the e-text to the book.

Our job in this phase is to fix all of the obvious scanning errors and double-check that we really do have all the text. Our aim here is to create an e-text that is ready for First Proof. In fact, since it's fairly clear what all the words are, this text could be considered ready for first proof.

1There Periniedes and Eurylochus held the victims, but l drew my sharp sword from my thigh, and dug a pit, as it were a cubit in length and breadth, and about it poured a drink-offering to all the dead, first with mead and there after with sweet wine, and for the third time with water. And 1 sprink-led white incal thereon, and entreated with many prayers the strengthless beads of the dead, and prornised that on my return to Ithaea 1 would offer in my halls a barren heifer, the best 1 had, and fill the pyre with treasure, and apart unto Teiresias alone sacrifice a black rarn without spot, the fairest of my flock. But when 1 bad besought the tribes of the dead with vows and prayers, 1 took the sheep and cut their throats over the trench. and the dark blood flowed forth, and lo, the spirits of the dead that he departed gathered them from out of Erebus.

Now we convert those numeral 1s to capital Is and to quotes, where appropriate, we straighten up the quotes and we deal with other obvious scanning errors, which brings us to . . .

Stage Three--The First Proof

At this point, we could hand over the text to an experienced proofer who doesn't have a copy of the book. This would be called a "first proof". An e-text is at first proof stage when there are still plenty of errors, but in each case it's pretty obvious what the correct word is. The excerpt now looks like normal text.

Unfortunately, in stage two above, we accidentally deleted a line.

'There Periniedes and Eurylochus held the victims, but I drew my sharp sword from my thigh, and dug a pit, as it were a cubit in length and breadth, and about it poured a drink-offering to all the dead, first with mead and there after with sweet wine, and for the third time with water. And I sprink-led white incal thereon, and entreated with many prayers the strengthless beads of the dead, and prornised that on my return to Ithaea I would offer in my halls a barren heifer, Teiresias alone sacrifice a black rarn without spot, the fairest of my flock. But when I bad besought the tribes of the dead with vows and prayers, I took the sheep and cut their throats over the trench, and the dark blood flowed forth, and lo, the spirits of the dead that he departed gathered them from out of Erebus.

Stage Four--Corrections from First Proof

We receive the first proof back from the proofer, and find that it has been mostly corrected.

The corrections made were "l/I", "there after/thereafter", "prornised/promised", "bad/had", and "rarn/ram".

We have also wrapped the lines--at 60 characters in this case, but it is commonly as much as 70 characters per line. Sentences which look wrong, but where it isn't clear what the right text should be, have been marked with asterisks (*).

'There Periniedes and Eurylochus held the victims, but I drew my sharp sword from my thigh, and dug a pit, as it were a cubit in length and breadth, and about it poured a drink-offering to all the dead, first with mead and thereafter with sweet wine, and for the third time with water. And I sprinkled white incal * thereon, and entreated with many prayers the strengthless beads of the dead, and promised that on my return to Ithaea I would offer in my halls a barren heifer, * Teiresias alone sacrifice a black ram without spot, the fairest of my flock. But when I had besought the tribes of the dead with vows and prayers, I took the sheep and cut their throats over the trench, and the dark blood flowed forth, and lo, the spirits of the dead that he departed gathered them from out of Erebus.

We look up the text where the first proofer has asterisked it, and make the corrections.

The text is now ready for second proofing. An e-text is ready for second proofing when you can skim through the text without noticing that there are errors.

We can either do a second proof ourselves, or send it out for second proofing.

Second proofing involves a very careful reading of the text, looking for small errors. In some ways, it's much harder than first proofing, since it's very easy to let your eyes run on auto-pilot and in doing so, miss subtle errors.

Having performed the second proof, which caught errors like "beads/heads", "Ithaea/Ithaca", "Periniedes/Perimedes" and "he/be", we now have our final e-text.

'There Perimedes and Eurylochus held the victims, but I drew my sharp sword from my thigh, and dug a pit, as it were a cubit in length and breadth, and about it poured a

drink-offering to all the dead, first with mead and
thereafter with sweet wine, and for the third time with
water. And I sprinkled white meal thereon, and entreated
with many prayers the strengthless heads of the dead, and
promised that on my return to Ithaca I would offer in my
halls a barren heifer, the best I had, and fill the pyre
with treasure, and apart unto Teiresias alone sacrifice a
black ram without spot, the fairest of my flock. But when I
had besought the tribes of the dead with vows and prayers,
I took the sheep and cut their throats over the trench, and
the dark blood flowed forth, and lo, the spirits of the
dead that be departed gathered them from out of Erebus.

Hooray! At long last we have an e-text to post, which can be downloaded, read and enjoyed by anyone in the world from now on.

About Net searching:

V.62. I've found an eligible text elsewhere on the Net, but it's not in the PG archives. Can I just submit it to PG?

You can submit it, but you can't "just" submit it.

We wish we could give a permanent home to all the etexts that people have produced and placed on the Net, but without proof of their public domain [C.10] status, we can't.

We need to be able to prove that the eBooks we publish are in the public domain, so, in order to use one of the many texts that are just floating around the Net, you need to find a matching paper edition that we can prove is eligible [V.18].

(By the way, please be sure that it isn't already in the PG archive. A lot of texts circulating on the Net originated at PG, and people quite often submit them back to us.)

Before you get into this, you should check whether the text you have found is likely to be in the public domain in the U.S. A quick way to verify this is to hit the Library of Congress Catalog site at <http://catalog.loc.gov> and search for the title or author. If you find no publications before 1923, then you should probably move on; the Library of Congress doesn't list every book, and in particular doesn't list all books published outside the U.S., but, if there isn't a pre-1923 copy there, it may be difficult to follow up on. If you're not dissuaded, do a search on the Net for used book shops that might have pre-1923 copies.

Sometimes, with a text on the Net, you know who typed it; it's on someone's website, or the transcriber is named in the text. Sometimes, the text has just been floating around Usenet or old gopher sites for years, with no attribution.

The first thing to remember is that we would like to give credit to the original transcriber if they want it, and if we can identify them.

The next thing to consider is that the original transcriber may well have an eligible copy of the book, and may be able to provide TP&V [V.25] for it.

So, if you can locate the original transcriber, it makes sense to e-mail them, explain what you propose to do, and ask them whether they can help with copyright clearance and whether they would like to be credited in the PG edition. Often, you will get no response, or a response but no prospect of material that will help with clearance, but sometimes you will get lucky.

If the transcriber can't help with TP&V, it's up to you to find a matching paper edition of the same book. This may not be as hard as it sounds. Libraries can help, and may get editions for you on interlibrary loan.

This is an ideal way for students, academics and librarians to contribute texts to PG, since you probably have access to a good library with stocks of old books to find matching paper editions.

If you find a matching paper edition, you then need to compare the etext you found with the book. Legally, what we're trying to prove here is that we have done "due diligence"--that we have done our best to prove that the etext is indeed a copy of a public domain work.

The minimum "due diligence" we can perform is to compare the first and last pages of each chapter, (or every 20 pages where the book is not neatly divided into chapters of about that size). You should list all of the differences between the book and the etext that you find on those pages. It is to be expected that there will be some minor differences of punctuation, spacing and spelling, and even perhaps of wording. Minor differences are OK, but we do need to list them, to prove that we did the comparison. When you have your lists, you can send in the TP&V as normal, accompanied by your lists, for clearance.

Many texts floating round without attribution, and indeed many with attribution, could do with a thorough checking, and another option you have is "comparative retyping", where you go through the whole etext, proofing it carefully against the cleared paper book, and changing everything that is different in the etext to match the paper edition. If you do this, you don't need to produce a list of differences, since there won't be any by the time you've finished; you can just submit it as a normal text--_and_ it may well be a lot cleaner! However, if you do take this path, please do a very thorough job on the proofing and comparison.

If the etext you find has been marked up, in HTML for example, you should remove all HTML for the PG edition, because, even though the text itself has been proved to be in the public domain, the original transcribers may hold copyright on the HTML markup, even if you can't find them. If you do want to make a HTML edition of it for PG, strip out all of the original markup and then re-add your own markup.

If you do find the producer and he or she wants to be identified, you may submit a double credits line like:

Transcribed by Sally Wright <theoriginaltranscriber@example.com>
Produced for PG by You <you@example.com>

V.63. I've found an eligible text elsewhere on the Net, but it's not in the PG archives. Why should I submit it to PG?

The first reason is file safety.

Yes, we accept that the file is already available to everyone today, but it may not be safe in the long term. We've seen college students who put books on their personal site, and then lose that site when they graduate. We've seen individuals who transcribe several books, and later lose interest, or move, or die, and the work they've done is lost. We've seen small projects with a few volunteers who produce and post books for a few years, but then break up or run out of funds to maintain their site. We've seen large institutions drop their collections as part of a cost-cutting exercise. We've even seen organizations lock public domain works up behind licenses, requiring users to commit to registration and a "no copying" agreement before downloading them.

Whenever a set of etexts is published and distributed by only one person or organization, there is a danger that their etexts will disappear from the Net sometime. We want _all_ etexts to be spread as widely as possible, copied as much as possible, so that no one event or loss, or whim of a sponsor, can obliterate them.

We think that the PG collection is, for that reason, the safest place to put a text for its long-term survival. There are copies of the PG archives all over the world, on public servers and private CDs. PG publications are widely converted, collected and read on PDAs. Other text projects copy works from PG.

The PG archive is so valuable, yet free and easily portable, that even if every current PG volunteer vanished overnight, people around the world would copy and preserve it. Even if PG itself decided to withdraw all our texts, we couldn't do it, because so many people have made copies.

The second reason is legal safety.

Unlike some other projects and individual efforts, PG retains documentary proof of the public domain status of its texts. This is more valuable than it might appear at first glance.

Publishers often claim a new copyright [C.17] on works that they republish, and as time goes on, it becomes harder and harder to prove that a particular book is in the public domain. Walk into your local bookstore and check out how many works by Shakespeare, Poe, Dickens, and Twain have copyright notices on them! People who want to translate these, or create derivative works like screenplays or lyrics or films must first prove that they are basing their work on a public domain edition, but the creeping copyright practices of commercial publishers make that difficult.

Here's a practical example: we were approached by a film student who wanted to make a short piece based on characters from James Joyce's "Ulysses". But before he could do that, he needed to confirm that the material on which he was basing his movie was in the public domain, and all the editions he could find were copyrighted. However, because PG had already established the public domain status of Ulysses, we could point him to our established PD version, and even tell him where to find a paper copy published in 1922. Without that evidence, he could not have made his project.

V.64. I have already scanned or typed a book; it's on my web site.
How can I get it included in the Gutenberg archives?

Great! We get these a lot, but it's always nice to see another!

You need to send us the TP&V [V.25] so that we can prove that your edition is in the public domain. If you don't have the TP&V, you will need to find a matching paper book with eligible TP&V for us to be able to use it.

V.65. I have already scanned or typed a book; it's on my web site.
The world can already access it. Why should I add it to the Gutenberg archives?

The Project Gutenberg archives are widely copied and searched, and much safer and more permanent that any individual website can possibly be. We aim to keep this collection together over not just years, but centuries. You took the trouble to transcribe this book. We can relate; that's what _we_ do, as well. We know you want this work to survive you and your ISP, and we believe we can do that. And it's not as if you have to take it off your website when we make a copy; you're just using your candle to light another!

If you want to let readers know that your site has other related

material, you can put that information in the Credits Line [V.47]. Taking a real-world example, you could ask us to add this to the Credits line for a C. M. Yonge text:

A web page for Charlotte M. Yonge will be found at www.menorot.com/cmyonge.htm

V.66. I have already scanned or typed a book, but it's not in plain text format. Can I submit it to PG?

Yes, of course. We'll be happy to discuss format options with you, and we're quite experienced in converting between multiple formats and deciding which formats work best and will have the longest life. All you need is to get us a copy of your TP&V [V.25].

About author-submitted eBooks:

V.67. I've written a book. Will PG publish it?

Maybe.

PG gets submissions from young people, for example, who just want to get a story they wrote published in PG. We wish them well with their writing, but that's not really why we're here.

If you are a published author, or perhaps an academic who wants to put a textbook into the archives, it's quite likely that we will publish it.

V.68. I have translated a classic book from one language to another. Will PG publish my translation?

Yes, if we can.

The book that you translated needs to be in the public domain, and we will need the same proof of eligibility that we would use if you were contributing the book in its original language.

For example, if you were translating Hesse's Siddhartha (published pre-1923 in German, but no pre-1923 English translation available), we would need to copyright clear [V.25] the original German edition from which you worked--it needs to be a pre-1923 or otherwise public domain edition. (We actually did this one, thanks to the hard work and scholarship of some volunteers.)

V.69. OK, this is one of the cases where PG will publish it.
What do I do next?

You need to decide about copyright issues. Do you want to release your work to the public domain, or do you want to retain copyright? If you want to retain copyright, what terms do you want to release it under? The next few questions deal with those issues.

Having decided that you want PG to publish it, and decided what restrictions (if any) you want to place on further distribution, you just need to write the appropriate letter and send the text to us.
[V.46]

V.70. I hold the copyright on a book. Can I release it to the public domain?

You can. All you need to do is put a statement into the released version of the text saying that you have.

If you want to release it into the public domain and distribute it through Project Gutenberg, you should send us a letter to that effect.

To: Michael S. Hart
Founder, Project Gutenberg
405 West Elm Street
Urbana IL, 61801-3231, USA

Dear Project Gutenberg:

I am the sole copyright holder for the book, "Wallaby Happiness." It gives me pleasure to release this work into the public domain, and I invite Project Gutenberg to publish this public domain edition.

Sincerely,

Gregory B. Newby

Once you have released it into the public domain, neither we nor anyone else needs your permission to publish it, but for us to be sure that it _is_ a public domain version, we do need a signed letter.

V.71. I hold the copyright on a book. Do I have to release the book into the public domain for Project Gutenberg to publish it?

Absolutely not! For example, many contributors of copyrighted material want to share it with the world, but do not want it commercially republished by other companies.

You can grant Project Gutenberg perpetual, non-exclusive, world-wide

rights to distribute your book on a royalty-free basis by sending a letter to Michael Hart. Your letter may be brief, but must be signed, and must include the name of the book and the assertion that you are the copyright holder or the agent for the copyright holder.

If you want some related information, like a link to your website, included in the text, we will be happy to oblige.

Once we have posted a text, many people will copy it. We have no effective mechanism for "recalling" texts that we have posted, so please be sure, before you commit to this, that you intend to follow through with it, because there is no way to change your mind later.

Here is a sample letter, including the address to send it to:

To: Michael S. Hart
Founder, Project Gutenberg
405 West Elm Street
Urbana IL, 61801-3231, USA

Dear Project Gutenberg:

I am the sole copyright holder for the book, "Wallaby Happiness." It gives me pleasure to grant Project Gutenberg perpetual, worldwide, non-exclusive rights to distribute this book in electronic form through Project Gutenberg Web sites, CDs or other current and future formats. No royalties are due for these rights.

Sincerely,

Gregory B. Newby

V.72. I hold the copyright on a book, and would like Project Gutenberg to publish it. Can I choose what rights to assign?

For PG to be in a position to copy it, we do need perpetual, worldwide, non-exclusive, royalty-free rights to distribute the book in electronic form. What rights you choose to assign to readers after that is a decision for you to make.

The Creative Commons site <http://www.creativecommons.org> may give you some ideas of what practical use you can make of your copyright to see that the work is used in the ways you intended.

About what goes into the texts:

V.73. Why does PG format texts the way it does?

PG texts are formatted as plain ASCII, with 60-70 characters per line, with a hard return [CR/LF] at end of line, and some people ask "Why do it _this_ way? You could omit the hard returns and let the reader's word processor or Reader software wrap the lines. You could use "8-bit" accented characters for non-English characters." "You could use ' - ' instead of '--' for an em-dash." And so on, through a different choice we could make for every formatting feature. And the answer, of course, is that we _could_ do it differently, and sometimes we do, but mostly we keep to one consistent style.

We'll be discussing each of the formatting decisions below, not only giving the summary PG answer, but also discussing the plusses and minuses of each, and the possible options.

Like any question beginning "Why does/doesn't PG . . . ?", the answer is "Because that's what the volunteers and readers want!". These conventions have been worked out over the years, largely by Michael Hart, our founder and chief volunteer, in conjunction with all of us volunteers, as the result of feedback from readers.

We are guided throughout by the principle that we want to produce texts in the simplest format that will adequately express the content. Quoting Michael Hart (1994):

Etext as developed and distributed by Project Gutenberg since 1971 was never intended to be a copy of a paper or a parchment [remember, first Project Gutenberg Etext was typed in from parchment replicas of the US Declaration of Independence].

The major purposes of Project Gutenberg have always been:

1. to encourage the creation and distribution of electronic texts for the general audience.

2. to provide these Etexts in a manner available to everyone in terms of price and accessibility [i.e. no special hardware or software], and no price tag attached to the Etexts themselves.

3. to make the Etexts as readily usable as possible, with no forms or other paperwork required, and as easily readable to the human eyes as to computer programs, and in fact, more readable than paper.

There is sometimes a conflict between "simplest format" and "adequately express the content"; further, different people have different views on what is "simple" or "adequate". You, the producer of the text, have spent the time and effort to make the eBook available to the world, you have thought more about it than anyone else, and we respect your informed judgment. However, please make sure that your judgment _has_ been informed, by studying the

precedents and reasons behind our guidelines.

Where a simple, standard PG-ASCII layout does not, in your view, "adequately express the content", you should think of making your text in another open format, perhaps HTML or XML or TeX, that allows you to use more characters, more formatting options, and images. We are always happy to accept these kinds of files. In these cases, you should also provide a standard PG-ASCII version, even if you feel it is unacceptably degraded, for those who cannot use your preferred format.

Just ten years ago, presentation as plain ASCII was not only a universal standard, it was effectively the only way that most people could view the books. The first version of the HTML specification had been drafted, but was unknown among the general public. XML did not exist. SGML was (as it still is) the province of specialists. Specialized eBook readers and PDAs had not yet appeared.

In 2002, plain vanilla ASCII is still readable everywhere, but people also want to convert our texts into other formats for more convenient loading on readers and web sites. We therefore have to keep in mind that our works will be processed by automatic conversion programs, none of which is perfect, and we have evolved some "defensive formatting" practices, which, while retaining the universality of plain text, also supply clues to automatic converters about how they should treat the layout. These do help to keep converters from making at least the worst mistakes. The most significant "defensive formatting" practices are indenting unwrappable text like quotations, and using _underscores_ rather than CAPITALS for italics. Different volunteers have different priorities: at one extreme, some people want to make the best plain text they can, giving no weight to conversion issues; at the other, some people emphasize the cues that will allow automatic reformatters to convert the texts well, even if that causes some ugliness in the plain text. Most of us operate somewhere between, making the choices we feel are best depending on the context. Getting a text on-line is the important thing; which choices you make in doing so is a matter of detail.

About the characters you use:

V.74. What characters can I use?

a) You should use plain ASCII for straight English texts.

b) When producing a text partly or completely in a language that requires accents, you should use the appropriate ISO-8859 character set for the language, and specify which you are using, and also

provide a 7-bit plain ASCII version with the accents stripped.

c) When producing a text in a language that doesn't use one of the ISO-8859 character sets, you should use the encoding most commonly used for that language. [e.g. Chinese--Big 5]

d) When producing a text containing more characters than can be found in any one of the ISO-8859 character sets, you should use Unicode.

You should use plain ASCII wherever possible--that is, the letters and numbers and punctuation available on a standard U.S. keyboard, without accented letters. The immediate and major exception to this is when you are typing a text written in a language like French or German that requires accents.

There is a problem with using non-ASCII characters. They do not display consistently on all computers; in fact, they do not even display consistently on the same computer! On my computer, for example, what looks like an e-acute in this editor just shows as a black box in another editor, or even using a different font in the same editor. And this is by no means confined to some theoretical minority; we have to deal with it all the time when posting texts.

Further, standards are changing: ten years ago, the character set Codepage 850 [MS-DOS] was very common; now it's rare except in some texts that have survived those ten years.

We want to preserve these texts over _centuries_, not just decades, and at the moment there is no single clear standard that we can use across all texts. Unicode may perhaps be a future standard, but, right now, it's not something that people use every day, and it's not supported by a lot of common software.

ASCII, while limited, is supported by almost all computers everywhere, so we make a point of always supplying an ASCII version where possible, even if the ASCII version is degraded when compared to the 8-bit original. When we get a text in, say, German, we post two versions of it--one with accents and one without.

V.75. What is ASCII?

Don't get scared by the computer jargon; ASCII (pronounced ASS-key) is just a name for the set of unaccented letters, numbers and other symbols on a standard U.S. keyboard.

ASCII (American Standard Code for Information Interchange) is a set of common characters, including just about everything that you can type in on an English-language keyboard. It includes the letters A-Z, a-z, space, numbers, punctuation and some basic symbols. Every character in this document is an ASCII character, and each character is identified with a number from 0 through 127 internally in the computer.

Just about every computer in the world can show ASCII characters correctly, which makes it ideal for PG's purpose of providing texts that can be read by anyone, anywhere, but ASCII does not include accented characters, Greek letters, Arabic script and other non-English characters, which causes some problems when we produce texts that need non-ASCII characters.

V.76. So what is ISO-8859? What is Codepage 437? What is Codepage 1252? What is MacRoman?

Today's computers mostly work on the basis of dealing with one "byte" at a time. A byte is a unit of storage than can contain any number from 0 through 255--256 values in all. It's very convenient for computers to associate one character with each of these numbers, so that we can have up to 256 "letters" viewable from the values stored in one byte. The first 128 values, zero through 127, are defined by ASCII--so, for example, in ASCII, the number 65 represents a capital "A", 97 represents a lowercase "a", 49 stands for the digit "1", 45 for the hyphen "-", and so on.

ASCII doesn't define characters for the values 128 through 255, and in early days computer manufacturers used these values to hold non-ASCII characters like accented letters and box-drawing lines. Of course, 128 wasn't nearly enough values to hold all of the characters that people needed to use for different languages, so they made the character sets switchable, so that a PC in France could use a different set of accented letters from a PC in Poland. Microsoft's version of this was called Codepages. Each Codepage held a different set of non-ASCII characters. Codepage 437, and later Codepage 850, were commonly used for English and some major Western European languages on MS-DOS.

MacRoman was Apple's first codepage, containing most of the accented letters in Latin-derived languages, and MacRoman is still in common use on Apple Macs today.

Later, the International Standards Organization ISO got around to looking at the problem, and defined ISO-8859-1, ISO-8859-2 and so on, as the standards for different language groups. These sets all define the characters 160 through 255 as accented letters and other symbols, and define the 32 characters from 128 through 159 as control characters.

Since Microsoft Windows has no use for the control characters 128 through 159, Windows fonts commonly use Codepage 1252, which has ASCII in the first 128 characters, ISO-8859-1 in characters 160 through 255, and other symbols in the characters 128 through 159. Just to make an already chaotic system worse, all characters can be defined differently in different fonts!

Of course, most of these codepages are incompatible with each other. For example, the byte value 232 shows as a lower-case "e" with a grave

accent in ISO-8859-1 and CP1252, a capital letter "E" with diaeresis in MacRoman, a Latin capital letter "Thorn" in CP850, a Cyrillic lower-case "Sha" in ISO-8859-5, a Greek capital letter "Phi" in CP437, and so on. So if you view a text intended for one of these character sets with a program that assumes a different character set, you see gibberish.

The good news, for mostly-English texts at least, is that ISO-8859-1, Codepage 1252 and Unicode agree on the numerical values of the accented characters and symbols to be represented by the values 160 through 255. And everybody accepts ASCII--a pure ASCII file is valid ISO-8859-anything, valid Codepage-anything, and valid Unicode UTF-8.

For more detail about the mappings between Unicode and other formats, you can view Unicode<-->ISO-8859 mappings at
ftp://ftp.unicode.org/Public/MAPPINGS/ISO8859/
Unicode<-->Windows mappings at
ftp://ftp.unicode.org/Public/MAPPINGS/VENDORS/MICSFT/
and Unicode<-->Apple mappings at
ftp://ftp.unicode.org/Public/MAPPINGS/VENDORS/APPLE/

If you're not confused enough by now, please read the excellent guide to the whole "alphabet soup" problem at <http://czyborra.com>.

V.77. What is Unicode?

Recognizing that no single set of 256 characters can hold all of the symbols necessary for true multi-lingual texts, ISO 10646 was created. This defined the Universal Character Set (UCS) using 31 bits, which has the potential for a staggering _2 billion_ characters.

The Unicode Consortium is a group of computer industry companies who agree the Unicode standard. Unicode accepts the ISO 10646 standards, and adds some restrictions and implementation processes. It plans for a modest million or so characters; however, this is enough for all living and extinct languages, and imaginable future ones too.

Using 4 bytes for each character is wasteful, though, when most characters need only one or two, and there are programming problems with implementing 4-byte characters, so Unicode provides Transformation Formats (UTF) which allow the characters to be encoded using fewer bytes where possible. UTF-8 and UTF-16 are common.

UTF-8, which is the most practical of these from the PG point of view, allows ASCII to be encoded normally, and usually uses two or three bytes for other non-ASCII characters.

Because of the extra work needed to support this extra space, and the fact that most people work mostly in one or maybe two languages, Unicode is being adopted only slowly, and most computer programs in 2002 do not

fully support it. But when you need to mix Arabic, Greek, Ogham and Sanskrit in one text, it’s the only possible answer!

For more about this, go straight to the source at <http://www.unicode.org>.

V.78. What is Big-5?

Big 5 is an encoding of a set of 13,000+ traditional Chinese characters.

V.79. What are "8-bit" and "7-bit" texts?

For practical purposes, 7-bit texts are plain ASCII; 8-bit texts have accented letters.

This comes from computer jargon. You can represent the 128 characters of ASCII using 7 bits--binary digits--but to represent the 256 characters needed for the various codepages and ISO-8859 standards, like accented letters, you need 8 bits. Hence, we call a text that uses non-ASCII characters in a character set like Codepage 850 or ISO-8859-1 an "8-bit" text.

When we post a text as both 8-bit and 7-bit, as we do when ASCII is not enough to render the text acceptably, we name the file with an "8" or a "7" at the start. So, for example, Crime and Punishment by Dostoevsky is named 8crmp10 for the 8-bit version with accents, and 7crmp10 for the 7-bit version without accents.

See also FAQ [R.35]: "What do the filenames of the texts mean?"

V.80. I have an English text with some quotations from a language that needs accents--what should I do about the accents?

If stripping the accents would unacceptably degrade the book, then submit two versions, one "8-bit" with the accents included and one "7-bit" plain ASCII, and we will post both.

This is a hard choice. What constitutes "unacceptable degradation"?

Clearly this is a decision that all of us in PG have to make. It’s a very common problem, and different people have different views. For that matter, different print publishers have different views; you will see the words "debris", "facade" and "cafe" printed with and without accents in different books, and even in different editions of the same book.

We don’t want to post two versions when we don’t have to. It doubles

the posting work, doubles the disk space needed, potentially confuses downloaders, doubles the maintenance when we need to correct the text. On the other hand, we don't want to degrade the text.

There is no clear line, no definitive answer to what level of degradation is acceptable. Most producers feel that there is no point in making a separate version when dealing only with a few foreign words thrown in among the English, but when, for example, some significant dialog between the characters is in French or Spanish, it's harder to say that stripping the accents is acceptable. You, the producer, need to decide this on a case-by-case basis. If you're not sure, discuss it with one of the Directors of Production or one of the Posting Team.

If you have made the text with accents, you can choose to make your own 7-bit version and send it to us, or just send the 8-bit version and we'll make the 7-bit version from it. Some people prefer to make their own 7-bit editions; some don't. Whether you use a Microsoft Codepage, one of the ISO standards or MacRoman doesn't matter--we can convert any of them for you.

V.81. I have some Greek quotations in my book. How can I handle them?

There is no way to show Greek letters in ASCII. You have three options:

You can just replace the Greek words with [Greek] to indicate to the reader that you have omitted it.

You can "transliterate" the Greek to ASCII. Greek letters do have a correspondence to plain "Latin" letters--for example, the Greek letter "delta" can be represented by the letter "d". There is a simple PG guide to transliteration at <http://www.promo.net/pg/vol/greek.html>. This practice has had a long and honorable history: words like "amphora" and "hubris", for example, are straight transliteration from the Greek. This is usually the best option.

If there is enough Greek to warrant it, and no other accented characters, you may be able to use the ISO-8859-7 character set, and submit both 7-bit and 8-bit versions [V.79]. ISO-8859-7 is for modern rather than classical Greek, but, if necessary, you will surely be able to express the Greek fully in Unicode. However accurate your Greek, that still leaves the issue of what to do with the 7-bit ASCII version, where transliteration is probably still your best bet.

V.82. I want to produce a book in a language like Spanish or French with accented characters. What should I do?

Use the appropriate ISO-8859 Character set [V.76] for your

8-bit version.

About the formatting of a text file:

This section of the FAQ goes into great detail about all kinds of formatting questions. However, looked at from a higher level, the only real issue is that we want to render texts clearly, with formatting that reflects the original, so that readers of the plain text format can read them easily, and people converting them to other formats can do so reliably. When you come across a case that is not covered by the detailed guidelines below, keep this ultimate aim in mind, and make the best decision you can. Don't get hung up for hours or days over a question of formatting--if you want advice, look at how other people have handled the same situation in previous texts, or ask other volunteers for their ideas.

V.83. How long should I make my lines of text?

For normal prose, such as you find in a novel, your lines should mostly be 60 to 70 characters long, not shorter than 55, not longer than 75 except where it can't be helped. Never, ever longer than 80, except where you're trying to render a non-text structure, like a family tree.

For poetry, make the text look as much like the book as possible. This also applies to some plays where the lines are clearly intended to be broken at specific points, whether blank verse or not.

V.84. Why should I break lines at all? Why not make the text as one line per paragraph, and let the reader wrap it?

We could either use 70-character lines and let readers unwrap them if they want to, or use infinite-length lines and let readers wrap them if they want to. We choose to wrap the lines so that they are readable on even the simplest of text editors and viewers.

V.85. Why use a CR/LF at end of line?

CR/LF can lead to double-spacing, notably on Mac and Unix, but at least there _is_ a CR in there for Mac users, and there _is_ an LF for *nix users.

If you don’t know or care what this is about, please skip blithely on.

There are three differing standards for how to represent the end of a line of text. In brief, Apple Macs use the CR character. Unix and its variants use the LF character. Microsoft systems, from MS-DOS through Windows, use both together.

If you want the history behind these:

CR stands for Carriage Return, and comes from the old typewriter / teletype idea of a command to move the print head from the right of the page back to the left when it reaches the end;

LF stands for Line Feed, and comes from the old typewriter / teletype idea of a command to move the print head down a line;

CR/LF together indicate moving down a line and back to the left of the page.

The history is not relevant to today’s computers in principle, but in practice they all use one of these legacy conventions, and there’s nothing we can do about it but pick one.

V.86. One space or two at the end of a sentence?

Whichever you prefer, but if using two spaces, please use them only at the end of a sentence, not after abbreviations like "Dr." and "per cent.", and not after non-sentence-ending punctuation like the question-mark in the sentence: "Must you go? when the night is yet so black!"

Many people have strong views on either side of the "one space or two?" question, and we’re not about to try and argue with them. Use whichever is most natural for you.

However, if using two, you take responsibility for deciding where the sentence ends. You can’t just place two spaces after every period, question-mark and exclamation mark, since periods are also used for abbreviations end ellipses, and question-marks and exclamation-marks don’t always end sentences.

V.87. How do I indicate paragraphs?

Just leave a blank line before each paragraph.

V.88. Should I indent the start of every paragraph?

No.

Printers do this when publishing paper books because they do not leave blank lines in the text, but there is no need for indenting in our eBooks.

V.89. Are there any places where I should indent text?

Yes. You should always make poetry look like the original, and that may mean indenting some lines, for example:

I was a child and she was a child,
    In a kingdom by the sea;
But we loved with a love that was more than love--
    I and my Annabel Lee;

Even when poetry doesn't have indented lines, it is a good idea to indent quotations embedded in prose. Remember, others will be converting your text later--to HTML, to PDA reader formats, to formats that don't even exist yet--and much of this conversion will be done automatically, by computer programs. It is very hard for a program to know when it can and can't re-wrap lines to fit a screen size unless it has a clear signal that _this_ line should not be wrapped. This is one of the biggest problems with auto-converting PG texts.

Just about all formatting programs "know" that lines that are indented shouldn't be wrapped, so by indenting lines just a space or two, you can prevent

I think that I shall never see
A poem lovely as a tree.

from turning into

I think that I shall never see A poem lovely as a tree.

in some future reader's eBook.

You don't really need to do this in texts where the whole book is poetry or blank verse, since these will probably be recognized as whole books that shouldn't be rewrapped, but when there are a few lines of quotation amid an acre of straight prose, a few spaces will be a life-saver. Even in the original plain text version, the extra spaces serve to set the quotation off from the main text.

You shouldn't get carried away and indent things 20 spaces for this reason, though. Anything up to four spaces is reasonable; more is excessive. If you're indenting many short verses in this way, keep your number of spaces for indentation consistent throughout the book.

There are some other times when you may judge it best to indent, where

text is indented in the paper book, like newspaper headlines or pictures of handwritten notes.

V.90. Can I use tabs (the TAB key) to indent?

No.

The problem with tab characters is that they act differently in different applications. Typically a tab will move the text to the next tab stop, which might be four spaces on your PC, but 20, or none, on someone else's. The effects are unpredictable.

V.91. How should I treat dashes (hyphens) between words?

In typography, there are four standard types of dashes: the hyphen, the en-dash, the em-dash, and the three-em-dash.

Originally, printers called these the "em-dash" because it was the same width as the capital letter M in whichever font they were using, the "en-dash" because it was the same width as the capital letter N, and the "three-em-dash" because it was as long as three capital Ms.

The hyphen is used for hyphenated words, like "en-dash" itself, or "to-day" or "drawing-room". For this, you just press the single dash or hyphen key on your keyboard.

In typography, the en-dash is a little longer than the hyphen, and is typically used for duration, where you could substitute the word "to". For example, if you were printing "1830-1874", or "9:00-5:30", you would use an en-dash instead of a hyphen. The en-dash is also sometimes used as hyphenation between words that are already hyphenated, for example, "bed-room-sitting-room" might use an en-dash as its central dash to emphasize that it is a different type of separator from the plain hyphens before "room". However, there is no ASCII character for an en-dash, and we use the hyphen in these cases. (HTML and some character sets do provide separate entities for en-dash and em-dash.)

The em-dash is shown in print as a longer dash, and for PG purposes, you should render it as two hyphens with no spaces around them.

You use the em-dash as a kind of parenthesis--as I am doing here--or to indicate a break in thought or subject within a sentence. There is no ASCII equivalent of the em-dash; there is no key on your keyboard that you can press to get one. For PG texts, we represent the em-dash as two dashes with no space between or around them--like this.

The em-dash can also be used at the end of a sentence or speech to indicate that the speaker stopped or trailed off. For example:

"When I saw you with Emily, I thought you were-- I thought she was--"

In a case like this, there may be a space following the em-dash, and the context may demand that there _should_ be a space following the em-dash, not because of the em-dash as such, but to make the break between the statements or sentences clear.

These two hyphens represent _one_ character, so you should never break them at line end, with one hyphen at the end of the first line and the other at the start of the second. If you have an em-dash near line end, you can break the line either before or after the em-dash, but never in the middle.

The fourth type of dash, the three-em-dash, is used to represent a missing word, or an undetermined number of missing letters. You will often see it in a sentence like:

Dr. P------ was known for his honesty.

or

Dr. ------ was known for his honesty.

where there is a convention that the character's name has been redacted. Logically, we should represent the three-em-dash as six dashes, but you may reduce that to four. Whichever you choose, do use it consistently in the text you're producing.

Unlike the em-dash, you should leave a space in such cases wherever a space would have been before the letters were replaced by dashes.

Here's a summary table of the dashes:

Name	ASCII	Used for
Hyphen	-	Hyphenated Words
En-dash	-	Durations, like "3:00-5:30"
Em-dash	--	Break in sentence or parenthetical comment
Three-em-dash	------	Indicating a word that was edited out.

V.92. How should I treat dashes replacing letters?

If the dashes obviously represent individual letters, use the same number of hyphens. Otherwise, you can use a three-em-dash (see above: 6 or 4 hyphens) in such places.

A common convention when a character in a novel is using bad language, or when reference is given to a character whose full name is not being used, is to replace the letters with dashes. For example,

"That D---l, Mr. C------s will regret his hasty actions!"

In this case, it is clear that "D---l" is meant to represent "Devil" and that there is a character whose name begins with "C" and ends in "s" whose name is not spelled out in full. Where the book makes it clear how many letters are represented by hyphens, just use that number of hyphens.

Where the number of letters omitted is not clear, you can decide how long you want to make your extended dash. Typographers often use the "three-em-dash" for this, so called because it is as wide as three capital Ms. Logically, since we represent an em-dash by two hyphens, we might represent a three-em-dash as six, but if you feel that six hyphens is too long, you can choose a shorter length, like four, but if you do, keep it consistent within your text:

It was in the town of S----, walking on M---- Street, that Sowerby came upon Dr. T---- taking the morning air.

V.93. What about hyphens at end of line?

Remove the hyphens from single words that were wrapped by the printer at line-end on the paper copy. Where two words are joined with a hyphen, you can leave the hyphen at end of the text line.

Books are usually printed with words broken at end of line to make the right side of the text perfectly even. You should remove all such hyphens. For example, in the sentence:

Mary’s mouth tightened as she saw the marks on the car-
pet, and her hands balled into fists.

you should remove the hyphen from "carpet".

Words which are strung together and hyphenated by the author pose a different question. It is perfectly OK from the point of view of a reader of the plain text version for such a hyphen to occur at end of line, for example:

Now that the guns were silent, convoys brought badly-
needed medical supplies and food.

However, be aware that if somebody later rewraps the text for use in a different format like HTML, it is possible that they will introduce a space where it should not be:

Now that the guns were silent, convoys brought badly- needed
medical supplies and food.

so there is still a small disadvantage to having a hyphen at line-end.

Sometimes it’s not entirely clear whether the hyphen is there because it has to be, or just because it happens to fall at the end of the line:

> Daisy rushed to the door, but there were no letters for her to-
> day, and she retreated sadly.

Sometimes "today" is written as "to-day", especially in older works. So which is this? Should we remove the hyphen or not? In this case, the best thing to do is search the rest of the text for the same word, and see whether it is consistently hyphenated or not in other places.

V.94. What should I do with italics?

There are three different ways volunteers currently render italics: like THIS, like _this_ and like /this/. Pick one, and use it consistently in your text.

There are really two questions here: "How should I render italics?" and "When should I render italics?"

The original PG standard for italics was to render emphasis italics as CAPITALS, using underscores for an italicized _I_, and do nothing for non-emphasis italics like foreign words and names of ships, and this is still the most common usage. For reading a plain-text file in a plain text editor, it is still arguably the most reader-friendly usage as well.

It has two drawbacks:

1. if you do want to preserve italics for non-emphasis words, you may end up with a very ugly text where there are too many capitals.

2. it is impossible to convert CAPITALS reliably back into italics, since the original text might have had a capital letter, or even been all capitals in the first place. This is especially true of automatic conversion for people who want to read PG texts on eBook readers.

To overcome these problems, many volunteers now use _underscores_ or /slants/ to render italics. These allow you to preserve all italics without creating an ugly plain-text, and to remove the ambiguity of CAPITALS. Underscores are more popular than slants, but some people feel that underscores should properly be reserved for underlined text. Since printers tend to avoid underlines, however, there aren’t many books where this causes a real conflict.

V.95. Yes, but I have a long passage of my book in italics! I can’t really CAPITALIZE or _otherwise_ /mark/ all that text, can I?

No, you really can't. On the other hand, if the author intended that section to stand out, you don't want to ignore that information and withhold it from future readers.

What you _can_ do is format it differently from the rest of the text. For example, if you're averaging a 68-character line throughout normal paragraphs, you could reasonably use shorter lines, like 58 characters, for the italicized section. Going a step further, you could shorten the lines and indent them a space or two as well. This will give a clear signal to future readers and converters that this section is to be treated specially.

V.96. Should I capitalize the first word in each chapter?

No.

Capitalization of the first word is often used in printed material to emphasize the break at the start of a section or chapter on the paper, but it is not necessary in an eBook, and leads to the same kind of ambiguity as does the capitalization of italics, and for far less reason.

If you feel you really _must_ capitalize the first word, we probably won't stop you, but if so, please do it consistently throughout the book, not just in one or two places, so that a future reader can be certain that these capitalized words were a chapter-head convention, and not otherwise intended for emphasis.

V.97. What is a Transcriber's Note? When should I add one?

A Transcriber's Note is a small section you can add to a text you produce to give the reader some information about changes you made to the book when rendering it into text.

A Transcriber's Note is not the same as a footnote--a footnote is part of the text you have transcribed; a Transcriber's Note is a note that _you_ add to the text, explaining something _you_ have done or omitted. If there is a Transcriber's Note, it may be at the top or the end of the text, and it should be clearly marked so that a reader cannot confuse it with the main text or an introduction.

The main thing is to ensure that a reader cannot confuse text that you have added with text that was in the original book.

Transcriber's Notes are rarely needed, but if, for example, you found misprints in the text, or things that might look like misprints even though they're not, you may note them here, if it seems relevant. If there is an image in the book that is important to the content, you may describe it in a note. If there was unusual typography that you

had to represent in some uncommon way, you might well explain that here.

You don’t need to add a Transcriber’s Note just for common conversions like italics, and you should not use such a note to add your own comments or views about the text or the author. It’s just there to let the reader know what decision you have made about rendering the text.

Here are some examples of Transcribers’ Notes:

Transcriber’s Note:

The irregular inclusion or omission of commas between repeated words ("well, well"; "there there", etc.) in this etext is reproduced faithfully from the 1914 edition . . .

Transcriber’s Note:

Inserted music notation is represented like [MUSIC--2 bars, melody] or [MUSIC--4-part, 8 bars]

[Transcriber’s Note: This letter was handwritten in the original.]

Transcriber’s Note:

The spelling "Freindship" is thus in the original book.

Transcriber’s Note: Some words which appear to be typos are printed thus in the original book. A list of these possible misprints follows:

If there is an image that is important to the content you may describe it at the point in the text where it appears, for example:

[Transcriber’s Note: Here there is a map of three islands just West of and parallel to a coastline running SW to NE, with a big X marked on the North of the middle island. A spur of land extends from the mainland, sheltering the islands from the north-east.]

Transcriber’s Notes that apply to the whole text should be placed at the start or end of the text--your choice. Notes that pertain to a specific point in the text, like the map example above, should be placed at the point where in the text where they are relevant, but not interrupting a paragraph except where it cannot be avoided.

V.98. Should I keep page numbers in the e-text?

No. But there are exceptional cases . . .

In general, the page numbers of the original book are irrelevant when making a reader's edition for PG; they are annoying and intrusive for anyone trying to read it, and if you did keep them, they would probably be removed by anyone converting it. Get rid of them!

But there are a few books where page numbers are appropriate. Non-fiction books that use page numbers as internal cross-references are the prime example; if, on page 204, the text reads

"Our studies of plants (see pp. 141-145) show that this is true."

and this kind of cross-reference is frequent throughout the text, then it is probably best to keep the page numbers, since it is otherwise very difficult to honor the author's intent.

In the more common case where cross-references exist, but are not frequent, and not essential to the text, you have several choices: leave the cross-references in, meaningless though the page numbers are, remove the cross-references, change the cross-references to something relevant (like "Start of Chapter 12" instead of "pages 141-145"), or, if you can make it work in context, insert references in the text for the cross-references to point to, like [Reference: Plants] and then reformat the cross-reference like "Our studies of plants (see [Reference: Plants]) show that this is true."

There are a few other cases, where the text you create is likely to be the subject of study or reference, in which it may also be desirable to retain page numbering.

When there are pages at the end of the book with notes referring to page numbers, the simplest answer is to change the page number references to chapter numbers, and add a quote from the page referred to if it's not already in the book's end-notes. That way, a reader can search for the phrase.

V.99. In the exceptional cases where I keep page numbers, how should I format them?

Within brackets of your choice, with one space either side, simply added to the text at the exact point of the page break. Unless there is some [142] special reason, you shouldn't insert a line break or new paragraph when indicating a page number; just insert it in the text, as I did with "142" above.

You should use whichever of round brackets, (143) square brackets, [144] or curly brackets {145} is not used (or least used) within the main text itself, and then use it consistently. Try to make sure that your page numbers cannot be confused with anything else.

Don’t run your[146]page[147]numbers right up against words with spaces omitted; this just makes the text hard to read. Use spaces before and after.

Where the page break is at the start of a chapter or headed section, you can put it on a line of its own, for example:

[148]

CHAPTER XI. PLANTS

Where a paragraph begins on a new page, you should put the page number at the start of the paragraph, as:

[149] With the extinction of the dinosaurs . . .

V.100. Should I keep Tables of Contents?

Yes, but just keep the contents themselves, and not the page numbers for each chapter or section, except where you have kept the page numbers in the whole text. When you have removed the page numbers from the book, it doesn’t make much sense to leave them in the TOC.

Here, for example, is a typical TOC. In the original text, each chapter had a page number beside it:

THE DUKE’S CHILDREN

CONTENTS

1 When the Duchess was Dead
2 Lady Mary Palliser
3 Francis Oliphant Tregear
4 It is Impossible
5 Major Tifto
6 Conservative Convictions
8 He is a Gentleman
9 ’In Media Res’
10 Why not like Romeo if I Feel like Romeo?
11 Cruel
12 At Richmond

Note that I have indented the lines here, to give a sign to automatic converters that these lines should not be wrapped into one paragraph.

V.101. Should I keep Indexes and Glossaries?

If you are working from a pre-1923 publication, then yes.

If you are working from a modern reprint, you must be careful not to take any of the text that might have been added by the modern publisher. If you have any doubt about whether the index or glossary was part of the original printing, you should leave it out. Often with reprints, under your Clearance Line [V.37], you may see an instruction not to use indexes. In such cases, or if there is any doubt at all, don't.

V.102. How do I handle a break from one scene to another, where the book uses blank lines, or a row of asterisks?

Use a blank line, followed by a line of 3 or 5 spaced asterisks or dashes, followed by another blank line.

In a printed book, where the point of view switches from one character to another, or some other break in the narrative is made without a new chapter or headed section, the publisher will often denote the break just by a couple of blank lines. This gives the reader a cue to notice that the point of view has switched, and avoids confusion.

However, a printed book cannot be edited or changed, while an eBook will be edited and converted over its lifetime, and it is likely that if you denote this break just by a couple of blank lines, as in the book, your break may be lost. For example, in automated conversion to a PDA reader format, it is common to merge multiple blank lines into one.

In making a PG e-text, you _may_ indicate this break by a couple of additional blank lines, but, if your text is later converted into another format such as HTML, the extra blank lines may get lost in the editing or rendering. Or the person doing the conversion may simply think that the extra blank line was a mistake, and remove it. To guard against this, you should add an unambiguous visual break such as a line of spaced asterisks:

* * * * *

The exact layout of your break is not really important, and you can use whatever format you prefer. Blank line followed by five spaced asterisks followed by another blank. Or you could use two blank lines, and dashes instead of asterisks. Just make sure that future readers can be in no doubt that you intended to indicate a break that was really in the original printed text.

V.103. How should I treat footnotes?

In a printed text, the most common treatment for footnotes is to put

them at the end of the page to which they refer. Sometimes, editors gather them all at the end of the book. Footnotes are a real formatting problem for an eBook without defined physical pages; there is no agreement between readers about which is the best way to render them.

There are three basic ways of rendering footnotes in an e-text:

You can insert them right into the text, in brackets, at the point in the paragraph where they occur, with or without an indication that they were originally footnotes. This is only reasonable in a text with very short footnotes.

You can insert them after the paragraph to which they refer, either contiguous with the paragraph or as a new "paragraph" of their own, as I am doing with this one. If the text contains any footnotes longer than a line, [1] you should not try to just append them to the paragraph; you should make a new "paragraph" of them, with a blank line before and after.

[1] Some footnotes can go on not only for several lines, but for several pages!

You can gather all footnotes at the end of the e-text, or to the end of the chapter to which they refer.

Of these three, gathering all footnotes to the end of the chapter or the end of the whole text is probably the friendliest option, since it preserves the original intention of allowing the reader to continue reading the main text without interruption. However, it may involve some renumbering and general note-keeping on your part, and may not be needed where there are only a few short footnotes. You can see an ideal example of this kind of footnote marking in our edition of Darwin's "The Voyage of the Beagle", file vbgle10.txt from 1997, Etext number 944, which you can get from:
<ftp://ftp.ibiblio.org/pub/docs/books/gutenberg/etext97/vbgle10.txt>

V.104. My book leaves a space before punctuation like semicolons, question marks, exclamation marks and quotes. Should I do the same?

No.

If you look closely at these "spaces", you will see that they are not as wide as a normal space--they tend to be half to three-quarters as wide. These don't actually represent spaces as such; they were just a convention used by typesetters to make the text feel less cramped, and they did not express any specific intent on the part of the author.

OCR software tends to see them as full spaces, and one of the jobs you typically have to do when editing a text that has been OCRed is to

remove them.

In some texts, this also happens following an opening quote, so your OCR might read a sentence as:

" Hello ! How are you to-day ? "

which you should correct to:

"Hello! How are you to-day?"

Samples of this can be seen in the images used for the FAQ "Why am I getting a lot of mistakes in my OCRed text?" [S.17]

V.105. My book leaves a space in the middle of contracted words like "do n't", "we 'll" and "he 's". Should I do the same?

Unlike the pseudo-spaces before punctuation, these really were intended as spaces indicating the break between words--that is, where we would nowadays contract two words into one, the author or editor has made the contraction, but left them as two separate words.

Since this effect was intended, it is usual to leave the spaces in. Some people who really do n't like this style of spelling do remove them, but generally volunteers want to preserve the text as printed.

V.106. How should I handle tables?

Just line up the information neatly in columns. If you use a non-proportional font [W.5] you will be able to do this reliably. You can also use the dash character "-" , the underscore "_" and the pipe character "|" to make borders if you really need to, but it's usually better to omit them. It is, though, often good to indent your table a little, to set it off from the main text, and to avoid the danger of having it automatically wrapped by some converter later. For example, from "The Albert N'Yanza, Great Basin of the Nile" by Sir Samuel White Baker:

TABLE No. 1.

Table for Increased Reading of Thermometer, using 0 degrees 80 as the Result of Observations for its Error.

Month.	1861.	1862.	1863.	1864.	1865.
January. . .	--	0'143	0'314	0'487	0'659
February . .	--	'157	'328	'501	'673
March . . .	0'000	'172	'344	'516	'688
April . . .	'014	'186	'358	'530	'702

May . . . . ’028 ’200 ’372 ’544 ’716
June . . . . ’043 ’214 ’387 ’559 ’730
July . . . . ’057 ’228 ’401 ’573 ’744
August . . . ’071 ’243 ’415 ’587 ’758
September. . ’086 ’257 ’430 ’602 ’772
October . . ’100 ’271 ’444 ’616 ’786
November . . ’114 ’285 ’458 ’630 0’800
December . . 0’129 0’300 0’473 0’645 --

V.107. How should I format letters or journal entries?

Make them look like they are in the printed book. If the signature is indented in the book, indent it in the letter. For example:

"Sir,
No consideration would induce me to
change my resolve in this matter, but I am
willing to engage your services as my agent
for a fee of 100 pounds.
"H. Middleton"

When a letter appears in the middle of lots of prose, using shorter lines for the letter is an effective way of making the letter stand out, without resorting to indenting the whole thing.

When the book is largely composed of letters or entries, as happens in an epistolary novel or the publication of somebody’s letters or journal, you might reasonably leave two or three (but whichever you choose, keep it consistent throughout the book!) blank lines between entries to give the reader a visual cue that the next is not just a new paragraph, but a new entry, for example:

10 pm.--I have visited him again and found him sitting in a corner brooding. When I came in he threw himself on his knees before me and implored me to let him have a cat, that his salvation depended upon it.

I was firm, however, and told him that he could not have it, whereupon he went without a word, and sat down, gnawing his fingers, in the corner where I had found him. I shall see him in the morning early.

20 July.--Visited Renfield very early, before attendant went his rounds. Found him up and humming a tune. He was spreading out his sugar, which he had saved, in the window, and was manifestly beginning his fly catching again, and beginning it cheerfully and with a good grace.

I looked around for his birds, and not seeing them, asked him where they were. He replied, without turning round, that they had all flown

away. There were a few feathers about the room and on his pillow a drop of blood. I said nothing, but went and told the keeper to report to me if there were anything odd about him during the day.

11 am.--The attendant has just been to see me to say that Renfield has been very sick and has disgorged a whole lot of feathers. "My belief is, doctor," he said, "that he has eaten his birds, and that he just took and ate them raw!"

11 pm.--I gave Renfield a strong opiate tonight, enough to make even him sleep, and took away his pocketbook to look at it. The thought that has been buzzing about my brain lately is complete, and the theory proved.

This is different from the case mentioned in the FAQ [V.102] "How do I handle a break from one scene to another, where the book uses blank lines, or a row of asterisks?". In that case, we added a row of asterisks because future reformatting or conversion could cause confusion about the scene break that was explicitly signalled by the blank lines on paper. In this case, each new letter or journal entry cannot be mistaken by a careful reader, so we don’t need asterisks or dashes to signal that; we’re just adding a bit of extra space to make it more readable.

V.108. What can I do with the British pound sign?

The British pound sign cannot be expressed in ASCII, but is very common in the works of English novelists. It evolved as a stylized version of the letter L (from the Latin "Librii"), and it’s entirely appropriate to represent it as such, either like:

The horse cost L8 12s. 6d.

or

The horse cost 8l. 12s. 6d.

This works particularly well where an amount is expressed in pounds, shillings and pence (Librii, soldarii, denarii).

Where there is a simple number of pounds, you may prefer just to use the word:

She was a handsome widow with 500 pounds a year.

V.109. What can I do with the degree symbol?

Just type out the word "degrees" or the abbreviation "deg."--for example:

By the time we reached Cairo it was 115 degrees in the shade.

Geographical degrees are more awkward, but should be handled the same way:

It was at 30 deg. 15’ E, 14 deg. 45’ N.

In general, any symbol can be represented in words.

V.110. How should I handle . . . ellipses?

Just as I did above . . . and here! Leave one space before and after each dot. Do not break an ellipsis over the end of a line. In principle, an ellipsis is one symbol, like an em-dash, and should not be broken at line end.

A special case arises when an ellipsis follows a sentence instead of being in the middle. . . . In this case, put the period after the last letter of the sentence, as you normally would, then follow the usual format for ellipses. You end up with four dots, with spaces everywhere except before the first.

V.111. How should I handle chapter and section headings?

For a standard novel, you can choose either four blank lines before the chapter heading and two lines after, or three lines before and one line after, but whichever you use, do try to keep it consistent throughout.

Normally, you should move chapter headings to the left rather than try to imitate the centering that is used in some books.

V.112. My book has advertisements at the end. Should I keep them?

Most people seem to think "no", and "no" is the safe choice, but opinions vary.

The typical arguments are: "The ads are not part of the author’s intent, so you should remove them." vs. "They give a flavor of the original book, so you should keep them". This latter is particularly cogent when the ads are for other books by the same author.

Decide which of these statements best fits your own views in the case you're looking at; after that, it's up to you!

V.113. Can I keep Lists of Illustrations, even when producing a plain text file?

Yes. As in the case of the Table of Contents, there is no point in including page numbers when your text doesn't have them, but the list of illustrations itself may go in.

V.114. Can I include the captions of Illustrations, even when producing a plain text file?

Yes.

You can format them as short paragraphs of their own, in brackets, with the word Illustration: followed by the caption, something like:

[Frontispiece: A Flash of Light]

or

[Illustration: Goldsmith at Trinity College]

Don't interrupt a paragraph to insert one, unless the reader really needs to know that the original illustration was in the middle of the paragraph; place the note between paragraphs instead.

V.115. Can I include images with my text file?

Yes, as I have done with the zipped version of the plain-text format of this FAQ, but in general it makes much more sense, if you want to include images, to make a HTML version of the book and include them there, where they are anchored into the text in a predictable way, and leave them out of the text version. But there are exceptional cases, such as this--I included images with this plain-text FAQ because I wanted you to be able to experiment with them using your own OCR package.

If you do include images with plain text, they will be included with the ZIP file, but not downloadable separately with the plain text file; for example, if your file gets named abcde10.txt, and you include images pic1.gif, pic2.gif and pic3.gif, then abcde10.zip will include all four files, but only abcde10.zip and abcde10.txt will be posted, so the images will be available only within the zip file, so, even if you are including images, don't assume that the reader will be able to see them.

If you do include images with plain text, be sure to mention them by filename in a note at the appropriate places in the text file; otherwise readers may not even realize they're there. For example:

[Illustration: Goldsmith at Trinity College--see goldtrin.gif]

If you do include images with a text file, don't make them too big. Readers downloading zip files of plain text expect them to be relatively small; don't burden them with huge downloads they don't want. Use the same kind of rules and processing that you would for a HTML file, or better still, include the images only with the HTML version.

About formatting poetry:

V.116. I'm producing a book of poetry. How should I format it?

Make it look like the original.

The only formatting change that you might consider is to limit the amount of centering. Often, in a poetry book, the title of a poem may be centered, when the body of the verse isn't. This can work on paper, particularly when the page is narrow, but "centering" the title on a 70-column line can mean that the title ends up far to the right of the body of the poem, which looks untidy. And even if you center the title correctly over the body of _this_ poem, the next poem may have longer lines, and so _its_ title may not have the same center as the first poem, and the title of one will be off-center with the title of the next!

If you have this kind of formatting in your book, you should consider moving all of the poem titles to the left margin rather than try to keep compensating for different line centers. It's more consistent, and easier to read, if you just left-align all titles. To see a not-quite-successful attempt at centering the titles over the poems, take a look at the Poems of Emily Dickinson, available from <ftp://ftp.ibiblio.org/pub/docs/books/gutenberg/etext00/1mlyd10a.txt>

In that case, it would have been better to left-align the numbers and titles. Centering isn't really an effective formatting choice in etexts.

V.117. I'm producing a novel with some short quotations from poems. How should I format them?

As nearly as possible like they look in the book, with the exception that you should indent the whole verse anywhere between 1 and 4 spaces

from the left. This is to give a signal to automatic conversion programs that these lines should not be wrapped.

For an example of a novel with many differently formatted quotations embedded, see the "a" version of Clotel, file clotl10a.txt, Etext number 2046, from the year 2000, which you can find at <ftp://ftp.ibiblio.org/pub/docs/books/gutenberg/etext00/clotl10a.txt>

Some of these quotations touch the left-hand column; today, we would think it better to insert at least one space before every line.

About formatting plays:

V.118. How should I format Act and Scene headings?

Pretty much like chapter headings. You can use 4 blank lines between acts, and 3 blank likes between scenes, or 3 between acts and 2 between scenes. If your book has "END OF ACT/SCENE" footers, leave them in the etext.

You may center act/scene headers and footers if they are centered in the book, but it's usually best to left-align them, for the same reasons it's usually best to left-align poem titles in poetry.

V.119. How should I format stage directions?

Generally, in brackets.

In printed texts, it is common to show stage directions as italics inside brackets. You don't have the option of italics in plain text, and you shouldn't need to use _underscores_ or /slants/, and certainly not CAPITALS, to indicate italics for stage directions. Normal text within the brackets is all you need. It will be immediately clear to a reader that bracketed text consists of stage directions.

[Square brackets] are most common for stage directions, but (round) or {curly} brackets will work too, if there's a reason why they are preferable in the case of your text. Just make sure that you use the same kind of brackets consistently and only for stage directions--don't use round brackets for stage directions if characters' speeches also contain text in round brackets.

Some printed plays follow the convention of not closing brackets when the direction is at the end of a speech or scene. For example:
[Exeunt.

Where the book doesn't close the bracket in a case like this, you

shouldn’t either.

V.120. How should I format blank verse?

Just like normal verse in poetry. Make it look like the printed book. Left-align it, and make one line of etext the same length as one line of print.

Sometimes in blank verse, a speech may start mid-line, and the print reflects that by leaving a space on the left, and starting mid-way. In a case like that, do the same in the etext.

About some typical formatting issues:

V.121. Sample 1: Typical formatting issues of a novel.

Look at the image novel.tif. It shows a page of a novel, with several typical formatting decisions to be made.

We note that there is no end-quote on the first paragraph, but that’s OK, since the second paragraph is a continuation by the same speaker, so the first paragraph doesn’t need a closequote. There is also an italicized "I", which will end up with underscores, but there is nothing else to give us any difficulty.

In the second paragraph, we have an ellipsis, an italicized French word with an accented letter, the British pound symbol, and an italicized "Here".

The ellipsis is simple.

Let’s assume we’re making this into a 7-bit text, so we’re going to convert the non-ASCII character a-circumflex and the pound sign. The a-circumflex just goes to an "a", but we have several choices we can make about the pound sign.

The italicized "Here" is clearly for emphasis, so we will mark that up. The word "flaneur" is italicized because it is not English, but possibly also for emphasis . . . if the sentence had read "The Major is a _fool_", with the word "fool" italicized, it would clearly be emphasis. As it stands, we don’t know whether emphasis is intended. This doesn’t matter if we are just using _underscores_ or /slants/ to render italics, but if we use CAPITALS, we’re going to have to impose our best guess on one side or the other.

The third paragraph shows some vaguely familiar squiggles--Greek

letters! We hit the PG transliteration guide at
<http://www.gutenberg.net/vol/greek.html> and spell it out . . .
rough-breathing upsilon = hu; beta = b; rho = r; iota = i; final
sigma = s. So the Greek word transliterates as "hubris". Since
hubris is a familiar word, we don't need to make a fuss about it,
though we may _italicize_ it.

We then have a note, which we will format a little differently from
the main text to help it stand out, and a new chapter heading.

We should certainly indent the second line of the Byron quotation to
preserve its original form, but we have the option whether or not to
indent the first line a little to signal to any future automatic
converter that this is not to be rewrapped.

In the first paragraph of the new chapter, we need to get rid of the
hyphenation of "Wentworth" at line-end and fix the two em-dashes.

In the second paragraph of the new chapter, we have a long dash
between "d" and "l", clearly meant to denote "devil", so we will fill
it in with three dashes, and we see a three-em-dash after "Lord H", so
we can use six, or possibly four, dashes for that.

Finally, we have a table, a list of money values against names.

Depending on the standards we've chosen to use throughout the book, we
could render these details in a variety of ways. For illustration,
here are two acceptable possibilities:

"I shall go down to Wokingham", said Middleton, "a few days
before the election, and the Major will stay here. I
understand that there will be no other candidate, and _I_
shall take the seat.

"The Major is a . . . _flaneur_. He has no interest beyond
his own advancement. I can buy him for a hundred pounds.
Here is his answer."

Wallace wondered at the _hubris_ of his friend, and
examined the note Middleton thrust upon him.

"Sir,
    No consideration would induce me to
change my resolve in this matter, but I am
willing to engage your services as my agent
for a fee of 100 pounds.
                H. Middleton"

CHAPTER XV

THE ELECTION

Now hatred is by far the longest pleasure;
Men love in haste, but they detest at leisure.
---- BYRON

On hearing of Middleton's visit, Mr. Wentworth began his preparations. Meeting with Thomas Lake and Riley at the back of the tap-room of The Bull--where the landlord saw to it that they remained undisturbed--he laid out their plan of campaign.

"That d---l Middleton shall not have the seat," he raved, "not for Lord H------; no, nor for a hundred Lords! We shall see to it that every man's hand is turned against him when he arrives."

Lake unfolded a paper from his vest-pocket and smoothed it on the table. "Here are the expenses we should undertake."

Doran	L13 10s.
Titwell	L 8 7s. 6d.
St. Charles	L25

* * * * *

"I shall go down to Wokingham", said Middleton, "a few days before the election, and the Major will stay here. I understand that there will be no other candidate, and _I_ shall take the seat.

"The Major is a . . . flaneur. He has no interest beyond his own advancement. I can buy him for L100. HERE is his answer."

Wallace wondered at the hubris of his friend, and examined the note Middleton thrust upon him.

"Sir,
No consideration would induce me to change my resolve in this matter, but I am willing to engage your services as my agent for a fee of L100.
H. Middleton"

CHAPTER XV

THE ELECTION

Now hatred is by far the longest pleasure;
Men love in haste, but they detest at leisure.
---- Byron

On hearing of Middleton’s visit, Mr. Wentworth began his preparations. Meeting with Thomas Lake and Riley at the back of the tap-room of The Bull--where the landlord saw to it that they remained undisturbed--he laid out their plan of campaign.

"That d---l Middleton shall not have the seat," he raved, "not for Lord H----; no, nor for a hundred Lords! We shall see to it that every man’s hand is turned against him when he arrives."

Lake unfolded a paper from his vest-pocket and smoothed it on the table. "Here are the expenses we should undertake."

Doran	13l. 10s.
Titwell	8l. 7s. 6d.
St. Charles	25l.

V.122. Sample 2: Typical formatting issues of non-fiction

While non-fiction is not in principle any more difficult to format than fiction, many non-fiction books have lots of features like illustrations, tables, section sub-headings and footnotes, that require some extra work on the part of the producer. If the illustrations are essential, you should consider adding a HTML format file to allow you to present them.

See the page image nonfic.tif. This presents many formatting changes: the centered title will go to the left; the italicized chapter contents will become regular text, and the em-dashes will become "--"; the degree symbol needs to be replaced with ASCII "deg.", and of course we need to render the table readably. After all that, we have to deal with the footnote.

Here is a reasonable rendering of this page:

CHAPTER XI

STRAIT OF MAGELLAN.--CLIMATE OF THE SOUTHERN COASTS

Strait of Magellan--Port Famine--Ascent of Mount Tarn--
Forests--Edible Fungus--Zoology--Great Sea-weed--
Leave Tierra del Fuego--Climate--Fruit-trees and
Productions of the Southern Coasts--Height of Snow-line

on the Cordillera--Descent of Glaciers to the Sea--
Icebergs formed--Transportal of Boulders--Climate
and Productions of the Antarctic Islands--Preservation
of Frozen Carcasses--Recapitulation.

An equable climate, evidently due to the large area of sea compared with the land, seems to extend over the greater part of the southern hemisphere; and, as a consequence, the vegetation partakes of a semi-tropical character. Tree-ferns thrive luxuriantly in Van Diemen's Land (lat. 45 degrees), and I measured one trunk no less than six feet in circumference. An arborescent fern was found by Forster in New Zealand in 46 degrees, where orchideous plants are parasitical on the trees. In the Auckland Islands, ferns, according to Dr. Dieffenbach [82] have trunks so thick and high that they may be almost called tree-ferns; and in these islands, and even as far south as lat. 55 degrees. in the Macquarrie Islands, parrots abound.

On the Height of the Snow-line, and on the Descent of the Glaciers in South America.
[For the detailed authorities for the following table, I must refer to the former edition:]

Latitude	Height in feet of Snow-line	Observer
Equatorial region; mean result	15,748	Humboldt.
Bolivia, lat. 16 to 18 deg. S.	17,000	Pentland.
Central Chile, lat. 33 deg. S.	14,500 - 15,000	Gillies, and the Author.
Chiloe, lat. 41 to 43 deg. S.	6,000	Officers of the Beagle and the Author.
Tierra del Fuego, 54 deg. S.	3,500 - 4,000	King.

In Eyre's Sound, in the latitude of Paris, there are immense glaciers, and yet the loftiest neighbouring mountain is only 6200 feet high. Some of the icebergs were loaded with blocks of no inconsiderable size, of granite and other rocks, different from the clay-slate of the surrounding mountains. The glacier furthest from the pole, surveyed during the voyages of the Adventure and Beagle, is in lat. 46 degrees 50 minutes, in the Gulf of Penas. It is 15 miles long, and in one part 7 broad and descends to the sea-coast. But even a few miles northward of this glacier, in Laguna de San Rafael, some Spanish missionaries encountered "many icebergs, some great, some small, and others middle-sized," in a narrow arm of the sea, on the 22nd of the month corresponding with our June, and in a latitude corresponding with that of the Lake of Geneva!

In this case, I made some decisions. I made the lines in the contents

at the top a bit shorter than usual, to help them stand out. I decided to use the full word "degrees" rather than "deg." where I could, but not in the table, where I shortened the entries as much as possible while preserving the sense. Since I was using the full word "degrees", I decided to go the whole hog and use the word "minutes" for the minutes symbol as well, (though the minutes symbol, a single quote, is in the ASCII set) since it seemed to make the text more readable than using the word degrees with the minutes symbol. I also made a choice about the table layout.

You might prefer different choices in some of these cases, and, as in our example of fiction above, there was more than one way to do it. However, this is a reasonable rendering.

What happened to the footnote? and how did it become [82] rather than the [1] of the original? In this case, I decided to put all footnotes at the end of the whole text, and renumber them accordingly. So the footnote on this page became number 82 in the overall text, and down at the end of the whole text, I would put:

[82] See the German Translation of this Journal; and for the other facts, Mr. Brown’s Appendix to Flinders’s Voyage.

I could also have transcribed this as:

. . .
Forster in New Zealand in 46 degrees, where orchideous plants are parasitical on the trees. In the Auckland Islands, ferns, according to Dr. Dieffenbach [*] have trunks so thick and high that they may be almost called tree-ferns; and in these islands, and even as far south as lat. 55 degrees. in the Macquarrie Islands, parrots abound.

[*] See the German Translation of this Journal; and for the other facts, Mr. Brown’s Appendix to Flinders’s Voyage.

if I chose to put each footnote with its own paragraph.

V.123. Sample 3: Typical formatting issues of poetry

Poetry is easy to format: just be sure to use a non-proportional font, and make it look as much like the text as possible. To avoid ragged-looking centering, left-align titles.

In a whole book of poetry, there is no need to leave an indentation before every line; unlike a verse lost in fields of prose, there is little danger that someone will wrap it by mistake.

Look at the image poetry.tif. On this page, we have an enlarged first letter to start each poem, and capitals following--we can remove all

that. The titles are centered, so we will move them left.

There are line-numbers at every fifth line, and these are common in poetry, especially where footnotes reference lines. We will keep these out on the right-hand margin.

The third poem obviously intends the centering of its last lines in each verse as a feature, so we will keep that as best we can.

The resulting etext looks like:

Mistress Mary

Mistress Mary, quite contrary,
    How does your garden grow?
With cockle-shells, and silver bells,
    And pretty maids all in a row.

Ozymandias.

I met a traveller from an antique land
Who said: Two vast and trunkless legs of stone
Stand in the desert. . . . Near them, on the sand,
Half sunk, a shattered visage lies, whose frown,
And wrinkled lip, and sneer of cold command,
Tell that its sculptor well those passions read
Which yet survive, stamped on these lifeless things,
The hand that mocked them, and the heart that fed:
And on the pedestal these words appear:
'My name is Ozymandias, king of kings:
Look on my works, ye Mighty, and despair!'
Nothing beside remains. Round the decay
Of that colossal wreck, boundless and bare
The lone and level sands stretch far away.

NOTE:
9 these words appear: in some editions : this legend clear.

The Rosary.

The hours I spent with thee, dear heart,
  Are as a string of pearls to me;
I count them over, every one apart,
        My rosary.

Each hour a pearl, each pearl a prayer,
  To still a heart in absence wrung;

I tell each bead unto the end--and there
A cross is hung.

Oh, memories that bless--and burn!
Oh, barren gain--and bitter loss!
I kiss each bead, and strive at last to learn
To kiss the cross,
Sweetheart,
To kiss the cross.

V.124. Sample 4: Typical formatting issues of plays

Look at the image play.tif. Stage directions are indicated by italics and square brackets. We don't have to do much special work with this--lose the italics, but keep the square brackets. The setting for scene I, act II is also italicized, but without square brackets. If we wanted to emphasize this, we could use shorter lines or add square brackets, but it probably isn't necessary here. We're using 4 blank lines between acts and 3 between scenes, so we mark these accordingly. We leave one blank line between speeches. And following these simple conventions, we get:

JACK. There's a sensible, intellectual girl! the only girl I ever cared for in my life. [ALGERNON is laughing immoderately.] What on earth are you so amused at?

ALGERNON. Oh, I'm a little anxious about poor Bunbury, that is all.

JACK. If you don't take care, your friend Bunbury will get you into a serious scrape some day.

ALGERNON. I love scrapes. They are the only things that are never serious.

JACK. Oh, that's nonsense, Algy. You never talk anything but nonsense.

ALGERNON. Nobody ever does.

[JACK looks indignantly at him, and leaves the room. ALGERNON lights a cigarette, reads his shirt-cuff, and smiles.]

END OF THE FIRST ACT

SECOND ACT

SCENE I

Garden at the Manor House. A flight of grey stone steps leads up to the house. The garden, an old-fashioned one, full of roses. Time of year, July. Basket chairs, and a table covered with books, are set under a large yew-tree.

[MISS PRISM discovered seated at the table. CECILY is at the back watering flowers.]

MISS PRISM. [Calling.] Cecily, Cecily! Surely such a utilitarian occupation as the watering of flowers is rather Moulton's duty than yours? Especially at a moment when intellectual pleasures await you. Your German grammar is on the table. Pray open it at page fifteen. We will repeat yesterday's lesson.

About problems with the printed books:

V.125. I found some distasteful or offensive passages in a book I'm producing. Should I omit them?

Please don't. Readers understand that books are works of their time and place, reflecting the opinions and prejudices of the people who wrote them, and the people they observed. We shouldn't try to pretend those prejudices out of existence. It may be, in a century or two, that our descendants are repulsed by _our_ prejudices.

It is perfectly normal, for all kinds of reasons, not to want to produce a particular book, but producing one while deliberately removing passages is censorship, and is unfair to our readers.

If you find it too disturbing to handle the content, you can of course abandon the book, or pass it along to some other volunteer.

V.126. Some paragraphs in my book, where a character is speaking, have quotes at the start, but not at the end. Should I close those quotes?

Probably not.

When one character is making a speech that spans more than one paragraph, it is usual _not_ to close the quotes until the speech is finished. This avoids confusion about whether the next paragraph is the same speaker or another--once a character has started speaking, there are no closequotes until the speech is

finished. However, there are openquotes at the _start_ of each new paragraph during the speech. This makes the quotes unbalanced, but it isn't a misprint; it's deliberate.

If this is not the case, if the same character is not continuing the speech in the next paragraph, then you may have found a typo in the book. [R.26]

V.127. The spelling in my book is British English (colour, centre). Should I change these to American spellings?

No.

Stay true to the edition you have. And this applies the other way, as well: if you have an American edition of a work by an English author, please leave the spelling as it is.

V.128. I'm nearly sure that some words in my printed book are typos. Should I change them?

The first thing to be aware of is that typos in books are not as rare as most people think. You may never have noticed typos in your normal reading, but under the kind of scrutiny that a book gets while being produced for PG, they often do become noticeable. It's quite common to find anything up to ten typos in a book.

Before you decide it's a typo, though, check that the same word doesn't occur elsewhere in the book with the same spelling. Often, the words or spelling used by pre-20th Century authors may just not be familiar to you.

When you find something that you believe to be a typo, you have four options: pretend you didn't see it :-), change the typo and add a transcriber's note [V.97], change the typo without a transcriber's note, or leave the typo as it is and add a transcriber's note. If you are adding a note, do it at the top or bottom of the file; don't try to work it into the text, and don't use the [sic] convention, since the reader won't know whether the [sic] was added by you or an earlier publisher.

In general, it's safest to leave the typo in place and add a note at the end of the file, listing the words you believe to be typos; that is the least contaminating and intrusive method. When adding the note, you don't need to leave a mark in the main text. You can just say something like:

[Transcriber's Note: "haw" near the end of chapter 15 appears to be a misprint for "hawk".]

The danger in making changes is that you may be wrong, and we really don't want to corrupt the text. This is particularly so in some old books where archaic usages, now obsolete, may look downright wrong to modern eyes. Sometimes, though, a typo is just so blindingly obvious that it warrants immediate replacement. Even in these cases, conscientious people will sometimes add a note, something like:

[Transcriber's Note: in chapter 12, I have changed "he stood on the tock", to "he stood on the rock".]

V.129. Having investigated what looks like a typo, I find it isn't. Do I need to do anything?

Often in PG work, you come across an odd word or usage. Might be a typo; might not. You check it out, and find that it is deliberate--perhaps a word from local dialect that just happens to resemble a different word, perhaps the author is using an odd word or spelling to make a point with the language. Especially if it's an isolated incident, and especially if it's not obvious, you can add a transcriber's note to the end noting that the word is thus in your edition, and that it is probably right. This may prevent some well-intentioned converter from changing it.

It's rare that you will need to do this; you may encounter such a case only once in a hundred PG books, but it is an option.

V.130. Aarrgh! Some pages are missing! Do I have to abandon the book?

No. It happens more often than you might think, and we're quite used to dealing with it.

Finish the book, and ask other volunteers to help by finding another copy of the book to fill in the missing section. For something like this, you can try asking on [V.12] the WebBoard, or gutvol-d, or ask Michael Hart to put a note in the Newsletter asking for assistance. We can post the book incomplete, and put a Transcriber's Note [V.97] in the header asking any future reader who has a copy to fill in the gap.

V.131. Some words are spelled inconsistently in my book (e.g. sometimes "surprise", sometimes "surprize"). Should I make them consistent?

No.

English spelling didn't really standardize until the start of the 20th Century (and even then it fractured; e.g. "standardize" vs. "standardise") and the further back you go, the more inconsistent it becomes. Shakespeare, for example, signed his own name with several

different spellings.

Where your printed edition genuinely uses alternate spellings of the same word, you should preserve them.

Word Processor FAQ

W.1. What's the difference between an editor and a word processor?

An editor shows you the characters you type, exactly as you type them. It puts new-line characters in when you hit the Enter key, and only when you hit the Enter key. Its ultimate aim is to give you exact control of plain text. EDIT in DOS, Notepad in Windows, vi and emacs in *nix, Tex-Edit Plus and BBEdit Lite in Mac, are all editors.

A word processor, in addition to entering the characters, also lets you change the font, the size of individual words, and whether they are italic or bold. It doesn't generally want individual line-ends put in on each line; it just rewraps the text as you change it. Its ultimate aim is to print your document on paper with full formatting facilities. WordPerfect for MS-DOS and Windows, MS-Word for Windows and Mac, AbiWord for Windows and Linux, and Nisus Writer for Mac are all word processors.

W.2. Should I use an editor or a word processor?

For dealing with plain text, which is what PG is about, you might expect a text editor to have the edge, since the formatting features of word processors can get in the way of making a clean text.

However, if you use a word processor, and you ignore all of the layout and formatting that have to do with fonts and paper, it will work equally well. There are a few common problems associated with Word Processors mentioned below.

W.3. Which editor or word processor should I use?

The one you like best!

Any of them will do the job. Even the most primitive editors of 1971 will do the job. The most feature-bloated word processor of tomorrow will do the job. No editor or word processor affects in the slightest the "quality" of the text produced.

For PG purposes, therefore, the only difference between them all is

how easy you find them to use, and what facilities they have for helping you--and those are decisions that only you can make.

If you already have a favorite editor or word processor, stick to it. If you don't, there's a huge selection available for you to consider, on any type of computer.

Sometimes, using a word processor, you may encounter some problems in saving your book as plain text. You have to figure out how to get it right just once, and then use that same method thereafter. If you have problems with this, ask other volunteers or one of the Posting Team for help.

W.4. How can I make my word processor easier to work with for plain text?

First, switch off _everything_ called "Smart ------" or "Automatic". Modern word processors commonly offer lots of typical typing support features--"Smart Quotes", "Auto Correct", automatically capitalizing the first word in each sentence, anything like that. By all means, leave on any informative highlighting of misspelled words or other errors that it offers, but switch off any feature that changes what you type without asking you. Older books contain text that doesn't sit comfortably with modern rules, and we don't want your word processor deciding what Chaucer really wrote!

Now, choose a non-proportional font, and apply it to the whole document. It's important to work in a non-proportional font, because you may have to line words up underneath each other and it is not possible to do this consistently in non-proportional fonts like Times or Arial.

If you work in Courier, size 10, 11 or 12, and your word processor is set for a normal page size, about 7 inches across excluding margins, then what you see in your WP is a pretty good approximation to how the text will look in PG plain text format. One formula, suggested by John Mamoun in the Volunteers' Voices section, is to Select All the text, choose Courier New font, 10 point size, and set the margins at 5.5 inches, then Save As "Text with layout".

W.5. What is the difference between proportional and non-proportional fonts?

A non-proportional, or "monospaced", or "typewriter" font, is one where all of the letters take up exactly the same amount of space on screen: a capital "W", a lower-case "i" and a space are all equally wide. The Courier family of fonts is commonly used for this.

A proportional font is one where each letter takes up just the amount of space it needs, so that a capital "W" is much wider than a small

"i".

Unfortunately, the different sizes of the letters in different proportional fonts means that it's not possible to line up letters consistently: a "W" may be equivalent to three "i"s in one proportional font, and to four "i"s in another. This means, for example, that it is not possible to use a proportional font to format plain text tables or poetry correctly--lining up the spaces and words using one proportional font will cause it to look skewed using another.

You should always look at PG texts in a non-proportional font, even if you prefer to work mostly using a proportional font, because readers and automatic converter programs will assume that you meant to your text to be viewed using a non-proportional font.

W.6. I can't get words in a table or poem to line up under each other.

You are using a proportional font. You should always use a non-proportional font like Courier for PG work. Change the font of the entire document to Courier and try again.

About using Microsoft Word:

PG volunteers use many different word-processors, but Microsoft Word is the one we hear most queries and problems about.

W.7. I've edited my book in Word--how do I save it as plain text?

First, make sure that all text is using Courier or Courier New and is at the same point size (usually 10-12). Move your right margin so that you see roughly the right number of characters per line (usually 65-70). Then choose File / Save As and then choose the format "Text Only with Line Breaks". Save your file with the extension ".txt" to distinguish it from your Word format file.

After saving, open your text file using Notepad or some other simple text editor and look at the results. You should see a typical PG layout of the text--lines up to 70 characters long, a blank line between paragraphs and no indentation at the start of each paragraph. If so, you're done.

W.8. Quotes look wrong when I save a Word document as plain text.

You may have left "Smart Quotes" on in Word options. This tells Word

to use left- and right-slanted quote marks at the beginning and end of a quote instead of the plain ASCII straight quotes. When you save a document that contains these angled quotes as plain text, they come out as non-ASCII characters that look wrong on most editors and viewers. The solution is to turn off Smart Quotes in Word and/or replace the ones it has already created.

W.9. Dashes look wrong when I save a Word document as plain text.

When Word recognizes an em-dash as such, it may try to use a special character for it. This may appear as a black square, an empty box, or a funny accented letter when you Save As text and look at it in a different editor.

You can usually do a Find and Replace on this character either in Word or in another editor after Saving As text to change it to two dashes.

For those interested, the "funny character" is character 151 (97H), and is specific to Codepage 1252 [V.76].

W.10. I saved my Word document as HTML, but the HTML looks terrible.

Yes. Word is not unique in having this problem, but HTML saved from Word is the case we hear most about. Microsoft themselves offer a free plug-in to Word that saves the file in "Compact HTML", which is a bit better. You can fix it by hand, or you can use Tidy <http://tidy.sourceforge.net>, a handy utility, which will do some clean-up on the HTML. If you're working with HTML, you really need a copy of Tidy anyway, because it's such a great way to do a check on the correctness of your HTML.

Tidy is also embedded in some Windows GUI tools, like Tidy-GUI, HTML-Kit and NoteTab.

Scanning FAQ

S.1. What is a scanner?

A scanner is a machine that makes an image, a picture of the page that is fed to it, and sends that image to your computer. It only makes an image, like a camera does; it doesn't turn that image into text.

S.2. What types of scanners are there?

The most common type of scanner, the kind you're likely to find in your local computer store, is a flatbed scanner. It has a glass bed usually a bit bigger than Letter paper size (or A4 if you live in Europe! :-) and most of the common models are optimized for typical office correspondence. One of these may cost anything from under $100 to $400, depending on its features, or you can pick them up cheaper second-hand. You use this by placing the paper or book face-down flat onto the glass, and scanning from there. This is the kind of scanner most commonly used by PG volunteers.

Some stores will call sheetfed scanners a different category. These are flatbed scanners with Automatic Document Feed (ADF), but they are fundamentally the same machine, and the ADF sheetfeeder unit may often be bought as an accessory to the flatbed scanner. Recently, a few sheetfed scanners have appeared that are very small, without a full flatbed, just a narrow strip that the paper rolls through. Avoid these for PG work; you often need to be able to scan the book flat.

Hand scanners, as their name implies, are much smaller, and typically very cheap, or even thrown in free. You use these by holding them in your hand and running them along the text like a brush. These are really not intended for PG work; you need a very steady hand movement to get them to scan a page of text into a readable image, and they shouldn't be considered as an option for a 400-page book--scanning and OCR is tough enough without that!

You can think of production scanners as industrial-strength flatbed scanners. The basic mechanisms are the same, but a production scanner will certainly have ADF (sheetfeeder), more features and speed, and be rated for very high volume scanning. Production scanners are used by publishers, businesses with high-volume paper processing needs, and print shops. This last is useful, because you may be able to get some scanning done by a print shop. It can't hurt to ask. If you're thinking about buying one of these babies (and who among us hasn't? :-), be sure you have $2000 or more to spend.

Drum scanners are mostly used by publishers for professional, high-quality artwork. The paper is placed on the surface of a drum that rotates past a fixed scanning head. The drum can be very large. Because the sensors don't have to move, the electronics and optics can be of higher quality, and produce very accurate, high-definition images. They are exactly what you would want for making professional quality scans of old movie posters, but they're expensive, and not very useful for scanning War and Peace to OCR.

Planetary scanners are a different breed to all the others. They are really not scanners at all, but a very high-end digital camera on a stand. You place the book face-up with the pages open, with the camera looking straight down on it. It takes a picture, and passes it on to the connected computer. Planetary scanners are ideal for old, fragile, valuable books that can't be exposed to the stress of normal scanning. They typically come supplied with specialized software, sometimes even

their own dedicated computer, and they are very, very expensive--$20,000+.

S.3. Which scanner should I get?

For most people, the answer is simple. Unless you have a lot of money and are sure you will be scanning a lot of books, you should get a normal, consumer-or-office type flatbed scanner, with or without an ADF sheetfeeder.

Having decided that, you're faced with the question of which scanner to buy. More good news! The market in scanners is very competitive, and there are many top-line vendors all watching each others' features like hawks, eager to deliver the highest-spec machine they can. There are only a couple of critical factors in this decision--most of it is about getting the best buy.

For PG work, you really _need_ an optical resolution no less than 300 by 300 dpi (dots per inch), and 600 by 600 is very desirable. Obviously, more is better, but it would be very rare to need more than 600 dpi for PG work. Pay no attention to the "interpolated" or "enhanced" resolution, where the software "guesses" what dots should fill in the gaps--you're only interested in the optical resolution. The good news is that it's very difficult to find modern scanners with a maximum optical resolution of less than 600 dpi, but if you're buying second-hand, you should check this out first.

You will also _need_ a scanning surface on the glass big enough to place your book with two facing pages flat. Again, the good news is that it's very hard to find a flatbed whose scanning surface is too small for PG work, since these scanners tend to be designed to handle office paper, which is about the right size. Most flatbed scanners have scanning surfaces of about 8.5" by 11.5", and this is standard for PG work. If you're working on books with very large pages, you may need to resign yourself to scanning one page at a time, but buying a scanner with a big flatbed for these rare occasions will be much more expensive.

You must make sure that you get a scanner that will connect correctly to your computer. There are currently (mid-2002) three main types of connections commonly available: SCSI, USB, and parallel.

SCSI (Small Computer Systems Interface) is the highest-quality option, but it means that you need a SCSI card in your computer, and be willing to figure out how to install it. If you're already a SCSI enthusiast, you don't need to read further; if you're not, I suggest you avoid it unless you enjoy tinkering. Production scanners mostly require SCSI.

Parallel-port connections used to be common, as a cheaper, easier alternative to SCSI. Since the introduction of USB they have become

rarer, but you will still see them for sale second-hand. These plug into your printer port, and don't require any further engineering skills.

Most new scanners hook up using a USB (Universal Serial Bus) interface, which is a no-muss, no-fuss "plug-in and go" option, but be sure, if you have an old PC, that it actually has a USB port and that your operating system supports it; some older Windows PCs and Macs may not. If your PC doesn't support USB, you should probably look at Parallel-port scanners.

By the time you read this FAQ, FireWire and USB 2.0 interfaces may also be common. For your purposes, these are like more advanced versions of USB. Just make sure that your computer has the right support to match the scanner.

If you're buying second-hand--and used scanners can be very cheap--make absolutely sure that you're getting the original software that came with the scanner, and that that software will work with your current operating system on your PC.

Having ensured that your choice of scanners passes these tests, you're now free to indulge your tastes for any extras you like. Color is nice, but rarely used, since we mostly transcribe older books that have no color printing. Higher resolutions are comforting to have, both since you may occasionally find them useful and because it shows that the optics are of higher quality than you actually need for your PG scans.

If you are nervous about your choice of scanner, or how easy it is to get one working, feel free to contact other PG volunteers for their opinions, as described in the FAQ "How do PG volunteers communicate?" [V.12].

S.4. What is ADF?

ADF stands for Automatic Document Feed, and it's just a jargon term for a sheetfeeder, where you put in a stack of pages to be scanned and go away while that's happening instead of putting in each page manually.

S.5. Should I get ADF?

That depends. Yes, ADF is a great idea, and can be a huge work-saver, and if you have the cash to spend, it may well be worth it. But ADF has a dirty little secret: like any other gizmo with moving parts, it occasionally jams. The sheetfeeders built into these low-cost machines are aimed at handling typical office paper straight from the laser printer--large, smooth, good quality, with perfectly-cut, perfectly-aligned edges. In your PG work, you will be dealing with

hundred-year-old pages of various thicknesses and textures, usually much smaller than the sheetfeeder was designed to work with. And you will have to have cut the pages, and may leave ragged edges in doing so.

Under these conditions, you may find that paper often jams in your sheetfeeder, and it defeats the purpose if you have to stand over the scanner while it works, or if you end up having to lift the cover and use your scanner as an ordinary flatbed, or, worse, if your paper gets scrunched up as if a dog had been playing with it.

And of course, in order to feed the pages through, you will have to cut them out of the book, destroying it. (It may be possible, with the help of a bookbinder, to have the pages professionally cut, and later re-bound.)

With ADF, you probably won't actually scan much faster than scanning flat, but you won't have to keep turning over the pages during that time.

So when you're making that choice, think carefully. If money isn't a problem, or you do expect to be working with cut sheets, then go ahead and get a sheetfeeder--it's great when it works! But don't be disappointed when it doesn't work all the time.

S.6. What's a "TWAIN driver" and why do I need one?

A TWAIN driver (see <http://www.twain.org>) is a piece of software that installs onto your Windows PC or Mac and controls your scanner from there. With any modern scanner, there will be a TWAIN driver included in its software package. Once installed, you shouldn't have to think about it again, or even know it's there.

A modern OCR package will usually find your TWAIN driver and use it to control the scanner. This is very handy. There may also be a small scanning package with your TWAIN driver, which will provide a screen where you can make fine adjustments to scanner settings, and start scans. You probably won't _need_ this, since your OCR package will probably do it for you, but it may be useful for semi-manual control of the scanner.

Unix-based systems like Linux use SANE <http://www.mostang.com/sane/> rather than TWAIN drivers.

S.7. How do I scan a book?

This depends on whether you have cut the pages out, or whether you are working with an intact book.

If you have cut the pages out, and you have an ADF, then you will obviously feed them through that.

If you don't have an ADF, there usually isn't much point in cutting the pages. Most modern OCR will recognize a "dual-page" or "two-up" scan, and, if yours does, then that's normally the best option. Scanning the uncut book, open and flat, is the most common scanning method used in PG.

Take the book and place it open, flat on the scanner glass. To fit both pages on the glass, you may need to position it lengthways, at 90 degrees to its natural angle. Most OCR software will recognize that the image has been rotated through a right-angle, and will correct it when it reads the text.

A common problem with scanning an opened book is "guttering", which happens when the spine of the book is not pressed flat enough, and the inside of each page, where it meets the spine, is curved against the glass. There's more about this, and an example, scan3, in the FAQ [S.17] "Why am I getting a lot of mistakes in my OCRed text?". To avoid guttering, make sure that the spine is held down throughout the scan. (Some people put a weight on the spine to hold the spine down on each scan; others just press their hand against it.)

Another common problem is light scattering, when too much light gets into the scanner. The scanner head detects light, and you want the only internal light source to be from the scanner itself, not ambient room light or sunlight. Scanners have covers, that are intended to be closed while scanning, for a controlled light level, but when you're scanning a book held open and flat, you can't close the cover fully. In a bad case, this can lead to a condition of the scan like overexposure of film and you can see an example in scan4 of the FAQ [S.17] "Why am I getting a lot of mistakes in my OCRed text?". If this happens, just make sure that your room is dim while you scan--don't have a ray of bright sunlight bouncing around the inside of the scanner!

Occasionally, when scanning cut pages with very thin paper, you may get a shadow of the text on the other side showing through. If this happens, you can try covering the inside of the scanner lid, which is normally white, with a piece of black paper.

Many modern OCR packages will control the scanner automatically, and you may be able to set your OCR so that it does an automatic timed scan every, say, 30 seconds. This is a great timesaver, since you don't have to go back and forth between the scanner and the screen. Just set your timer, hold down the book for the scan, take the book up, turn the page, put it down again, and wait for the next scan to start. Set the timer for whatever interval you are comfortable with. Highly recommended, if your OCR or scanning package can do it.

By default, most scanners will always scan the entire area of the flatbed, but usually, your book will occupy only about half of it.

Look for a setting on your OCR or scanning package which allows you to reduce the area that the head scans. Just scan enough to get the image of your pages. This makes the time for each scan and subsequent OCR recognition shorter, and in a really good case can cut your total scanning and OCR time in half.

Scanning all pages together is usually fastest, but you may prefer to scan each double-page, then correct it in your OCR package's editor, then scan the next. This is a more leisurely approach favored by some volunteers.

S.8. My book won't open flat enough for a good scan, and I don't want to cut the pages.

Well, then, you have a difficult choice to make, but you do still have several options:

You can accept a poor-quality scan, and spend a lot of time fixing up the guttering on the margins.

You can bite the bullet, and cut the pages.

You can type the book, or find a typist who will work on it for you.

You can find a print shop or bookbinder who will cut the pages professionally, and re-bind the book when you're done. You may even replace it with a fresh new binding that will give the book a new lease of life.

Take your choice.

Most books will open flat enough for an adequate scan, though you may have to put stress on the spine to do it.

If you have a really precious book, and you can't find a typist, you might consider the options of a digital camera [S.11] or finding someone with a planetary scanner [S.2] to scan it for you.

Michael Hart said: "I would give up every book I own, including my first edition of the OED, my Civil War edition of the Merriam Webster's Unabridged, etc., etc., etc., so everyone could use it any time they wanted rather than that only I or my friends could use it . . . and obviously _I_ could use it too."

Fortunately, it rarely comes to that.

S.9. How long does it take to scan a book?

Putting the book flat on the glass means that you scan two pages at a

time. A reasonable modern scanner will scan the area of two typical pages at 400dpi in anywhere from 20 to 40 seconds--let’s call it 30 seconds for two pages. That’s four pages a minute, or 240 pages an hour. You could reasonably get through a 400 page book in two hours, even allowing for an occasional break or glitch.

Of course, you should also allow time for scanning a few trial pages with different settings before you start, to decide which settings to use. Ten minutes spent here can save you hours of proofreading time.

There are two big tips that can save you a lot of scanning time:

If your OCR or scanner control package has a timer setting, that automatically keeps scanning without user intervention, you can forget about the screen and just keep turning the pages as needed.

You should set your scanner just to scan the area the book covers on the glass. By default, your software will probably scan the full area of the glass, and usually, your book won’t need that. By scanning only what you need, you may typically save anything from 20% to 70% of the time taken to scan the full area. If your book is small enough to open flat _across_ the scanner instead of "down" the side, 400 pages an hour is not out of the question with this trick.

S.10. What scanner settings are best?

For a given book, scanner, PC and OCR software, there must be some "ideal" scanner settings, but if you change any of these components, the ideal scanner settings will change with them. Some OCR packages recognize greyscale better than black and white; some don’t like greyscale at all. Some books have small print needing higher resolution; some are speckled so that higher resolution leads to more errors.

Obviously, the best settings also depend on the individual book, and some books will require you to get downright creative with the settings, but most PG books are scanned in Black and White or greyscale, somewhere between 300dpi and 600dpi.

This decision is a trade-off between speed and accuracy, and an illustration of the difference between principle and practice. In principle, a true-color, 9600dpi scan is a much better rendering of the page than a B&W 400dpi scan. In practice, all that extra information doesn’t usually help the OCR make better distinctions between letters, and the larger and more detailed the scan, the longer it takes to make the scan, the more disk space the image file takes, and the more processing time and memory the OCR package needs to recognize it.

A further paradox emerges when considering higher vs. lower resolutions: depending on the paper and ink quality, you may see

more errors start to appear on very high resolution scans. These are caused by small imperfections in the paper or ink spots that show up on the high-res scan, and that the OCR tries to interpret as letters or punctuation.

So, in summary, bigger is better, but only up to a point.

Brightness is a setting often neglected, that can make quite a big difference to your results. Look at the scanned image: if you see lots of dark patches, make your scan lighter; if your letters appear thin and faded, make your scan darker.

See the FAQ [S.17] "Why am I getting a lot of mistakes in my OCRed text?" for some typical scans and results.

S.11. Can I use a digital camera in place of a scanner?

Digital cameras are getting better resolution all the time, and some volunteers have experimented with making a kind of home-made planetary scanner from a digital camera and a stand. So far, the results don't quite match a dedicated scanner, but as digital cameras improve, this may become a common option. One problem, which planetary scanners use specialized software to correct, is that the natural curve of the pages near the middle of the book tends to give a foreshortened aspect to the letters there, which can cause problems for OCR software, like guttering.

Whatever the current problems, the prospect of using digital cameras is exciting, because it will mean that non-typists will be able to produce old books borrowed from libraries without worrying about scan quality vs. damage to the spine.

S.12. What is OCR?

OCR stands for Optical Character Recognition. This is very important software that looks at the picture of the page that your scanner has supplied, and turns it into text.

When the scanner delivers the image of the page, that image is only a picture. You can't, for example, search for text in it, or edit the text to add a blank line. Your editor or word processor can't work with it. The OCR program does the job of "reading" and "typing" the image for you. OCR packages call this "reading" or "recognizing".

S.13. What differences are there between OCR packages?

One word: huge. All OCR packages do the same job, but they do it in

different ways, with different features, and with different levels of accuracy. OCR can save you a lot of time, or cost you a lot of time. It's really worth putting some effort into making sure you get the right OCR package, and, once you have it, into understanding how to use it. It'll save you time in the long run.

S.14. How accurate should OCR be?

OCR packages commonly say that they are "99%+" accurate, or something like that. Let's analyze what that actually means: say there are 1,000 characters (letters) on each page, then with 99.9% accuracy, you would expect to have to make 1 correction per page. With 99% accuracy, that would be up to 10 corrections per page. And in a 400-page book, this all adds up.

But there's a "Your Mileage May Vary" clause built into that. Typically, the manufacturers test their OCR on fresh, laser-printed or press-printed copy with perfect scans, and this is fair, since they are aiming their products primarily at businesses that process these kinds of materials. _You_ are not dealing with fresh print; you're dealing with old books, yellowed, spotted, marked, imperfectly printed in the first place, and possibly using unfamiliar fonts. And it's unlikely that you will have the patience to get a perfect scan on every page. The result is that the accuracy of OCR for typical PG work doesn't match the accuracy on images of perfect, fresh paper.

Apart from the scan quality, OCR also has to contend with different fonts and sizes for the letters.

However, if you're getting more than 10 errors per page, you should look at some examples of OCR in the FAQ [S.17] "Why am I getting a lot of mistakes in my OCRed text?".

S.15. Which OCR package should I get?

The accuracy of OCR software has improved enormously in the last few years, and OCR technology looks likely to keep improving even faster than software in general. Further, there is competition in this area, and products leapfrog each other with new versions regularly. The brands most commonly mentioned by PG volunteers (mid-2002) are Abbyy, OmniPage and TextBridge [P.1], and trial versions of all three have been available for download over the Web, and may still be when you read this. [Warning: these are big downloads--40MB or more.]

Most common OCR packages will offer two main working options: to scan a page and view/edit the resulting text on the spot before saving, and to scan a whole batch of pages together and view/edit them all later. Some people like to fix up one page at a time; others prefer to get all of the OCR work done at once, then get the whole text into their

editor. Most OCR software will cater for both, and if this is important to you, you should check that the OCR you're buying supports the way you want to work.

If you intend to work in a language other than English, make sure that the OCR you buy supports the characters in your language.

Some OCR software has a "training" or "learning" mode. Using this mode, it scans and "reads" or "recognizes" a page, then you correct that page, and the OCR "learns" from its mistakes and tries to do better on the letters it misread when it recognizes the next page. If you're dealing with a very rare font, this can make a difference to your OCR quality, but modern OCR packages come with enough inbuilt font knowledge for most languages, and you probably won't need this.

If possible, try a couple of OCR packages before you decide. If you want opinions on specific versions, contact other PG volunteers and ask for their opinions, as described in the FAQ "How do PG volunteers communicate?" [V.12].

S.16. What types of mistakes do OCR packages typically make?

Each text has its own peculiarities, but there are a number of well-known scanning errors you will be dealing with all the time.

Punctuation is always a problem. Periods, commas and semi-colons are often confused, as are colons and semi-colons. There are also usually a number of extra or missing spaces in the e-text.

The problem of quotes can assume nightmarish proportions in a text which contains a lot of dialog, particularly when single and double quotes are nested.

The numeral 1, the lower-case letter l, the exclamation mark ! and the capital I are routinely confused, and often, single or double quotes may be mistaken for one of these.

Lower-case m is often mistaken for rn or ni.

The letters h and b and e and c are commonly mis-read, and these are probably the hardest of all to catch, since ear/car, eat/cat, he/be, hear/bear, heard/beard are all common words which no spell-checker will flag as problems.

For example:

" Hello1' called jirnmy breczily. 11Anyone home ? "

There seemed to he no-oneabout. Only tbe eat beard him."

should read:

"Hello!" called Jimmy breezily, "Anyone home?"

There seemed to be no-one about. Only the cat heard him.

S.17. Why am I getting a lot of mistakes in my OCRed text?

If you're new to OCR, you may have come with the idea that OCR is almost perfect, and just makes a few mistakes now and then. No. It's slightly amazing that OCR works at all, and when it does, it isn't perfect.

You might reasonably expect to average anything up to 10 errors per page for typical PG work; if you're seeing more, then there is a problem with

a) your printed book
b) your scan, or
c) your OCR package

Problems with the printed book fall into three categories: bad printing, age, and unusual fonts. Bad printing consists of problems like too much or too little ink on the press at the time the book was printed, and irregularities in the print where the metal type was damaged. Age causes yellowing--even browning--of the paper, and faded print. Unusual fonts may be hard for OCR to recognize, and very tightly-spaced print may make adjacent letters seem to touch, which confuses OCR software.

There are many ways for you to have problems with your scan. Obviously, if your scanner is defective or the glass is dirty, you will notice it immediately, but there are many mistakes you can make that will result in a poor-quality image, and cause later problems for your OCR.

You may not be able to control the quality of the paper you have to work with, but there is a lot you can do about the quality of your scan.

The two mistakes that people inexperienced with scanners most commonly make are not holding the spine down firmly enough to get a flat image of the paper, and not setting the brightness correctly, or letting too much light get in. In your early scans, watch out for these problems.

First, if you haven't already, read the FAQ "How do I scan a book?" [S.7] and check that you're following the basic recommendations there.

Now let's look at some samples, and see the kinds of problems you might encounter.

A disclaimer about these samples: specific OCR packages are named, but

you should _not_ take these as a fair and comprehensive comparative review of the software. The object of this exercise is to show typical scanning conditions and problems, and the resulting OCR output. OCR packages have quite a range of variance within themselves, may work better on some texts than others, may improve with "training" or different settings, and I have even seen the same OCR package produce different text from the same image with the same settings! Further, since OCR quality is improving rapidly, and packages leapfrog each other in quality, the next version of a particular brand may be vastly better than any of the software mentioned here. Of particular interest in this context is the leap in quality between OmniPage 10 and OmniPage 11.

* * * * *

Scan 1--A perfect Scan

Scan1 is as near to a perfect scan as you can expect in PG work. It comes from "The Founder of New France" by Charles W. Colby. It is only a 300 dpi image, but given the quality of the print and of the scan, 300dpi is all we need. Ironically, it comes from Gardner Buchanan, who complains about the age and infirmity of his scanner in his description of how he produces a text. The moral is that you don't have to have the latest equipment to get good results!

The actual scan is in the image file scan1-3.tif

It doesn't really need any comment, and all of the packages except gocr rendered it perfectly. Note the fake "space" before the semicolon--if you look closely at the image, you will see why the OCR packages mistook it for a full space, as discussed in the FAQ [V.104] "My book leaves a space before punctuation like semicolons, question marks, exclamation marks and quotes. Should I do the same?"

> Champlain was now definitely committed to
> the task of gaining for France a foothold in
> North America. This was to be his steady
> purpose, whether fortune frowned or smiled.
> At times circumstances seemed favourable ;
> at other times they were most disheartening.
> Hence, if we are to understand his life and
> character, we must consider, however briefly,
> the conditions under which he worked.

gocr 0.3.6 converted this as:

> Champtain was now definitely committed to
> the task of gaining for France a foothotd in
> _orth America. This was to be his steady
> purpose, whether fortune frowned or smiled.
> At times circumstances seemed favourable .,
> at other times they were most disheartening.

_ence, if we are to understand his life and
character, we must consider, however brieRy,
the conditions under which he worked.

* * * * *

Scan 2--A Typical Scan

Scan2 is a paragraph from Baroness Orczy's "Castles in the Air". Notice the ink-splotch above the capital "I" in the first line, which will give our OCR some problems. The page is also unevenly inked elsewhere, and I have scanned it with the brightness level a bit too high.

I have made two separate scans, one at 300dpi and one at 400dpi, both Black and White, named scan2-3.tif and scan2-4.tif respectively. The page was cleanly cut, and carefully placed straight onto the scanner glass with the cover down. The original print is somewhere between the size of Times New Roman 10 and 11, with capital letters about 2.2 millimeters high, but better and more clearly spaced. These scans are fairly typical for PG work. Because of the relatively large letters, and the reasonable scan, there isn't much difference between the text produced from the 300 dpi scan and the 400 dpi scan.

I actually cut this book to get the pages out so that I could feed it through my ADF, but the paper is so thick and textured that it sticks together, and jams when feeding through. The thick, absorbent paper, combined with the uneven inking, means that, no matter how good the scan, any OCR has to contend with the irregular edges of letters, which are clearly visible even at 300dpi.

Here is the output for these scans from some OCR software packages. I changed just one thing: Abbyy recognized the em-dashes as such, and output them as a special character in Codepage 1252 for em-dashes, which isn't available in ASCII, so I converted that to the PG standard 2 dashes.

Abbyy FineReader 6:

Yes, indeed, I was on the track of M. Aristide Fournier, and of one of the most important hauls of enemy goods which had ever been made in France. Not only that. I had also before me one of the most brutish criminals it had ever been my misfortune to come across. A bully, a fiend of cruelty. In very truth my fertile brain %vas seething with plans for eventually laying that abominable ruffian by the heels: hanging would be a merciful pun-ishment for such a miscreant. Yes, indeed, five thousand francs--a goodly sum in those days, Sir--was practically

assured me. But over and above mere lucre there was the certainty that in a few days' time I should see the light of gratitude shining out of a pair of lustrous blue eyes, and a winning smile chasing away the look of fear and of sorrow from the sweetest face I had seen for many a day.

Yes, indeed, Twas on the track of M. Aristide Fournier, and of one of the most important hauls of enemy goods which had ever been made in France. Not only that. I had also before me one of the most brutish criminals it had ever been my misfortune to come across. A bully, a fiend of cruelty. In very truth my fertile brain was seething with plans for eventually laying that abominable ruffian by the heels: hanging would be a merciful punishment for such a miscreant. Yes, indeed, five thousand francs--a goodly sum in those days, Sir--was practically assured me. But over and above mere lucre there was the certainty that in a few days' time I should see the light of gratitude shining out of a pair of lustrous blue eyes, and a winning smile chasing away the look of fear and of sorrow from the sweetest face I had seen for many a day.

gocr 0.3.6:

__e_, indeed, f___as on_the track of h_. hristide Fournier, 3nd of one of the most im__ant hau1s of enem)_ goods ___hich had e__er been made in France. h?ot onl3_ that. I had a1so before me one of the most brUtish crimînat_s it h__4 e___er been m31 misfortune to co_me acro__3. A bu113_, a tiend oí cruelt__. In very truth m3_ fertiIe brain ___as s_e_1_::_g __-ith planS for e__entua113_ _ay:ng that abominab1e ru_iin b.__ t1_e hee1s . hanginig __ou1d be a n_erciful pun-i;__,i__gnt íor such a miscreanf. yes, in_i__ee3, fj_1e thou3and francî-a b_ood13_ sum in those days, _ir-_vas practica1l3_

a3_ured me. _ut o___er and above n_ere lucre there was the certaint_v that in a few_ da3_s' ti_e I shou1d see the lib_ht of gratitude shininb_ out of a pair _f _usLtrous btue e3_e3_, and a ___inning smi1e chasing a__ay the Ioo_ of _ear and of sorrow from the s__eetest iace T had Seen fof man)_ a day.

Yes, indeed, f___as on the track of h__. Ariseide Fournier, and of one of the most important hau1s _f enemy goods ___hich had ever been made in France. NoEUR on1y that. I had also before me one of the most brutish crimina1s it h_ad ever been my misfo__tune to come acros__. A bu11y, a fiend of crue1ty. _n very truth my fertib brain _vas seeî3_:i_g __ith plans for e__entua11p 1aying _at abom_in_ ab1e ru_an by the heels. hanging _____ou1d _ a merciful pun-

iï_h_ment for such a miscreant. Yes, indeed, five thou__and
f_ancs-a b_ood1y sum in those days, _ir-_vas practica1ly
a3îured me. But over and above mere _ucre th.ere was
th_e certainty that in a few days' ti_e _ shou1d see the
1i__t of gratjtude shining out of a pair o_, _userous b1ue
 b .
e__es, and a __inning smi1e chasing away the l_k of
_,ear and of sorrow from the s___,eetest face _ _ad _.een _o_
many a day. . .

Recognita Standard 3.2.7AK:

~'es, indeed, ~w-as on the track of ltT. Aristide Fournier,
and of one of the most important hauls of enemy goods
"=hich had ever been made in France. ~Tot only that. I
ha~i also before me one of the most brutish criminals it
had ever been my misfortune to come across. A bully-, a
fiend of cruelty. In very truth my fertiIe brain was
s; ething w-ith plans for eventually iaying that abominable
ruffian by the heels : hanging ~-ould be a merciful pun-
ishment for such a miscreant. ires, indeed, five thousand
franes-a goodly sum in those days, Sir-was practically
as~ured me. But over and above mere lucre there was
thP certainty that in a few days' time I should see the
light of gratitude shining out of a pair of lustrous btue
ey•es, and a winning smile chasing away the hk of
fear and of sorrow from the sweetest face I had seen for
many a day.

Yes, indeed, l~was on the track of h~i. Aristide Fournier,
and of one of the most important hauls of enemy goods
w~hich had ever been made in France. lVot only that. I
had also before mP one of the most brutish criminals it
had ever been my misfortune to come acrass. A bully, a
fiend of cruelty. In very truth my fertile brain was
seething with plans for ez~entually laying that abomin_ able
ruffian by the heels : hanging ~~.-ould be a merciful pun-
ishment for such a miscreant. Yes, indeed, five thousand
f:ancs-a goodly sum in those days, Sir-was practically
assured me. But over and above mere lucre there was
the certainty that in a few days' time I should~ see the
light of gratitude shining out of a pair of iEustrous blue
eyes, and a w inning smile chasing away the look of
fear and of sorrow from the s"-eetest face ~ had seen ~'or
rr~any a day.

OmniPage Pro 10:

Yes, indeed, twas on the track of 11T. Aristide Fournier,
and of one of the most important hauls of enemy goods
which had ever been made in France. Not only that. I

ha(i also before me one of the most brutish criminals it had ever been my misfortune to come across. A bully, a fiend of cruelty. In very truth my fertile brain was seething with plans for eventually laying that abominable ruffian by the heels: hanging would be a merciful punishment for such a miscreant. Yes, indeed, five thousand francs-a goodly sum in those days, Sir-was practically assured me. But over and above mere lucre there was the certainty that in a few days' time I should see the light of gratitude shining out of a pair of lustrous blue eyes, and a winning smile chasing away the look of fear and of sorrow from the sweetest face I had seen for many a day.

Yes, indeed, fwas on the track of h-I. Aristide Fournier, and of one of the most important hauls of enemy goods which had ever been made in France. Not only that. I had also before me one of the most brutish criminals it had ever been my misfortune to come across. A bully, a fiend of cruelty. In very truth my fertile brain was seething with plans for eventually laying that abominable ruffian by the heels: hanging would be a merciful punishment for such a miscreant. Yes, indeed, five thousand francs-a goodly sum in those days, Sir-was practically assured me. But over and above mere lucre there was the certainty that in a few days' time I should see the light of gratitude shining out of a pair of lustrous blue eyes, and a winning smile chasing away the look of fear and of sorrow from the sweetest face I had seen for many a day.

OmniPage Pro 11:

Yes, indeed, twas on the track of AT. Aristide Fournier, and of one of the most important hauls of enemy goods which had ever been made in France. Not only that. I had also before me one of the most brutish criminals it had ever been my misfortune to come across. A bully, a fiend of cruelty. In very truth my fertile brain was seething with plans for eventually laying that abominable ruffian by the heels: hanging would be a merciful punishment for such a miscreant. Yes, indeed, five thousand francs-a goodly sum in those days, Sir-was practically assured me. But over and above mere lucre there was the certainty that in a few days' time I should see the light of gratitude shining out of a pair of lustrous blue eyes, and a winning smile chasing away the look of fear and of sorrow from the sweetest face I had seen for many a day.

Yes, indeed, fwas on the track of h-I. Aristide Fournier, and of one of the most important hauls of enemy goods

which had ever been made in France. Not only that. I had also before me one of the most brutish criminals it had ever been my misfortune to come across. A bully, a fiend of cruelty. In very truth my fertile brain was seething with plans for eventually laying that abominable ruffian by the heels: hanging would be a merciful punishment for such a miscreant. Yes, indeed, five thousand francs-a goodly sum in those days, Sir-was practically assured me. But over and above mere lucre there was the certainty that in a few days' time I should see the light of gratitude shining out of a pair of lustrous blue eyes, and a winning smile chasing away the look of fear and of sorrow from the sweetest face I had seen for many a day.

Textbridge Millennium Pro:

Yes, indeed, rwas on the track of M. Aristide Fournier, and of one of the most important hauls of enemy goods which had ever been made in France. Not only that. I hail also before me one of the most brutish criminals it had ever been my misfortune to come across. A bully, a fiend of cruelty. In very truth my fertile brain was seething with plans for eventually laying that abominable ruffian by the heels: hanging would be a merciful punishment for such a miscreant. Yes, indeed, five thousand francs-a goodly sum in those days, Sir-was practically assured me. But over and above mere lucre there was the certainty that in a few days' time I should see the light of gratitude shining out of a pair of lustrous blue eyes, and a winning smile chasing away the look of fear and of sorrow from the sweetest face I had seen for many a day. - - -

Yes, indeed, f was on the track of M. Aristide Fournier, and of one of the most important hauls of enemy goods which had ever been made in France. Not only that. I had also before me one of the most brutish criminals it had ever been my misfortune to come across. A bully, a fiend of cruelty. In very truth my fertile brain was seething with plans for eventually laying that abominable ruffian by the heels: hanging would be a merciful punishment for such a miscreant. Yes, indeed, five thousand francs-a goodly sum in those days, Sir-was practically assured me. But over and above mere lucre there was the certainty that in a few days' time I should see the light of gratitude shining out of a pair of lustrous blue eyes, and a winning smile chasing away the look of fear and of sorrow from the sweetest face I had seen for manyaday. -

* * * * *

Scan 3--Guttering and Smaller Print

Scan3 is a paragraph from "The Egoist" by George Meredith. It was scanned in a dim room, with the scanner cover open and the book held open, flat against the scanner glass. However, the spine was not pressed firmly enough against the glass, and as a result you can see that the words on the left-hand edge (which were near the spine) appear to be slanted, a bit distorted, and not well lit. This problem is familiar to people who scan for PG--everybody gets distracted sometimes, and fails to keep enough pressure on the spine. As you see from the results below, it caused problems for all of the OCR packages on the words affected. If you find this kind of "guttering" regularly in your own scans, where the characters near the spine are not being recognized correctly by your OCR, you need to make sure that your book is down as flat as possible before making a scan. Because of the smaller size and the guttering problem, the 400dpi scan made for better quality text in this case.

Here’s the output from the sample OCR:

Abbyy FineReader 6:

> NEITHER Clara nor Vernon appeared at the mid-day table,
> n Middleton talked with Miss Dale on classical matters,
> like a good-natured giant giving a child the jump from
> stone to stone across a brawling mountain ford, so that an
> uncdified audience might really suppose, upon seeing her
> over the difficulty, she had done something for herself. Sir
> \Villoughby was proud of her, and therefore anxious to
> soltlo her business while he was in the humour to lose her.
> He hoped to finish it by shooting a word or two at Vernon
> before dinner. Clara’s petition to be set free, released from
> him, had vaguely frightened even more than it offended hia
> nrido.

> NEITHER Clara nor Vernon appeared at the mid-day table.
> Dr. Middleton talked with Miss Bale on classical matters,
> like a good-natured giant giving a child the jump from
> stone to stone across a brawling mountain ford, so that an
> unedified audience might really suppose, upon seeing her
> over the difficulty, she had done something for herself. Sir
> "VVilloughby was proud of her, and therefore anxious to
> settle her business while he was in the humour to lose her.
> He hoped to finish it by shooting a word or two at Vernon
> before dinner. Clara’s petition to be set free, released from
> him, had vaguely frightened even more than it offended his
> pride.

gocr 0.3.6:

__,,,____,_ Cl,_I._c nor Vernon a__e_Ped _t tl_le _id_da_ tab1e_
_, _ii_(__etoiI f,,_lk(;cl with _MiSs _ale _U_1d_ abS8iG_l I_i_t_t_l.__
i,_i,;,_ .,, _(_u_-i,L_t_ii.e(l 6iiLIblt 6'7_V. ill_ _ C 'll . tf e__Ul__b rU_l
gt(),ii_, tu _fj(),I(, ,_uruSS.,__ T__ llll_ g UlOUUt_lU o_ _ 8O .t _' t_ail
u,,_,_ifj(;il ;,_i((ic,IGG l_i_' lt re_ y 8UE)_OB_'_ U_Oll 8eelll6 lttr
_,__i. t_ic (li__icu1ty, SIIe t1_d iluI_e 8ol_eth_ng_ fo_ be_.Self. _i__
_ji___()_i___lll)y w,,s prui_il of heT_ and k__eTefope an_iouS to
_(_(.__u l___i. i)i__, ii,ess wIlile he Wa8 in the hU_ouT to luse Iier_
j__ l_()_)(_(l t() tiilish it b_ ShOOtiltg a WOTd o__ t_O &t Verno_
_o__(),__ (li,_iIci._ Cl__T_'S _eti_tio_ tO be Set fTee_.Te1ea8ecl fro_
)ii))),, lIL_Ll v_b__uely f_.ighteUe eVen _OTe kba_ lt OfEe_ded hi_
pi_i..(l_u- . _ , , --.___ _ _,- - -__-

_____ Cl__i.a nop Vernon appeared &t t'h_e _id_day t__le_
D_. _id(lle_oi_ t_lked with Miss _ale ,on _ _Ssi__l __i tt_r_'_
iij_e _ 6ood-n___tLi_.ed 6iai_t 6_i_ing & Ghild the ___np _'_.on_
_tune to _tone aGro_S a braWlin(__ inOU__taiß_foPd_ So t2_at a__
u__p,(_ified ___idiei_Ge _ni62it real y 8uppO.8e_ upon _seeii_6 l_e_
o_____ the difhculty_ she had done _o_neth_n6 fop ber_elf_ _i_
_viljoli____k)y w__s proud of heT, and the_efo_e an_iouS to
___.tle li__i. i)u__inesS Whike he W_S î_ the hum'ou_ to_ lose her_
__e l_op(_d to finish it by 8hooting a wopd o_ tWo ak Verno__ _
eforR _(in_icr_ Clara's petition to _ Set _free, releaSed fro_
)ii__, h_d va6uely frigbte_ed eve_ _ore tban it o_e_ded hiD
pi.icle. -. - - - - - - '

Recognita Standard 3.2.7AK:

~rFr~rrmx Clara nor Vernon apneared at the mid-da~'table.
Dr. bLidrlleton talkc;d wi.th Miss Dale vn elassieal matters,
like a ~n~a-mZtured giant gi.ving a child th' jucnp frvm
stonc to stone across a brawling mounta,in ford, so that au
uiicilificd .ruciicucc mil;•ht really suppasc, upon seeixig hor
•n~er thc ciillicul.ty, she had clouo something for herself. Sir
~Villcm;;lrlry wvs proua of her, and therefors angiaus to
sct.tla lrur tn~sincss while he was in the humoar to lose her.
lle lu,hcot to iinish it by shooting a word ar two at Vernon
bol'ore ~linncr. Clara's petition to bo set froe, released £rom
JGGnt., hvd vagucly frighteued even more than it offended hia
ri~le.
p

NEITfi~R Clara nor Vernon appeareci at the xnid-day table.
Dr. Middleton talked with Miss Dalo on classics,l rnatters',
like a good-natured giant giving a child the jtimp from
stone to stone across a brawling mountain ford, so that an
unedified audience might really suppose, upon ~ seeing her
over the difficulty, she had done something for herself. Sir
yillon ;hby was proud of her, and therefore anxiotis to
scttle luer business while he w~as in the hurxiour to lose her:
He hoped to finish it by shooting a word or two at Vernon

before dinner. Clara's petition to be set free, released from jcLm, had vaguely frighteued even more than it offended his pride.

OmniPage Pro 10:

NF r~rn,Px Clara nor Vernon appeared at the mid-dap table. Dr. Middleton talked with Miss Dale on classical matter, like .t good-natured giant giving a child the jump from stone to stone across a brawling mountain ford, so that an uneVified audience might really suppose, upon seeing her over the difficulty, she had done something for herself. Sir jV;llo,r;;lrl>y was proud of her, and therefore anxious to set.tlo lror Uusiness while he was in the humour to lose her. Ile. lropcol to finish it by shooting a word or two at Vernon bol'ore dinner. Clara's petition to beset free, released from)zinc, had vaguely frightened even more than it offended his pride.

NEITHER Clara nor Vernon appeared at the mid-day table. Dr. Middleton talked with Miss Bale on classical matters', like a good-natured giant giving a child the jump from stone to stone across a brawling mountain ford, so that an unedified audience might really suppose, upon ~ seeing her over the difficulty, she had done something for herself. Sir yillou ;hby was proud of her, and therefore anxious to settle her business while he was in the humour to lose her. He hoped to finish it by shooting a word or two at Vernon before dinner. Clam's petition to be set free, released from him, had vaguely frightened even more than it offended his pride.

OmniPage Pro 11:

NF f,rnMR Clara nor Vernon appeared at the mid-day table. Dr. Middleton talked with Miss Dale on classical matters, like .t good-natared giant giving a child the jump from stone to stone across a brawling mountain ford, so that an une(lifie(l audience might really suppose, upon seeing her over the difficulty, she had done something for herself. Sir jVillon;hl)y was proud of her, and therefore anxious to setale leer business while he was in the humour to lose her. lle hoped to finish it by shooting a word or two at Vernon bofore dinner. Clara's petition to beset free, released from)lint, had vaguely frightened even more than it offended his pride.
-.2 ..1_ - ____

NEITHER Clara nor Vernon appeared at the mid-day table. Dr. Middleton talked with Miss Dale on classical matters', like a good-natured giant giving a child the jump from

stone to stone across a brawling mountain ford, so that an unedified audience might really suppose, upon,seeing her over the difficulty, she had done something for herself. Sir Willoughby was proud of her, and therefore anxious to settle her business while he was in the huniour to lose her. Il"e hoped to finish it by shooting a word or two at Vernon before dinner. Clara’s petition to be set free, released from hint, had vaguely frightened even more than it offended his pride. - -

TextBridge Millennium Pro:

NErr’!’~~ Clara nor Vernon appeared at the mid.day table. pr. ~1id(IIeto11 talked with Miss Dale on classical matters, like a good-natured giant giving a child the jump from stone to stone across a brawling mountain ford, so that au ~1edifi~ tLU(llCIICC might really suppose, upon seeing h er over the (hjiheulty, she had done something for herself. Sir wiflouighby was proud of her, and therefore anxious to settle her business while he was in the humour to lose her. lie ho1)ed to finish it by shooting a word or two at Vernon before dinner. Clara’s petition to be set free, released from him, had vaguely frightened even more than it offended his prœ~t~.

NEITHER Clara nor Vernon appeared at the mid-day table. Pr. Middleton talked with Miss Dale on classical matters, like a good-natured giant giving a child the jump from stone to stone across a brawling mountain ford, so that an une(lified audience might really suppose, upon - seeing her over the difficulty, she had done something for herself. Sir Willoughby was proud of her, and therefore anxious to settle hier l)uSifleSS while he was in the humour to lose her. lie hoped to finish it by shooting a word or two at Vernon before dinner. Clara’s petition to be set free, released from hirn~, had vaguely frightened even more than it offended his pri(le.

* * * * *

Scan 4--A Really Bad Case!

Scan4 is a paragraph from Pope’s translation of Homer’s "Odyssey". This is a very, very tough one. It was obviously a cheap printing to begin with, using thin, poor-quality paper in a page size of 6" by 4.5", with capital letters about 1.5 mm high, a little bigger than Times New Roman size 8. Text this small really needs a higher-resolution scan. The book was falling apart when I got it, the ink was fading and flaking, and there was no point in even thinking about trying to scan it flat, so I cut the pages. To add an extra challenge, I scanned the sample with the cover open in a medium-lit

room for the 300 and 400dpi scans, but closed the cover for the 600dpi to show the best quality I could possibly get. (I was pleased to note that Abbyy, while recognizing the page in the 300dpi and 400dpi images, flashed up a suggestion that I should lower the brightness of the scan.)

This particular book was one I sporadically tried to produce, without success, on an older scanner and a bundled OCR program over a period of two years, back in 98/99. Eventually, in 2000, it was the first book processed through Charles Franks' Distributed Proofreaders site. The initial text produced by the OCR was very poor, but the human volunteers made up for it! Thanks, guys! Today, just two years later, with a better scanner and better OCR, I could have done it myself, as you will see from the best of the results of the 600dpi scans. That's how much things have improved recently.

A separate point to note here is that you can see the "three-quarter space" effect before the exclamation mark and semi-colon that was discussed in [V.104].

The results of the OCR are:

Abbyy FineReader 6:

" Ah me ! on what inhospitable coast,
On Tvh.it new region is Ulysses toss'd ;
Possess'd by wild barbarians fierce in arms ;
Or men. whose bosom tender pity warms ?
What sounds are these that gather from the shores ?
The voice of nymphs that haunt the sylvan bowers,
The fair-hair'd Pryads of the shady wood ;
Or azure daughters of the silver flood ;
Or human voir-e? but issuing1 from the shades,
AVhv cease I straight to learn what sound invades?"

" Ah me ! on what inhospitable coast,
On what new region is Ulysses toss'd ;
Possess'd by wild barbarians fierce in arms ;
Or men, whose bosom tender pity warms '?
"What sounds are these that gather from the shores ?
The voice of nymphs that haunt the sylvan bowers,
The fair-hair'd Dryads of the shady wood ;
Or azure daughters of the silver flood ;
Or human voice? but issuing from the shades,
Why cease I straight to learn what sound invades?"

" Ah me ! on what inhospitable coast,
On what new region is Ulysses toss'd ;
Possess'd by wild barbarians fierce in arms ;
Or men, whose bosom tender pity warms ?
"What sounds are these that gather from the shores ?
The voice of nymphs that haunt the sylvan bowers,
The fair-hair'd*Dryads of the slrady wood ;

Or azure daughters of the silver flood ;
Or human voice? but issuing from the shades,
Why cease I straight to learn what sound invades?"

gocr 0.3.6:

[The 300 and 400 dpi scans produced nothing recognizable.
The result of the 600 dpi scan is below.]

'' _hh i_3e ! o_1 ___l_at_ i__l__sl__ it_nble CoaSt_
On ___l_,__ _)e_v i_e_io__ i__ ___ _._____ses toss'd ;
_(3s3gs3_d l3.__ ___iiíi l3_3__b___i_c_i3_ fie_Ce in il__S- _
Or i11pn, __-i)c3se l_osonl te_1de_ _it____ __ai_n3__ ?
___l_at __o__i1ds Qre tlipse tliat g__tl_p_r fE_oi33 the shoTes ?
'_ilie __oi__e of i)____ E1)l3l3s tl3nT 1i_n__nt the s__l__inn bo_Ye_5_
3'l_e fni___i____ir'd _____-ads of' il_e sli__d__ i___oOd _
Op az(_pe da_____litc__s of _tlie sil __?r t1ood ;
Or l___i31_nn ___)i___? l3__t i3____ii_6 fi_oi11 tlie __hiade__ _
__'l3.__ _ea___e _ s_rai__li.t to l_ar_i1- i_--li__t so_nd- in__ad_S___''

Recognita Standard 3.2.7AK:

.: lh nt"'. on w-hat inlu,;y:t, I,:e co;;~t,
On ~cli^t ne~- re~ion i.. 1= 1-.-:.:e~ tm:'d ;
Possea'd 1n- wil~l L;,rba~:c, .~ fierce in arm~ ;
Or u.~u. w-Ln.e bossum tender pit~- warna'?
~l-u:lt .<,:~;;:::3s are tll~ce that ~atl:er from the shnre~ ?
'l'l.e -;;o'.re :,; nwtthil: tW ,t l:aa;nt the s~-l:c 1llJOR'er5,
'lhe :a,:~-h ~;r'd~It.wa~i~ ot' tl:e ~Il;;dv vood;
Or az.lre dau~~l.ts~: oY tl:c •:iv-~~r floo;:3 ;
C?r humnn ~-<:i: e'? l,~:tt i~~; from tl:c• ~had~~,
11-lts- cea~e I ctrai rlit to learn ~s-l:, t socud incades %"

" ~h me ! ou "-Mat iuMospita~le coast,
On ~i-lmt ne~c reyion is L 1~-~ses to~s'd ;
Pos:e;s'd 1"~ w-iMl lrvrbaria:ns fiet~ce in arms ;
Or m~ n, "-hose hosom tender pit~- warm5 ?
~~~hat ~ounds are tlmse tMat ~;atMer from t:he shores ?
~t'l~e ~-oi~~e of n~-Inhhs t.hat liaunt the s~-l~~a n howers
.
Tlie fair-hnir'd D~ vads ot tl:e shad~- "-ood ;
Or aznre dau~liters of tMe sil~-~r fiood ;
Or lmman ~-oi:~e'? but iauin~ frotn the shades, a
lVly cea.~e I straibht to learn "-Mat souud in~ad°s?"

" Ah me ! on what inhospitable coast
On ~~-hat new r e~ion is L;1 ~-sses toss'd ~
~~~

,
Possess’d 1J~- "-ilil I:OII’uai’la ils fierce in arms_ •
Or men, whose hosom tender pit~l ~varn~s ?
~’G’l~at somnds are these tliat ~atl~er from the shores ?
~I’lie v oice of n~-mpl~S that ~munt the sy Ivan bowers,
Tlie fair -hair’d D~~~-ads of tl~e slmdy wood ;
Or azure daylltcrs of tlle silver flood ;
Or lm:nan voice? uut issL~ing from the shades,
~~’lm cea~e I strai~ht to Iearn ~~-lmt so~nd inv ades ?"

OmniPage Pro 10:

,. _lh in- ’ on "-hat inh-slit al.:e coast,
On "M.^t new reion is 1=1;-a:e~ to-s’d ;
P"::e:~’d hw "ild Larba.:an~ fierce in arms ;
Or inn. "-hnse bo.,om tender pity warms
What <m-,n ds are thFSe that gather from the shores?
’1-l.e vo_,e o2 u~vnhit: thm hn,,-,nt The sylvan bowers,
The is ;r-ha;r’d h.-;-ads of the liz-Ay iNood
Or azure dau_ht;- of tl:c o=1 cr flooj ;
Or hnnmn wire? l,11t i--rii:g from the shadP3,
Al-ly cease I straiAlit to learn what sound invades?"

’Wh me ! on what inhospitable coast,
On what new region is L fusses toss’d ;
Possess’d br wild barbaric ns fierce in arms ;
Or men, whose bosom tender pith- warms
AN-hat sounds are these that gather from the shores ?
The voice of nymphs that Haunt the sylvan bowers,
The fair-hair’d IWvads of the shady -wood ;
Or azure daughters of the silver flood ;
Or human voice? bat iauina from the shades,
Why cease I straight to learn what sound invades?"

" Ah me! on what inhospitable coast,
On what new region is Ll ysses toss’d ;
Possess’d bv -wild barbarians fierce in arms ;
Or men, whose bosom tender pity warnis ?
AVlia– sounds are these that gatller from the shores
The voice of nYI11pliS that haunt the -sylvan bowers,
The fair -hair’d D.-yads of the shady wood ;
Or azure daughters of the silver flood ;
Or human voice? lout issuing from the shades,
Why cease I straight to learn what sound invades?"

OmniPage Pro 11:

.‘ lh in-’ on what inhospital,le co-st,
On xclznt near region is t 1:-sse~ toss’(: ;
Possess’d bY Mild barbarians fierce in aims ;
Or inn. whose boson tender pity warms

What <m-,n ds are tlipse that gather from the shores ?
’_I-I.e 1-o=,- of nv:npii? that haunt the sylvan bowers,
She ra;r-ha;r’d 1):, ads of the shad- wood ;
Or az.ire dau_lit~- of tl:e silo-:-r flood ;
Or human voice? l,,tt i?snina from the shadpq,
Al-lry cease I straiAit to learn shat sound invades?"

’”’ :Ah me ! on what inhospitable coast,
On iyhat new region is Ulysses toss’d ;
Possess’d br wild barbarimis fierce in arms ;
Or men, whose bosom tender pity warms
AN-hat sounds are tliese that gather from the shores ?
The voice of nymphs that haunt the sylvan bowers,
The fair-hair’d D~ yads of the shady -wood
;
Or azure dau.L-hters of the silver flood ;
Or human voice? but issuing from the shades,
Why cease I straight to learn what sound invades?"

" Ah me! on what inhospitable coast,
On what new region is Ulysses toss’d ;
Possess’d by -wild barbarians fierce in arms ;
Or n1en, whose bosom tender pity warnis ?
AVliat sounds are these that gather from the shores
The voice of nyniplis that haunt the sylvan bowers,
The fair-hair’d Dryads of the shady Wood ;
Or azure daughters of the silver flood ;
Or human voice? but issuing from the shades,
Why cease I straight to learn what sound invades?"

TextBridge Millennium Pro:

no on what inhe~ptaEie coast,
On what new realun is hivs,e’ to5sd
,s~s ˜-~d liv wild lie il)~m.ihI fir see in al-rn~
Or u~,-n. w’linse bo,uuiu tender pity warnls
Wl at ~ are t1ie~e that ~atler from the shores ?
’n.e a oro of imvntpirs tint he~nt the sad van bowers,
’flie tah’-ha~r’d D~vahs ct the shady wood
1)1’ az Ire dauul~t ~ of tl,e shvr flood
Or liunian vi i ’I ? h’tt is- eng from the shades,
\VIiv cea-~e I straight to learn w hat sound invades 1"

Ah me on what inhospitable coast,
On what new region is U vases toss’d
Possess’d by wild barbarians fierce in arms
Or men, whose bosom tender pity warms ~
What sounds are these that gather from the shores?
The voi’e of nymphs that haunt the sylvan bowers,

The fair-baird Prvads of tl~e shady wood
Or azure daughters of the silver flood
Or human vuiae? but issuing fi'om the shades,
Why cease I straigl~t to learn what sound invades?"

Ah me on what inhospitable coast,
On what new region is Ulysses toss'd
Possess'd by wild barbarians fierce in arms
Or men, whose bosom tender pity warms?
What sounds are these that gather from the shores?
rfhe voice of nymphs that haunt the sylvan bowers,
The fair-hair'd Dtyads of the shady wood;
Or azure daughters of 'the silver flood
Or human voice? but issuing from the shades,
Why cease I straigl~t to learn what sOund invades?"

What can we conclude from this?

Small mistakes in scanning, like letting too much light in, getting your scanner settings wrong for the page, or not pressing the paper flat enough, can make a major difference to the final quality of the text that you will have to correct.

Sometimes, no matter what you do with your scanner, problems with the paper or the print will make it difficult for your OCR package to give good output.

Generally, bigger is better within the range 300dpi-600dpi, but you only need higher resolution with more difficult material.

Different OCR packages will produce widely differing texts from the same images. Given a really good image, most OCR software will work acceptably, but when you have lower quality material to work with, the gap between OCR packages shows clearly.

S.18. I got an OCR package bundled with my scanner. Is it good enough to use?

That depends on how well your package performs on the actual scans that you do, and how much you value your time vs. money. Most scanners are bundled with OCR software, but these OCR packages are often older or "brain-damaged" versions, with their functionality deliberately lowered. It's unlikely that you'll get a current-version, top-of-the-line OCR package thrown in for free.

You may have to pay extra for better OCR, but it means that you spend less time making corrections. The question is how much better you want your OCR to be.

Save the images from the FAQ "Why am I getting a lot of mistakes in my OCRed text?" [S.17] and try processing them with the OCR you have. Compare the quality of the text produced with the quality of the samples. This should give you some idea of how your OCR compares to others.

Try a few pages from your book with your OCR. How many mistakes do you see on each page? Do you find that acceptable?

S.19. I want to include some images with a HTML version. How should I scan them?

We don't often see color prints in our books, but if you do have one, then scan it in color. Otherwise, try both greyscale and B&W, and see which gives you the best image.

It's usually better to scan images in a higher resolution than you're going to use, and then use an image manipulation package to reduce them [H.10] to a size appropriate for your HTML file. An initial scan at 600dpi is often good. Image manipulation programs will also allow you to "clean up" the pictures, by increasing contrast, despeckling, or other filtering.

S.20. I want to include some images with a HTML version. What type of image should I use?

GIF, JPEG and PNG images are supported by current browsers, and you should stick with those unless you have a specific reason not to.

GIF and PNG tend to be more efficient--provide better quality at a given file size--for simple line-drawings; JPEG is usually better for photographic images.

S.21. Will PG store scanned page images of my book?

No. Or, at least, not yet.

The idea has been kicked around a bit. There's no question of replacing etexts with page images, but many volunteers who have already scanned the book anyway like the idea of saving page images as well--for general information, and as a means of checking future correction suggestions against the original. Some volunteers already keep their page images, stored for possible future use.

Working some back-of-the-napkin figures: a page of text might take up 1KB of space on a computer as plain text or HTML or XML. The same page

might take 70KB if stored as a black-and-white image, of just enough quality to serve as a reliable guide to making corrections. Pages with pictures, or stored with enough resolution to allow some future researcher to write a paper on the changing shape of serifs in the 18th and 19th centuries, would start at around 350KB per page, and go up from there.

A 300 page book thus becomes

about 300KB as plain text (and around 150K zipped)
about 20,000KB as minimal-quality images
about 100,000KB as high-quality images

and with the images, we won't save much space on the zipping, because they're already compressed.

On a normal "56K" modem, getting about 4KB / second, it would take:

75 seconds to download the text file (40 for the Zip)
80 minutes to download the minimal images
over 5 hours to download the high-res images.

Someday, the disk and bandwidth capacities that we will take for granted will be such that uploading images, when we have them, will be quite natural, just for the few people who will want them. But we're not quite there yet.

Late flash! As of late 2002, the Internet Archive is providing space to volunteers for storing page images. To see the images, and find out more, go to <http://texts01.archive.org/gutenberg-images/>

HTML FAQ

H.1. Can I submit a HTML version of my text?

Yes.

H.2. Why should I make a HTML version?

Well, you can make one just because you want to, but on some texts there is special reason to.

If you want to preserve the pictures that accompany the text, making a HTML version means that you can specify where and how those images appear.

If there is particular meaningful information in the layout of the

text that can't be expressed in ASCII, like special characters or complex tables or fonts, HTML may offer an open format alternative.

H.3. Can I submit a HTML version without a plain ASCII version?

You can submit it, but the Posting Team will then consider whether we should also make an ASCII, or perhaps ISO-8859 or Unicode version of it. We really do want our texts to be viewable by everybody, under every circumstances, and we do not want to start posting texts that are in any way inaccessible to anyone.

See also the FAQ [G.17] "Why is PG so set on using Plain Vanilla ASCII?"

H.4. What are the PG rules for HTML texts?

1. The only absolute rule is that the HTML should be valid according to one of the W3C HTML standards.

You can verify that your HTML is valid at the W3C's HTML Validator at <http://validator.w3.org/>

For a more convenient and friendly, though less official, check of the correctness of your HTML, you should use Dave Raggett's Tidy program at <http://tidy.sourceforge.net>, which not only points out any messiness in your HTML code, but also has some neat modes to clean it up and standardize the formatting.

After that, we have some requirements and recommendations. Compliance with the requirements might be waived if there is a really good reason to make an exception in this case.

2. Requirement: File names and extensions

If you want your text to work within 8.3 filename conventions, you may use .htm as the extension for your HTML files; otherwise, use .html as the extension. If you are working to 8.3 conventions, all of your images as well as your HTML files should have 8.3-compliant filenames.

All file names and extensions should be in lower-case throughout. Yes, we know this is not strictly necessary, but we don't want to have to correct every file that comes with "image.gif" referenced in the HTML accompanied by a file IMAGE.GIF.

3. Requirement: HTML and plain-text

Project Gutenberg does publish well-formatted, standards compliant

HTML. However, we insist that a plain text version be available for all HTML documents we publish (even if images or formatting are absent), except when ASCII can't reasonably be used at all, for example with Arabic, or mathematical texts.

4. Requirement: Archive format for posting

If the HTML book contains more than one file (including images), create a ZIP (preferable) or TAR archive containing all of the files in the book. The ZIP file may, if you wish, unzip to a subdirectory named for the book. For example, a book called 'The Humour of Mark Twain' might unzip in a directory called 'mthumor'. Make sure directory names contain only alphabetic and numeric characters, no spaces, and are 8 characters or less, even if you're not sticking to 8.3 conventions for filenames.

5. Recommendation: Simplicity

Make your HTML as simple as possible. HTML is an evolving standard, and one that may be completely obsolete in the long term. Use of advanced features may just mean that your version will be obsolete or unreadable that much faster.

6. Recommendation: Images

Images included with your HTML should be in a format that Web browsers can read: GIF, JPEG or PNG. Images should be edited for high quality in a reasonably small file size. Make the best decision you can concerning the image size and placement in the text. Every image included must be linked into (referenced by) the HTML.

7. Recommendation: Line lengths

If it is reasonable to do so, try to wrap paragraphs of text at around the normal PG margin of 70 characters. Ideally, your HTML should be as near as possible identical to your text version except for the HTML tags and entities. People who open your HTML won't all be using browsers, people will need to make corrections, not all editors can handle very long lines, and even with editors that can handle long lines, it's easier to work with short lines.

Apart from these rules and recommendations, we also have a rule about the PG header, but that will normally be handled by the Posting Team. Where your HTML is all in one file, the header text will be inserted within PRE tags in that file. Where the HTML is split into multiple pages, the header will be put into a separate file named index.htm or index.html, and will link to the first page of your HTML.

H.5. Can I use Javascript or other scripting languages in my HTML?

No.

We don't want our readers to have to worry about any potential for malicious or just plain buggy code.

H.6. Should I make my HTML edition all on one page, or split it into multiple linked pages?

For a typical novel, one page or HTML file is appropriate, but when that single HTML file gets up around 2 megabytes in size, it may be worth considering a split because of the difficulty of loading it in some browsers.

In some other cases, where the content requires different styles on different pages, or different pages need different character sets, or the page, with images, just gets too heavy, you may need to split the HTML even if the HTML itself isn't technically too big.

When we post a HTML eBook containing multiple files, whether they contain text or images, we post them only in zipped format, so if you don't have images, and want your text to be directly accessible, you should stick to one file where possible.

H.7. How can I check that I haven't made mistakes in coding my HTML?

There are two kinds of mistakes you can make in coding HTML: you can produce invalid HTML, or you can produce HTML that doesn't do what you want.

Checking for invalid HTML is straightforward. The W3C site <http://validator.w3.org> will formally validate your file and point out any mistakes, and this is the official standard. However, it is not always convenient to use, especially when you're in a cycle of fix-and-retest. For this, you should try the program Tidy <http://tidy.sourceforge.net>, which runs on your computer, tells you about errors, and has other useful functions as well. Tidy is available for just about every operating system, and there are several Windows utilities that include Tidy. The links on the main Tidy page will lead you to the right version for you. Tidy is fast and friendly, compared to validation over the web, but it is not the last word. The W3C Validator may find formal errors, such as DOCTYPE mismatches with HTML tags or entitles, that Tidy may not. The best solution is to complete your HTML tests using Tidy, and then, when Tidy finds nothing further to

gripe about, submit it to <http://validator.w3.org> for the official seal of approval. Please run these checks before submitting your HTML; we can generally fix it for you, but it may take us a lot of work.

Producing HTML that actually does what you want is equally important. If you've converted the eBook from text, you may have created inconsistencies, or closed an italics tag in the wrong place, or used the wrong tag at some points. The only way to check this is by reading through the HTML in a browser.

H.8. Can I submit a HTML or other format of somebody else's text?

Maybe.

This question has several complications. First, you must understand that it is quite possible, even likely, that your HTML file will eventually be overwritten by better information.

The value of a HTML file, as opposed to a plain text file, lies in its ability to capture elements of the original that have been lost in the plain text. A plain text file, using extended character sets like ISO-8859 [V.76] or Unicode [V.77] and _underscores_ for italics, can capture all of the author's intent in almost all cases. Sometimes, images and other important features of the original cannot be captured in plain text alone, but can be captured in HTML, or other markup.

When Michael Hart stopped posting books, in September 2001, we had HTML formats of about 1.6% of all our eBooks. At the end of 2002, that has risen to nearly 11% of all our eBooks. If you have a clearable copy of an existing posted book, with extra features not included in the original plain text, we would encourage you to make a new edition, or version, or format, correcting any errors in the original, and adding any new information not included there.

If, on the other hand, you just want to make a "blind format change"--making your best guess at what the HTML, or other format, layout should be for a book you've never seen, based on the original producer's work--your best bet is to get in touch with the original producer, and ask whether they can supply more material for you to work with. Otherwise, you are at best just rearranging information rather than contributing something new.

A blind format conversion can be done in anything from 2 minutes [R.33] to an hour. It just doesn't make sense for us to keep posting these files when they contain nothing new, and especially when two people may want to convert the same text. It is likely that, at some time in the next couple of years, we will start on a large-scale conversion project, to add some form of markup to all of the existing text files for ease of serving, and having a mish-mash of existing

markup styles to deal with at that point won't help either.

H.9. How big can the images be in a HTML file?

The images should be as big as necessary, and no bigger.

Sorry, but there is no clear number to give here. Web page designers sweat blood to save an extra 20K on a page; so should you. If you're an experienced HTML maker, you know this stuff; if you're not, take it as a guideline that you should generally aim to keep your images in the 30K to 50K size range, with occasional forays into 70-80K territory. That's generally big enough for a clear picture, unless you're reproducing fine artwork.

H.10. The images I've scanned are too big for inclusion in HTML. What can I do about it?

This is a common problem, where images from the book occupy a full or half page. Your images should be of an appropriate size for downloading, and 2 megabytes of high-quality scan per image is not really an appropriate size for most PG texts!

You should reduce the size, and maybe the quality, of the original scan for simple viewing purposes. There is lots of image-manipulation software to do this. For Windows, you might look at the freeware Irfanview, and for both *nix and Windows there is ImageMagick [P.1]. Look for the words "resize" and "resample" in the Help.

Apart from simple converters, which do enough for this purpose, you can also manipulate the images in full imaging creation and editing packages like Paint Shop Pro, Adobe Photoshop and The Gimp [P.1].

Different image encoding methods can make a huge difference to the filesize. Any of the packages mentioned above can encode images as GIF, JPEG or PNG, and, particularly for black and white line drawings, these can encode to very different sizes. So, for example, a 60K JPEG may save as a 30K GIF, because the GIF encoding works better for that particular image. Try your images out, and see what works.

When manipulating images, always work from your original. Don't convert your original to a JPEG, and then shrink that and convert it to a GIF. Depending on the format, images may lose definition as they are converted (search for "lossy compression" in your favorite search engine to find out more about this), and they certainly lose definition as they are resized, and you end up with the "imperfect copy of an imperfect copy of an . . ." effect. When you're experimenting, take your original, resize and Save As GIF, then go back to your original, resize and Save As JPG, and so on.

You can also use an image optimizer. These are specialist software programs that try to make image files smaller without sacrificing resolution or detail.

H.11. Can I include decorative images I've made or found?

No.

Please include only the images you got from the book. If you want to make an edition of the book for your own web site, you can of course use whatever you like there, but for PG purposes, we want the book, the whole book, and nothing but the book.

H.12. How can I make a plain text version from a HTML file?

You can edit out the HTML by hand, of course, but there are several easier ways to convert.

You can view the HTML in a browser, Select All text, and just Copy and Paste into your editor. This is easiest, but doesn't handle formatting like tables very well.

You can use the Lynx [P.1] browser to convert your text with the command

```
lynx -dump myfile.html > myfile.txt
```

Bruce Guthrie's HTMSTRIP for MS-DOS [P.1] is very configurable.

<http://www.w3.org/Tools/html2things.html> has a list of other HTML to plain text converters.

H.13. How can I make a HTML version from my plain text file?

This is not a course in HTML, but, for most books, you don't really need a course in HTML. Making a HTML format of most books is very easy, and doesn't take long, once you have mastered basic HTML. Let's assume you have your completed PG plain text file ready, and walk through the steps commonly needed to make a HTML version. We'll do this by successive approximation, doing the major things first, and then dealing more and more with the detail.

There are lots of specialized HTML editors out there, but you don't actually need any of them. The same editor that you used to create your text will also create your HTML. HTML is just text, with two types of special instructions added: tags and entities.

A _tag_ is an instruction to the browser, usually to display something

with specific rules. Tags are shown within angled brackets: for example, <p> is the instruction to start a new paragraph.

An _entity_ is a named special character that might not be available in your character set. Entities are shown starting with an ampersand "&" and ending with a semi-colon ";" : for example, — is the representation of an em-dash.

I'm marking up a made-up short text as I write these steps, loosely based on the sample page from question [V.121]. You can see the changes made at each stage by looking at the files

htmstep0.txt (text before starting)
htmstep1.htm (after adding the HTML header and footer)
htmstep2.htm (after adding paragraph marks)
htmstep3.htm (after marking main headings)
htmstep4.htm (after adding special line breaks and indents)
htmstep5.htm (after adding italics and bold)
htmstep6.htm (after adding accents and non-ASCII characters)
htmstep7.htm (after adding an image)
htmstep8.htm (showing some extra techniques)

Before you start, make sure that you can see these files both in your browser and in your editor. In your editor, you should see the HTML codes; in your browser, you should see the text as it is intended to be viewed.

Note for people who already know HTML: yes, this example omits lots of possible ways to do things, and lots of refinements. You already know how to do what you want to do--skip onwards, and give the beginners room to learn in peace! :-)

Step 1. Add the HTML header and footer information

Add the following lines at the top of your text file:

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=ISO-8859-1">
<title>The Project Gutenberg eBook of My Book, by A. N. Author</title>
</head>
<body>
```

Let's explain these one by one:

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">

says that your file is HTML 4.01 Transitional, which is the latest version, allowing the widest range of tags and entities.

<html>

denotes the start of the HTML

<head>

denotes the start of the HTML header information.

<meta http-equiv="Content-Type" content="text/html; charset=ISO-8859-1">

says that the characters are text, using ISO-8859-1 encoding.
If you need to use a different character set, you should change
ISO-8859-1 to whatever you intend to use. ISO-8859-1 is good for
lots of PG books in English that use French or German words.

<title>The Project Gutenberg eBook of My Book, by A. N. Author</title>

You should obviously change this to the actual title and author
you're producing. The

</head>

denotes the end of the HTML header information and

<body>

denotes the start of the actual text itself - the body of the book.

At the very end of the file, you should append these two lines

</body>
</html>

these denote the end of the body of the book,
and the end of the HTML.

At this point, you actually have a valid HTML file! OK, if you view it
with a browser, it doesn't look anything like the way it's supposed to,
but it _is_ HTML. Save it with a name like MYFILE1.HTM or STEP1.HTM and
get a copy of Tidy for your DOS, Unix, Mac or Windows system from
<http://tidy.sourceforge.net>. Run Tidy on your file, telling it just
to look for errors (tidy -e if running from a command-line; if you're
using a GUI version, there should me a menu option or tickbox for
showing errors only). Tidy should tell you that there are no errors.
Yay!

If it does say that there are errors, deal with them now, before you continue. Make sure, at each step, that you have cleaned up any errors; it's a lot easier now than later. Also, when you've finished each step, save your file with a number in its name, so that if you run into problems later and get confused, you can, at worst, drop back to the correct version at the end of the previous step.

The most likely error you might have at this point relates to the characters "<", ">", or "&". These are the characters used by HTML to indicate tags and entities. If these characters are used in the text of your file, (and ampersand is likely to be), you should replace them with entities, so that HTML will know that they are to be displayed as characters, not interpreted as commands.

Replace & with &
< with <
> with >

There is an example of this in the file htmstep1.htm

Step 2. Add paragraph marks.

For novels and general prose, paragraphs are the main logical and display unit. Paragraphs are marked in HTML with the sign at the start, and at the end. You don't actually need the at the end, but adding these is a good habit to get into. You do, very much, need the at the start.

The line-lengths within a   pair are irrelevant; the browser in which the text is viewed will ignore extra spaces and line-ends, and will wrap text to fit the screen. This is bad for poetry and tables, but we will discuss those later. For this step, all you need to know is that you can leave your text exactly as it is, and just add the paragraph marks.

Put a at the start of the line before the first letter of every paragraph, and a just after the last letter or punctuation of every paragraph. If you can do macros in your editor, this will just take a minute; otherwise, it may be rather boring, but at least it is simple. For this step, put the paragraph marks around _everything_ that has a blank line after it, even poetry or chapter titles. We'll come back and change that later.

Now save your text as something like MYFILE2.HTM or STEP2.HTM. Again, run Tidy to check for errors, and fix them before continuing.

If you now look at the file htmstep2.htm in your browser, you will see that it is starting to take shape. Look at it in your editor, and you will see the paragraph marks.

Step 3. Add marks for headings.

We want to indicate to the reader that certain lines are for chapter or other headings. HTML provides the tags <h1>, <h2>, and so on for this. <h1> is for the biggest heading, and usually, you will reserve this for the title, and use <h2> for chapter headings. If you find these too big, you could choose <h2> for main headings, and <h3> for chapters. Whenever you use one of these header tags, you must close it with its equivalent end tag. So a chapter heading might look like:

<h2>Chapter XI</h2>

Since there won't be many headers, and most headers are only on one line, this is usually not hard. Look at the file htmstep3.htm to see how our sample is improving, and if you're working along with me, don't forget to save your file under a new name and check it.

In our example, we have marked some lines with paragraph marks where we now want to put headings, so we will change those <p>s into <h2>s, since we don't need or want to mark a line as both.

Step 4. Line up verse, tables of contents, and other lists.

The HTML tag
 tells the browser to force a line break without starting a new paragraph. We use this when we don't want text all wrapped together, but not separated with blank lines either, for example in verse and tables of contents.

In our sample, we add the
 tag to the end of each line in the table of contents and the end of each line of the verse. If we were working on a whole book of poetry, the same principle would apply, but we'd be using the
 tag a lot more.

Where we want to indent a line of poetry, we can use " " at the start of the line. Normally, however many spaces you leave between words, HTML condenses them to one space, so normal indentation doesn't work. But the "non-breaking space" entity will cause the browser to show one space for each character, so that you can indent as much as you need.

The file htmstep4.htm shows the effect: this is now an entirely readable HTML text!

Step 5. Add back in italics and bold.

The HTML tag <i> tells the browser to start displaying italics,

and the </i> tells it to stop. Similarly, the tag tells it
to display bold, and marks the end of the bold text. See
htmstep5.htm for the changes.

Step 6. Restore accents and special characters.

Since we declared our HTML file to use ISO-8859-1 back at the start,
we can use any of the common accented characters for Western European
languages, but we may also use HTML entities. For example, for the
"a circumflex" in "flaneur", we can use either the ISO-8859 character
directly, or the HTML entity name "â" or number "â".

There is a trade-off between characters and entities: entities do not
limit you to any particular character set, but characters are directly
readable when looking at the HTML source.

Within entitles, there is also a trade-off between entity names and
numbers: older browsers may not recognize some of the entity names, but
the entities do make the text work in multiple character sets. Which you
choose is entirely up to you, but it's best to be consistent; if you
like entities, use them everywhere. Entities can be represented by their
names--for example, —--or by their number, derived from their
ISO-10646 (see Unicode) number--for example, —.

There are other special character entities you may choose, to replace
the ASCII equivalents in the main text. Here are some of the common
ones:

We've already seen

& & ampersand replaces "&"
< < less than replaces "<"
> > greater than replaces ">"
      space replaces a space when you want to indent

and these are also very useful for many PG texts:

— — em-dash replaces "--"
° ° degree replaces "deg." or "degrees"
£ £ British pound replaces "L" or "l" or "pounds"

There are many others. <http://www.w3.org/TR/html4/sgml/entities.html>
has a fuller list. Please note that you don't _have_ to use these
entities in your HTML; if you're happy with the text reading
"500 pounds", there is no need to make that "£500".

I've made a couple of entity changes in htmstep6.htm.

Step 7. Link Images into the text.

First, you need to have your image ready. You should already have resized your image to the size you want it to be viewed at. You should also have saved it as a GIF, JPG, or PNG image, since those are the formats most supported by current browsers.

If your image is named front.gif, and it is a picture of the frontispiece of the book, you should add the line

<img src="front.gif" alt="Frontispiece">

to your HTML at the place where you want it displayed.

The "alt" text gives a label to the image, and is displayed if the image can't be shown, or in the case of a browser for visually impaired people.

You don't _have_ to add images with your HTML file, unless you want to. In many older books, there are no images at all to be added.

My final HTML text is now in htmstep7.htm. You need to have the image front.gif in the same directory in order to see it. When your HTML text is posted, the images will be zipped with it, so that future readers can see them.

Step 8. Over to you!

This is enough to make a reasonable HTML format of most PG texts, but it doesn't begin to cover everything that can be done in HTML. If you've gone this far, I recommend the W3C's tutorials:

<http://www.w3.org/MarkUp/Guide/>

and

<http://www.w3.org/MarkUp/Guide/Advanced.html>

which cover the ground we've just crossed, and go a bit further.

Here are a few more things you might want to know, but don't go nuts adding tags just because you can! Use them only when you really need them. The file htmstep8.htm shows some of these techniques. Personally, I think that this is a bit overdone, and I prefer the effect of htmstep7, with left-aligned chapter headings, but that's a matter of taste.

Once you're used to the basic HTML needed for most PG eBooks, you'll probably be able to convert one in under an hour.

How do I force more space between specific paragraphs?

Insert a blank paragraph like this: or
use an extra
 tag.

How do I make text, or image, or headings centered?

Put the and tags around what you want centered,
like:

```
<center><h2>Chapter 12</h2></center>
```

How do I make some text bigger or smaller?

Put the <big> and </big>, or <small> and </small> tags around it.

How do I lay out tabular information?

The simplest way to do it is with the <PRE> and </PRE> tags.
These will cause whatever is within them to be displayed as
plain text, just as it was in the original, so that spaces
separate the entries just as they did in the text version.
You can also use this for poetry, though you usually won't
need to. It's not entirely satisfactory, but it will work.

Making a full HTML table requires you to use the <table>,
<tr> (table row), and <td> (table detail) tags, among others,
and a full exposition of tables is beyond the scope of this FAQ.

Briefly, you start a table with the <table> tag.

```
<table>

</table>
```

For each row you want in the table, you open and close a table
row <tr> tag, like:

```
<table>
   <tr>
   </tr>

   <tr>
   </tr>
</table>
```

and then for each cell within a row, you specify a <td> tag and
the contents of that cell:

```
<table>
   <tr>
```

```
        <td>This is the Top Left cell</td>
        <td>This is the Top Right cell</td>
      </tr>
      <tr>
        <td>This is the Bottom Left cell</td>
        <td>This is the Bottom Right cell</td>
      </tr>
    </table>
```

This only scratches the surface of tables. However, there are many guides available on the Web, and they're easy to find, once you know which tags you're looking for. A brief discussion of tables is provided by the W3C as part of the HTML 4.01 spec at <http://www.w3.org/TR/html4/struct/tables.html#h-11.5> and the tutorial at <http://www.w3.org/MarkUp/Guide/Advanced.html> also shows how to make HTML tables.

Step 9. Some common problems

When you're just starting to code HTML, it may seem that errors are coming at you from all sides. Tidy may spew out a stream of complaints that you don't recognize or understand. If it's any consolation, this is normal!

Just take the error list one line at a time, starting at the top. Often, one actual mistake, like not closing a tag, may cause many errors, since an unclosed tag can cause many subsequent tags to be reported as errors.

Common errors include:

1. Simple typos in tags, like <h2Chapter 3</h2> instead of <h2>Chapter 3</h2>
2. Unclosed tags, like forgetting to add the </h2> in the sample above, or forgetting the slash in the closing tag so that you type italics instead of italics.
3. Not nesting tags correctly. Get used to thinking of tags as brackets; the first one opened should be the last one closed. For example, you should type: This is centered. instead of This is centered.

One option for making a HTML version is to use GutenMark <http://www.sandroid.com/GutenMark/> to create the basic HTML straight from your text, and then edit the resulting HTML to add the features you want. If you're having a lot of problems with your main conversion, this is worth a try.

Programs and programmers FAQ

P.1. What useful programs are available for Project Gutenberg work?

These suggestions came largely from a poll of volunteers in June, 2002. The programs listed are a summary of the programs we actually use. There are many other programs out there that can do the same jobs, so don't limit your search just to these.

1. OCR

Abbyy <http://www.abbyy.com>
OmniPage <http://www.omnipage.com>
TextBridge <http://www.textbridge.com>

These are the three main commercial packages that volunteers bought specifically for the purpose. In a few cases, people had got older versions of these bundled with their scanners.

Clara OCR <http://www.claraocr.org/>
Gocr <http://jocr.sourceforge.net>

These are Free Software packages. Some people who responded to the survey had tried them, but nobody had actually used them to produce a text.

DocMorph -- a free, web-based OCR <http://docmorph.nlm.nih.gov/docmorph/>

This one is interesting--you can just submit your image through a web page, and the service will return OCRed text. However, the process of submission, waiting for your text, and then cutting and pasting into your document is slow.

Other volunteers use various OCR software that came bundled with their scanner.

2. Editing

The main answers, given by more than one person, were:

AbiWord <http://www.abiword.org>
emacs
Microsoft Word
vi
Windows WordPad
Word Perfect

Other editors mentioned included:

Crisp for Windows <http://www.crisp.demon.co.uk/>
EditPad <http://www.editpadpro.com>
Editplus for Windows <http://editplus.com/>
Foxpro 2.6 for DOS
Metapad <http://www.liquidninja.com/metapad/>
Windows Notepad

Programs recommended by Apple Macintosh users included:

AppleWorks
BBEdit Lite <http://www.barebones.com/products/bbedit_lite.html>
Microsoft Word
Nisus Writer <http://www.nisus.com/>
Text-Edit Plus <http://hometown.aol.com/tombb>
TextSpresso <http://www.taylor-design.com/textspresso/>
Add/Strip <ftp://mirrors.aol.com/pub/info-mac/_Text_Processing/>

3. Checking and proofing

For spelling, most people just use the spellchecker built into their editor or word-processor. The *nix users running emacs or vi tended to use variants of the standard Unix spell command, such as ispell or aspell. Mac users have the free spelling checker Excalibur, available from <http://www.eg.bucknell.edu/~excalibr/excalibur.html>.

Gutcheck <http://gutcheck.sourceforge.net> was used for format checking, and a few people had written some checking procedures of their own.

4. Working with HTML

In the survey, most volunteers preferred to handcraft their HTML using their normal editor. Those using a word processor edited the HTML as text, rather than composing a word processor file and then Saving As HTML. There was remarkable unanimity on this.

Specific HTML editors that were mentioned for occasional use were:

Adobe PageMill (no longer available)
Mozilla Composer <http://www.mozilla.org>

HTMLKit <http://www.chami.com/html-kit/>
HTMLPad <http://www.intermania.com/htmlpad/>

However, not all HTML work is about editing, and the following packages were honorably mentioned for other functions. Especially important is Tidy, which is pretty much necessary for all but the most experienced people for quick HTML checking. <http://tidy.sourceforge.net> has the original, and links to versions of Tidy for Windows (Tidy-GUI) and just about all other platforms.

GutenMark:
Converts Project Gutenberg texts to HTML and TeX.
<http://www.sandroid.com/GutenMark/>

HTMSTRIP by Bruce Guthrie:
MS-DOS. Converts HTML to text
<http://users.erols.com/waynesof/bruce.htm>

Lynx (lynx --dump):
Converts HTML to text
<http://www.lynx.org>

Dave Raggett's HTML Tidy:
Checks HTML for correctness, reformats and fixes
<http://tidy.sourceforge.net>

W3C html2txt (web-based):
Converts HTML to plain text.
<http://cgi.w3.org/cgi-bin/html2txt>

W3C Validator (web-based):
The Last Word on the correctness of HTML.
<http://validator.w3.org>

wget:
A very neat utility for getting web pages
<http://www.wget.org/>

5. Working with images.

There are two main applications of images in PG--images to be used within texts, like illustrations in HTML, and the management of page images for scanning. These packages are used by volunteers variously for both of those purposes. Their typical use within PG is indicated. "Advanced image processing" packages will permit you to edit and restore damaged images, but for PG work, we mostly just need to manage, convert, resize and crop them.

ACDSEE for Windows

For image reviewing
<http://www.acdsystems.com>

Adobe Photoshop
For advanced image processing
<http://www.adobe.com/products/photoshop/main.html>

ImageMagick for *nix, Mac and Windows
Resizing and format conversion
<http://www.imagemagick.org/>

Irfanview for Windows
Image viewing, conversion, cropping and resizing
<http://www.irfanview.com>

The Gimp
For advanced image processing
<http://www.gimp.org/>

Picture Publisher
For advanced image processing
<http://www.micrografx.com/mgxproducts/picturepublisher.asp>

VuePrint Pro
For viewing images
<http://www.hamrick.com/>

Proofreaders' Toolkit (PRTK)
For splitting batches of image files into individual pages
<http://robertrowe.dns2go.com/>

P.2. What programs could I write to help with PG work?

Look at the programs listed above in [P.1]. Can you write a better version of any of them? Improving OCR and editors constitutes a major challenge, unless you're a world-class expert, but checking and reformatting texts is an area not addressed by large scale programs, and you might contribute there.

Formats FAQ

F.1. What formats does Project Gutenberg publish?

In principle, there's no format that we won't publish, but, in practice, we prefer formats that are open and editable.

An open format is one whose structure is publicly defined and

documented, and not burdened with patent or trade secret or copy-protection (a.k.a. "DRM") restrictions. Anyone can write a reader or creator for an open format, and in 500 years' time, anyone interested will still be able to write a program to display the file. Closed formats, by contrast, will almost certainly be unreadable in just a few decades, when the companies now promoting them disappear, or lose interest, or decide to stop supporting them because they want to sell a replacement.

Being able to edit the file is also important. We make corrections to our editions constantly, and it is important to us that we should be able to update our files easily. If adding one word to a sentence involves a complete re-marking of the whole text and a complete rebuild of the file, we have to ask ourselves whether this format is really necessary for this text. Further, the people who re-use our texts should also be allowed to copy and reformat them freely, and non-editable formats restrict their ability to do this in various ways.

F.2. What is, and how do I make or use:

[Note: Character sets and formats are both listed here. Character sets refer to the characters you can use; formats describe how those characters are put together. For non-text formats such as music files, there is no exact equivalent to a character set.]

ASCII (Character Set)

ASCII (American Standard Code for Information Interchange) is a set of common characters, including just about everything that you can type in on an English-language keyboard. It includes the letters A-Z, a-z, space, numbers, punctuation and some basic symbols. Every character in this document is an ASCII character, and each character is identified with a number from 0 through 127 internally in the computer.

You can view or edit ASCII text using just about every text editor or viewer in the world.

Big-5 (Character Set)

Big-5 is a set of 13,494 traditional Chinese characters. You will need to use an editor or viewer that supports the character set.

Codepage 437, 850, 1252, etc. (Character Sets)

These codepages are Microsoft-specific character sets which allow the

display of accented characters and other symbols. To view a text that uses one of these, you will have to use a Microsoft application that supports them. Many of the fonts supplied with Word for Windows will display and edit CP-1252 correctly. For Codepages 437 and 850, you may have to open a Command Prompt and use a DOS editor like EDIT. A search form <http://www.microsoft.com> should bring up information about the codepage you're interested in, or you can read the excellent overview at <http://czyborra.com/charsets/codepages.html>. For Unix users, iconv and recode provide translation facilities from one character set to another, and support many or all of the MS codepages.

DVI

DVI stands for DeVice Independent, and is commonly used to store text and instructions for displaying it involving complex mathematical symbols and expressions, though it can be used for any content. Given a DVI file, you need a viewer to render it on the specific device you're using. Specifically, DVI is used as the standard output format for TeX, discussed below.

HTML/HTM (Format)

HyperText Markup Language defines the standard format of web pages. You should be able to view these with any web browser, and edit them with any text editor or a specialized HTML editor. <http://w3.org> is the definitive reference.

ISO-8859/ISO-Latin (Character Sets)

ISO-8859 is a series of character sets used to represent the accented characters most commonly used in European languages. There's ISO-8859-1, ISO-8859-2, and so on. ISO-Latin is just another name for the same thing. You can read the overview at <http://czyborra.com/charsets/iso8859.html>

LIT (Format for PDA-based eBooks)

This is a proprietary, closed format for files that can be displayed only by the Microsoft Reader. Search <http://www.microsoft.com> for more information. It is not possible to edit or correct files in this format; it is not possible to export files from this format; they have to be made in another format and converted.

MacRoman (Character Set)

MacRoman is an 8-bit Apple Mac-specific character set which allows the display of accented characters and other symbols. To view a text that uses MacRoman, you will have to use an application that supports it, and there are few outside the Apple fold. However, iconv and recode are programs that convert between many character sets, and MacRoman is supported by both.

MID/MIDI (Format for music)

Musical Instrument Digital Interface is a music description language, encompassing not only file formats but definitions of interfaces. A MIDI file contains instructions for sending messages to a musical instrument to recreate the sounds. <http://www.midi.org/> has much more on this.

MP3 (Format for any audio file)

MPEG-1, Level 3, was defined by the Moving Pictures Expert Group as a means for encoding sounds. Many, many MP3 players exist for all platforms, and can be found easily with a Net search. The official home page of the MPEG is <http://mpeg.telecomitalialab.com/> and copies of the specification can be purchased from the ISO at <http://www.iso.ch>

MPEG/MPG (Format for moving pictures)

The Moving Pictures Expert Group have released a series of formats for encoding video and audio. MPEG (pronounced EM-peg) formats are published and widely used. The official home page of the MPEG is <http://mpeg.telecomitalialab.com/> but you will find information about MPEG formats, and software to play MPEG files, all over the Net. You can also purchase specifications through <http://www.iso.ch>

MUS (Format for music)

MUS from Coda Music <http://www.codamusic.com/> is a proprietary, closed format for editing and replaying sheet music. However, we do post music files in this format because of its many features. We hope to be able to post these also in more open standards at some point in the future, but at the moment, there is no open format with similar capabilities. You can find out more about this at <http://www.ibiblio.org/gutenberg/music/music_helpex.html#what-software>

PDB (Format for PDA-based eBooks)

The Palm Data Base format can actually be used for purposes other than eBooks, and there are many possible variants of formats for Palm-based readers all using the extension PDB on PCs, and they're not all entirely compatible. Some of them are proprietary, and it may not be possible to edit them directly, or export files from these formats; they have to be made in another format and converted. Some can be converted back to text. The most common, though, is the "Palm-DOC" format, which is an open format and can be edited on the Palm itself.

PDF (Format for eBooks)

Portable Document Format is a format for storing texts, containing any fonts or graphics. It is copyrighted by Adobe, <http://www.adobe.com> but is well and publicly documented. It is sometimes referred to as a kind of compiled Postscript (see PS below). It is viewable using the Adobe Acrobat Reader. It is not possible to edit files in this format.

PRC (Format for PDA-based eBooks)

This is a proprietary format for files that can be displayed only by the MobiPocket Reader. See <http://www.mobipocket.com> for more information. It is not possible to edit or correct files in this format; it is not possible to export files from this format; they have to be made in another format and converted.

PS (Format for text and graphics)

Postscript is technically a programming language, not just a format. It has conditional statements, procedures and program flow control. However, it is commonly referred to as a format. Adobe <http://www.adobe.com> holds copyright on the Postscript specifications (there have been three "levels" published) but Postscript is well and publicly documented and has wide support, not only in printing, but in screen display as well. Apart from Adobe's official version, you can also render Postscript files with Ghostscript, a Free Software package. Postscript can be edited directly, but any complex editing may present difficulties.

RTF (Format for text)

Rich Text Format was originally a Microsoft specification, but it is an open format that is used by many word processors to exchange text and format information in an application-independent way. Nearly all current word processors will read and edit an RTF file, and, like HTML, it can also be edited as plain text.

TXT

TXT is a generic extension used for any plain text file, regardless of the character set. Thus, while most of our .TXT files contain ASCII, some contain ISO-8859 or Big-5 or Unicode.

TeX (Format for typesetting, printing and viewing)

TeX (pronounced "tech"--the "X" is actually the Greek letter chi) is a public domain format created by Donald Knuth for typesetting, though it can also be used for normal printing and viewing. TeX consists mostly of the plain text, with instructions for how it is to be displayed. This is compiled into DVI format (see above) which can be rendered onto any device, like a printer or screen, by a program that is aware of the device's capabilities. The Comprehensive TeX Archive Network <http://www.ctan.org/> is the best place to start looking for TeX-related programs for your platform.

Unicode/UTF-8, UTF-16, UTF-32 (Character Set)

Unicode is intended to be a single character set that can handle all of the characters in all of the languages that ever were, or ever will be. It accords with the ISO-10646 standard for the characters, but, in addition, imposes rules of implementation. UTF-8, UTF-16, UTF-32 and their variants are ways of expressing Unicode using different rules for transforming bytes into characters. Unicode is steadily gaining ground, with at least some support in every major operating system, but we're nowhere near the point where everyone can just open a text based on Unicode and read and edit it. Check <http://www.unicode.org> for more.

XML (Format for . . . well, just about anything :-)

eXtensible Markup Language looks a bit like HTML, but whereas tags such as <p> have a standard meaning in HTML, XML allows anyone to define their own set of tags and meanings using a Document Type Definition (DTD) file. Add a CSS (Cascading Style Sheets) file to that, and you have the ability to display the text according to predefined rules. In principle, this seems to make it ideal for the

storage and processing of etexts, since a suitable DTD and CSS, together with the right programs, should make it possible to produce any format of eBook automatically from an XML original. Some PG volunteers have looked at, and are looking at, ways to convert the entire archive using a satisfactory DTD; however, meantime we aren't actually producing much XML, since most volunteers aren't working with it, and nobody wants to start producing many XML texts until we have agreed on a DTD. <http://www.w3.org/XML/> is the definitive source for more information about XML.

Volunteers' Voices

In this section, we asked volunteers to talk about their practical experiences with Project Gutenberg, how they joined, why they give up their hours to work for Free Etexts, how they get down to the nitty-gritty of producing texts.

Some people chose an interview format for their responses, with pre-set questions; others just wrote.

Amy Zelmer

I stumbled across Project Gutenberg a couple of years ago--can't remember just what I was looking for on the web but the idea of PG intrigued me. I was also looking for something to get me reading materials which I wouldn't ordinarily read, so didn't particularly want to find a book in which I was interested--and the whole process of finding a book, finding out if it was already "in progress" and then checking out copyright clearance seemed just a little daunting from what I was able to gather from the info on the web.

Furthermore, I live in a small regional city in Australia, so the possibilities of finding something in either the local library or in a second-hand bookshop was next to nil.

Fortunately I also found Sue Asscher's name and figured that I'd ask a fellow Aussie how to get started. Sue seems to have an inexhaustible stock of books waiting to be entered -- and got me started on Thomas Huxley's "Essays and Lectures". I've now done five other books and am currently working on Darwin's "The Power of Movement in Plants"--quite a variety, but it's at least met my goal of reading something different.

Fortunately Sue was also patient about answering my beginner's questions about formatting dilemmas and has been able to co-ordinate other aspects of the process, like getting scans of diagrams and final

proof-reading. That means all I have to do is put in the text.

I'm a reasonably good typist -- and the practice with PG is certainly improving both my speed and accuracy! (That's meant as a word of encouragement to others.) I generally type for about 20 minutes at a time, then take a break; both my concentration and desire to prevent RSI (repetitive strain injury or occupational overuse syndrome) mean that it's better to do shorter sessions more frequently than to carry on for too long a time. I generally use Microsoft Word 2001 for Macintosh for the first entry and spell check, then save the material in "text only" and do a final read through, removing page numbers and correcting errors which the spell-checker missed as I go.

I've also done some data input for another ebook collection. However, they separate the text and send out small batches of pages to many volunteers. I find that rather frustrating since it's impossible to see how your piece fits until the whole thing is finally posted.

I've done some scanning, OCR and proof-reading of material, but generally find the close proof-reading which is required very frustrating. To each his own method.

Ben Crowder

I've been a book lover ever since the day I learned to read. Several years ago I discovered Project Gutenberg while surfing the net and was delighted to find so many good books freely available. I downloaded all the etexts I was interested in and read quite a few of them. After a few years, I decided to get more involved, so I started proofing with Distributed Proofreaders. I liked that a lot -- I was a newspaper editor in high school for two years -- but I felt an itch to try to produce etexts on my own. I didn't have a scanner, however, so the only solution I could see at the time was to find a book and start typing it in by hand. I'm a relatively fast typist and I figured it wouldn't take that long.

So, I went to my university library, found a pre-1923 edition of G.K. Chesterton's _The Ball and the Cross_ (Chesterton is one of my favorite writers), and began typing. It took much longer than I expected -- certainly over 30 hours, perhaps even close to 50. When I finished, I came across a page on the PG site that mentioned there should be two spaces between sentences. I looked at the etext I'd just typed in and realized in horror that I'd used single spaces the whole way through. :) [1] I had been *sure* that PG used single spaces, convinced that I'd read it in one of the PG docs, which had taken a little while to get used to since I normally use two spaces. But all the PG etexts I checked had two spaces between sentences, so I began the monotonous task of adding an extra space between each sentence (and being very careful not to add spaces in where they shouldn't be). Several hours later the book was finally done. I'd gotten copyright clearance before I started, so I soon submitted it

and within a few days I saw those lovely words in my inbox, "Posted (#5265, Chesterton)".

[1] Ben was right both times: people have posted advocating both one space and two. Either would have been accepted!--jt

Since then, I've been addicted to producing etexts. Languages interest me greatly, so I found an Old Icelandic primer that someone had scanned in, OCRed the images using DocMorph (it didn't take as long as I thought it would, and the output was decent enough to work with), and realized I would have a problem entering in the foreign characters (o's with hooks underneath, etc.). Thank heavens for Unicode. Vim (my editor of choice) has fairly good Unicode support and it didn't take long to make a list of the Unicode codes for the Icelandic characters.

As noted, I use Vim for all my editing. I can rewrap lines to 65 characters by typing "gq", I can use regular expressions for search and replaces (*very* handy), I can edit in Unicode when I need to, and I can speed things up greatly by making keyboard mappings for repetitive tasks. (On one text I was working on, I had to add a blank line between each paragraph. Each was numbered, but the blank lines had somehow been taken out before I got the text, so I started going through and adding them in by hand. The file was 30,000 lines long, however, and I quickly realized it would take a *long* time. I then noted which keys I was pressing to add the blank line between each paragraph, mapped them to <F9>, and held the key down while Vim zipped through the rest of the file. It sped it up by a factor of over a hundred.)

My university library is well-stocked and has lots of old books, so I usually rely on it when I need to get TP&V's for texts I'm not typing in myself. I still don't have a scanner, so I either find already-existing texts on the Internet and reformat them for Project Gutenberg (after getting permission, of course), or find page images on the net and OCR them myself, or type the books in by hand. Typing in by hand takes a long time and so I prefer the first two methods.

Volunteering with Project Gutenberg has been extremely satisfying. The people are wonderful to work with, the work is fun, and it feels very good to know that one is making a difference in the world.

Col Choat

How I got started

People sometimes ask me how I got started in preparing etexts for Project Gutenberg, and while they probably ARE interested in my story often they are really more interested in finding out whether it is

something that they might want to get involved with. Jim Tinsley, a colleague at PG, recently prepared a "questionnaire" as a way of stimulating existing volunteers to document their PG experiences. Answering the questionnaire seems as good a way as any to answer the question, "how did you get started".

HOW DID YOU LEARN ABOUT PG?

I think it was probably from a newspaper or a computer magazine. I can't really recall, now.

WHAT WAS YOUR FIRST CONTACT LIKE.

Initially, I visited the site to search for books I was interested in, to see if they had been posted at PG. That was quite a straightforward process. I downloaded a few texts and either read them at my computer or, occasionally, printed them out to read later.

When I became interested in volunteering, I visited the site to get some information about how to go about it. I found it a bit daunting, really. There was a lot of information but it was difficult for me to get it sorted out in my mind. There were copyright issues, editing rules, and procedures for lodging etexts. There was a question and answer page and some background and information for those wanting to subscribe to the PG mailing lists. In the end, I just sent an e-mail to Michael Hart, whose e-mail address was listed on the site, and said "what can I do?" I notice that volunteers still sometimes do that.

WHAT WAS THE FIRST PG JOB YOU DID? HOW DID IT GO?

I decided to prepare an etext from a book I had in my home library, titled "UNDER THE NORTHERN LIGHTS". It is a series of short stories about the Canadian North by Alan Sullivan. I had a small "hand" scanner at home, which I hadn't used much before. I didn't know any better, so I would scan in about ten pages and save them as "tif" files. Then I would use the OCR (Optical Character Recognition) software supplied with the scanner to convert the image to text for subsequent editing. I recently purchased an A4 scanner with state-of-the-art OCR software and I can't believe how I persevered with that hand scanner for so long.

I tried to apply the editing rules outlined on the PG site, though they weren't as prescriptive as I would have liked. I wanted certainty, as I felt that I didn't know enough to apply own editing rules. I didn't have a good text editor, either, so I probably made the job more difficult than it needed to be. More about the "tools of the trade" later, though.

When I submitted the title pages of the book to PG for copyright clearance it was rejected because the book was published in 1926. I

don’t know what I was thinking about when I chose it. It must have just LOOKED old enough. I had scanned and proofed about half of it, so I just abandoned it and looked for something else. Interestingly, Australians and residents in other countries with similar copyright laws, can now read it as it is in the public domain in Australia and is now on the Project Gutenberg of Australia site. I was able to finish it and post it at PG, after all.

HOW DID YOU DEVELOP YOUR PG EXPERIENCE FROM THERE?

I think that one of the most valuable things I did was to join the volunteer discussion group. I found that I didn’t need to take part, but could just take note of all the different issues raised by other volunteers. Some days there was no activity by the group, but then a hot topic would be raised (e.g. whether some books, such as Mein Kampf by Adolf Hitler, should not be accepted by PG, even if eligible) and there would be plenty of comments. I realised also that I could ask for help on specific questions regarding preparation of texts and receive prompt informative answers. Once, when I thought that I was sending to ONE of the members of the group an e-mail with a large attachment, I was quickly made aware that EVERYONE had received it. Some weren’t amused, but I am a quick learner--I didn’t do it again.

Subscribing to the weekly newsletter is also worthwhile. There is a link on the main page of the PG web site to allow people to subscribe to the mailing list and discussion group. I also found a few people who I began to e-mail privately, outside the discussion group. That helped a lot, too. Perhaps there is merit in instigating a mentor scheme, whereby a new volunteer can refer to another more experienced one for help, guidance and encouragement. I would be interested in taking part in that.

CAN YOU TELL US ABOUT THE FIRST TEXT YOU PRODUCED.

As I mentioned earlier, my first attempt was abortive (initially, at least). However, as I had realised that there was not much Australian content on PG, I decided to go in that direction. Then I found that there were many eligible Australian titles already on the internet, mostly in HTML format. These can only be read using a web browser, so I decided that it would be worthwhile to download them, convert them to text files, compare them with a book of the same title which was eligible for PG copyright approval, and then have them posted at PG. I had learned my lesson, so from then on I always got the approval BEFORE I started work on the conversion.

I prepared a number of etexts using this method and quickly increased the amount of Australian content at PG. However, I still wanted to create an etext from a book. My sister had given me, as a gift, "Australia’s Greatest Books" by Geoffrey Dutton, which reviewed approximately one hundred books and I decided to work my way through them. I had already converted a number from HTML, as outlined above,

so the first on the list to be scanned turned out to be the journal of Charles Sturt who explored south-eastern Australia between 1828 and 1831. I was quite pleased with myself when the two volumes were finally posted at PG.

WHY DO YOU SPEND YOUR HOURS CONTRIBUTING TO PG?

The simple answer is "because it is FUN". It is easy to make up justifications, but since there is no necessity to do it, it must be because I enjoy it. I get a sense of achievement that the work I do will be "out there" for a long time. We haven't begun to realise where technology will lead us. The books I prepare will be able to be read by people anywhere on earth, and even beyond, by astronauts travelling to Mars. "Send up THE ODYSSEY will you Scottie, I have always meant to read it."

I have had some unexpected pleasures, too. I have "met" some wonderfully generous and interesting people and I have read some wonderful books that I would not have taken the trouble to read if I weren't preparing them for PG.

DO YOU SPECIALISE IN ANY PARTICULAR KIND OF WORK, OR TEXTS?

I started out thinking that I would stick to books with an Australian flavour. But I can't help myself. If I see something that I am interested in, and it is already on the internet, but not at PG, I have to do it. I have submitted etexts of James Joyce's "Ulysses", and works by D. H. Lawrence, and Norman Douglas. I also have a long list of books I would like to scan in myself, not all of which are about Australia--one day.

WHAT DO YOU LIKE ABOUT MAKING A PG ETEXT?

I think I have covered that already. I like the sense of achievement, the fun of reading the book, and the thought that it will be available to many people who would not otherwise have access to it, possibly in a form which has not yet been invented.

WHAT DO YOU DISLIKE ABOUT MAKING A PG ETEXT?

Sometimes the going is not easy. Occasionally I get impatient with the length of time it is taking and sometimes I get bored with the subject matter. I recently purchased a new scanner with excellent OCR software, which converts the page image to text, and that has given me a new lease of life because less proofing is required. I sometimes remind myself that I don't have to do it, then I find that I want to anyway.

WHERE DO YOU GET YOUR ELIGIBLE BOOKS

Local libraries have a surprising amount of eligible material. The main difficulty is finding books with a publication date of 1922 or earlier, for PG in the US anyway. I have found a number of "facsimile" editions which are direct reprints of the original, and these are acceptable. I also look around second-hand bookshops. I recently found a battered copy of "A short history of Australia" published in about 1910, and bought it for $A1.50. For books eligible for posting at the PG Australian site, cheap paperbacks are readily available. I am working on one now, and have ripped all the pages out of it to make it easier to scan. It only cost a few dollars. There are also a number of sites on the internet which list second-hand books for sale.

DO YOU TYPE OR SCAN? WHAT SCANNER/OCR/EDITOR/WORD PROCESSOR DO YOU PREFER?

This section might as well cover all of the "tools of the trade". I have noticed that volunteers have many favourite tools, and from what I can make out most will do the job. The list below covers what _I_ have settled on. I should note that I work in the Windows environment, and tools are readily available for all the things I need to do.

Scanner

I recently purchased a Canon A4 flatbed scanner without a document feeder for under $A200. It has a hinged lid for scanning books and comes bundled with image enhancing software and OCR software for converting image to text.

OCR (Optical Character Recognition) Software

'Omnipage Version 9' came bundled with the scanner. I find that I don't need any of the other software which came with the scanner--Omnipage does it all for me. I can scan, proof, spellcheck and save the output to a text file with very little effort.

Editor

I use Editplus which is available as shareware on the internet. It enables me to read in the file produced by the Omnipage OCR software and reformat it to a line length suitable for PG texts (about 70 characters). It also allows one to display guide lines vertically on the page to help with checking for "long" lines. I have loaded James Joyce's "Ulysses" into Editplus and it handled it, so I presume that it will handle files of any size. As with everything one wants to do at PG, there is always someone more than willing to help with problems encountered, just by posing questions to the volunteer discussion.

FTP (File Transfer Protocol) Software

Some volunteers e-mail their submissions to PG as an attachment to an

e-mail. However, it is also possible to place them at the PG site for processing, using FTP. Microsoft Windows Explorer has an FTP facility which can handle this and that suits me. I know that there are many others and SmartFTP is an excellent freeware product for those who need Windows-based FTP software.

Other Tools

I use Microsoft Word to convert HTML files to text files. Firstly, I cut and paste the html document into word, then I convert any italics to upper case, since italics are not supported in plain text files; then I save the document as a text file. Then I use Editplus, mentioned above, to reformat the line length. Sometimes it is necessary to add an extra "carriage return" at the end of each paragraph, to comply with the preferred style for PG texts. This can be done from within Word or Editplus by replacing characters. New volunteers may need to ask for information about this process.

HOW DO YOU CHECK YOUR TEXT? ANY SPECIAL TOOLS? SPELLCHECKER? DO YOU PRINT IT OUT AND READ IT? PUT IT ON YOUR PDA AND READ IT? HAVE A VOICE SYNTHESIS PROGRAM READ IT ALOUD TO YOUR FROM YOUR PC?

I have tried a few different methods. I don't have a notebook computer or etext reader so I must either read it on a PC or print it out. There is a spellchecker with Editplus, which allows one to add new words, so I use that to begin with. I also use GUTCHECK, a program developed by Jim Tinsley, which picks up many errors. One would need to contact him via PG, if one wanted a copy. I travel by train to work, so I often make a printout and read that for the final proof, or co-opt my wife if it is something I can interest her in. I have a checklist, which I have developed over time, that I use to ensure that I have covered all that I need to--but then I AM one for lists.

DO YOU HAVE ANY TIPS 'N' TRICKS OR SPECIAL ROUTINES YOU GO THROUGH WHEN PREPARING A TEXT?

I think I have covered most of my methods already. I sometimes find that "dashes" within sentences need attention. I like to show them as "--" so I try to be consistent and not let them slip through as " - ". I think we at PG could get together a more or less prescriptive list of editing rules for new volunteers to follow. Once they gained experience they could change them if they wanted to. I do like to place an end marker ("THE END") at the end of my progressing work, so that I don't inadvertently lose any of it and I make several rotating backups of the file I am working on. I have "lost" computer files once or twice over the years and don't want to get that sick feeling in my stomach EVER again.

As I said earlier, I do have a checklist, and it could help if PG (that includes me, as PG is "us") provided a downloadable list of things which need to be done to get an etext posted e.g. copyright

approval, scanning, editing, proofing, placing relevant information at the beginning of the etext, etc. All the information is there already, it just needs bringing together into one document.

HOW LONG DOES IT TAKE YOU TO MAKE A TEXT?

Obviously it depends on the number of pages, efficiency of the scanner and the number of hours one puts in. The two volumes of Sturt mentioned above probably took me six months, but I was doing many other things in the meantime. To scan in and edit, say, "The Prophet" by Kahlil Gibran would only take a fraction of that time as it is quite thin and easy to read. If one were concerned about getting an idea of the time it would take to complete an etext, I would suggest that he/she do a little casual proofing at the "Distributed Proofreaders" site first, to get an idea of what is involved.

DO YOU WORK ALONE, OR DO YOU SHARE THE WORK OF EACH TEXT? DOES ANYONE REGULARLY HELP YOU PROOF THE TEXT?

I generally work alone, however my wife will proof sometimes. She has become interested in the book that I am working on at present and is waiting for me to supply her with more pages. When I was getting started, a new volunteer agreed to proof something for me (she approached me) but then she never did any of it and didn't even e-mail me to advise that she had changed her mind. Editing and proofing is not for everybody and one needs to find out if one likes doing it. However, courtesy costs nothing.

DO YOU DO SOME PG WORK REGULARLY, OR DRIFT IN AND OUT AS OPPORTUNITY PERMITS, OR WHEN YOU FEEL LIKE IT.

All of the above at different times. I am not an avid television watcher and would rather do some "work" (or should I say "pleasure") for PG much of the time.

HOW MANY DIFFERENT KINDS OF WORK, OR DIFFERENT BOOKS, HAVE YOU DONE?

Because I have converted many books from work already on the internet, I have covered quite a range, though I haven't actually scanned and proofed too many books. Those that I have done have been Australian historical works. But I have rounded up books on philosophy, aboriginal legends, and several novels. Since many internet sites come and go, I am interested in "grabbing" etexts and posting them at PG in case the site disappears from the internet. It has become a pastime in itself. I recently discovered "South Wind" by Norman Douglas, a book which caused quite a sensation when it was first published because it portrayed a bohemian lifestyle. Ironically, I used to have the book in my home library, but dispensed with it when I needed space. Now it is at PG and I can get it whenever I want it.

WHAT DO YOU LIKE ABOUT THE PG PROCESS?

The democratic, helpful, friendly approach of all the people involved is one of the things I like best. I have "met" so many wonderful people, without having to "live" with them, if you know what I mean. Not long after I started associating with PG, Michael Hart posted an e-mail to the volunteer discussion group, advising of the death of a long-time volunteer. It seemed like she had been one of the "family".

One really needs to be indifferent to praise and the prospect of reward to start volunteering for PG. There is certainly no money in it. However, one quickly finds that there is a community of people out there with a common interest, and with the same outlook and the same interest in doing a job well, without tangible reward. There is no lack of praise though, and one soon finds that one is not indifferent to it.

WHAT DO YOU DISLIKE ABOUT THE PG PROCESS?

There isn't much that I don't like. Nothing worth mentioning, anyway.

IS THERE ANYTHING YOU'D LIKE TO SEE PG DOING DIFFERENTLY?

There are a few things, however since I don't know all the reasons for some things being done the way they are, and because everything is done by volunteers anyway, I wouldn't like to canvass them here. To have produced nearly 5,000 etexts over more than 30 years is testament to the fact that most things are being done "right".

IF ONE OF YOUR FRIENDS APPROACHED YOU TO ASK ADVICE ABOUT HOW TO GET STARTED CONTRIBUTING TO PG, WHAT WOULD YOU TELL THEM?

I would spend some time with him/her and work through some of the issues. I know that I would have benefited from that approach. I would gradually introduce her(him) to the different issues which need to be addressed and find out exactly what her expectations were, and try to help her in fulfilling them.

WHAT WOULD YOU EXPECT PG TO BE LIKE IN FIVE YEARS? TEN YEARS?

Much the same as it is now, I hope. After all, the goal will continue to be to provide "fine literature digitally re-published". Though I expect that, like other organisations, it will continue to evolve in response to new challenges and opportunities. Ten years ago, who would have thought that there would be 5,000 etexts posted; that there would be volunteers operating an online proofreading site; and that there would be a volunteer writing free software to read PG etexts? The

rapid growth of PG over the last few years will present many challenges for the future.

Writing of etext readers, I am reminded that I recently joked to a volunteer that I wanted him to write software for reading etexts, whereby a hologram would appear on the inside of my eyelids so that I could read etexts with my eyes closed. Who knows, it might be possible. However, whatever advances in technology occur over the next ten years, one thing is certain: the work of all the volunteers to date will ensure that there is an amazing library of ebooks available covering creative works by some of the greatest minds who have ever lived. Future readers of PG ebooks will have been given a wonderful gift by the many volunteers who have contributed to PG over the decades.

Project Gutenberg of Australia

On the wall in a colleague's office was pinned a piece of paper on which was written a quotation. I don't recall now what it was and the colleague has been gone for some time and has taken the paper with him. However under the quotation the author was acknowledged as "Prince Machiavelli". I had a vague idea that the quote actually came from "The Prince" by Nicolo Machiavelli, and wondered how I could satisfy my curiosity. Then I remembered reading about Project Gutenberg and decided to see if the book was posted on the PG site, though I didn't really expect that it would be. Needless to say, the etext WAS there and I was able to download it and read it in its entirety, due to the time spent by John Bickers and Bonnie Sala (their names appear at the beginning of the etext) in preparing it for PG. Interestingly, there were other works by Machiavelli there, which I hope to get back to one day.

Later, when I e-mailed PG and expressed an interest in volunteering I was, because I said that I was Australian, referred to Sue Asscher, the Australian Production Director for PG. Sue asked me to proofread "A Vindication of the Rights of Women" by Mary Wollstonecraft. Also, about this time, a journalist had contacted Sue with regard to a story being prepared for PG. He wanted to contact some volunteers to ask why they were interested in PG. Sue referred the journalist to me, with my permission of course, and one of his first questions was "Is there much Australian content on PG?" After I had checked the PG etext list I could only reply "not much".

So I decided to start creating etexts by Australian authors, for PG. Sue Asscher pointed out that there were many eligible Australian works already in the public domain as etexts, so I started rounding up etexts and matching them with books which had been published before 1923, so that they could be posted at PG. Then I started creating etexts myself, for works I could not find already on the internet. My sister had given me, many years ago, a book by Geoffrey Dutton titled "Australia's Greatest Books", so I decided to start working my way

through the eligible titles from the list of about one hundred books reviewed by Dutton. I had already found a number of them on the internet and some were already at PG. But there were still a "few" to be done. There still ARE a few to be done, if anyone is interested in helping.

Then Sue Asscher again had a hand in setting the direction I would take by asking me to proof an etext of "Animal Farm" by George Orwell, whose work had recently entered the public domain in Australia. We didn't know where we would post it, as it is not in the public domain in the US, but I agreed to proof it as I had read it many years ago and enjoyed it.

About this time, I also decided to make up a personal web site. Being a software developer, people were always asking me about the internet and web sites, in the mistaken belief that I knew ALL about computers. I decided to get an idea of how web page design and web site management worked by creating a site that listed all of the "Australian" content at PG. When I couldn't find anywhere to put the Orwell, which I had recently proofed, I decided to create a page on my site for etexts in the public domain in Australia, so that Australians and internet users in other countries with similar copyright laws, could read and/or download them.

Michael Hart, the founder of PG, was quick to interest me in creating an "official" PG site in Australia. After registering a business name, getting a domain name and finding a sponsor to host the site, Project Gutenberg of Australia was up and running.

It all happened very quickly, and as with many things which happen in one's life, it all seems to have come about by serendipity. Even the site's motto "A treasure-trove of literature" was stumbled upon by chance when I looked up, in connection with another unrelated matter, the word "treasure-trove" in a dictionary, to ascertain if the word was hyphenated. Imagine my surprise to find treasure-trove defined as "treasure found hidden with no evidence of ownership". That EXACTLY defined the literature found on PG.

My own association with PG resulted from the culmination of a life-long interest in books and literature and an equally strong interest in computers. Every volunteer brings his/her own particular interests and skills to PG and that, together with the democratic approach taken by the small executive team, is what makes PG the strong, co-operative organisation that it is. My interests and skills, and a generous dose of serendipity, led to the creation of Project Gutenberg of Australia.

Dagny

I discovered Project Gutenberg in 1996 and immediately wanted to help because I love books and wanted everyone to have access to all the wonderful books that, even today with Internet searching, are difficult to find or very expensive when you do locate them.

I began by proofing a few works but what I really wanted to do was share my Balzac collection with other fans. I discovered Balzac in the 1970s and recall my frustrations in trying to find more than a dozen stories of the over one hundred Balzac wrote. It was over a decade before my husband discovered a complete set at a used bookstore while on vacation. Unfortunately, not everyone is so lucky.

With the first few stories I typed for Project Gutenberg I worried about everything: should I correct a type-setting error, leave it, footnote it, etc. This took a long time and involved a lot of correspondence. Now, my idea is to make the text as readable as possible. For me that means correcting type-setting errors I notice. Others prefer to leave them intact. In the end, I don't believe the readers care. I have found them generally to be very grateful to have found some treasure they had been seeking. In some cases of an author's more obscure works, they didn't even know the book existed, a rare find indeed for them.

It is so satisfying to receive an e-mail from someone thanking you for all your hard work. Most readers don't take the time to write but true fans often do and they make it all worthwhile. I have even met people in this way that went on to become a Project Gutenberg volunteer themselves because they wanted to give something back to the Project from which they had received so many pleasurable hours.

Gardner Buchanan

SOURCE MATERIAL

First of all, there is the issue of what texts I choose to do. For me, this is fairly simple. I'm a bit of a small-time book collector already, and have a personal theme: "Canadian English Literature" and "Canadian English-Language History". I have no trouble whatsoever in coming up with submissible editions of works that fit this theme somehow. Nevertheless there are specific authors and works that I'm not having luck with, so I'm still making the rounds of the used book shops regularly and picking up all sorts of stuff.

Eligible volumes have typically cost me $10.00-$150.00 for a collectable edition, or $0.50-$15.00 for a recent paperback edition or garage-sale item. I paid $0.50 for a eligible, but not very collectible copy of Glengary School Days by Ralph Connor at a garage sale. As it turns out someone has beaten me to it--it has been in the collection since 2001. Sometimes if I'm contemplating picking up a

more expensive book that I don't already have a personal interest in, I'll go back and double-check The Online Books page to see if someone has already submitted the book.

Another way I obtain texts is from the Early Canadiana Online archive. They host page images of quite a large collection of old books written in or about Canada, or written by Canadians. The page images are reasonably well suited to OCR.

I tend to produce E-texts two different ways. One way is to submit page images to Charles Franks who runs Distributed Proofers and let him worry about bulk-OCR'ing. I then manage the distributed proofing, which is a fairly low-effort business. The other way is to scan, OCR and proof all by myself. I'm currently averaging two of my own projects to every Distributed Proofer one.

SCANNING AND OCR

I have an very slow parallel-port scanner, a UMAX Astra 2000P. It sucks mightily. I'd rate it a 2 out of 5, if it wasn't acting up--creating a black bar across the page, part way along--so I have to scan books a certain way around to avoid having the bar land in the text. As it sits now, it's in 0.5-1 territory. It is glacially slow at the best of times, and due to being a parallel port model, locks up my whole computer during the scan.

Nevertheless, it is completely adequate to my needs for PG work. I've scanned more than a dozen books on it, and it's done yeoman service--despite its warts. Scanners like this one can be picked up used for $30, and are worth the money.

The way I work when I'm producing a book myself, is scanning and proofing page by page. I do the scans two-pages-up, then OCR, proof and copy the pages to a working document, before going on to scan the next pair of pages.

My scanner came with two OCR "packages": Omnipage something-or-other which I was never able to install, and Recognita Standard 3.2.7. I use Recognita, and for 300dpi scans I do, it is adequately fast and accurate. It is a no-frills package, and DOES make many mistakes, but it is entirely useable for my purposes. I rate it 2 of 5.

I've used the Abbyy FineReader 5.0 try & buy. This is a magnificent OCR system. It handles huge batches and is fast and astoundingly accurate. I rate it 5 out of 5. Unfortunately it costs about $million to patriate a web-bought item into Canada, and while priced at a very reasonable US$100.00, would cost me about CAN$600 after exchange-rate, brokerage fees, shipping, more fees, taxes, service charges and more taxes (on the fees).

I could buy Omnipage off-the-shelf here, but frankly if I can't get Abbyy, I'll stick with Recognita.

As I scan each page, I paste it into Windows-95 Wordpad. Sometimes I also do some proofing in Wordpad, but mainly I proof, fix quotes, M-dashes and paragraph breaks in the OCR program before copying to Wordpad. I like to keep the page boundaries intact, and I mark them in my Wordpad document like this:

:
:
kjdk ldjd ll;llkj dklj dklj
kjdk ljd llllkj klj dklj

page 354

kjdk ldjd lll;;llkj dklj dklj
kjdk ldd lll;;llkj dklj dklj
kjdk ldjd ll;llkj dklj dklj
kjdk ljd llllkj klj dklj

page 355

kjdk ldd lll;;llkj dklj dklj
kjdk ldjd ll;llkj dklj dklj
kjdk ldd lll;;llkj dklj dklj
kjdk ljd llllkj klj dklj
:
:

At this point I also fix-up hyphenated words that straddle page-boundaries. I note paragraphs that start in a new page and mark them with <p>, and I note indented or block-quoted sections and mark these with <in>..</in>. This helps when I go back to format it since I can easily see where the special cases are.

Wordpad handles large documents reasonably well and will grok UNIX files (ie: <LF> only, not <CR><LF>). For this it rates 3.

PROOFING AND FORMATTING

When the whole text is assembled, whether by myself or by Distributed Proofers, I use about the same process for formatting and final proofing.

I use MS-Word 95 to do a spellcheck. This I rate 3 out of 5. I do a select-all, and language appropriately - for me, usually UK rather than American English. I wish I had a Canadian English dictionary for Word 95, but have not needed one badly enough to actually look. Word has a pretty good spell checker and the custom dictionaries are easy to muck around with. I use a custom dictionary for any big project - I have one for Chronicles of Canada, and different one for all the John Richardson books I’ve done.

At this point in my personal process, I abandon Windows and go over to FreeBSD.

I use vi (rated 9 out of 5) to do a number of hacks. I search for and fix up hyphenations that were broken (peer- less) and such like. I also search for and fix some OCR special case errors like 'you'->'yon' and 'be'->'he'. This latter sometimes requires a while, just to step through all the be and he's to see if they're right.

Still in vi, I next use some incantations to run the UNIX 'fmt' command on each paragraph to get it reformatted. I use:

```
fmt -55 60
```

Fmt gets a 3 out-of 5 for what I need it for. It double spaces after sentences, which--although it is probably the right thing to do--is not the PG convention (for me at least). It also adds a space when joining lines with an M-dash. I go back and fix both of these using vi. I take into account the <in></in> tags and manually format accordingly at this point.

As I reformat, I give the text it's final proofing. I'll have the original text in-hand at this point, and will use the page markers (remember them) to figure out where I am. As I reformat, I delete the page markers and other markup. When I'm finished this step, the book is almost done.

Next, I use Gutcheck 0.2 (5 of 5, for intended purpose - way to go Jim!) to check for all the things it checks for. At this point I usually get something like 50 hits, of which 30 are real. I'm then back in vi, and fix up all those problems. Finally, I'm done.

As I go along, I tend to keep various versions of the document. I'm at version 27 of 'The Imperialist' right now. Each scanning editing, spell checking or whatever type of session gets a new version: imperialist_12.txt, imperialist_13.txt,... At various times I might find it useful to use 'wc', 'grep' and 'diff' to figure out what is going on, where a word appears or whether I deleted something I didn't mean to.

HARVESTING PAGE IMAGES

I mentioned above that I sometimes work from page images that I obtain from the web. There are several archives around that hold eligible materials as page images that you can easily download and OCR. I personally have worked mainly with the Early Canadiana Online archive.

After a bit of poking around with the web interface to this collection, I have been able to work out how the individual pages are numbered and organized. I have written some shell scripts that I can use to fetch all the pages of a volume and convert them from GIF to

TIFF format. Harvesting a 200 page book takes a few hours.

Once I have all the pages, I have to do some work with an image editor to get them ready for OCR. I use Corel PhotoPaint 7 to crop each image to just the text area and to remove the black bands at the sides due to the spine or whatever. The page images are often made from microfiche, and dust marks are common as well. These I can sometimes edit out with PhotoPaint.

Because some of the page images, or certain sections thereof, can be completely unreadable, I often find myself either tracking down a modern edition or visiting a local university library to find a copy of the book to look up a few paragraphs or passages that are not readable in the images. Even having to do this, I find that the capture of images from the archive is still a big time saver, and allows me access to an edition that would otherwise be totally inaccessible.

Having gathered the images and prepared them for OCR, I next submit them to Charles at Distributed Proofers, or handle them myself, using the same process as if I were scanning them.

DISTRIBUTED PROOFERS

I've done several books using Charles Franks' most excellent Distributed Proofers web application. I tend to choose DP when I don't have the personal time to read and proof a volume myself, or when the poor quality of the text defies the ability of my (not very good) OCR package.

When scanning for DP, I still scan images two-up. I then have a collection of shell scripts that cut the page images in half to produce single-page TIFF files. I then use a manual procedure with Corel PhotoPaint 7 - if required - to fix up skewed pages or ones with black margins. For the most part, page images that I scan myself are registered exactly enough in my scan area that the page images don't need to be edited.

Page images that I've harvested from a web archive do have to be fixed up before they can be used by DP.

Charles, I believe, prefers that as a project manager I would deal with my own OCR. He has, however, been kind enough to run several batches of page images through his OCR setup for me to good effect. I believe he uses Abbyy Finereader, and my procedure for submitting pages to Charles is to run a subset of the pages I intent to send him through a demo copy of Finereader to make sure that the results are vaguely acceptable. If everything looks good, off it goes.

When the project has run its course with DP, I download the completed text and proceed to format and re-proof it, for the most part, as if I'd scanned and OCR'd it myself.

Jim Tinsley

How I (eventually) got started.

Five years ago, I was the most clueless newbie ever to try volunteering for PG. If you're feeling lost about how to help PG, you can be sure that you're not alone! And if I can write PG's first complete FAQ after my bad start, you can surely do better! :-)

Back in 1997, the web site existed, but there were no FAQs, no Volunteers' Board, no gutvol-d, no Distributed Proofing sites. I started by making a donation and e-mailing Michael, suggesting that I could help out with small jobs, or programming. I didn't get any, and I had no idea what, if anything, I could usefully do by myself.

I looked up the in-progress list at the time, and e-mailed a few people who were listed as working on books, offering to help. None of them were still working on the books. (We no longer show people's e-mail addresses on the InProg list.) I still had no idea how to get eligible books, no scanner, and no idea how to approach producing an etext.

I subscribed to the monthly Newsletter, and just read it for a year. In a "Project Gutenberg Needs YOU" edition, Dianne Bean, the U.S. Director of Production at the time, was given as a contact. I e-mailed her, and finally things started happening.

She sent me a short piece to second-proof, and explained that I should just fix whatever needed fixing. I returned it, and she introduced me to Bill Brewer, who was, at the time, scanning Wisters like they were going out of style. He and I formed a scanning/proofing team for a while.

How I began producing, and my problems with scanning and OCR.

I had some ideas for books I wanted to produce, but I couldn't find them locally, so I turned to the Internet, and discovered how easy it is to find and buy used books on-line.

I bought a HP flatbed scanner. It came with freebie OCR software--"PrecisionScan"--with images and OCR all in the same interface.

I scanned my first book, which fortunately had large, clear text, and the OCR made a reasonable job of it, according to my standards at the time, which were that getting any text at all without typing was a form of magic :-)

I now know that I could have made a better job of it if I had pressed the spine down hard, either closed the top to keep out ambient light or darkened the room, and made each scan a bit more exact. I'm much better at flatbed scanning now.

My PrecisionScan software _did_ recognize two facing pages, and dealt with them correctly, though IIRC it put some garbage characters between the pages that I had to remove by hand.

It did require a lot of editing, though, and recently I've gone back over my original text and found lots of mistakes. Partly because of the scan, partly because of my inexperience.

Throughout the editing, I kept having to make formatting decisions in a vacuum, reinventing wheels and applying rules from a HowTo. Now, having read and formatted and proofed and produced so many texts, I just _know_ how to format a text without thinking, and just reading or even skimming a few texts before producing my own would have given me a lot of background and saved a lot of time. I had proofed several books, but never thought to look closely at formatting decisions.

That text took me a month of working most evenings, and a lot of sticktoitiveness. I can really appreciate the effort that a volunteer has to put in to produce their first text by casting my mind back to that month. I think it's the not-quite-knowing-what-you're-doing that's the worst part. I remember being soooo relieved when I sent it off for second proofing.

The guy who took it for second proofing didn't get back to me for a month, and then said that he wasn't going to do it. This was disappointing. I sent it to another guy for proofing. He came back after a few weeks asking some questions. I answered them. After a few more weeks, I followed up with another e-mail. No answer. A few weeks after that, I gave up, and just submitted the file for posting.

The next book I produced didn't have such nice, clear, large type, and the scan was what I would today call abysmal. I'd guess that I retyped a quarter of the book. The less said about that one, the better.

My third book just _would not_ OCR sensibly. The print was very small and faint, and the OCR produced gibberish. Even with my low standards, I couldn't kid myself that this was working. I tried 400dpi, 600dpi. No dice. I might get 10 complete words on a page.

It was at this point that I bought TextBridge. I really had no idea about the difference between the freebie OCR programs they give away with scanners and a genuine commercial product, but I was trying in desperation to get _something_ different that would read this image.

Textbridge was an eye-opener for me. It still didn't make a good job of the bad images, but it made a decent shot at maybe half of them, and having bought it, I tried it on the two books I had worked so hard at before--it gave hugely improved results. The book that had only

been about 75% OCRed became 100%, but with some errors. I cursed the time I had wasted making up for the deficiencies of my freebie package.

Since then, I've kept upgrading my TextBridge (I think I started on version 8, now on Millennium) and bought OmniPage and Abbyy as well. I mostly use Abbyy 6 now.

Last time I looked, there were downloadable trials of Abbyy, TextBridge, and OmniPage. Big downloads though.

Last year, I got a new Epson Perfection 1640 scanner to replace my old HP Scanjet. I never had any complaint about the Scanjet itself--it served me well--but the new Epson is faster, has higher resolution, and ADF.

Even better, I now know how to scan. I know how to process 200+ pages an hour while scanning the book flat, two pages at a time. I know how to adjust the settings to scan only the area covered by the book. I try different settings for each new book to see what works.

So much for scanning and OCR. I was a _very_ slow learner in this area.

How I prepare a text now.

I was never quite so bad on the proofing end of things. As an editor, I use Brief in DOS and Crisp (a Brief clone) on Windows. (I mostly use vi on *nix, but I do very little-to-no PG work on *nix apart from an occasional scripting thing that I can do in one line of Perl, but would be annoying on MS).

Now, I'm all for tolerance and equality and respect for the faiths of other people, :-) but I gotta say that for someone who has used a powerful editor, editing with Word or any standard Windows editor is like scratching your nose with a rake.

When I first get the text off the OCR, I have many pages with breaks between them, and usually no line-spacing between paragraphs, but each paragraph indented.

I whip out Crisp, and run a macro to search and destroy all page-breaks and page-numbers and blank lines between, and then another to put line breaks between paragraphs and unindent them. Since I watch this process carefully to avoid messing up quotations, it takes me maybe 15 minutes.

Now I have a basically formatted text. The line-lengths are usually too short, and there are hyphenated words at line-ends that I will need to rejoin, and some that I need _not_ to rejoin. Another macro fixes up the hyphenation. At each hyphen, I just decide whether to rejoin or not. Say 20 minutes, max. Then I rewrap. Another 15 minutes.

So in maybe an hour I have a proofable text, and the really nice part about it is that I've had a flying tour of the text three times, so I've already noticed any peculiarities.

If I've noticed any unusual features like letters or poems that need special treatment, I do it at this point.

To prepare the text for proofing, I just flick through it in Crisp with spellquery on, in US or UK English as needed. This puts a red line under queried words, just as Word does. I spend maybe 5 or 10 seconds per 50-line screenful. I don't expect to catch them all; this is just a quick pass to thin 'em out. I may also catch some formatting issues, but I'm not looking for them.

Now I proofread.

I've tried lots of ways of proofreading. Often it's just sitting at the screen. Sometimes I print out the texts or parts of it, and mark errata with a pen. Occasionally, I get the computer to read the text to me, and I follow along in the book, noting any errors. (This is good when you want very high accuracy - do a replace of ":" with "colon", "," with "comma" and so forth before you start the reader.) Recently, I've tried reading the text on a PDA, and bookmarking the problems.

Whatever way I do it, it takes time. I'm better at it now than I was, but I still tend to miss things like he/be.

Some people swear by particular fonts for proofreading, saying that font X shows "1"/"l" differences more clearly than font Y. I just use Arial or Verdana for printouts and Courier or Fixedsys on screen; the special fonts don't seem to make a difference to me.

So I've finished proofing and made my corrections. Now I leave it sit for a few days. I need to get my mind off it, so that I won't miss the same errors I missed before.

When I come back to it, I'm looking at what software people would call a Release Candidate, and something changes in my head . . . I'm thinking of it in a different mode, not as a work-in-progress, but as a potential finished project. This makes me much more critical, and less willing to accept mistakes.

Usually there are dash-problems to fix up (emdashes as " - " instead of "--") and other minor stuff like that. I do global searches for " -" and "- " and "...".

I do a quick skim though it, sampling paragraphs here and there as a test of its quality. I make any formatting adjustments like chapter line spacing or indenting letters that I might notice.

Then I run gutcheck. Gutcheck is a little program I wrote / write /

will-write over the years that complains about common problems in a PG text . . . bad line-lengths, common typos, numbers within words (like the "1" in "wor1d") unbalanced quotations, spaced or unspaced punctuation, non-ASCII characters. I fix the problems that Gutcheck points out.

Again, I switch spellquery on in Crisp, and skim through, more slowly than the first time. This time, I'm looking for _anything_ that shouldn't be in a PG text.

I run gutcheck again, just to be sure.

And off it goes!

The Posting Team

For a couple of years, I churned out a text regularly every two months, spending about 40 hours on each, and took on some occasional proofing, but after I became moderator of the Volunteers' Board, people started referring texts to me for checking or reformatting. This took up more and more of my available PG time, and my own production slowed accordingly.

It was in response to these requests that I wrote gutcheck, which embodies all the standard non-spelling checks I would run on a file. Gutcheck allowed me to spend less time on each text, but still feel reasonably sure that there was nothing glaringly wrong with it.

When Michael formed the Posting Team last year, I volunteered, and it was a natural progression for me, since I was already used to doing a lot of last-minute work on texts.

I found posting to be disorienting and confusing at first; people bombard you with half-scraps of information about books to be posted; some texts need serious work; some texts haven't been cleared, and need to be referred back; some people want special treatment for their texts, which may conflict either with my views or with PG precedents, or both; there are lots of questions. But like every other new job, it just takes time to learn the ropes.

The actual process of posting now takes very little time: I can go through the necessary steps in 3-5 minutes. But posters are the last line of defense against errors, and even the most careful volunteers make them (and yes, we do too!). It takes a minimum of 15 minutes to run standard checks on a perfectly clean file, and it can take several hours to fix up a file that needs help. On average, it takes me about an hour to do my reasonable best for every text submitted.

Apart from posting proper, there are a lot of queries to be answered, many of which I hope I've dealt with in this FAQ, "special cases" that eat as much time as I'm willing to give them, corrections to be made to existing texts, and interminable debates about whether PG

should do _this_ or _that_.

Now that the learning curve is past, the problem with posting is that it generates a lot of e-mail and discussion, and eats a lot of time, and is a 7-day-a-week commitment. Having posted over a thousand texts, I'm now particularly interested in ways to improve text quality.

John Mamoun

How to create an e-text efficiently or automatically is an interesting logistical problem. Here is my procedure, which I recently used to make an e-text in about a week, with maybe 6 man-hours of work on my part:

I take the book, and use an x-acto blade to cut out all of the pages. I then feed the pages into an HP 4C scanner with an automatic document feeder accessory attachment that I got from e-bay for $200. I feed it up to 50 pages at a time, and it automatically scans them in.

I work the scanner using software called scan2000, from www.informatik.com (30-day shareware trial period, $50 to register). This program automatically works with the scanner to save each image as a CCITT4 standard format TIFF file. Most importantly, it automatically numbers each page, starting with an initial value you specify (typically 001.tif) and increasing the number of the file name by an increment you specify (typically by 2 pages, since you scan double sided pages; you scan the evens first, then flip the pages over and scan the odds, but you want the page numbers in order, right?). So the scanner outputs, say, 001.tif, 003.tif, 004.tif, etc., then you flip the pages over and re-feed them into the scanner; the even pages are saved as 002.tif, 004.tif, etc., after you tell the program to begin the first of the even page files with 002.tif.

So now I have a bunch of consecutively numbered CCITT4 TIFF files. At this point, I could use a freeware program called cc42 (search for it at www.pdfzone.com) to combine all of the sequentially numbered CCITT4 TIF files into a single PDF file with the pages in order.

Or, if making e-texts, not PDF files, I OCR the pages and save them as corresponding pages like 001.txt, 002.txt, etc. I also use Paint Shop Pro (shareware 30 day trial) to batch-convert the tiff files into GIF file format. I can then upload the GIF files and the correspondingly numbered text files to the Distributed Proofreaders page (http://texts01.archive.org/dp/) to have them rapidly proofread by numerous proofreaders, who finish the task at a rate of 50-100 pages a day per book, very roughly speaking. When done, I then download the text files as a single text file combining all of the files. The upload function on the DP site is tedious, requiring one to upload

each file one-by-one, but I spoke to the webmaster recently, and he said there are, with special arrangements, ways to FTP them or even e-mail them to him on CD.

Now, hard returns. It was once a grave problem to fix hard returns so that the text outputted to 65 characters per line. Then I got a freeware program called Clipcase at www.shareware.com. With Clipcase, you select a body of text (about 20 pages or so; any more, and the program crashes) in your word processor, copy the text to the clipboard, then load up Clipcase, paste the text into the Clipcase window, the process the text.

When this happens, all of the hard carriage returns within the text are eliminated, EXCEPT for returns between paragraphs. Then, you select the text, copy it, and paste it into any word processor to process it. I use Microsoft Word. After pasting all of the text into it, I select all of the text, choose Courier New font, 10 point size, and set the margins at 5.5 inches. With this setup, when the text is saved as "Text with layout," the resultant text is 65 characters per line, every line. Setting hard returns is automatic.

Then I spell-check the text, and also skim through it to look for typos and "categories" of errors to tend to occur repeatedly within the text. One common error is having a single dash instead of two dashes, for example:

He lingered-slowly.
as opposed to: He lingered--slowly.

Another common error is a space between a period, exclamation mark or other punctuation mark, and the letter that came before it, such as:

Hey !
instead of Hey!

or " Hey, "
instead of "Hey,"

I then use the "Find/Replace" command within Microsoft Word to efficiently get rid of these. For example, I might tell it to look for ^w", where ^w means "a white space" and " is a quote. This looks for white spaces before quotes. "^w looks for white spaces after quotes. ^w! means a white space before an exclamation mark. I can also have it look for "any letter"-"any letter," so that it finds single dashes between letters, and then I can decide if I want to replace these with double dashes. By using these kinds of find/replace tricks, it becomes easier to remove typos.

When done, I save as "text with line breaks" and it is done.

That’s basically my procedure. 1 week turnaround time and 6 man-hours on my part for a 190k text file...

Ken Reeder

The Story of My Life (as pertains to PG) by Ken Reeder
June, 2002

I am currently finishing up my fourth etext, with two more etexts in process, another seven books sitting on the shelf waiting, and a lot of additional books that I would like to do when those are done.

Sixteen months ago I was blissfully unaware of PG and of the world of online books. A couple of things seemed to come together to lead to my involvement with PG. I spent some time helping one of my sons, for a school project, in an unsuccessful search for an online English translation of Pliny's Historia Naturalis. About a year before that I had been tinkering, for no particular reason, with trying to type one of my favorite older sci-fi books into a text file. And I had been thinking, occasionally over the course of a few years, about a series of books to which I was avidly devoted when I was about twelve or fourteen years old, which was widely available then but is relatively scarce now. It was a web search on the name of that author, Joseph Altsheler, which happened to lead me to some couple-year-old messages on the PG volunteers' bulletin board.

I poked around the PG web site a little and thought, hey, I think I could be interested in this. Only a few months before I had, for no particular reason, picked up a clearance-model parallel flatbed scanner (for which I paid $36, including shipping). The scanner package included some OCR software, so I already had the basics needed to scan a book to produce an etext.

So I rummaged around on the PG web site a good bit more, and lurked on the volunteers' board, and figured out that I could find the books that I wanted on Ebay or ABEbooks, and bought a couple of books for $10 or $15 each. I scanned a chapter or two and tried out the OCR, which worked very well. (The OCR software that came with my scanner is TextBridge Pro, which it turns out is one of the more highly-regarded OCR packages, so I was just lucky in that respect because I had no clue. I could see that the OCR software was clearly much better than some DOS software that I had used at work about 15 years ago.)

What appealed to me was that, firstly, it seemed like this was a worthwhile thing to do, with a big plus being that you can do the work from your own home, in your pajamas if you want, in whatever time you can spare. And I thought that, being a detail-oriented software-developer geek kind of guy, that I would kind of enjoy it and also be pretty good at it - actually, I've always had an aptitude for proof-reading.

So I went ahead and mailed in a couple TP&V for copyright clearance,

and set out to actually produce my first etext, a 348-page book which I completed in about 10 weeks, start to finish.

For a book with nice clear, good-sized print, I figure that it averages out to about 7 or 8 minutes per page to go through my complete production process. Some of the books that I am working on, with smaller or less-perfect print (and/or other complications) take a little (or a lot) longer.

I feel that I've got my process pretty well set by now. I've put together several little home-made utility programs, written in FoxPro, which assist me. (I've put in some effort to try to adapt some of these for possible use by others, but the problems are that it takes a lot more work to polish software to the point that I feel comfortable letting somebody else pound on it, and the scope of what I think the software ought to do gets bigger every time I work on it, and it's not nearly as enjoyable - for somebody who develops software at work every day - as producing etexts.)

My complete production process, with rough time breakdown, is as follows:

1. Scan the book, 2 pages at a time, about 1 minute per scan (30 seconds per page). (I do not cut the pages out of the book, I just lay it flat on the scanner and press down on the spine.)

2. Run the BMP file through TextBridge Pro, about 30 seconds per page. (Again, when working with clear, good-sized print.) I save the output as text with no line breaks.

3. Run a little FoxPro utility that I wrote that massages and formats the file a little bit.

4. Do my first-pass proof-read, about 2 minutes per page, combining the pages into chapters.

5. Run another little FoxPro utility, which checks for some things that I might have missed during proof-reading.

6. Use MS Word to perform a spelling and grammar check, another 30 to 60 seconds per page.

7. Run another little FoxPro utility (number 3), which inserts line breaks, then run another one (number 4) which does some more exception-checking.

8. Do my second-pass proof-read, about 2 minutes per page.

9. Combine the chapters into one big file. Run a couple more little FoxPro utilities (numbers 5 and 6) which do some final formatting, checking and analysis.

10. Send the file to Jim Tinsley, who will graciously run it through

his GUTCHECK program which scans for a lot of common errors.

11. Call it an etext and send it in for posting.

My primary goal is to produce a quality etext - I don't particularly care about trying to speed things up. I mean, I don't want to needlessly waste a lot of time, but I look at this as a hobby and I enjoy working on it, so I don't get out my stop watch to see if I can get 20 pages done faster today than yesterday. (When I go out running, then I'm concerned about whether I'm faster today than yesterday.) I generally put in maybe 5 hours a week on PG - actually, it's often easier for me to fit in some PG work on weekday evenings than on the weekend. And it is definitely gratifying when the etext is done and not only does it get posted on PG, but then links and copies pop up in different places like the "Online Books Page", and DMOZ.org, and Blackmask.com and Bookshare.org.

I have not encountered any real stumbling blocks so far. There were a few things that took some time to figure out. For example, when my first etext was ready, I was pretty sure that it was expected that I would put the PG header on myself, but I looked all over the web site and could not find a "master" copy. (Actually, I think the master, such as it was/is, is available on Lyris, but I was not subscribing to Lyris then.) So I just pulled the header from a very-recently posted etext, but then after I sent the etext in it was posted with a different header anyway. (Nowadays, my understanding is that the PG "staff" prefers to put the header on.) I also spent some time researching 8-bit code pages, but I expect that the new big-FAQ will provide easy access to all the answers that I had to hunt down then. There's a lot of good information buried in past messages on the volunteers' board, but no good way to search out information on a particular topic.

So far I've been able to fill all my book needs without spending much money. I find my books through ABEbooks, or from Ebay, plus I've gotten a few at Ohio Book Store downtown on Main Street. I've rarely paid as much as $20 for a book, even including shipping. There's one book that I've purchased (but not yet started work on) which costs $1000 or more for the original edition, but which is also available in paperback reprints for about $10. There are some other books in my future plans which look like they will be more expensive, but we'll worry about that when the time comes.

My wife still cannot understand why I spend my time scanning books, whereas my kids (and, I guess, most other people I know) seem to think it's a little eccentric but basically acceptable behavior. Personally, I definitely enjoy producing etexts and hope to keep doing so for a long time. My thanks to Michael Hart, Jim Tinsley, Greg Newby, and untold others who devote so much effort to nurture the project and grease the skids for the rest of us. Long live Project Gutenberg.

Lynn Hill

I have been involved with PG since 1994, when I first began reading texts on-line during slow times at the office where I worked. (I once got into trouble with a co-worker when she found me "processing" Little Women instead of the week's payroll report.) I was surprised to find, even then, such a wide variety of material in the PG archives. I found myself re-reading favorite books from my childhood, and delighting in finding "new" ones--Little Lord Fauntleroy, The Secret Garden, Heidi, the Oz stories. They were not at all like the sugary old films I had seen on television. They were funny, heartwarming, and utterly charming. After some years as a reader of the texts, I found myself thinking, "I'd like to try this."

When I first checked out the web page for volunteers, I felt overwhelmed. There were all sorts of FAQ's, but when I read them, I was baffled by all the information about file types, fonts, and other details. I didn't even know where to get books, let alone what to do about jagged rights edges or indented lines. It was frustrating -- I had all this enthusiasm but didn't know where to apply it. I dawdled for some months, then came back and turned to the PG Volunteers' message board for help.

Help came from many sources. I found someone who needed a file proofread, so I offered to read it. This worked out well, and I even found a couple of typos in it. I proofed some more files for this person, and then some for other people on the board.

After a while, I was ready to try a whole book -- and from Dianne Bean came my first PG book, "The Golden Slipper" by Anna Katharine Green. When I opened the box, a stale smell floated out, and then I found a chunky book with the ugliest green cover I've ever seen on anything. The date was 1915, and the book was starting to crumble all around the edges. My first reaction was "Who would ever want to read this???" But since I had promised to do it, I dutifully started scanning and reading as I went along. The book was a collection of mystery/suspense stories about a teenage crime-stopper named Violet Strange. (I always felt as if Scooby Doo and his friends might turn up at any moment.) As I read, I began to like Violet, and to notice how different her world seemed from ours. By the time I reached the end of the book, I felt proud of myself for "saving" some good stories for the future, and ready to try another book.

My suggestion to new PG'ers is to jump in and not be shy about volunteering. PG is a big group of great people who care, but they do not know you are out there until you say something. Once you speak up, they will do anything short of triple backflips to help you.

There are many ways new folks can join in, from scavenging old books at yard sales all the way up to proofing files or scanning and typing in whole books. When you send in your first copy of title page and

verso, be patient -- it takes time for your copyright research to be done. This is a great time to do proofing on-line at one of the distributed proofreading web sites.

I get my books from library sales, yard sales, friends I met on the PG Volunteer board, and even from elderly neighbors who wanted to lend me favorite books they have saved. When you want old books, tell everybody you know. They may come up with a lot of eligible books you wouldn't have expected.

When you find an old book, my second piece of advice is not to be too hasty in deciding whether you want to read it or not. Old books are dated, naturally, but they can show you things about life in the past which you can't pick up from an A&E documentary. I am especially interested in the way women and children are portrayed in these old books--every woman is not necessarily a lady, and every child is not a sweet little angel. (If you haven't read Little Lord Fauntleroy, you are missing a lot of laughs.) These insights and ideas can keep you going through a lot of long dark winter evenings, and they're handy to think over when you hit the occasional dull chapter or scene.

My hardest text to do was See America First, by Orville Heistand. The author invites readers to join him on a trip from Ohio to Massachusetts, in which he visits several landmarks and historical sites and entertains you all the way with obscure poetry, proverbs, and little moral lectures about each rock and robin he encounters. I told my husband, Chris, that the author's (literally) rambling style was driving me crazy. Chris proofread some chapters for me, then commented, "Boy, you never see anybody these days have such a fun time going nowhere!"

By now, I've done nine complete texts, and have boxes of other books to do. I have found that children's books are my favorites, but I will try anything if it is clear enough to read. I don't work on PG every day, or even every week if I get too busy with other things, but I keep coming back. I find PG projects to be very relaxing, a way to use my computer and writing/proofing skills, and also a refreshing change from my daily work. It's also a great excuse and motivation to read lots of books!

Sandra Laythorpe

HOW I STARTED AS A GUTENBERG VOLUNTEER

I first learned about Project Gutenberg from a Computer magazine, so I searched for it on the Internet, and found all these classic books I had wanted to read for years, and they were free! At that time, I read a paperback copy of The Heir of Redclyffe by Charlotte M Yonge. I thought it was a wonderful book - indeed I still think it is the best

novel to come out of the nineteenth century. After reading the 'How To' files on the Gutenberg site, I thought maybe I could produce Miss Yonge's books with the equipment I had. I wrote to Michael Hart and asked him, and got a very positive reply and lots of information from him.

I jumped in the deep end! I bought a very old copy of The Heir of Redclyffe, sent the photocopies of the title pages to Michael, and sat down at the computer, learned to use my OCR facilities, and got on with it, learning by my mistakes. The Instruction files told me most of what I needed to know, and Michael gave me an introduction to David Price, an experienced Gutenberger, who would be able to help me. He has been invaluable in explaining things; I don't think I could have produced my first attempt without his guiding hand.

I buy my books off the Internet, or from local dealers. Most of Miss Yonge's work is still available from second-hand bookshops, and I am happily living in a location where they are not too scarce. I have Gutenberg colleagues, now, helping with CMY, and I post books to them snail-mail, if they can't buy them in their own countries.

THIS IS HOW _I_ DO IT.

I use PrimaPage OCR program; it was on the disc which came with my Primax Colorado Direct scanner, and I do the work on my PC. Before I start, I open my scanner program, and adjust the settings to take black and white photos, and the brightness to about minus 35 or 40. This is crucial, as I won't even be able to _see_ the page until I get it right. When I first began, it took many adjustments to get it right. There should be as few mistakes as possible on the OCR result. If the photograph is too light, the OCR reads words wrongly. If the photograph is too dark, there are shadows which create black patches on the pages. If I can't get rid of these black patches, I have to tear the pages out of the book and do them one at a time. Important: don't buy first editions!

I use the scanner to take a photograph of two pages. The photograph appears on the screen. Then I close the photograph, which my computer calls 'untitl1'. Next I open my OCR program, and search for file 'untitl1', and open that. Then I ask the program to clean it, and then I click onto the button that 'reads' the photograph and converts in from pixels into letters = Optical Character Recognition!

When I get the OCR result (which takes only a few seconds), I save the 'read' text file into my own documents, numbering the file the same as the number of the page of the book. I have created a folder called 'Gutenberg', and I save it in there in a text-only format. So I go to my Gutenberg folder, open this new file, and visually correct the mistakes. I save the finished page, create a Chapter 1 file, and save it and subsequent pages that I have prepared, to build up the whole book. After I have proofed the OCR result, I paste the finished text into a Microsoft Word document, setting the font at Courier New size

10. This sets the lines at the right length for Gutenberg. When I have finished the whole book in Word, I save it as text-with-line-breaks, to get the final text file, which I send to be posted on the Gutenberg site. I proof my work two or three times, depending on the quality of the OCR result, and do a final spelling check with MS Word. I don't ask other people to proof my texts, because Miss Yonge's idiosyncrasies are liable to get edited out, unless the proofer has the book to hand.

It took me 6 months to prepare my first text, The Heir of Redclyffe, but I can do 10 pages an hour now.

In my Gutenberg folder, I have other useful files for reference, mostly downloaded Gutenberg Instructions files. So if I need to find something out, I can look in these files--it is much easier than searching on the Internet. If I need to know something I can't find in these files, I may ask a question on the Volunteers WWW Board, although I try not to, because the answers are nearly always in the files.

I try to process 2 sheets of 16 octavo pages a day, taking about 3 or 4 hours. I do my housework & gardening in the morning, then settle down to an afternoon's happy Gutenberging :-).

WHY DO I GUTENBERG?

When I became semi-retired, I wanted to do some voluntary work on the Internet. Coincidentally I began reading the works of Charlotte M Yonge, and discovered that most of her works are out of print now. I felt that they deserved a much wider audience, so I decided that my voluntary job would be to do just that. Miss Yonge lived in a village only a couple of miles away from me, so I had a local interest, too. On my web page, http://www.menorot.com/cmyonge.htm, you will find out a little about her, and Otterbourne, the village she lived in all her life, and find links to other web sites about her.

I discovered the Charlotte M Yonge Fellowship http://www.cmyf.org.uk/ and am now in contact with other people who appreciate her work, including academics who write clever things about her. Her books are about families, their interactions with each other, and how they, in Christian terms, grow in grace. I don't think there is another writer who can write so well about families. She was a Tractarian, a Christian who, in the nineteenth century, believed that people could be influenced for good by what they read. For this reason, 20th century people found her characters too moralistic, and her prose too turgid. I think her novels are delightful, her characters lovable, and her prose is minutely descriptive. It was said about her that she was 'able to make goodness exciting'. This is a rare talent, perhaps only found in other Christian writers like John Bunyan or Charles Kingsley.

Through the Gutenberg site, Miss Yonge's works are more easily available than ever. She originally wrote for upper and middle class

young women. Even though I live a century and a half later, I can recognise her characters in their 'descendants' who live around me, but I sometimes wonder what Chinese, African, or even modern American readers think of her, their own backgrounds so different from the English Victorians.

I enjoy making Gutenberg texts, the work is simple, once you know how to. I would prefer, however, to see them presented in HTML. The modern ebooks all need to be in HTML format to present nicely on their tiny pages. I believe Gutenberg is going to publish HTML files, I would like to learn how to do it. Eventually, I think Gutenberg files will be available in a format that will work on all PCs, handhelds, palms, and ebooks;--but I don't know what that format is yet, I don't think standards have even been worked out among the ebook publishers.

Finally, yes, I do find mistakes in my published texts. When I have finished all 200+ of Miss Yonge's books, I am going to go through them all for the second time, and remove the mistakes. So, my work is cut out for many years to come. . . .

Suzanne Shell

Over the past several years, I visited the Project Gutenberg website occasionally, looked at what was involved in making a significant contribution to the effort, and left after downloading a few books--PG was a project that would need to wait until I retired.

In the summer and fall of 2002, I was doing research on e-books (sources, devices, costs) for my library, and ran across Distributed Proofreaders. I discovered Blackmask.com at about this time, and also followed a link from there to Distributed Proofreaders. Serendipity! After backing away a few times, I took the plunge and registered on November 5, then began proofing. The however-many-pages-I-wanted-to-proof commitment was just right for letting me get a feel for the process, and to start me thinking of the ways I could exploit all this free labor to get the books _I_ wanted into PG.

I was feeling quite virtuous about proofing my 10-20 pages per day, when I visited the site on November 8, and NONE of the books I was working on were available. Also there was this perfectly absurd number listed for number of proofers having proofed at least one page (it had roughly quadrupled). I KNEW the site had been hacked. Actually the site had been slash dotted. The DP discussion forums were so active, it was hard to find time to read all the messages, questions, suggestions, and complaints; these rapidly led to new documentation and more detailed proofing guidelines. Books moved through the site so rapidly that they brought out the "hard stuff"

from the bottom of the to-do stack, and were STILL desperate for content. I was a relative "veteran" after just a few days, and helped out a little by answering questions, but I was still a beginner. I had some PG dreams that DP could make reality, but I needed to learn the ropes first.

Some of my ambitions revolved around professional goals--there are some public domain titles, which, if available in electronic form, would be extremely useful to my library's patrons. There are also some standard reference books and indexes--Granger's Index to Poetry is one example--that have pre-1923 editions that could still be important resources. In order to learn what I needed to know about providing content, though, I decided to start with something less overwhelming (wanting to read it on my e-book reader was just a coincidence). I went to my bookshelves and pulled out my P. G. Wodehouse reprints. I downloaded and read the scanning and submitting FAQ from the DP site, requested and received clearance for the first book (_Uneasy Money_) in late December, and got to work mastering my scanner. I tried Omnipage Pro first, but decided that ABBYY Finereader Pro did a significantly better job of the OCR. I offered to be a "behind the scenes" manager for the book while it worked its way through the site, but was made an official "Project Manager" instead. Although the first frenzy following the slash dot invasion had calmed down, DP was still feeling a need for more content and more hands to manage projects.

On January 5, _Uneasy Money_ started proofing; it went through 2 rounds of proofing in less than 20 hours. I felt a like a hick marveling at a traffic light changing colors, but I sat at my PC and watched the page count go down. By this time, I had also scanned and OCR'd a couple more Wodehouse reprints and a short book of poetry. I was hooked! Juliet Sutherland and the other admins had recruited some experienced DP'ers to help train new post-processors in the job of preparing final PG texts. I was handed over to one of them. After several projects, I "graduated" and was given permission to upload my own projects. My intent was to do 3 or 4 projects a month, no more than I could handle post-processing by myself. I planned to process an occasional reference book in addition to all the Wodehouse I could get my hands on. So much for plans...

One ongoing concern of many Distributed Proofreaders was how to train new volunteers in the DP style of proofreading. (It is somewhat idiosyncratic because of the distributed nature of the process.) We were still coping with the aftereffects of the massive influx of slash dotters--quantity benefited, but quality suffered. Super7, one of the highest volume proofreaders, suggested setting aside a project without complex formatting for "Beginners" and asking that the second round proofers (all of whom should be veterans) send feedback and encouragement to the newcomers. This was tried successfully, and with a couple of variations. Since I had been planning to start running a variety of genre fiction through the site, I then volunteered to manage these as beginners' projects for as long as the supply held out. All of a sudden, starting in

February 2003, the amount of time I needed to spend locating, scanning, OCR'ing and managing books increased drastically, and the amount of time I could devote to post-processing decreased. Luckily, "veterans" stepped in to answer newcomers' questions, and to serve as "Mentors" in the second round of proofing. Recently, others have provided "beginners' projects", to help keep up with the demand of a steadily increasing flow of new volunteers. These projects are also useful for helping new post-processors learn the job.

I still have some ambitious projects planned; Granger's _Index to Poetry_, the unabridged edition of _The Golden Bough_, Curtis' _The North American Indian_, and the _Book Review Digest_ (volumes for 1905-1921). A couple of volumes are already waiting to be proofed, others are waiting to be scanned on the PG tabloid scanner. But, in the meantime, there are 23 new Wodehouse books in PG thanks to Distributed Proofreaders, not to mention such remnants of early 20th century popular culture as _The Sheik_.

I believe that a major accomplishment of Distributed Proofreaders has been the creation of way to provide on-the-job training for PG volunteers. Steady improvement in the quantity and quality of training techniques and documentation, enhancements to the user-friendliness of the site, and ready access to the collective experience and advice of a wide range of volunteers in the Forums have resulted in a growing core of active and experienced volunteers in all the facets of e-book production. I'm sure that I could not have progressed from a total newbie to a regular PG contributor within a 5-month period without this support structure. Regular communication and collaboration with book-lovers from around the world has enriched my life. The fact that it is easier to get leave from my job than from DP, is perhaps beside the point...

Tony Adam

How did you learn about PG?

It's been so long, I don't really remember! I probably read about it on a library listserv (I'm a librarian), and since making old texts accessible has always been a concern of mine, I jumped right in.

What was your first contact like?

Great! Mike Hart has always been easy to deal with via e-mail, although we've never talked. He and the "crew du jour" directed me to the FAQ and I took it from there.

What was the first PG job you did? How did it go?

My first job might have been Henry James' _Turn of the Screw_ (I just found a note from September 1993 on copyright clearance for it). Since in a former incarnation I was editorial assistant for the _Henry James Review_, I thought that would be a good start. I've always typed the files (I'm a fast typist), and I think we had few problems along the way.

How did you develop your PG experience from there?

Helter-skelter, much like my reading habits. I work at a historically black university, so getting 19th C African-American works posted is a central concern. I've done _Clotelle_ (the first A-A American novel) and the autobiography of Henry O. Flipper, the West Point cadet, and I'm always looking for something new in that area. Somewhere along the way I got sidetracked into essays by Whittier and other U.S. poets, and I've collaborated on early American historical documents and Sir Walter Scott with a fellow PGer up in Ohio and Chinese documents with another contact in Japan. A couple of years ago, I saw that someone in San Francisco needed help with the Shakespeare Apocrypha, and that has occupied my time on and off since. It's always something!

Can you tell us about the first text you produced?

I think it was _The Turn of the Screw_, which was a good starting point--not too long, a good read, etc. Just plugging away at the text a few pages a day made the process go quickly.

Why do you spend your hours contributing to PG?

I love the idea of making all of this print knowledge available to anyone anywhere. Working in a library that has suffered budget problems over the years opened my eyes to the need for acquisition of as much free stuff as possible for our students and faculty. Besides, in a perverse way, it's fun!

Do you specialize in any particular kind of work? of texts?

I've probably focused more on plays, historical documents, and 19th C U.S. works than anything else.

What do you like about making a PG text?

Having a project come to fruition--finally seeing an almost forgotten text come to life again.

What do you dislike about making a PG text?

The work can be tedious at times, depending on the author. But sometimes you have to plow through to get something significant processed. For example, we probably should have more philosophers represented, but what a horrible thing it would be to scan Kant!

Where do you get your eligible books?

Mostly from my library's collection, although I finally purchased my own copy of the Shakespeare Apocrypha (it's very hard to find, which makes it very suitable for posting). I've interlibrary loaned some items, but that's also been unusual.

Do you type or scan? What Scanner / OCR / Editor / WP do you prefer?

I still type everything--it's easier when working with a play, I've discovered. But I'm purchasing a scanner in the very near future and will do more with that.

How do you check your text? Any special tools? spellchecker? Do you print it out and read it? Put it on your PDA and read it? Have a voice synthesis program read it aloud to you from your PC?

I usually run it through the spellchecker, although depending on the work, I read it line by line a second time.

Do you have any tips'n'tricks or special routines you go through when preparing a text?

The best thing to do is put yourself on a schedule--do a set amount of pages every day, and you'll be surprised how quickly you get to the end. I also make a pencil mark in the book at a stopping point and even read back a paragraph to double check what I last entered.

How long does it take you to make a text?

Depends on my work schedule, other assignments, time of year, etc. A play might take a couple of weeks, but a Walter Scott novel could take six months. I think my record is probably one day for an essay, but that's unusual.

Do you work alone, or do you share the work of each text? Does anyone regularly help you proof the text?

I've worked alone and on teams, depending on the text. No one regularly helps to proof the text, but occasionally someone else does.

Do you do some PG work regularly, or drift in and out as opportunity permits, or when you feel like it?

I consider myself a regular, as time permits. In other words, I haven't dropped out of the picture, but sometimes I might not enter anything for up to a month.

How many different kinds of work, or different books, have you done?

Not sure how many different books I've done, but it's been a wide variety: James' and Scott's novels, Whittier's essays, a whole collection of early American documents (mostly New Netherlands), Shakespeare (accepted canon and the apocryphal works), some odd works (_The Psychology of Beauty_ comes to mind)--the list goes on and on. I've even forgotten that I've done some titles!

What do you like about the PG process?

That it's open-ended--if I think I have something that should be posted, I don't have to jump through hoops and ladders to get permission (other than copyright clearance).

What do you dislike about the PG process?

Can't think of anything offhand.

Is there anything you'd like to see PG doing differently?

I know it's a bone of contention, but we probably need to explore moving away from ASCII.

If one of your friends approached you to ask advice about how to get started contributing to PG, what would you tell them?

Start with something fun, that's close to your heart, and keep plugging away a little bit at a time.

What do you expect Project Gutenberg to be like in 5 years? 10 years?

We'll probably be a whole lot bigger (texts and personnel), with a different look to the texts. Maybe we'll even have more audio versions of texts, using some of the new software that's coming out.

Tonya Allen

I discovered Project Gutenberg in about 1997. After several years of enjoying PG's texts, in June of 2002 I decided it was time to start contributing. Via the PG web site I learned that the easiest way to

do this would be to help out with proofreading via Charles Franks' Distributed Proofreaders web site. The day I signed on I proofed nine whole pages of a children's book called _Curly and Floppy Twistytail_ and felt very proud to be contributing.

At that time, there were probably only about 40 active volunteers on the site each day. Often I proofed an entire book almost all by myself over the course of a week or so. Things moved at a leisurely pace; guidelines were few and simple; and I had fun reading old books and discovering new authors.

After a few months a request was made for volunteers to post-process texts in French. I volunteered to help with this, and that was how I became a post-processor (PPer). Shortly afterwards, the web page listing texts available for post-processing and sign-out was unveiled. I remember several times checking and being disappointed because there was nothing currently available (hard to imagine now when there are always at least 40 texts waiting).

One day in November, I picked out a likely-looking text from the proofing page, and settled down for an hour of reading. As I recall, it was _The Greek View of Life_, a sizeable text of which only a few pages had been proofed so far, and which I thought would last for several days at least. At about that time, someone emailed me to say that DP had been "/.ed." "What does that mean?" I replied. I soon found out.

I had been proofing away peacefully for awhile when suddenly instead of the next page, I got a page about twenty pages further on. The same thing happened again and again, and suddenly all the pages were gone; the whole text had been completed. DP had indeed been slashdotted.

Since then, a lot of amazing things have happened. The number of active volunteers per day has increased almost 1000%. The number of texts that go through the site has increased exponentially. All kinds of proofing and processing tools have been developed. I now spend most of my time checking texts that others have PPed, and submitting them to PG, at an average rate of one to four per day--quite a leap from nine pages of _Curly and Floppy Twistytail_. And I'm looking forward to everything that lies ahead as DP continues to evolve.

Walter Debeuf

Quite by chance I became aware of PG when I was surfing and looking for interesting sites. I vaguely knew the name because I had heard of the Project a long time ago. After reading the "History and Philosophy of PG", I immediately became wildly enthusiastic about it. This was

what I had been looking for for years, a meaningful use of my PC, and because I am a fervent lover of good literature, I didn't hesitate to contact the founders of the Project. I made a suggestion that I should work on French and Dutch e-texts. The very same day I received an answer from PG in which they told me they were very pleased with my contribution but that I had to keep in mind that all books must be free of copyright and published before 1923.

This wasn't so great. . . . After I browsed in the "Help And FAQ" of the PG site, I read that I didn't have to worry about all that, because they are willing to do all the clearance!

On my own bookshelf I found an old book of Jules Renard, "Poil de Carotte". It seemed old enough to me, but I couldn't find any copyright notations. So, I mailed to Mr Hart all the information I found on the title page and the verso, and asked him what he thought about it. The next day I received his answer, he wrote: "We still have to prove this edition was pre-1923, so I am forwarding to our authority on such copyright research." This authority is Ms. Dianne Bean who mailed me a few days later very pleasantly that I could start typing, because the copyright issues had been resolved. She asked me to send a "TP&V" (a photocopy of the title page and verso) of the book to Mr. Hart, because they need that for legal reasons.

But something wasn't very clear to me concerning the format I had to use. In the "FAQ" they spoke about "plain vanilla ASCII", something I never had heard about in my life! In "How to Volunteer, PG Volunteers' Board" Mr. Jim Tinsley answered all kind of questions about all kinds of problems people have when they start volunteering. So I did the same and sent him my question. I received an extensive answer about all kind of formats in the "ISO 8859 Alphabet Soup" and he recommended me to use "Codepage 1252" which is very common in Windows. Here are the addresses which Jim sent to me:

"If you are interested in the differences, I recommend the excellent web page

http://czyborra.com/charsets/codepages.html

in the excellent reference site http://czyborra.com"

I chose a French book, first because I had it already on my bookshelf, and secondly because I wanted to perfect my knowledge of the French language and typing seemed the right way to do it. When copying an author's text, you are very close to it. You also have to pay full attention to the spelling of the words. Gradually you come under the spell of the story and you forget that you are typing . . .
Nevertheless, it is hard work, especially when it is not your native language, and therefore you shouldn't try to rush it. At first I started with two or three pages a day, which means that you would need about two months typing for an average book. But good typists can do it more quickly.

I can only applaud the aim of PG, to put books available on the net as much as possible and without cost, for every one in the whole world. I love to co-operate with it.

In the meantime there are thousands and thousands of books in the PG-collection, and that makes it a little difficult to find other examples which are free of copyright, because they must be from before 1923. Since I've got the "PG-bug" it's a challenge for me to find suitable copies, and I look for them high and low. I can buy a few books for a song and I take them home as a trophy, looking forward to the work which is waiting for me . . .

In libraries you can find old publications which you can find nowhere else.

It's amazing how fascinating old books can be and how much you can learn from them. For the moment I'm working on "Pecheur d'Islande" by Pierre Loti, in which I get acquainted with an old tradition of fishermen, very interesting. Without PG I would probably never have read this. There must be still a lot of little treasures in some old and dusty attics, waiting to be born again by the magic touch of a PG-volunteer.

If you do it, no compensation or payment is waiting, but . . . doing something disinterested and unselfish gives you a good feeling.

Bookmarks:

B.1. Project Gutenberg:

Home Page and Search <http://www.gutenberg.net/>
Contact Information <http://www.gutenberg.net/contactinfo.html>
Donations <http://www.gutenberg.net/donation.html>
List of FTP sites <http://www.gutenberg.net/list.html>
Web Browse to texts <http://www.ibiblio.org/pub/docs/books/gutenberg/>

Mailing Lists <http://www.gutenberg.net/subs.html>
Volunteers' Board <http://www.gutenberg.net/vol/wwwboard/>
Copyright Rules <http://www.gutenberg.net/vol/pd.html>
Books In Progress <http://www.dprice48.freeserve.co.uk/GutIP.html>
(The InProg List)

Greek Transliteration <http://www.gutenberg.net/vol/greek.html>

Music <http://www.ibiblio.org/gutenberg/music/music_helpex.html#what-software>

GUTINDEX.ALL <ftp://ibiblio.org/pub/docs/books/gutenberg/GUTINDEX.ALL>
(Complete list of posted eBooks)

B.2. Distributed Proofing Sites:

Charles Franks <http://www.pgdp.net/>
JC Byers <http://www.wollamshram.ca/1001/index.htm>
Dewayne Cushman <http://www.metalbox.net/dcushman/pgroot.htm>

B.3. Other On-Line eBook Pages:

The On-Line Books Page <http://onlinebooks.library.upenn.edu/>
/In Progress List <http://onlinebooks.library.upenn.edu/in-progress.html>
Internet Public Library <http://www.ipl.org/>

B.4. Lists of Suggested Books to Transcribe:

PG Books In Progress <http://www.dprice48.freeserve.co.uk/GutIP.html>
On-Line Requested List <http://onlinebooks.library.upenn.edu/in-progress.html#requests>
Steve Harris' "To-do"s <http://www.steveharris.net/PGList.htm>

B.5. Finding Paper Books On-Line:

Advanced Book Exchange <http://www.abebooks.com>
Alibris <http://www.alibris.com>
Trussel BookSearch <http://www.trussel.com/f_books.htm>
Library of Congress Catalog <http://catalog.loc.gov>

B.6. Character Sets

Overviews <http://czyborra.com>
<http://www.cs.tut.fi/~jkorpela/chars/index.html>
ISO-8859 <http://czyborra.com/charsets/iso8859.html>
Microsoft & Other Codepages <http://czyborra.com/charsets/codepages.html>
Unicode <http://www.unicode.org>

*** END OF THE PROJECT GUTENBERG EBOOK THE PROJECT GUTENBERG FAQ 2002 ***

This file should be named pgf2002.txt or pgf2002.zip

Project Gutenberg eBooks are often created from several printed
editions, all of which are confirmed as Public Domain in the US
unless a copyright notice is included. Thus, we usually do not

keep eBooks in compliance with any particular paper edition.

We are now trying to release all our eBooks one year in advance
of the official release dates, leaving time for better editing.
Please be encouraged to tell us about any error or corrections,
even years after the official publication date.

Please note neither this listing nor its contents are final til
midnight of the last day of the month of any such announcement.
The official release date of all Project Gutenberg eBooks is at
Midnight, Central Time, of the last day of the stated month. A
preliminary version may often be posted for suggestion, comment
and editing by those who wish to do so.

Most people start at our Web sites at:
http://gutenberg.net or
http://promo.net/pg

These Web sites include award-winning information about Project
Gutenberg, including how to donate, how to help produce our new
eBooks, and how to subscribe to our email newsletter (free!).

Those of you who want to download any eBook before announcement
can get to them as follows, and just download by date. This is
also a good way to get them instantly upon announcement, as the
indexes our cataloguers produce obviously take a while after an
announcement goes out in the Project Gutenberg Newsletter.

http://www.ibiblio.org/gutenberg/etext03 or
ftp://ftp.ibiblio.org/pub/docs/books/gutenberg/etext03

Or /etext02, 01, 00, 99, 98, 97, 96, 95, 94, 93, 92, 92, 91 or 90

Just search by the first five letters of the filename you want,
as it appears in our Newsletters.

Information about Project Gutenberg (one page)

We produce about two million dollars for each hour we work. The
time it takes us, a rather conservative estimate, is fifty hours
to get any eBook selected, entered, proofread, edited, copyright
searched and analyzed, the copyright letters written, etc. Our
projected audience is one hundred million readers. If the value
per text is nominally estimated at one dollar then we produce $2
million dollars per hour in 2002 as we release over 100 new text
files per month: 1240 more eBooks in 2001 for a total of 4000+
We are already on our way to trying for 2000 more eBooks in 2002
If they reach just 1-2% of the world's population then the total
will reach over half a trillion eBooks given away by year's end.

The Goal of Project Gutenberg is to Give Away 1 Trillion eBooks!

This is ten thousand titles each to one hundred million readers,
which is only about 4% of the present number of computer users.

Here is the briefest record of our progress (* means estimated):

eBooks Year Month

1	1971 July
10	1991 January
100	1994 January
1000	1997 August
1500	1998 October
2000	1999 December
2500	2000 December
3000	2001 November
4000	2001 October/November
6000	2002 December*
9000	2003 November*
10000	2004 January*

The Project Gutenberg Literary Archive Foundation has been created
to secure a future for Project Gutenberg into the next millennium.

We need your donations more than ever!

As of February, 2002, contributions are being solicited from people
and organizations in: Alabama, Alaska, Arkansas, Connecticut,
Delaware, District of Columbia, Florida, Georgia, Hawaii, Illinois,
Indiana, Iowa, Kansas, Kentucky, Louisiana, Maine, Massachusetts,
Michigan, Mississippi, Missouri, Montana, Nebraska, Nevada, New
Hampshire, New Jersey, New Mexico, New York, North Carolina, Ohio,
Oklahoma, Oregon, Pennsylvania, Rhode Island, South Carolina, South
Dakota, Tennessee, Texas, Utah, Vermont, Virginia, Washington, West
Virginia, Wisconsin, and Wyoming.

We have filed in all 50 states now, but these are the only ones
that have responded.

As the requirements for other states are met, additions to this list
will be made and fund raising will begin in the additional states.
Please feel free to ask to check the status of your state.

In answer to various questions we have received on this:

We are constantly working on finishing the paperwork to legally
request donations in all 50 states. If your state is not listed and
you would like to know if we have added it since the list you have,
just ask.

While we cannot solicit donations from people in states where we are
not yet registered, we know of no prohibition against accepting
donations from donors in these states who approach us with an offer to

donate.

International donations are accepted, but we don't know ANYTHING about how to make them tax-deductible, or even if they CAN be made deductible, and don't have the staff to handle it even if there are ways.

Donations by check or money order may be sent to:

Project Gutenberg Literary Archive Foundation
PMB 113
1739 University Ave.
Oxford, MS 38655-4109

Contact us if you want to arrange for a wire transfer or payment method other than by check or money order.

The Project Gutenberg Literary Archive Foundation has been approved by the US Internal Revenue Service as a 501(c)(3) organization with EIN [Employee Identification Number] 64-622154. Donations are tax-deductible to the maximum extent permitted by law. As fund-raising requirements for other states are met, additions to this list will be made and fund-raising will begin in the additional states.

We need your donations more than ever!

You can get up to date donation information online at:

http://www.gutenberg.net/donation.html

***

If you can't reach Project Gutenberg,
you can always email directly to:

Michael S. Hart <hart@pobox.com>

Prof. Hart will answer or forward your message.

We would prefer to send you information by email.

The Legal Small Print

(Three Pages)

START**THE SMALL PRINT!**FOR PUBLIC DOMAIN EBOOKS**START
Why is this "Small Print!" statement here? You know: lawyers.
They tell us you might sue us if there is something wrong with
your copy of this eBook, even if you got it for free from
someone other than us, and even if what's wrong is not our

OR

[*] You provide, or agree to also provide on request at no additional cost, fee or expense, a copy of the eBook in its original plain ASCII form (or in EBCDIC or other equivalent proprietary form).

[2] Honor the eBook refund and replacement provisions of this "Small Print!" statement.

[3] Pay a trademark license fee to the Foundation of 20% of the gross profits you derive calculated using the method you already use to calculate your applicable taxes. If you don't derive profits, no royalty is due. Royalties are payable to "Project Gutenberg Literary Archive Foundation" the 60 days following each date you prepare (or were legally required to prepare) your annual (or equivalent periodic) tax return. Please contact us beforehand to let us know your plans and to work out the details.

WHAT IF YOU *WANT* TO SEND MONEY EVEN IF YOU DON'T HAVE TO?
Project Gutenberg is dedicated to increasing the number of public domain and licensed works that can be freely distributed in machine readable form.

The Project gratefully accepts contributions of money, time, public domain materials, or royalty free copyright licenses. Money should be paid to the:
"Project Gutenberg Literary Archive Foundation."

If you are interested in contributing scanning equipment or software or other items, please contact Michael Hart at:
hart@pobox.com

*END THE SMALL PRINT! FOR PUBLIC DOMAIN EBOOKS*Ver.02/11/02*END*

ss permission.]

*END THE SMALL PRINT! FOR PUBLIC DOMAIN EBOOKS*Ver.02/11/02*END*

er this listing nor its contents are final til

midnight of the last day of the month of any such announcement.

The official release date of all Project Gutenberg eBooks is at

Midnight, Central Time, of the last day of the stated month. A

preliminary version may often be posted for suggestion, comment

and editing by those who wish to do so.

Most people start at our Web sites at:

http://gutenberg.net or

http://promo.net/pg

These Web sites include award-winning information about Project

Gutenberg, including how to donate, how to help produce our new

eBooks, and how to subscribe to our email newsletter (free!).

Those of you who want to download any eBook before announcement

can get to them as follows, and just download by date. This is

also a good way to get them instantly upon announcement, as the

indexes our cataloguers produce obviously take a while after an

announcement goes out in the Project Gutenberg Newsletter.

http://www.ibiblio.org/gutenberg/etext03 or

ftp://ftp.ibiblio.org/pub/docs/books/gutenberg/etext03

Or /etext02, 01, 00, 99, 98, 97, 96, 95, 94, 93, 92, 92, 91 or 90

Just search by the first five letters of the filename you want,

as it appears in our Newsletters.

Information about Project Gutenberg (one page)

We produce about two million dollars for each hour we work. The time it takes us, a rather conservative estimate, is fifty hours to get any eBook selected, entered, proofread, edited, copyright searched and analyzed, the copyright letters written, etc. Our projected audience is one hundred million readers. If the value per text is nominally estimated at one dollar then we produce $2 million dollars per hour in 2002 as we release over 100 new text files per month: 1240 more eBooks in 2001 for a total of 4000+
We are already on our way to trying for 2000 more eBooks in 2002
If they reach just 1-2% of the world's population then the total will reach over half a trillion eBooks given away by year's end.

The Goal of Project Gutenberg is to Give Away 1 Trillion eBooks!
This is ten thousand titles each to one hundred million readers,
which is only about 4% of the present number of computer users.

Here is the briefest record of our progress (* means estimated):

eBooks Year Month

1 1971 July

10 1991 January

100 1994 January

1000 1997 August

1500 1998 October

2000 1999 December

2500 2000 December

3000 2001 November

4000 2001 October/November

6000 2002 December*

9000 2003 November*

10000 2004 January*

The Project Gutenberg Literary Archive Foundation has been created

to secure a future for Project Gutenberg into the next millennium.

We need your donations more than ever!

As of February, 2002, contributions are being solicited from people

and organizations in: Alabama, Alaska, Arkansas, Connecticut,

Delaware, District of Columbia, Florida, Georgia, Hawaii, Illinois,

Indiana, Iowa, Kansas, Kentucky, Louisiana, Maine, Massachusetts,

Michigan, Mississippi, Missouri, Montana, Nebraska, Nevada, New

Hampshire, New Jersey, New Mexico, New York, North Carolina, Ohio,

Oklahoma, Oregon, Pennsylvania, Rhode Island, South Carolina, South

Dakota, Tennessee, Texas, Utah, Vermont, Virginia, Washington, West Virginia, Wisconsin, and Wyoming.

We have filed in all 50 states now, but these are the only ones that have responded.

As the requirements for other states are met, additions to this list will be made and fund raising will begin in the additional states. Please feel free to ask to check the status of your state.

In answer to various questions we have received on this:

We are constantly working on finishing the paperwork to legally request donations in all 50 states. If your state is not listed and you would like to know if we have added it since the list you have, just ask.

While we cannot solicit donations from people in states where we are not yet registered, we know of no prohibition against accepting donations from donors in these states who approach us with an offer to donate.

International donations are accepted, but we don't know ANYTHING about how to make them tax-deductible, or even if they CAN be made deductible, and don't have the staff to handle it even if there are ways.

Donations by check or money order may be sent to:

Project Gutenberg Literary Archive Foundation

PMB 113

1739 University Ave.

Oxford, MS 38655-4109

Contact us if you want to arrange for a wire transfer or payment

method other than by check or money order.

The Project Gutenberg Literary Archive Foundation has been approved by

the US Internal Revenue Service as a 501(c)(3) organization with EIN

[Employee Identification Number] 64-622154. Donations are

tax-deductible to the maximum extent permitted by law. As fund-raising

requirements for other states are met, additions to this list will be

made and fund-raising will begin in the additional states.

We need your donations more than ever!

You can get up to date donation information online at:

http://www.gutenberg.net/donation.html

***

If you can’t reach Project Gutenberg,

you can always email directly to:

Michael S. Hart <hart@pobox.com>

Prof. Hart will answer or forward your message.

We would prefer to send you information by email.

The Legal Small Print

(Three Pages)

START**THE SMALL PRINT!**FOR PUBLIC DOMAIN EBOOKS**START

Why is this "Small Print!" statement here? You know: lawyers.

They tell us you might sue us if there is something wrong with

your copy of this eBook, even if you got it for free from

someone other than us, and even if what’s wrong is not our

fault. So, among other things, this "Small Print!" statement

disclaims most of our liability to you. It also tells you how

you may distribute copies of this eBook if you want to.

BEFORE! YOU USE OR READ THIS EBOOK

By using or reading any part of this PROJECT GUTENBERG-tm

eBook, you indicate that you understand, agree to and accept

intellectual property infringement, a defective or damaged

disk or other eBook medium, a computer virus, or computer

codes that damage or cannot be read by your equipment.

LIMITED WARRANTY; DISCLAIMER OF DAMAGES

But for the "Right of Replacement or Refund" described below,

[1] Michael Hart and the Foundation (and any other party you may

receive this eBook from as a PROJECT GUTENBERG-tm eBook) disclaims

all liability to you for damages, costs and expenses, including

legal fees, and [2] YOU HAVE NO REMEDIES FOR NEGLIGENCE OR

UNDER STRICT LIABILITY, OR FOR BREACH OF WARRANTY OR CONTRACT,

INCLUDING BUT NOT LIMITED TO INDIRECT, CONSEQUENTIAL, PUNITIVE

OR INCIDENTAL DAMAGES, EVEN IF YOU GIVE NOTICE OF THE

POSSIBILITY OF SUCH DAMAGES.

If you discover a Defect in this eBook within 90 days of

receiving it, you can receive a refund of the money (if any)

you paid for it by sending an explanatory note within that

time to the person you received it from. If you received it

on a physical medium, you must return it with your note, and

such person may choose to alternatively give you a replacement

copy. If you received it electronically, such person may

choose to alternatively give you a second opportunity to

receive it electronically.

THIS EBOOK IS OTHERWISE PROVIDED TO YOU "AS-IS". NO OTHER

WARRANTIES OF ANY KIND, EXPRESS OR IMPLIED, ARE MADE TO YOU AS

TO THE EBOOK OR ANY MEDIUM IT MAY BE ON, INCLUDING BUT NOT

LIMITED TO WARRANTIES O