



A Brief History of Project Gutenberg

by

Michael S. Hart

A Brief History of Project Gutenberg

by

Michael S. Hart
(Founder of Project Gutenberg)

Licensed under the Creative Commons Attribution-NonCommercial 2.0 License
(<http://creativecommons.org/licenses/by-nc/2.0/>)

Foreword

Michael S. Hart and several other members the eBook-Community group were kind enough to send me information about the history of ebooks and other information to help me with a discussion I was preparing for the 2005 Read an Ebook Week.

Not only did Michael send me valuable information, he also sent me three emails containing a history of Project Gutenberg.

In keeping with the occasion, I asked if I could turn the emails into an ebook that would be distributed free of charge in celebration of this special week for ebooks.

Michael agreed, and I put together a cover and inserted the text. The ebook went through a few small modifications, but the understanding was that this would be just the beginning of the book. There was a fourth part to be written and other work to be done on the first three parts.

This seems a natural thing, considering that the history of ebooks is just beginning to unfold. I look forward to seeing what this brief history will look like for the 2006 Read an eBook Week.

Biff Mitchell

2005 Spokesperson for Read an eBook Week

March 7, 2005

THE HISTORY OF PROJECT GUTENBERG

It seems I have been remiss in keeping everyone up to date on the history of Project Gutenberg over the past few years, so I am taking this opportunity to write A Brief History of Project Gutenberg in several segments that will, hopefully, make amends for this lack on my part.

[The author's more personal commentaries are in brackets.]

[Please note that corrections are still being made as this is only the second draft.]

PART ONE: THE FIRST 10 YEARS

1971-1980

In terms of actual page production, many people dismiss the opening decade of Project Gutenberg nearly completely. Now in terms of space, the published eTexts, as we called those at the time, will all fit on a modern floppy disk.

Because of stringent storage allocations our eText files on the million-dollar mainframe were just barely allowed. The struggle to put even these small files online was enormous, as it was a totally revolutionary idea to put up a file for a non-predetermined time period. This idea of something an entire future could download had never been brought up, thus, it was very hard to get permission to post even a file as small as the Declaration of Independence, because it was going to take up permanent space on the computer. Files on the following list were perhaps the first inkling of a kind of permanence the early Internet pioneers did not consider:

Dec 1979 Abraham Lincoln's First Inaugural Address

Dec 1978 Abraham Lincoln's Second Inaugural Address

Dec 1977 The Mayflower Compact

Dec 1976 Give Me Liberty Or Give Me Death, Patrick Henry

Dec 1975 The United States Constitution

Nov 1974 Gettysburg Address, Abraham Lincoln

Nov 1973 John F. Kennedy's Inaugural Address

Dec 1972 The United States Bill of Rights

Dec 1971 The United States Declaration of Independence

These first nine files were collected into all7011.txt and all7011.zip for easy redistribution in upper and lower case in later years. The original files were all upper case, as there was no lower case on those early machines we were using.

You may see that we skipped two years between the US Bill of Rights and the US Constitution; we were originally going to try to include the complete Constitution in just a year after the Declaration, but we were told that would take too much space, and we were given just enough space for the Bill of Rights. The next year we asked again, but room was still very scarce, so we asked again the next year and the year after. By then I was able to make a convincing argument that waiting any longer might delay it so long that people wouldn't have access to it long enough before Bicentennial year of 1776. We were finally given room at the end of 1975.

This may not sound very exciting to you from 30 years later, but it was very exciting to us, being able to put these files online for a whole country to use during the United States Bicentennial.

We finished out the decade with more of those "Freedom Celebration" documents, as they were called, which were placed on the walls of a of schools, malls, and other places.

During this period, the greatest struggle was just to talk operators, even those that were very good friends, into giving us enough space to store anything but the smallest files. It was one thing to have \$100,000,000 in "computer money" that could be used to run programs and send emails, but it was quite another thing to be granted space to store files that people from around the country could download.

Here's just one early example:

When I completed the Declaration of Independence, I wanted to email it to everyone on the Net [DARPA Net, as we called it], but I found, to my great surprise, that if I had done this, even with such small files as the Declaration of Independence [5K], that it would create a complete network crash, since most of our wires were 113 baud, or 11 characters per second.

Luckily, I asked for help in sending it, and avoided becoming quite well known as the first person to bring the Net to its knees; and a "Morris Worm" would have only been an asterisk, and so would I!

In the end, we simply posted a message to what later became comp.gen so people could get the file on request. My recollection is that six people downloaded it, other than the other four on our site, so the greatest penetration would have been about 10%, which sounds big by today's standards. But I had been hoping for more.

A word about the computer operators of the day: we used to joke in many ways that the computer operators were the current priesthood--you handed in your offering through the stainless steel window, and prayed that they would be worthy enough for the computer to run it. If you understand this, then perhaps you can also understand how it was the computer operators had so much power. Not only should they be considered as the entire force of computer security of that day, but they could also save you hours, if not days of time by telling you just where and/or why your program wasn't running. I was quite seriously lucky that my brother's best friend was the operator from midnight to 8AM, when most of the free computer time was available, and that he gave me the account I used to start Project Gutenberg--and as lucky that my best friend became the 8AM-5PM operator.

I should add that even at such an early date, I had help from those anonymous contributors who so often appear when you need them. In this case, I never was able to find out who typed in the first U.S. Constitution versions. I asked and asked, and even though there weren't that many people, I never could find or thank the one who did it. That version was a print version in what served as a sort of markup of the day, so all I had to do was take out all the markup, backspace/underscores etc. to create a version that looked good onscreen. If anyone knows, it would still be nice to find out today, and send our thanks. For now, I would just like to include a general thanks to all the volunteers who have helped Project Gutenberg over more than a third of a century.

That's the story of the first decade of Project Gutenberg. I should add here that even though the Apple II was out, I had none of the kind of money it would have taken to buy one, so my computer ownership starts not in this segment, but in the next one.

Postscript:

For those interested in counting ye olde Project Gutenberg eBooks--please note that there is no growth curve for this period; a growth graph would simply be a

straight line, 1 title = 1 year, so it is a trivial point to say that at this growth rate it would take ~15,000 years to do ~15,000 titles, and that I would have been dead long before we ever got to eBook #100.

Hence, we do not talk about doubling rates for this period since the years required doubled at the same rate as the index entries did.

Nevertheless, you will, from time to time, see people manipulate an army of statistics in such a way as to include these in patterns of growth, even though it is common knowledge that the earliest growth figures of any such pattern are quite linear. Just look at a curve of the population of the earth for a perfect example. Such curves, if studied in detail, yield a wealth of such growth information.

PART TWO: THE SECOND 10 YEARS

1981-1990

Of course, the second 10 years of Project Gutenberg started with the bang of the introduction of the IBM PCs, something not affordable to me, so I started out the decade with some CP/M machines from NEC and OKI that I combined into a small network in my cheapy downtown loft. I managed to borrow an Apple II now and then, but never really got into their O/S, and didn't really get into CP/M all that much. I was a hot target for MS-DOS, and finally built an IBM out of parts of others that were replaced, and a few that I finally bought. In the end, even at junk prices, the total cost of my first IBM, complete with a 5 meg Seagate St-506 was \$3300.

[CP/M = Control Program for Microprocessors]

[MS-DOS = MicroSoft Disk Operating System]

[By the way, the ST-506 drive was from an Apple II, but was transferable to an IBM running DOS, before Apple became the more proprietary company it is today.]

If I had known how much it was going to cost in the end, it is possible I would never have started building it. It was great, however, and I switched from CP/M to MS-DOS in a day or so. It was much easier than CP/M and I loved the idea of batch files, so I was a quick convert. In addition, the idea of BBSs in addition to the Internet was also a popular event of the 1980's and I became the system operator of the local Champaign County Computer Club BBS [the only BBS in a 100 mile radius. We got callers from all over the U.S. and even from other countries. I should probably add that BBSs were first started in 1978 just 100 miles north of here, in Hinsdale, IL, by Ward Christensen and Randy Seuss, who were just a couple of years older than I was].

The fact that I now *owned* my own computers meant that I was no longer constrained by the limitations of the "operators" I mentioned so much in the previous segments of PG history, and I started out on my quest to put the Bible and Complete Works of Shakespeare online.

By the late 80s, I had a computer always set up on a dining room table that my friends could use to type in Shakespeare, only it turned out that U.S. copyright

law had changed in 1976 with no publicity whatsoever [and it turned out the edition I was using would not be in the public domain] so a massive number of hours of labor turned out to be in vain.

[Little did I know at the time that another extension of an already doubly extended copyright law was already planned!]

However, the advent of the BBS world attracted a large number of churches around the world to this new medium [and thus it was not long before they were on an entirely new track of putting the Bible online with persons from all over the country willing to contribute to projects such as my desire to prove something as large as the Bible could be transmitted via the Internet.] Thus, anonymous work on various eBook projects continued to be the norm, as most of the portions of the Bible and "The History of Democracy" items I found from other sources were unsigned. [I say most, but the truth is that I can't recall anyone ever doing one of these, or even a portion, and signing a name to it.]

[In fact, some very intensive questioning of the mainframe operators of the day never did turn up any names.]

Indeed, this has seemed to be the norm for most development in the major portions of Project Gutenberg's history. Thus most of the first 5,000 Project Gutenberg eBooks were be in the category of having been mainly created by "anonymous."

As far as I know Project Gutenberg was the first place from which you could download the entire Bible with one command. Of course we started with the New Testament, as it was much shorter than the Old Testament, but the entire Bible was on our servers in test versions by the end of the 1980s.

Here is how the 1980s were listed at the time:

1980-1990 Various Editions of Shakespeare and The Bible

[The Shakespeare was never released due to copyright problems.

Hence the changed file names and number from older index entries.]

Aug 1989 The Bible, Both Testaments, King James Version

[kjb10xxx.xxx] 10

Dec 1984 The Bible, The New Testament, King James Version
[biblexxx.xxx] xx

The Bible and Shakespeare represented the entire effort for the 1980s, and the Bible alone is about 1,000 times larger than our first file, the U.S. Declaration of Independence; and so is the Complete Shakespeare.

Notes of the period:

By 1992 we had added some Biblical Apocrypha, and had plans in hand for using a copyrighted edition of Shakespeare with permission from the World Library, whose eText CD we should mention that we were glad to announce at the 1990 MidWinter American Library Association meeting in Chicago. January 6 would go down as the release date for the first eText/eBook CD on the market.

Project Gutenberg has always been more than glad to help in any way we can "to encourage the creation and distribution" of all forms of eTexts, eBooks, etc., whether commercial, academic, or non-profit in orientation. We never required that anyone receiving our assistance would have to release their work via our sites.

However, we were adamant about not counting books in ways a few others did, such as counting each one of AEsop's Fables as an individual title. In response to this sort of thing, we waited many years before we finally released Shakespeare and the Bible as individual plays and books, as a method of refusing to "pad our bibliography," so to speak.

Even so, you will see in the next section of this history a remarkable growth curve, even without such "padding."

If anything, the following postscript from Part One of this history is even more appropriate for this Part Two, as only one actual title was counted for the entire decade.

Postscript 1971-1990:

[1970s]

For those interested in counting ye olde Project Gutenberg eBooks--please note that there is no growth curve for this period; a growth graph would simply be a straight line, 1 title = 1 year, so it is a trivial point to say that at this growth rate it would take ~15,000 years to do ~15,000 titles, and that I would have been dead so long before we ever got to eBook #100 that no one would have remembered.

[1980s]

[Obviously, in terms of the number of titles, this was even more of an extreme situation than the 1970s, and thus should not be used in any growth curve projections, as the growth rate was negative.]

Hence we do not talk about doubling rates for this period since the years required doubled at the same rate as the index entries did or even less, as in the 1980s.

Nevertheless, you will, from time to time, see people manipulate an army of statistics in such a way as to include these in patterns of growth, even though it is common knowledge that the earliest growth figures of any such pattern are quite linear. Just look at a curve of the population of the earth for a perfect example. Such curves, if studied in detail, yield a wealth of such information.

PART THREE: THE THIRD DECADE

1991-1999

The Alice in Wonderland stories changed Project Gutenberg's image forever, from something dedicated only to the classic "History of Democracy" series and the equally classic works of Shakespeare and the Bible to something of a more general interest to a wider range of readers. The Alice stories of this new focus were something people of all ages will read, again and again: parents showed it to their children, and children showed it to their parents and grandparents.

Alice was a major turning point for Project Gutenberg.

1991 was a banner year for Project Gutenberg: the following 12 eBooks were released one per month; the first time I had proposed a regular schedule, and also the first time I said we should attempt to keep up with Moore's Law until we were at the level of 10,000 free eBooks. [Some say we should be using 1993 as the base year for Moore's Law, but I resisted the temptation to rewrite history.]

Dec 1991 Roget's Thesaurus

[rogetxxa.xxx] 22

Dec 1991 Roget's Thesaurus

[rogetxxx.xxx] 22

Nov 1991 AEsop's Fables

[aesopxxx.xxx] 21

Oct 1991 Paradise Lost, John Milton [Milton #1]

[plbossxx.xxx] 20

Sep 1991 The Song of Hiawatha

[hisongxx.xxx] 19

Aug 1991 The Federalist Papers

[federxxx.xxx] 18

Jul 1991 The Book of Mormon

[mormonxx.xxx] 17

Jun 1991 Peter Pan [for US only]**, James M. Barrie

[peterxxx.xxx] 16

May 1991 Moby Dick [From OBI]*, Herman Melville

[mobyxxxx.xxx] 15

Apr 1991 The 1990 CIA World Factbook, [CIA Factbook #0]
[worldxxx.xxx] 14
Mar 1991 The Hunting of the Snark, Lewis
Carroll[Carroll#3][snarkxxx.xxx] 13
Feb 1991 Through the Looking Glass, Lewis
Carroll[Carroll2][lglassxx.xxx] 12
Jan 1991 Alice in Wonderland, Lewis Carroll [Carroll #1]
[alicexxx.xxx] 11
[These two Roget's are not exactly the same]
*Moby Dick is missing Chapter 72

We established several trends in 1991, and continued others:

1. Our first regular production schedule.
2. We reformatted the CIA World Factbook to facilitate jumps to country headings.
3. We started the "GUTINDEX.ALL" format.
4. We accepted Moby Dick from the Online Book Initiative (not to be confused with the Online Book Pages of later fame).

This eBook turned out to be missing significant page numbers and needed help in other ways, but we wanted to signify the fact that other eBook producers could donate materials and know they would be kept online, and eventually brought up to higher standards, as opposed to donated eBooks already being completed up to our standards on first release.

5. We had two versions of Roget's Thesaurus, both from the same source and person, so we used only one number to index both. [An additional statement that we were not going to "pad our bibliography," as mentioned in Part Two re: Shakespeare and the Bible.]

In addition, in 1992 we released Sophocles' Oedipus Trilogy using a single number, even though we indexed each individual play, for ease of searching the index. However, we also did two Dr. Jekyll and Mr. Hyde editions from separate sources, so we used two numbers to index those two. New editions from the

same source didn't get new index numbers; new editions from different sources did. Some of these editions were quite different; some merely an edition of similar content from a different publisher. We always planned on doing several editions of the greatest works, each with different index numbers, to reflect the great editions of those works. The idea of not padding our bibliography extended to very short works we would later include in anthologies, such as "Gift of the Magi" in 1993, which went unnumbered, as did President Clinton's speech at his first inauguration, which aroused so much White House talk that we received two phone calls from Jock Gill, the first "White House Internet Guru," asking us to delete the files [we had three editions out within an hour or two of the speech]. We declined.

Over the next few years we doubled production levels every year:

Year	# For Monthly Goals	Yearly Goals	Actual eBooks Numbered
1991	1 eBook Per month	12 Per Year	22 Total eBook Numbers
1992	2 eBooks Per month	24 Per Year	47 Total eBook Numbers
1993	4 eBooks Per month	48 Per Year	95 Total eBook Numbers

In 1994 we reached the official date of our 100th eBook [it was actually released three weeks early, as a tribute to one fallen Project Gutenberg volunteer without whom we may never have even been able to get far enough to have been noticed.]

We also announced that we were going to do the 11th Britannica, and reserved one eBook number for each volume at year's end: I must confess that this was mostly just to prove the feasibility of doing such large works, not that we expected to complete the 11th Britannica before 10-20 years had passed. We reserved the spaces 181-200 for this purpose, and still keep a running total of the reserved numbers in our Newsletters. [And subtract these to achieve an honest total.]

Year	# For Monthly Goals	Yearly Goals	Actual eBooks Numbered
1993	4 eBooks Per month	48 Per Year	95 Total eBook Numbers
1994	8 eBooks Per month	96 Per Year	191 Total eBook Numbers [-11 Brit]
1995	16 eBooks Per month	192 Per Year	383 Total eBook Numbers [- 8 Brit]

[The Britannica numbers were split between the end of 1994 and the beginning of 1995, and we posted the first portion in 1995]

[Subtract the numbers in [brackets] for an accurate count.]

Year # For Monthly Goals Yearly Goals Actual eBooks Numbered

1996 32 eBooks Per month 384 Per Year 768 Total eBook Numbers [- 19 Brit]

[We did one more than scheduled, as we released two Wizard of Oz eBooks in February, hence the change from odd to even numbers at the end of a year.]

Year # For Monthly Goals Yearly Goals Actual eBooks Numbered

1997 32 eBooks Per month 384 Per Year 1152 Total eBook Numbers [-19 Brit]

[Due to a lack of resources, we did not double our goal this year]

1998 40 eBooks Per month 480 Per Year 1580 Total eBook Numbers [-19 Brit]

[Due to a lack of resources, we did not double our goal this year]

[However, we did 28 more than the 480 that were scheduled = 428]

1999 40 eBooks Per month 480 Per Year 2018 Total eBook Numbers [-19 Brit]

[Again we scheduled 40 per month, this time adding 38 more = 438, in our effort to pass #2,000 by December 10, 1999, as we had with #100, 6 years earlier.]

The postscripts used in Part One and Part Two are important in increasingly specific aspects concerning this Part Three you have just read, as most of the various manipulations of the statistical predictions are made using two major points of Project Gutenberg history from this period, eBooks #10 & #100 & 1,000, which are discussed below, along with #1.

Postscript:

For those interested in counting ye olde Project Gutenberg eBooks--please note that there are several historical points that have been used from the 1990s to establish the baseline for Moore's Law.

However, these have not been well thought out, and it takes only an instant of looking at the various predictions made by each historic point that has been misused to understand why I say "mis"-used.

As you can see by comparing the table below with the actual history of Project Gutenberg as presented in this series, the idea of eBook #1 being used to start a Moore's Law doubling rate is silly in the extreme for several reasons:

1. There was no attempt to double the rate of production until 20 years later, so any such predictions would be the result of including 20 years of statistical doubling, when there was no doubling of the schedule in reality.
2. Since using the 18 month Moore's Law doubling rate is thus the same over 30 years as 20 doublings, it is thus trivial to see that such a prediction would be over a million eBooks, when all predictions being made were for ten thousand in a similar time frame.
3. In fact, the rate of production as indicated by the eBooks and their numbering would actually indicate a DECLINE in a statistical prediction based on the years 1971-1990.

The candidates for the baseline most often used and misused are:

	Start	Finish	Total	Total			
#####	Year	Year	Years	Doubles	$x2^y$	=	Grand Total in Year
#1	1971	2001	30	20	$1*2^{20}$	=	1,048,576 in 2001
#10	1990	2005	15	10	$10*2^{10}$	=	10,240 in 2005
#100	1994	2003	9	6	$100*2^9$	=	51,200 in 2003

A simple look at the actual growth periods immediately after the stated base year of 1991 equally thoroughly indicates why anyone using eBook #100 as a baseline would be equally silly, and would get equally silly predictions.

In each of the years, 1991, 1992, 1993, 1994, 1995, a production schedule was in effect that would, and did, approximately double the total number of Project Gutenberg eBooks in the given year.

####	####	Growth	Growth	Moore's	Moore	Actual
------	------	--------	--------	---------	-------	--------

Year	Start	End	Rate	in %	Law %	Total	Total
1991	10	22	22/10	120%	58%	158%	220%
1992	22	47	47/22	114%	58%	250%	470%
1993	47	95	95/47	102%	58%	394%	950%

Thus you see that the numbers were more than doubling every year: instead of growing by 100% every 18 months, they grew an average of 115+% over that three year period.

Moore's Law would have predicted 40 eBooks after 4.5 years [three doubling periods] and instead there were 100 eBooks after 3 years [two doubling periods].

If you want to go for total accuracy, rather than official dates, you would have to include #100 in the count for 1993, as follows:

Year	Start	End	Rate	Growth in %	Moore's Law %	Moore's Total %	Actual Total%
1991	10	22	22/10	120%	58%	158%	220%
1992	22	47	47/22	114%	58%	250%	470%
1993	47	100	100/47	113%	58%	394%	1000%

Postscript:

For those interested in counting ye olde Project Gutenberg eBooks--please note that there is no growth curve for this period; a growth graph would simply be a straight line, 1 title = 1 year, so it is a trivial point to say that at this growth rate it would take ~15,000 years to do ~15,000 titles, and that I would have been dead so long before we ever got to eBook #100 that no one would have remembered.

[Obviously, in terms of the number of titles, this was even more of an extreme situation than the 1970s, and thus should not be used in any growth curve projections, as the growth rate was negative.]

Hence we do not talk about doubling rates for this period since the years required doubled at the same rate as the index entries did or even less, as in the 1980s.

Nevertheless, you will, from time to time, see people manipulate an army of statistics in such a way as to include these in patterns of growth, even though it is common knowledge that the earliest growth figures of any such pattern are quite linear. Just look at a curve of the population of the earth for a perfect example. Such curves, if studied in detail, yield a wealth of such information.

Author's Note:

Part Three is still not complete and Part Four of this brief history is in progress.

Author's Bio:

I've lived a VERY full and interesting life...

Thanks!!!

So Nice To Hear From You!

Michael

hart@pglaf.org

hart@pobox.com

hart@login.ibiblio.org