How Much Information?

About the Project

Senior Researchers: Peter Lyman and Hal R. Varian
Research Assistants: James Dunn, Aleksey Strygin, Kirsten Swearingen

This study is an attempt to measure how much information is produced in the world each year. We look at several media and estimate yearly production, accumulated stock, rates of growth, and other variables of interest.

If you want to understand what we've done, we offer different recommendations, depending on the degree to which you suffer from *information overload:*

Heavy information overload: *the world's total yearly production of print, film, optical, and magnetic content would require roughly 1.5 billion gigabytes of storage. This is the equivalent of 250 megabytes per person for each man, woman, and child on earth.*

Moderate information overload: read the *Sound Bytes* and look at the *Charts* illustrating our findings.

Normal information overload: read the *Executive Summary.*

Information deprived: read the detailed reports by clicking on the contents to your left. Or download the entire Web site as a PDF file. (It is about 100 pages long.)

This study was produced by faculty and students at the School of Information Management and Systems at the University of California at Berkeley. We gratefully acknowledge financial support from EMC. We have put "[???]" in the text where we had to make "questionable" assumptions. If you have suggestions, corrections, or comments, please send email to how-much-info@sims.berkeley.edu. We view this as a "living document" and intend to update it based on such contributions.

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Executive Summary

Abstract

The world produces between 1 and 2 exabytes of unique information per year, which is roughly 250 megabytes for every man, woman, and child on earth. An exabyte is a billion gigabytes, or 10^{18} bytes. Printed documents of all kinds comprise only .003% of the total. Magnetic storage is by far the largest medium for storing information and is the most rapidly growing, with shipped hard drive capacity doubling every year. Magnetic storage is rapidly becoming the universal medium for information storage.

Introduction

The cost of magnetic storage is dropping rapidly; as of Fall 2000 a gigabyte of storage costs less than $10 and it is predicted that this cost will drop to $1 by 2005. Soon it will be technologically possible for an average person to access virtually all recorded information. The natural question then becomes: how much information is there to store? If we wanted to store "everything," how much storage would it take?

We have conducted a study to answer this question. In particular, we have estimated yearly US and world production of originals and copies for the most common forms of information media. We have also attempted to estimate the cumulated stock of information in various formats. Finally, we have described the magnitudes of some communication flows that are currently not stored but may well be

in the future.

Information produced by medium

Most information is stored in four physical media: paper, film, optical (CDs and DVDs), and magnetic. There are very good data for the worldwide production of each storage medium, and there are reasonably good estimates of how much original content is produced in each of these different formats.

We have identified production of content by media type, translated the volume of original content into a common standard (terabytes), determined how much storage each type takes under certain assumptions about compression, attempted to adjust for duplication of content, and added up to get total estimates.

Table 1 depicts yearly worldwide production of original stored content as of 1999. In general, the upper estimate is based on the raw data, while the lower estimate reflects an attempt to adjust for duplication and compression. We discuss these adjustments below and in the medium-specific documents. Note that the growth rate estimates are very rough. See the "Qualifications" section and Appendix A for further discussion; the details of the calculations are presented in the accompanying documents.

Table 1: Worldwide production of original content, stored digitally using standard compression methods, in terabytes circa 1999.

Storage Medium	Type of Content	Terabytes/Year, Upper Estimate	Terabytes/Year, Lower Estimate	Growth Rate, %
Paper	Books	8	1	2
	Newspapers	25	2	-2
	Periodicals	12	1	2
	Office documents	195	19	2
	Subtotal:	**240**	**23**	**2**
Film	Photographs	410,000	41,000	5
	Cinema	16	16	3
	X-Rays	17,200	17,200	2
	Subtotal:	**427,216**	**58,216**	**4**
Optical	Music CDs	58	6	3
	Data CDs	3	3	2
	DVDs	22	22	100
	Subtotal:	**83**	**31**	**70**
Magnetic	Camcorder Tape	300,000	300,000	5
	PC Disk Drives	766,000	7,660	100
	Departmental Servers	460,000	161,000	100
	Enterprise Servers	167,000	108,550	100
	Subtotal:	**1,693,000**	**577,210**	**55**
TOTAL:		**2,120,539**	**635,480**	**50**

Three striking facts emerge from these estimates. The first is the "paucity of print." Printed material of all kinds makes up less than .003 percent of the total storage of information. This doesn't imply that print is insignificant. Quite the contrary: it simply means that the written word is an extremely efficient way to convey information.

The second striking fact is the ``democratization of data.'' A vast amount of unique information is created and stored by individuals. Original documents created by office workers are more than 80% of all original paper documents, while photographs and X-rays together are 99% of all original film documents. Camcorder tapes are also a significant fraction of total magnetic tape storage of unique content, with digital tapes being used primarily for backup copies of material on magnetic drives.

As for hard drives, roughly 55% of the total are installed in single-user desktop computers. Of course, much of the content on individual user's hard drives is not unique, which accounts for the large difference between the upper and lower bounds for magnetic storage. However, as more and more image data moves onto hard drives, we expect to see the amount of digital content produced by individuals stored on hard drives increase dramatically.

This democratization of data is quite remarkable. A century ago the average person could only create and access a small amount of information. Now, ordinary people not only have access to huge amounts of data, but are also able to create gigabytes of data themselves and, potentially, publish it to the world via the Internet, if they choose to do so.

The third interesting finding is the "dominance of digital" content. Not only is digital information production the largest in total, it is also the most rapidly growing. While unique content on print and film are hardly growing at all, optical and digital magnetic storage shipments are doubling each year. Even today, most textual information is "born digital,'' and within a few years this will be true for images as well. Digital information is inexpensive to copy and distribute, is searchable, and is malleable. Thus the trend towards democratization of data---especially in digital form---is likely to continue.

Qualifications

It goes without saying that the numbers in Table 1 can only be taken as rough estimates. We have had to make various assumptions in order to construct our these figures, and some data sources are contradictory or simply not available. Here we list some of the most serious methodological qualifications, each of which offers interesting challenges for those who would seek to refine these estimates.

Duplication.

It is very difficult to distinguish "copies" from "original" information. A newspaper, for example, is published on paper, often published on the Web as well, and is generally archived on microfilm. In fact, most printed materials are produced and/or archived magnetically. There is also lot of duplication within each medium: many newspapers reproduce stock prices, wire stories, advertisements and so on. Ideally, we would like to measure the storage required for the *unique* content in the newspaper, but it is very hard to measure that number. As indicated above, the duplication issue is particularly serious for digital storage, since little of what is stored on individual hard drives is unique. We've tried to adjust for this the best we can, and documented our assumptions in the detailed treatment of each medium.

Compression.

Unlike print or film, there is no unambiguous way to measure the size of digital information. A 600 dot per inch scanned digital image of text can be compressed to about one hundredth of its original size. A DVD version of a movie can be 1000 times smaller than the original digital image. We've made what we thought were sensible choices with respect to compression, steering a middle course between the high estimate (based on ``reasonable'' compression) and the low estimate (based on highly compressed content). It is worth noting that the fact that digital storage can be compressed to different degrees depending on needs is a

significant advantage for digital over analog storage.

Archival Media.

Should information stored as "backup" be included in the total? This question arises for microfilm, rewritable CD ROMS, and even with print, but digital magnetic tape is the most difficult case. Tape's most common use is to archive material on hard drives and therefore should not count towards the stock of ``original information'' produced each year. Industry rules of thumb suggest that there is about 10 times as much storage on tape as on hard drives. This fraction has been falling as more and more data is stored on arrays of hard drives, which are much more convenient to use. We've omitted most tape storage for this reason. However, we should also note that vast quantities of original scientific data are stored in tape libraries; we describe a few such repositories in the detailed treatment of magnetic storage.

World and US production.

The US produces about 25% of all textual information and about 30% of the photographic information, a significant fraction of the world's total. We don't have good data on magnetic storage, but it seems plausible that the US produces at least half of the content stored on magnetic media. We've used numbers for world production when available, but in some cases have had to extrapolate from US production. Little data is available about information production in the Third World.

Growth rates.

The production of unique content in books, photos, and CDs is barely growing. DVD content is growing rapidly, but that's because it is a new medium and a significant amount of legacy content is being converted. By contrast, shipments of digital magnetic storage are essentially doubling every year.

TV and Radio.

Original TV content produced each year is generally stored on magnetic camcorder tapes, and so is counted in that category of storage media. Much radio content is simply broadcast music, which we have already captured with the CD statistics. See Table 3 for information on how much storage it would take to back up all TV and radio broadcasts, with minimal adjustment for duplication.

Digital Communication

Our project is primarily concerned with content that is stored, either by institutions or by individuals. But there is a lot of material that is communicated, without being systematically stored. Some of this material is born digital, such as email, Usenet, and the Web. Some of it is non-digital, such as telephone calls and letters.

We expect that digital communications will be systematically archived in the near future, and thus will contribute to the demand for storage. Table 2 shows how much storage would be required to archive the major forms of digital communication.

Table 2: Summary of yearly unique computer-mediated information flows.

Content	Terabytes
Email	11,285
Usenet	73

In 2000 the World Wide Web consisted of about 21 terabytes of static HTML pages, and is growing at a rate of 100% per year. Many Web pages are generated on-the-fly from data in databases, so the

total size of the "deep Web" is considerably larger.

Although the social impact of the Web has been phenomenal, about 500 times as much email is being produced per year as the stock of Web pages. It appears that about 610 billion emails are sent per year, compared to 2.1 billion static Web pages. Even the yearly flow of Usenet news is more than 3 times the stock of Web pages. As Odlyzko (2000) puts it, "communication, not content, is the killer app."

Non-digital Communication

We also estimated the storage requirements if one attempted to archive all the non-digital communication flows in the United States. We consider only the US since we didn't have very good data for worldwide communication. The results are shown in Table 3.

Table 3: Summary of yearly non-digital communication flows in the United States 1999.

Content	Terabytes
Radio	788
TV	14,150
Telephone	576,000
Postal	150,000

The striking thing here is the volume of voice telephone traffic, most of which is presumably unique content. Radio and TV, by contrast, have a huge amount of duplication from station to station, since many of the broadcasts are reusing the same content.

Consumption of Information

Though the main focus of our report is on the supply of information, it is interesting to look at data measuring the consumption of information as well. Table 4 depicts hours per year of time spent on various media in US households in 1992 and in 2000. We do not have good data on information use in the workplace.

Table 4: Summary of yearly media use by US households in hours per year, with estimated megabyte equivalent. (Hours from Statistical Abstract of the United States, 1999, Table 920, (projected)).

Item	1992 Hours	2000 Hours	2000 MBytes	% Change
TV	1510	1571	3,142,000	4
Radio	1150	1056	57,800	-8
Recorded Music	233	269	13,450	15
Newspaper	172	154	11	-10
Books	100	96	7	-4
Magazines	85	80	6	-6
Home video	42	55	110,000	30
Video games	19	43	21,500	126

Internet	2	43	9	2,050
Total:	**3,324**	**3,380**	**3,344,783**	**1.7**

The notable features of this table are 1) the hours spent on TV and radio consumption and their consistency over time; 2) the reduction in time spent on printed information; and, 3) the dramatic increase in home video, video games, and Internet usage. However, it is important to note that the latter three categories are still very small in terms of total hours.

It is also noteworthy that total time spent in media access has hardly changed in eight years. Even while information supply is growing dramatically (especially in electronic media) the actual consumption of information is barely changing: a smaller and smaller fraction of what is produced is actually consumed, on average, a trend noted by Pool (1984). Census data indicate that over 40% of the US population has access to the Internet, so this trend is likely to increase.

Individual and Published Information

We remarked above that technological advances have allowed for a "democratization of data:" individuals can now generate a huge amount of information on their own. Table 5 summarizes the yearly production of information by and about individuals.

Table 5: Yearly production of individual information

Item	Amount	Terabytes
Photographs	80 billion images	410,000
Home videos	1.4 billion tapes	300,000
X-rays	2 billion images	17,200
Hard disks	200 million installed drives	13,760
Total:		**740,960**

The production of individual information can be compared to the amount of "published" information in Table 6. Note that the amount of "individual" information is over 600 times larger as the amount of published information.

Although the Web, Usenet, and email include a great deal of individual information, they have been omitted from both of these tables, since it is difficult to know whether to classify this material as "individual" or "public." In the future we expect the distinction between "individual" and "public" to become increasingly blurred.

Table 6: Yearly production of published information

Item	Titles	Terabytes
Books	968,735	8
Newspapers	22,643	25
Journals	40,000	2
Magazines	80,000	10
Newsletters	40,000	.2
Office Documents	7,500,000,000	195
Cinema	4,000	16

Music CDs	90,000	6
Data CDs	1,000	3
DVD-video	5,000	22
Total:		**285**

Conclusion

The world's total production of information amounts to about 250 megabytes for each man, woman, and child on earth. It is clear that we are all drowning in a sea of information. The challenge is to learn to swim in that sea, rather than drown in it. Better understanding and better tools are desperately needed if we are to take full advantage of the ever-increasing supply of information described in this report.

About this Report

Financial support for this study was provided by EMC. We view this report as a "living document" and intend to revise it based on comments, corrections, and suggestions. Please send such materials to how-much-info@sims.berkeley.edu.

About the School of Information Management and Systems

UC Berkeley's School of Information Management and Systems is the first school in the nation to explicitly address the growing need to manage information more effectively.

With respect to education, we are training a new type of professional: "information managers". Our graduates are familiar with the latest and most powerful techniques for locating, organizing, retrieving, manipulating, protecting, and presenting information. They study not only technology, but also the institutional, legal, economic and organizational factors necessary for creating information systems that meet peoples' needs.

With respect to research, we are examining ways to build more effective tools and systems for managing information. This effort is inherently multidisciplinary, involving computer science, information science, social science, cognitive science, and legal studies.

Appendices

- **A. Powers of Ten**
  The Powers of Ten table is helpful in illustrating the relative size of gigabytes, terabytes, petabytes and the like.

- **B. Upper and lower estimates**
  The upper estimate is a reasonably "hard" number; based on published data. The lower estimate is an attempt to adjust for duplication and compression. Here is a quick summary of some of those adjustments.

  - **Paper.**
    There is some duplication with ISBN numbers due to paperback, hardback, different editions, etc. There is duplication with financial papers, ads, and so on in newspapers. We used CPC compression, which captures images; conversion to ASCII eliminates images, but compresses text dramatically.

  - **Film.**
    If we used JPEG compression, rather than PhotoCD, we get a much smaller number

for the storage requirements for images.

- **Music CDs.**
  If we use MP3 compression we get a much smaller number for the storage requirements of audio files.

- **Magnetic.**
  We assume that about 20 percent of magnetic storage is unique.

- **C. Reading the data**
  The left-side navigation and Site Map provide links to summary reports on each medium. The summaries provide links to detailed reports and spreadsheets containing the raw data.

  Within each media type, we have distinguished between originals and copies, and between the yearly flow of production and the accumulated stock. We've also described growth rates and compression issues for each medium.

- **D. Acknowledgements**
  Gray and Shenoy (2000) provides useful information on trends in magnetic storage. Lesk (1997) conducted an earlier study that attempted to estimate the total stock of information. Pool (1984) examined the flow of information in the US circa 1980. See the individual acknowledgements for the names of people who helped us.

Bibliography

- Jim Gray and Prashant Shenoy.
  Rules of thumb in data engineering.
  in *Proceedings of 16th International Conference on Data Engineering, pages 3-12. IEEE, 2000.*
  http://www.research.microsoft.com/~Gray/.

- Michael Lesk.
  How much information is there in the world?
  Technical report, lesk.com, 1997.
  http://www.lesk.com/mlesk/ksg97/ksg.html.

- Andrew Odlyzko.
  Content is not king.
  Technical report, AT&T Labs, 2000.
  http://www.research.att.com/~amo/doc/networks.html.

- Ithiel De Sola Pool, Hiroshi Inose, Nozomu Takasaki, Roger Hurwitz.
  Communications flows : a census in the United States and Japan.
  Elsevier Science, New York, 1984.

- Ithiel De Sola Pool. "Tracking the Flow of Information".
  Science (12 August), 1983, 221:4611, 609-613.

- U.S. Census Bureau.
  Statistical Abstract of the United States, 1999
  Washington, D.C., 1999.
  http://www.census.gov/prod/www/statistical-abstract-us.html

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Print Media - Summary

Originals

Flow

Approximately **240 terabytes** (compressed) of unique data are recorded on printed media worldwide each year, as shown in the following table:

Table 1: World Flow

Media Type (Sources and Year Cited)*	Unique Items per Year	Conversion Factor	Total Terabytes (Annual Worldwide)
Books (UNESCO 1996)	**968,735**	Scanned image (600 dpi): 40 MB/book	39
		Digital compression: 8 MB/book	8
		Plain text: 1 MB/book	1

Newspapers (ISSN 1999)	**22,643**	Scanned image (600 dpi): 5,475 MB/year	124
		Digital compression: 1095 MB/year	25
		Plain text: 110 MB/year	2.5
Scholarly journals (Ulrich's 2000)	**40,000**	Scanned image (600 dpi): 225 MB/year	9
		Digital compression: 45 MB/year	2
		Plain text: 4 MB/year	.2
Mass-market periodicals (Ulrich's 2000)	**80,000**	Scanned image (600 dpi): 650 MB/year	52
		Digital compression: 130 MB/year	10
		Plain text: 13 MB	1
Newsletters (Oxbridge Directory 1997)	**40,000**	Scanned image (600 dpi): 20 MB/item	.8
		Digital compression: 4 MB/item	.2
		Plain text: .4 MB/item	.02
Archivable, original office documents (National Archives 1998)	**7.5 X 10^9 pages**	Scanned image (600 dpi): 130 KB/page	975
		Digital compression: 26 KB/page	195
		Plain text: 2.5 KB/page	19
		Totals:	**Scanned: 1200 TB**
			Compressed: 240 TB
			Text: 24 TB

* Detailed source information listed at end of this document.

Table 2: United States Flow

Media Type (Sources and Year Cited)*	**Unique Items per Year**	**Conversion Factor**	**Total Terabytes** (Annual Worldwide)
Books (US Statistical Abstract 1999)	**64,711**	Scanned image (600 dpi): 40 MB/book	3
		Digital compression: 8 MB/book	.5
		Plain text: 1 MB/book	.05
Newspapers (Newspaper Association of America)	**2,386**	Scanned image (600 dpi): 5,475 MB/year	13
		Digital compression: 1095 MB/year	3
		Plain text: 110 MB/year	.3
Scholarly journals	**10,500**	Scanned image (600 dpi): 225 MB/year	2

(Tenopir and King)		Digital compression: 45 MB/year	.5
		Plain text: 4 MB/year	.04
Mass-market periodicals (Ulrich's 2000)	**20,000**	Scanned image (600 dpi): 650 MB/year	13
		Digital compression: 130 MB/year	2.6
		Plain text: 13 MB	.26
Newsletters (NEPA)	**10,000**	Scanned image (600 dpi): 20 MB/item	.2
		Digital compression: 4 MB/item	.04
		Plain text: .4 MB/item	.004
Archivable, original office documents (National Archives 1998)	**3 X 10^9 pages**	Scanned image (600 dpi): 130 KB/page	390
		Digital compression: 26 KB/page	78
		Plain text: 2.5 KB/page	7.5
		Totals:	**Scanned: 421 TB**
			Compressed: 84 TB
			Text: 8.2 TB
* Detailed source information listed at end of this document.			

Notes on Conversion Assumptions

Books. Estimate 300 pages per book. (*Source: Robert M. Hayes, UCLA, "The Economics of Digital Libraries" www.usp.br/sibi/economics.html*)

Newspapers. Estimate 30 pages per newspaper, then multiply by 365 days per year. (*The page number is low, to reflect the number of small and non-daily newspapers published around the world.*)

Scholarly Journals. Estimate 1,700 pages per periodical per year. (*Source: Donald W. KLing and Carol Tenopir. "Economic Cost Models of Scientific Scholarly Journals," 1998. www.bodley.ox.ac.uk/icsu/kingppr.htm*)

Mass Market Periodicals. Estimate 5,000 pages per periodical per year. (*Source: Robert M. Hayes, UCLA, "The Economics of Digital Libraries" www.usp.br/sibi/economics.html*)

Newsletters. Estimate 150 pages per newsletter per year. (*Source: Oxbridge Directory of Newsletters - 1997*)

Office documents. The estimate above is limited to documents that an organization might retain permanently such as documents comparable to those retained by the National Archives in Washington D.C., which estimates that they retain 2% of US government documents produced each year.

More detail on the conversion factors used for the above estimates appears in the Print Detail Report.

Stock

United States

According to a press release from January 2000, booksinprint.com 2000 includes **3.2 million titles** - about **26 TB** total. This figure is supported by online booksellers such as Amazon.com and Barnes&Noble.com who claim to offer access to 3 to 4 million titles.

If one wished to more fully address the universe of book titles in the United States, including those that are no longer in print, one could look to the holdings of the larger national libraries and copyright repositories - for example, the Library of Congress print media collection includes almost **26 million books (208 terabytes).**

World

To estimate the international stock of books currently available for purchase, we extrapolate from the United States production figures. The US engages in the world's largest trade in printed products, producing about 40% of the world's printed material, according to the US Industry and Trade Outlook 2000. The world stock of original titles might be about **8 million titles** - equivalent to **64 TB.**

Using the same 40% rule of thumb, we can also estimate the worldwide stock of books (including those out of print). The national library and copyright repository of the United States - the Library of Congress - contains about 26 million books. Therefore, the world stock of books might be approximately **65 million titles.**

Rate of change

The number of titles within most print media forms have increased each year worldwide - between 2 and 10%. Within the US, the number of book titles increased every year until 1996, when there was a 5% downturn.

Copies

Flow

If all of the writing paper and newsprint produced each year were used to store printed information, this would be equivalent to about **980,000 terabytes** worldwide.

World

As of 1997, the world was producing 90 million metric tons of printing and writing paper and 36 million metric tons of newspaper. In equivalent bytes, this translates to **540,000 TB** (world) for printing and writing paper, and **432,000 TB** for newsprint.

The number of books sold worldwide may be estimated using the 40% rule of thumb (cited above) and the US book sales statistics (cited below): about **2.75 billion books**, equivalent to **22,000 TB**.

United States

The US produces about 30% of the world's paper and paperboard output (*Source: US Industry & Trade Outlook 2000*). In 1999, the US produced 23.8 million metric tons of printing and writing paper and 6.4 million metric tons of newsprint. In bytes, this translates to **142,800 TB** for printing and writing paper and **76,800 TB** for newsprint. These figures provide an upper bound on the total number of bytes required to digitally store all the information produced in printed format each year.

About **1.1 billion books** were sold in the United States in 1999. Using the 8 MB/book estimate, this is equivalent to **8,800 TB**. (*Source: Wall Street Journal, July 17, 2000, "A New Chapter: Independent Booksellers Hope to Find Strength in Numbers" by Scott Eden.*)

In the United States **55,979,332 daily newspapers** and **59,894,381 Sunday newspapers** circulate each year. (Source: *Newspaper Association of America, citing Editor and Publisher.*)

The total number of US magazines circulated annually exceeds **500 million**. (*Source: US Industry and Trade Outlook.*)

Each year, almost **500 billion copies** are produced on copiers in the US; nearly **15 trillion copies** are produced on copiers, printers, and multi-function machines. (*Source: XeroxParc*). For specific information on fax printing, see the Telecommunications Summary.

Stock

According to the 1999 US Industry and Trade Outlook, the United States produces more printed products than any other country in the world. NAFTA countries have a 50% world market share. Estimating that the US produces 40% of the world's printed materials, we can estimate that each year the world produces **7.5 billion archiveable pages**, which would be equivalent to **195 terabytes** (compressed). (*Source: National Archives and Records Administration.*)

Rate of Change

As with the other print industries, growth in paper production is expected to be incremental but fairly consistent, both within the United States and internationally. Globally, paper and paperboard production capacity is forecast to grow from 333.6 million metric tons in 1998 to 348.1 million metric tons in 2001, an increase of 14.5 million metric tons (about 4%) over those three years. (*Source: U.S. Industry and Trade Outlook, 2000*).

According to the American Forest and Paper Association, US capacity to produce paper will increase by an average of only 0.7% annually over the next three years (2000-2002).

References

- Bogart, Dave, ed. *The Bowker Annual Library and Book Trade Almanac*, 44th edition. New Jersey: R.R. Bowker, 1999.
- Cummings, Anthony M., Marcia L. Witte, William G. Bowen, and Laura O. Lazarus. *University Libraries and Scholarly Communication: A Study Prepared for the Andrew W. Mellon Foundation*. The Association of Research Libraries, 1992
- Hayes, Robert M., UCLA School of Information Science, "The Economics of Digital Libraries"
- King, Donald W. and Carol Tenopir. "Economic Cost Models of Scientific Scholarly Journals," 1998.
- American Forest and Paper Association, *1999 Statistics for Paper, Paperboard and Wood Pulp*. Washington, DC, 1999. To order a copy, see www.afandpa.org/about/about.html or call (800) 244-3090.
- *Annual Report of the Librarian of Congress*. Washington, DC: Library of Congress, 1999.
- ArchiveBuilders.com White Paper, "Computer Storage Requirements for Various Digitized Document Types"
- Association of Research Libraries Statistics
- *Books in Print, 1999-2000.* New Jersey: R.R. Bowker, 1999.
- International Standard Serial Number Register
- International Standard Book Number Register
- JSTOR digital library
- Magazine Publishers of America, Information Center, (212) 872-3745.
- National Directory of Magazines.

- Newsletter and Electronic Publisher's Association (NEPA)
- *Newspaper Association of America, Facts About Newspapers*.
- Newspaper Project, National Endowment for the Humanities.
- *Oxbridge Directory of Newsletters 1997*. New York: Oxbridge Communications, Inc., 1997.
- *Ulrich's International Periodical Directory*. New York: Bowker Publishing, 2000.
- *UNESCO Statistical Yearbook 1999.* Paris, UNESCO, 1999.
- U.S. Census Department, 1999 Statistical Abstract of the United States
- *U.S. Department of Labor, Bureau of Labor Statistics*
- U.S. Government Printing Office, *Prepared Statement before the Committee on Rules and Administration, U.S. Senate on Public Access to Government Information in the 21st Century* July 1996.
- *U.S. Industry and Trade Outlook*. Available in print from McGraw Hill/U.S Department of Commerce Washington, D.C. or download from www.ntis.gov/products
- Chaskas, Eric. US National Archives and Records Administration.
- *Walden's Paper Report*. Twice-monthly newsletter, published by Walden-Mott Corporation, Ramsey, NJ. Available by subscription only. See www.walden-mott.com/PaperReport/PAP_RPT.HTM or call (201) 818-8630.

Charts

Click here to see charts supporting the above estimates, with time-series data.

More Discussion

Click here to read additional discussion of the conversion factors and related issues and to obtain detailed bibliographical information.

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Film - Summary

Originals

Flow

Photographs

There are over 2700 photographs taken every second around the world, adding up to well over 80 billion new images a year taken on over 3 billion rolls of film, according to estimates published by the United States Department of Commerce.

Photo CD, a format for digitized photography introduced by Kodak in 1992, has been widely adopted both by professional and amateur photographers. Kodak reports that the typical photograph can be digitized in this format in 5 megabytes without loss of picture quality. Utilizing this conversion factor, then, the world's 82 billion photos store 410 petabytes of data every year in photographs.

Motion Pictures

Apart from still photography, film is also used to store moving pictures. In the years from 1990 to 1995, UNESCO reports that there were 4,250 films produced annually throughout the world. The Motion Picture Association of America reports that for the year 1998, its members released 221 movies (compared to 219 in 1997), while releases by all U.S. companies, including independent film companies, rose from 461 in 1997 to 490 in 1998.

It takes approximately 2 gigabytes to store an hour of motion picture images in digital form using the MPEG-2 compression standard. If the images in 4500 full length movies were converted into bits, the world's annual original cinematic production would, therefore, consume about 16 terabytes.

X-Ray Film

The other major use for film is the storage of x-ray images for medical, dental and industrial purposes. Approximately 2 billion radiographs are taken around the world each year, including chest x-rays, mammograms, CT scans, and so on. (Traditionally, 8% of x-ray film is used in dentistry and industrial applications.) When x-ray films are converted to digital format, it is important that there is no important clinical information lost. The University of Pittsburgh Clinical Multimedia Laboratory suggests that an average conversion of a chest x-ray to digital storage with lossless compression will require 8 megabytes. To store all the world's x-rays to a computer file of this size would, therefore, require 17 petabytes each year.

Table 1: Original Data On Film Annually Worldwide

	Units	Digital Conversion	Total Petabytes
Photography	82,000,000,000	5 mb per photo	410
Motion Pictures	4,000	4 gb per movie	0.016
X-Rays	2,160,000,000	8 mb per radiograph	17.2
		All Film Total:	427.216

Stock

Photographs

The number of original photographs stored around the world is not a widely reported topic. There are commercial firms that collect professional photographs for resale and the largest of these report collections in the tens of millions. For example, Getty Images reports holding 70 million images mostly on silver halide film and its principal rival, Corbis Images, claims 65 million images.

As for amateur photographs, in 1997 it was estimated that there were 150 billion photographs stored in the United States. This is approximately 8 to 10 years production of photographs as of that time. Assuming that similar storage rates apply worldwide, it is likely that on the order of 750 billion photographs now exist. Using the same PhotoCD estimate of 5 megabytes per photo, there are 3.5 exabytes of original photographic data.

Motion Pictures

The number of motion pictures made around the world from 1895 to 1990 was approximately 242,000. *The International Film Index, 1895-1990*, lists entries of that many titles of films. (There is no guarantee that some of these movies still exist, however.) The overall total is broken down into these categories of film types:

Table 2: Film Breakdown by Type

Type	Number

FEATURE	110,586
SHORT	50,056
ANIMATED	9,787
DOCUMENTARY	7,277
TELEVISION	4,197
SERIAL	553
SILENT	59,680
TOTAL	242,136

The "Television" category refers to films that were originally made for television broadcast rather than theatrical release. "Serials" are films made primarily in the United States between 1914-1936 and were made to be shown in weekly installments.

As discussed in the precedng section, in the decade since 1990, there have been around 4500 movies made each year, adding another 45,000 to the total world stock of original motion pictures. Accordingly, a good estimate of the world's original pictures is approximately 300,000.

X-Rays

The clinical and legal uses of medical x-rays continue for an indefinite time and, therefore, prudent practice is to preserve x-rays and medical records generally for as long as possible. The same principle applies to dental x-rays. The only use of x-ray that may result in regular destruction of the resulting images is industrial testing, but even there it is likely that images are retained for a substantial period of time. Therefore, it is believed that there is little systematic destruction of the flow of new x-rays and virtually all of them are added to the stock. For the sake of calculation, it is assumed that a full ten years of x-ray images will constitute the stock. This is equivalent to approximately 21.6 billion images or 172.8 petabytes.

Copies

Flow

Photographs

There has been very little history of the large mass of photographs being copied. Kodak estimates that only about 2 percent of photographs are ever copied or modified in any way after they are originally developed. Of course, some photographic images are widely distributed in newspapers and magazines. But these represent a miniscule fraction even of the professional photographer's work, itself a small minority of all photographs.

One copy of 2 percent of the annual new photographs would be 1.6 billion photographic copies. With PhotoCD compression, this would represent 8 petabytes per year of photographic copies.

Motion Pictures

The Wolfman Report on the Photographic and Imaging Industry in the United States states that the average number of prints per original motion picture is 700. The Silver Institute, however, reports that 6000 release prints are made for each feature movie. A figure of 1000 copies for motion pictures will be used on the assumption that many of the world's motion pictures have more limited releases than the typical Hollywood blockbuster. American studios only account for 400 to 500 movies per year. 45,000 copies of motion pictures per year at 4 gigabytes per copy is 18 petabytes of copies.

X-Rays

The clinical requirements for medical x-rays demand that originals be used in almost any situation. There is no siginificant use of copies of x-rays at all.

Stock

Photographs

The stock of copies of photographs can be calculated by reference to the assumptions made for the storage of the originals. Ten years of copies of photographs would be 16 billion. If these were all digitized at the rate of 5 megabytes per picture, there are 80 petabytes of photographic copies on hand.

Motion Picture

The copies on film of motion pictures made for distribution are short-lived. Most of these copies are deliberately destroyed when the general theatrical release of the movie is over. If even ten of these survive for the approximately 4,000 motion pictures made annually for the last twenty-five years, this would be equivalent to about 1,000,000 copies. If each copy is equivalent to 4 gigabytes of data, this stock of copies of motion picture film is 4 petabytes.

X-Rays

There is no significant stock of copies of x-rays.

References

- United States Industrial and Trade Outlook 2000
- UNESCO Statistical Yearbook
- 1999 Photo Marketing Association
- Silver Institute

More Details About Film

How Much Information?

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Optical - Summary

Originals

For optical media, we focused on the three major industry categories: CD audio, CD-ROM, and DVD.

Flow

Annual world title production of the 3 media types appears in the following chart. (All figures are rounded to the nearest hundred or terabyte.)

Table 1: Annual world title production of the 3 media types.

Media Type (Source* & Year Cited)	Unique Items per Year (US)	Unique Items per Year (World)	Conversion Factor	Total Terabytes (Annual US)	Total Terabytes (Annual Worldwide)
CD - Music (1998)	33,100	90,000	Uncompressed: .650 GB/item	22	58
			Compressed (to MP3): .065 GB/item	2	6
CD- ROM (1999)	500	1,000	Uncompressed (to MP3): .650 GB/item	1	3
			Compressed: .065 GB/item	.3	.6
DVD-video (1999)	3,000	5,000	4.38 GB/item	13	22

Totals:	**Uncompressed**	**36**	**83**
	Compressed	**15.3**	**29**
*Detailed source information listed at end of this document.			

World

To estimate how many CD-audio originals are created each year worldwide, we used RIAA statistics regarding the US market share and US record releases (see below). The United States holds a 37% share of the world music market and releases about 33,100 items per year. Therefore, the world produces roughly **90,000 originals** per year, equivalent to **58 TB** (uncompressed).

Between 1998 and 1999, 1,000 new CD-ROM data titles were added to *CD-ROMs in Print*, an international directory published by Gale Research. This equals about **3 TB** of new information in one year (uncompressed).

The United States currently produces about 60% of the DVD titles available worldwide. Using US title production statistics, we estimate that about **5,000 new DVD titles** are produced internationally each year - this is about **22 TB**. (Source: Jim Taylor's *DVD FAQ).*

United States

The US produces approximately 37% of the world's CD-audio titles, 50% of the world's CD-ROM titles and 60% of the world's DVD titles. (Sources: Recording Industry Association of America, International Recording Media Association, and Jim Taylor's *DVD FAQ*).

The Recording Industry Association of America (RIAA) reports the number of new releases and album re-releases each year. In 1998, **33,100 titles** were released, roughly equivalent to **22 TB.**

The US share of the CD-ROM replication market is 52%, according to the International Recording Media Association. If we assume that the US holds a similar share of CD-ROM title production, then about **500 titles** are produced by the United States each year, equivalent to about **1 TB** per year.

For the past three years, new DVD titles have been added at a rate of about **3,000 per year** (**13 TB per year)** --a tremendous rate of content growth, but that's because it is a new medium and a significant amount of legacy content is being converted. This rate should decrease as the medium becomes more well-established. (Source: *DVD Entertainment Group*)

Stock

The All Music Guide (a comprehensive database tool used by industry leaders) reports a total of **523,363 titles** (445,735 popular music and 77,628 classical music albums). If each work were stored on a 650 MB CD, this would be equivalent to **340 TB.** We can assume that this figure represents the US portion of original works, then extrapolate the world stock to be about **1,400,000 titles**. [???]

According to the 1999 edition of *CD-ROMs in Print,* internationally there are about 16,200 unique CD-ROM titles (**about 11 TB**)--business applications (such as word processing and spreadsheet packages), games, reference tools, and instructional programs. This figure is consistent with other CD-ROM directories, such as the *Multimedia & CD-ROM Directory*, which lists 17,000 titles.

The United States produces about two-thirds of the total DVD titles. As of June 2000, there are about 8,500 titles available in the United States, 13,000 worldwide. This is equivalent to **37 TB** (US) and **57 TB** (world).

Rate of Change

Production of CD-Audio and CD-ROM originals is not increasing dramatically. However, the production of DVD originals is growing at a tremendous rate (about 100% per year), due to the fact that material previously available in another format is being reissued on DVD.

Copies

Flow

Replication

In 1999, there were about 4.7 million audio CDs and 3.6 million CD-ROMs replicated worldwide, according to the International Recording Media Association (IRMA). In addition, in 1999 194 million DVD-Video units, 12 million DVD-ROM units, and 2 million DVD-Audio units were replicated.

A total of 1.583 billion recordable CDs (CD-Rs) were sold around the world in 1999, according to Santa Clara Consulting Group. (Source: News Release, *"Record sales drive Memorex up to third in world CD market" www.lewisvpr.com/Releases/UKReleases/000620_memorex.html*) Estimates of growth in demand for CD-R vary--one consulting firm projects global demand to reach 4.74 billion in the year 2000, while others project demand to be between 2.5 and 3 billion units. (Photonics Industry & Technology Development Association, cited in www.taiwanheadlines.gov)

Retail

During 1999, according to the Recording Industry Association of America (RIAA), **938.9 million CDs** were shipped to retail by U.S. producers. The US has a 37% share of the world's sales, according to the International Federation of the Phonographic Industry (IFPI). Therefore, one can extrapolate that CD shipments worldwide are about **2,500 million units**. This is equivalent to **1,625 TB**.

According to the DVD Entertainment Group, more than 130 million DVD video movies and music video titles were shipped to retail between spring 1997 (when the format launched) and January 2000. **100 million discs** were shipped during 1999 alone.

Stock

The stock of audio CDs in the United States can be estimated by summing the CD unit sales since the format became popular. The RIAA only provides statistics going back to 1990 - these 10 years of shipments add up to **6,200 million units**, equivalent to about **4,030 TB**. Ten years of worldwide shipments are approximately **15.2 billion units**, equivalent to **10,937 TB**.

Since the format was launched in 1997, more than **1.5 billion DVDs** have been replicated worldwide, almost **1 billion** in North America alone.

Optical Media Bibliography

- All Music Guide
- The CD Information Center
- *CD-ROMs in Print, 13th Edition*: An International Guide to CD-ROM, CD-I, 3DO, MMCD, CD32, Multimedia, Laserdisc, and Electronic Products. New York: The Gale Group, 1999.
- *CD-ROM Finder: The World of CD-ROM Products for Information Seekers*. Medford, NJ: Learned Information, 1993
- The International Recording Media Association
- DVD Channel News
- DVD Entertainment Group
- DVD FAQ
- DVD Insider
- International Federation of the Phonographic Industry

- Medialine
- Optical Storage Technology Association
- SUN CD-ROM FAQ

Charts

Click here to see charts supporting the above estimates, with time-series data.

More Discussion

Click here to read additional discussion of the conversion factors and related issues and to obtain detailed bibliographical information.

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Magnetic - Summary

Originals

Flow

Tape

Analog

Video

In the year 2000, 1.4 billion blank VHS video tapes will be produced for the entire world. If all of these tapes were filled to their 120 minute capacity and then converted to digital using MPEG-2 compression, there would be approximately 4 gigabytes of data per tape. One year's production of blank videotape, therefore, provides storage space adequate for 5600 petabytes of data.

Assuming twenty percent [???] of this tape is used for the storage of original data, the flow of new data stored on analog VHS videotape per year would be 1120 petabytes.

Video camcorder tapes (all formats except VHS) are produced at the rate of 150 million per year according to the Japan Recording Media Industry Association. Almost all of this tape is used for the storage of original data. Assuming one hour per tape in MPEG-2 format yields 300 petabytes.

Total original analog video production worldwide runs at about 1420 petabytes annually. [???]

Audio

In the year 2000, 921 million blank audio tape cassettes will be produced for the entire world according to British research firm, Understanding & Solutions. If all of these tapes were filled to their 120 minute capacity and then converted to digital using the common CD audio format, there would be approximately 1 gigabyte of data per tape. One year's production of blank audiotape, therefore, provides storage space adequate for 921 petabytes of data.

Assuming twenty percent [???] of this tape were used for the storage of original data, the flow of new data stored on analog audiotape per year would be 184.2 petabytes.

Digital

There are 25 million computer tape drives installed in the world at present. These drives provide storage capacity for all range of computers - from desktop personal computers to the most mammoth supercomputers. Fred Moore estimates that the amount of data stored on tape is between 4 and 15 times the amount of enterprise data on disks and that there is about $1 billion per year of computer tape media sold worldwide.

In order to estimate the amount of data stored on tape, we will first estimate the number of enterprise storage systems and then multiply by 10. IDC estimates that 250 petabytes of RAID storage capacity will be shipped worldwide in 2000. RAID storage systems are taken as most closely approximating the storage capacity deployed to data-intensive corporate, government and scientific uses, the largest consumers of magnetic tape backup.

A midrange estimate of the total amount of data stored on magnetic tape would, therefore, be 2.5 exabytes.

In all but the largest computer applications, however, tape is generally used solely for backup of data already stored on hard disk drives. Quantum, the manufacturer of DLT tape, the most popular format for enterprise storage, estimates that 90 percent of the tape capacity in that format is used for backup. Fred Moore also points out that it is more and more common for multiple copies of data to now be stored on tape.

If it is assumed that ten percent of the total amount of data stored on tape is original data of the sort generated by scientific experiments in high-energy physics or by observational earth satellites or archival storage on tape where the data is no longer stored on disk, original magnetic tape data is roughly the same as all the RAID capacity shipped annually - 250 petabytes in the year 2000.

250 petabytes is also generally consistent with estimates derived by use of forecasts of producer revenue of around $1 billion for tape media and an average cost of around $5 per gigabyte of tape storage.

Disks

Floppy Disks

In the year 2000, 1 billion 3.5 inch floppy disks, each capable of storing 1.44 megabytes, will be produced for the world. This is an aggregate storage capacity of 1.4 petabytes. If [???] five percent of this is original data, new data per year on floppy disk would be 0.07 petabytes.

Removable Disks

In the year 2000, 88 million removable 100 megabyte disks and 25 million removable 1 gigabyte disks will be produced for the world. Together, these two varieties of removable disks provide 33.8 petabytes of storage capacity. If five [???] percent of this is original data, new data per year stored on removable disks would be 1.69 petabytes.

Hard Disks

In the year 2000, hard disk drives capable of storing 2500 petabytes of data will be produced for the world.

The amount of original data stored on hard disks is most likely to vary according to the computing environment in which the disks are deployed. It is possible to divide all hard disk storage into three categories:

1. Personal computer, laptop, or workstation. This type of computer is responsible for approximately 55% of the computer disk storage capacity currently shipped.
2. Departmental server. This class of computing environment is responsible for about 30% of the overall storage market.
3. Enterprise server. These big computers account for about 15% of the hard disk storage.

As with all such broad categorizations, there may be quite a bit of blurring around the edges. However,

consideration of these three different environments leads to the conclusion that the amount of original data stored on the computers in each is probably substantially different.

Single user computers and the applications software usually found on them is not suited for the production of large amounts of original data. A recent study (McKenzie, "Microsoft's 'Applications Barrier to Entry': The Missing 70,000 Programs") found that most people used only a few applications other than those found in the Microsoft Office application suite. These applications are usually text-based, such as word processing or spreadsheets, and so require minimal storage space. Most personal computers now sold come with hard disk storage capacity in the range of 10 gigabytes. 100 megabytes of original data constitutes 1 percent of disk capacity, which is the estimate for this category of computer disk. [???]

Departmental servers would be commonly found in business, government, educational or other organizational settings. These servers provide disk space for a group of users, who all contribute to the production of organizational data. Aside from the databases and spreadsheets, there may be product catalogs and other graphic intensive marketing material, PowerPoint presentations, and so on. An estimate of the original data stored in these hard disks is 35%. [???]

Enterprise servers are the large-scale computing environments where "big iron" traditionally has reigned. The applications here are corporate or governmental transaction processing on a large scale or the generation of huge data sets from scientific research missions. The amount of original data stored on these computers is estimated at around 65%. [???]

Table 1: Original Data Stored On Hard Disk By Computing Environment (1995-1999) [In Petabytes]

	Original Data	1995		1996		1997		1998		1999	
		Total	Original	Total	Original	Total	Original	Total	Original	Total	Original
Personal	1%	58	0.6	101	1	189	2	399	4	766	8
Departmental	35%	32	15	55	19	114	40	239	84	460	161
Enterprise	65%	16	8	27	17	41	27	86	56	167	109
Total:		105	23.6	183	37	344	69	724	144	1,393	278

Table 2: Original Data Stored On Hard Disk By Computing Environment (2000-2004) [In Petabytes]

	Original Data	2000		2001		2002		2003		2004	
		Total	Original	Total	Original	Total	Original	Total	Original	Total	Original
Personal	1%	1,405	14	2,553	26	4,466	45	7,165	72		
Departmental	35%	843	295	1,532	536	2,680	938	4,299	1,505		
Enterprise	65%	306	199	557	362	974	633	1,563	1,016		
Total:		2,554	508	4,642	924	8,120	1,616	13,027	2,593		

The amount of original data compared to the total amount of data stored on magnetic hard disks works out to twenty percent over all the years under consideration. Accordingly, in the following table summarizing the annual flow of original data in various media, 20% of all hard disks are assumed to store original data, the actual amount will vary depending on the capacity of the hard drive itself.

Table 3: Original Data Flow Storage Estimates

Media Type	Unique Items per Year	Conversion Factor	Total Annual Petabytes (World)
Blank Audio Tape (2000)	184,200,000	1 gb per tape (CD audio format - no compression)	184.2
Blank Video Tape (2000)	355,000,000	4 gb per tape (MPEG-2 compression)	1420
Computer Tape Drives (2000)	5,000,000	Varies	250
Floppy Disk (2000)	5,000,000	1.44 mb per disk	0.07
Removable Disks (2000)	4,400,000 1,250,000	100 mb (low capacity) 1 gb (high capacity)	1.69

Hard Disks (2000)	37,400,000	Varies	500

Stock of Originals

Tape

Analog

Video

[???] 10 billion video cassettes, both prerecorded and recorded at home, have been accumulated. This is equivalent to about five years production of these tapes. One billion camcorder format tapes have also been added over the same time. This would be equivalent to 3 billion hours of stored original video, which if digitally encoded would produce a stock of about 6000 petabytes of original videotaped data. [???]

Audio

The total stock of original audio content stored on tape may be estimated by assuming twenty [???] percent of five years production of blank cassette tapes contains original content. This equals 1 billion cassette tapes. The digital equivalent of this audio information is 1000 petabytes.

Digital

The stock of original data on magnetic tape may be approximated by adding the yearly flow of original data over the course of the expected lifetime of the medium. Some unfortunate experiences with the loss of computer data stored on magnetic tape has led to the practice of continuous migration of this data to new media every five to ten years. This process leads to the reduction in the number of tape cartridges that need to be managed as tape capacity inexorably rises as well as insuring modernization of the tape format.

Therefore, the stock of original data on magnetic tape may be taken as five year's worth of original data flow. [???]

Disks

Floppy Disks

Floppy disks useful life is estimated to be three years. The amount of original data stored on the 4.5 billion floppies produced over the course of the past three years is around 5 percent of the total data on those disks, or 0.32 petabytes.

Removable Disks

The amount of original data stored on removable disks over the past three years is approximately 5 petabytes.

Hard Disks

Over the past three years, hard disk capacity of 4625 petabytes has been produced. If twenty [???] percent of that capacity has been used to store original content, the stock in that format is now 925 petabytes.

Copies

Flow

Tape

Analog

Video

In the year 2000 1.8 billion prerecorded video tapes will be distributed worldwide according to the International Recording Media Association. This entire production will be copies, principally of feature films. If converted to digital using MPEG-2 compression, prerecorded videotape would consume 7,200 petabytes per year.

If [???] 80 percent of the blank videotape distributed per year were used for copies, this would constitute a digital equivalent of 4,480 petabytes.

The total yearly world production of copied data on analog videotape, therefore, is 11,680 petabytes. [???]

Audio

In the year 1999, 125 million prerecorded audio cassette tapes were distributed in the United States according to the Recording Industry Association of America. This entire production was copies, principally of music. If converted to digital using audio CD format, and assuming about one hour of music per tape, prerecorded audio tape would consume 62.5 petabytes per year.

If 80 percent [???] of the 921 million blank audiotapes distributed per year worldwide were used for copies, this would constitute a digital equivalent of 737 petabytes.

The total yearly production of copied data on analog audiotape, therefore, is 800 petabytes.

Digital

If ninety percent of the computer tape distributed annually were used for copies, this would constitute 2.25 exabytes of copied data for the year 2000. The amount of data on tape, however, is calculated by reference to the RAID storage capacity shipped, which has been rising rapidly. Similar trends will most likely be seen in the amount of data stored on tape, particularly as it becomes more common to make multiple tape copies of data.

Disks

Floppy Disks

If 95 [???] percent of the 1.4 petabytes of annual floppy disk storage were used for copies of data, this would add 1.33 petabytes to the stock of digital data stored on floppy disks.

Removable Disks

If 95 [???] percent of the 33.8 petabytes of annual removable disk storage were used for copies of data, this would add 32.1 petabytes to the stock of digital data stored on removable disks.

Hard Disks

If 80 [???] percent of the 2500 petabytes of annual hard disk storage were used for copies of data, this would add 2000 petabytes to the stock of digital data stored on hard disks.

Stock of Copies

Tape

Analog

Video

Five years production of blank T-120 VHS videotapes is approximately 7 billion units. If 80 [???] percent of this were used for storage of copied data, the total stock of such data in MPEG-2 conversion would be 2240 petabytes. Five years production of prerecorded videotapes is approximately 8 billion units. This is the digital equivalent of 3200 petabytes. The total world stock of copied data on videotape is 5440 petabytes.

Audio

The world production of blank audiotape for the past five years is approximately 5 billion units. There has also been sales of prerecorded audiotapes during that period of 6 billion units. Assuming that 80 [???] percent of the blank audiotape has been used for the storage of copied data, the overall stock of copies of audio data on tape is 4000 petabytes on blank tape and 6000 petabytes on prerecorded tape. Therefore, there is a total stock of 10,000 petabytes of copies of audio data on magnetic tape.

Digital

Ninety percent of the data stored on tape is copies and tape is estimated to store 4 to 15 the amount of enterprise data on hard disk. IDC estimates that approximately 400 petabytes of RAID have been shipped in the past three years. If all data on tape is calculated as ten times that amount, or 4 exabytes, and copies as 90 percent of that overall amount, copies of data on magnetic tape worldwide would be 3.6 exabytes.

Disks

Floppy Disks

There have been 4.5 billion floppy disks produced in the last three years, creating a total storage capacity of 6.5 petabytes. If 95 percent of that storage were used for copies, the total stock of copies on floppy disk would be 6.1 petabytes.

Removable Disks

Over the past three years approximately 100 petabytes have been stored on removable disk. Assuming that 95 [???] percent of that is backups and copies, the total stock of copied data in this media format is 95 petabytes.

Hard Disks

Over the past three years, hard disk capacity of 4625 petabytes has been produced. If eighty [???] percent of that capacity has been used to store copies of data, the stock of copied data on hard disk is 3700 petabytes.

More Details

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Internet - Summary

The Internet is one of the youngest and fastest growing media in today's world. Internet growth is still accelerating, which indicates that the Internet has not yet reached its highest expansion period [1]. It should be noted, however, that while the Internet is a completely new kind of medium, by separating it into a distinct category, we are allowing for a certain amount of double counting, because all the Internet-based stock of information is already accounted for under "magnetic" or "tape" categories. Furthermore, we should make clear the distinction between the stock and the flow of information. While web sites and some portion of email messages are being stored and accounted for under different storage categories, there are other "components" of what we know as "Internet," such as Internet Relay Chat (IRC) or Telnet, which exist only as a flow of communication. What makes the Internet extremely successful is that it is one of a handful of media (such as radio and TV), where one unit of storage might generate terabytes of flow, as opposed to books and newspapers, where one exemplar is usually read by one or two people, and the flow of information is relatively low.

World Wide Web

There are two groups of Web content. One, which we would call the "surface" Web is what everybody knows as the "Web," a group that consists of static, publicly available web pages, and which is a relatively small portion of the entire Web. Another group is called the "deep" Web, and it consists of specialized Web-accessible databases and dynamic web sites, which are not widely known by "average" surfers, even though the information available on the "deep" Web is 400 to 550 times larger than the information on the "surface." [2]

The "surface" Web consists of approximately 2.5 billion documents [1 and 5], up from 1 billion pages at the beginning of the year [3], with a rate of growth of 7.3 million pages per day [1]. Estimates of the average "surface" page size vary in the range from 10 kbytes [1] per page to 20 kbytes per page [4]. So, the total amount of information on the "surface" Web varies somewhere from **25 to 50 terabytes** of information [HTML-included basis]. If we want to obtain a figure for textual information, we would use a factor of 0.4 [4], which leads to an estimate of **10 to 20 terabytes** of textual content. At 7.3 billion new pages added every day, the rate of growth is [taking an average estimate] 0.1 terabytes of new information [HTML-included] per day.

If we take into account all web-accessible information, such as web-connected databases, dynamic pages, intranet sites, etc., collectively known as "deep" Web, there are **550 billion web-connected documents**, with an average page size of 14 kbytes, and 95% of this information is publicly accessible [2]. If we were to store this information in one place, we would need **7,500 terabytes** of

storage, which is 150 times more storage than we would need for the entire "surface" Web, even taking the highest estimate of 50 terabytes. 56% of this information is the actual content [HTML excluded], which gives us an estimate of **4,200 terabytes** of high-quality data. Two of the largest "deep" web sites - National Climatic Data Center and NASA databases - contain **585 terabytes** of information, which is 7.8% of the "deep" web. And 60 of the largest web sites contain **750 terabytes** of information, which is 10% of the "deep" web.

When we look at the distribution of the web sites, the most apparent trend is that English loses its dominant position. Currently, only 50% of all Internet users are native English speakers, though English web sites continue to dominate with approximately 78% of all web sites and 96% of e-commerce web sites being in English [6]. It's hard to estimate what percentage of web sites have their origins in the United States, because .com domains can be registered in virtually any country, English-language web sites are often created in countries like Japan, and many international web sites are hosted in the United States. 17 million out of 27.5 million domains registered worldwide are .com, and 2 million are .uk, making Great Britain's domain the biggest country domain in the world [7].

More Details

Email & Mailing Lists

Email has become one of the most widespread ways of communication in today's society. A white-collar worker receives about 40 email messages in his office every day [8]. Aggregately, based on different estimates, there will be from 610 billion [9] to 1100 billion [10] messages sent this year alone. With the average size of an email message 18,500 bytes [11] and growing, the amount of flow becomes surprisingly gigantic, somewhere between **11,285 and 20,350 terabytes**. Of course, not all of this email gets stored. Mail.com has 14.5 million email boxes and uses 27 terabytes of storage; with approximately 500 million mailboxes worldwide, the required storage space is more than **900 terabytes**, which means that only one in 17 messages is kept for some period of time.

Mailing lists can be viewed as a subcategory in email. It is hard to determine the number of mailing lists in existence, but we can approximate it based on some available statistics. One of the most frequently used mailing list managers - LISTSERV - is used to send 30 million messages per day in approximately 150,000 mailing lists [12]. A sample of mailing lists has shown that 30% of them are managed using LISTSERV. Using this information, we would estimate the total number of mailing list messages at 36.5 billion per year with aggregate volume of **675 terabytes**.

Distribution of mailboxes has the same pattern as the distribution of web sites. While in 1984, 90% of the world's e-mailboxes were located in the U.S., at the end of 1999 this number dropped to 59%, and is expected to decrease even further. [13].

More Details

Usenet

Most of the statistics in this category are vague, so the numbers we have should be regarded with a certain skepticism. Cidera, which is the 14th biggest news provider on the Internet [14], gets approximately 0.150 terabytes of Usenet feeds per day. We would estimate the total amount of original news feeds at 0.2 terabytes per day, which leads to **73 terabytes** of original Usenet postings per year, which are redistributed by local ISPs and news servers an endless number of times.

FTP

We are missing any significant data on this sector, but we know that Walnut Creek CD-ROM archive contains a total of **0.412 terabytes** of data on two servers [ftp.cdrom.com and ftp.freesoftware.com] and the amount of storage was expanding at 100% every year over the past 6 years [15]. It should be noticed that the distinction between FTP and HTTP becomes more blurred, as more and more file archives become available through HTTP.

IRC, Messaging Services, Telnet...

These categories mostly represent a flow of information as opposed to the stock. Liszt.com has one of the biggest directories of IRC channels - 37750 channels on 27 networks, with 150,000 users, all of them typing text as fast as they can. [16]

References

[1] "Sizing the Internet," *Cyveillance*, http://www.cyveillance.com/resources/library.asp
[2] "The Deep Web: Surfacing Hidden Value," *BrightPlanet LLC*, http://www.completeplanet.com/Tutorials/DeepWeb/index.asp
[3] "Web Surpasses One Billion Documents," *Inktomi Corp.*, http://www.inktomi.com/new/press/billion.html
[4] "Accessibility of Information on the Web," *Nature Magazine, Volume 400, Number 6740, Page 107*
[5] "Size of the Web: A Dynamic Essay for a Dynamic Medium," *The Censorware Project*, http://censorware.org/web_size/
[6] "State of the Internet 2000," *United States Internet Council & ITTA Inc.,*http://usic.wslogic.com/intro.html
[7] "Domain Statistics," *DomainStats.com*, http://www.domainstats.com
[8] "Sending AOL a Message," *Newsweek, Aug 9, 1999, p.51*
[9] "Email Facts," *24/7 Media*, http://www.247media.com/research/trends/email.html
[10] "Like It Or Not, You've Got Mail," *BusinessWeek*, http://businessweek.com/1999/99_40/b3649026.htm
[11] UC Berkeley Email Stats
[12] "LISTSERV Statistics," *L-Soft*, http://www.lsoft.com/news/default.asp?item=statistics
[13] "Year-End 1999 Mailbox Report," *Messaging Online*, http://www.messagingonline.com/
[14] "Top 1000 Usenet Sites" *Freenix*, http://www.freenix.org/reseau/top1000/
[15] David Greenman, Walnut Creek CD-ROM Archive
[16] *Liszt.Com*, http://www.liszt.com/

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Broadcast - Summary

Originals

World

Table 1: World

Media Type	Number of Stations	Unique Items per Year	Conversion Factor	Total Terabytes (Annual)	
				Lower Bound	Upper Bound
Radio (CIA Factbook 2000)	43,973	65.5 million hours of original programming	.05 GB/hour	3,274	3,274
Television (broadcast only) (CIA Factbook 2000)	33,071	48 million hours of original programming	1.3 GB - 2.25 GB hour	62,769	108,638
			Total:	**66,043**	**111,912**

Table 1: United States

Media Type	Number of Stations	Unique Items per Year	Conversion Factor	Total Terabytes (Annual)	
				Lower Bound	Upper Bound

Radio (FCC, 1999)	12,600	15.8 million hours	.05 GB/hour	788 TB	788 TB
Television (broadcast and cable networks) (FCC, NCTA, 1999)	1,884	3.4 million hours of original programming	1.3 GB- 2.25 GB/hour	4,470 TB	7,736 TB
			Total:	**5,258**	**8,524 TB**

Radio

Conversion Factor

Each hour of audio requires about **50 MB**, if stored at MP3 quality. (Different sources cite different figures, depending upon assumptions made about compression and sound quality.)

World

There are **43,773** active radio stations in the world, according to the CIA World Factbook 2000: about 16,500 AM stations, 26,000 FM stations, and 1,500 shortwave stations.

We estimate that FM radio stations broadcast 20 hours per day, AM stations 16 hours per day, and shortwave stations 12 hours per day. Therefore, there are approximately **290 million** hours (188 million FM, 98 million AM, and 6 million shortwave) of radio programming per year. Applying the 50 MB/hour rule of thumb, one may estimate an annual storage requirement of about **14,500** TB if one were to record everything broadcast on the radio.

United States

As of 1999, there are **12,615** radio stations in the United States, according the Federal Communications Commission: 4,783 AM, 5,766 FM and 2,066 FM Educational stations. As noted above, the two formats broadcast different numbers of hours each day: 20 hours for FM stations, 16 hours for AM. Total US broadcasting hours would therefore be roughly **85 million hours** per year. Again, each hour of broadcasting would require 50 MB of storage, using the MP3 format. Total storage required for all US radio broadcasts is about **4,300 TB.**

About 84% of US radio stations have music as their primary focus, and provide little original content. (Source: Radio Marketing Guide & Fact Book, Radio Advertising Bureau 2000-2001) The remaining 16% are news, talk and religious stations, providing, presumably, almost 100% "original" information each day. Regardless of format, a percentage of most stations' broadcast time includes some commentary, weather reports, news updates, and traffic reports - perhaps an average of 5 minutes per hour. In addition, radio stations average between 12 and 16 minutes of commercials per hour. We can use this information to estimate how much "original" programming appears on the United States radio airwaves:**15.8 million hours**. (This estimate excludes advertisements and music.) The equivalent in bytes is **788 TB**.

Television

Conversion Factor

Satellite TV transmits between 3 and 5 mb per second. Therefore, one hour of broadcasting will require 10.8-18 Gbits of storage (compressed to MPEG-2). We can estimate about **1.3 - 2.25 GB** per hour of TV broadcasting. Source: Internet Archive.

World

There are **33,071 television stations** in the world, according to the CIA World Factbook 2000. If these stations broadcast about 16 hours per day, this would equal about 193 million hours total programming. We estimate about 1/4 of the programs are "original," - this is **48 million hours** each year. Estimating that one hour of video requires 1.3 GB of storage, then worldwide, program storage would be about **63,000 TB**; using the 2.25 GB estimate, it would be about **109,000 TB**.

United States

As of 1999, there are 1,616 broadcast television stations in the United States, according the US Federal Communications Commission. This figure includes the major networks (ABC, CBS, NBC, FOX, PBS and newcomers WB, UPN and PAX), the networks' affiliates, as well as local and public broadcasting stations. In addition, the National Cable Television Association reports that there are 210 national cable networks and 54 regional networks, as of August 2000. (54 more cable networks are planned but not yet operational.)

If all 1,884 of these stations broadcast 20 hours per day, that would equal just under **14 million hours** per year. We estimate that about 1/4 of the television programs broadcast are "original" - this is **3.5 million hours** each year, equivalent to **between 4,400 and 7,800 TB**.

Stock

In 55 years of programming, the networks have accumulated the following stock of material (*Source: Library of Congress Report, Television/Video Preservation Study: Volume 1: Report, October 1997*).

Table 3: Stock of Material Accumulated by the Major Networks.	
ABC	1,037,000 films/tapes
CBS	1,045,000 tapes and more than 150,000,000 feet of film
NBC	600,000 film reels (currently estimated at 100,000,000 feet) and 1,600,000 videotapes

Meanwhile, some of the major studios have accumulated original materials as well:

Table 4: Materials Accumulated by the Major Studios.	
Disney	6,500 television programs on 80,000 reels and tapes
Fox	54,000 television programs on 780,000 reels and tapes
MCA/Universal	18,000 (through 1994) television programs on 217,000 reels and tapes
Paramount (Viacom)	8,000 television programs on 1,200,000 reels and tapes
Sony/Columbia	35,000 television programs on 600,000 reels and tapes
Turner Entertainment	20,000 television programs on 337,000 reels and tapes
Warner Brothers	28,000 television programs on 1,000,000 reels and tapes

These figures overlap, of course, with those we have compiled for magnetic tape.

As of 1998, there are well over **18,000 hours** of programs in syndication available to be aired. This is equivalent to **18 TB** of information. (*Source: Television and Video Almanac 1998*)

Rate of Change

The number of radio stations and broadcast television stations in the United States has increased slightly - between 1 and 2% - every year since 1990.

While the number of cable television systems has decreased since 1993, the number of cable networks (i.e. channels) has increased annually. 54 new cable networks are scheduled to be launched in the coming year.

Copies

Radio

World

We estimate that FM radio stations broadcast 20 hours per day, AM stations 16 hours per day, and shortwave stations 12 hours per day. Therefore, there are approximately **290 million hours** (188 million FM, 98 million AM, and 6 million shortwave) of radio programming per year. Applying the 50 MB/hour rule of thumb, one may estimate an annual storage requirement of about **14,500 TB** if one were to record everything broadcast on the radio.

United States

As of 1999, there are **12,615 radio stations** in the United States, according the Federal Communications Commission: 4,783 AM, 5,766 FM and 2,066 FM Educational stations. As noted above, the two formats broadcast different numbers of hours each day: 20 hours for FM stations and 16 hours for AM. Total US broadcasting hours would therefore be roughly **85 million hours** per year. Again, each hour of broadcasting would require 50 MB of storage, using the MP3 format. Total storage required for all US radio broadcasts is about **4,300 TB**.

Television

World

There are about 33,000 television stations in the world (including some but not all cable stations), according to the CIA World Factbook 2000. This means that there are approximately **193 million hours** of television programming per year (assuming each station broadcasts 16 hours per day). Estimating that one hour of video requires 1.3 - 2.25 GB of storage, then worldwide, television would require between **250,000 and 435,000 TB**.

United States

As of 1999, there are 1,616 broadcast television stations in the United States as well as about 260 cable networks operating on nearly 10,500 cable systems. Each station broadcasts an average of 7,300 hours per year (again estimating 20 hours of broadcasting per day). If one wished to capture all of the programming generated by every station and cable system, regardless of duplication, it would be about **88 million hours**, requiring between **114,000 and 198,000 TB** of storage.

Fun Facts about Broadcast Media.

Television

- For many years, most large TV stations and the major networks subscribed to the Code of Good Practices of the National Association of Broadcasting, which established limits on the number of commercial minutes that could be telecast each hour. The limits were voluntary but widely followed: 9 1/2 minutes of commercials during primetime; higher amounts during other times of night and day. In 1992, however, the guidelines were ruled a violation of Federal antitrust law. Throughout the industry, most pledged to continue the limits - but gradually that eroded, as networks added more ad time. Prime time today has an average of 15 minutes of ads per hour. The FCC regulates advertising only during children's programming: 10.5 minutes/hour on weekends, 12 minutes/hour on weekdays. (*Source: Gerald Baldasty, University of Washington* *http://faculty.washington.edu/baldasty/Feb3.htm*)
- Approximately 7 in 10 television households, more than 65 million households, subscribe to cable. (*Source: National Cable Television Association*)

Radio

- There are currently more than 4,500 streaming radio stations on the Internet, distributed as follows: Africa (16); Asia (109); Europe (970); North America (2,786); Oceania (235); South America (166) and Internet Only (313). (*Source: RadioDirectory,* *http://www.radiodirectory.com/Stations/*)

- "Talk radio," in which celebrities and experts from various fields answer listener "call-in" questions and offer their advice on various topics, has grown spectacularly in recent years. The "call-in" format is the fastest-growing in radio, accounting for nearly 1,000 of the 10,000 commercial radio stations in the United States. (*Source: U.S. MEDIA IN THE 1990s: THE BROADCAST MEDIA By Fredric A. Emmert, US Information Agency*)
- 6,108 new radio stations have been started in United States since 1970 (*Source: Arbitron Ratings, http://www.arbitronratings.com*)
- The average US listener 12 and older hears approximately 1,100 hours of radio each year. (*Source: Arbitron Ratings, http://www.arbitronratings.com*)
- Since 1996, commercial spot loads have averaged 12-16 minutes per hour. These averages can be even higher during morning and afternoon drive hours. (*Source: "An Analysis of the Effects of Consolidation on the Radio Industry" http://mmstudio.gannon.edu/~gabriel/rapela.html.*)

Bibliography

- Arbitron Ratings http://www.arbitronratings.com/
- Castleman, Harry. Podrazik, Walter J. *The TV schedule book : four decades of network programming from sign-on to sign-off*. New York: McGraw-Hill, 1984.
- Central Intelligence Agency World Factbook 2000 http://www.odci.gov/cia/publications/factbook/
- Federal Communications Commission Audio Services Division, *Broadcast Station Totals*, 1990 - 1999 http://www.fcc.gov/mmb/asd/totals/index.html
- Fletcher, William H. Making Instructional Digital Video and Audio Work, http://miniappolis.com/mpeg/mpeg.html. April 1998.
- Internet Archive. Phone 415-561-6767 or see http://www.archive.org/
- National Cable Television Association http://www.ncta.com/directory.html
- *International Television and Video Almanac*. New York: Quigley Publishing, 1998.
- UNESCO, *Latest statistics on radio and television broadcasting*. Paris: UNESCO [United Nations Organization for Education, Science and Culture] Division of Statistics on Culture and Communication, 1987.
- U.S. Census Department, 1999 Statistical Abstract of the United States, www.census.gov/prod/www/statistical-abstract-us.html

See chart for more details

How Much Information?

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Phone - Summary

The material in this section is drawn from Coffman and Odlyzko (1998, 2000).

The International Telecommunications Union (ITU) database provides estimates of telephone traffic for 207 countries for 1997-98, which total 2.5 x 10^{12} minutes per year. Adding in an estimate for the missing countries brings us to 7.5 x 10^{12} minutes per year, or roughly 600,000 terabytes per month. Compression would reduce storage requirements by a factor of 6 to 8.

The US accounts for about 250,000 terabytes per month, of which roughly a third is modem calls.

We have somewhat better estimates for long-distance traffic in the US, including voice, Internet, public data networks and private lines.

Table 1: Traffic on U.S. long distance networks in terabytes, year-end 1999.

Network	Traffic (terabytes/month)
US voice	48,000
Internet	10,000 - 16,000
Other Public Data Networks	2,000
Private Line	5,000 - 8,500

Notes

- Internet traffic has, on average, been doubling every year for 30 years, though the growth rates varied significantly during that period.
- The current rate of growth of Internet traffic is roughly 100% per year.
- By 2002 data traffic will surpass phone traffic.
- Residential Internet use in the US is growing at about 30% a year. When residential users switch to broadband, the total volume of data accessed increases by a factor of 5-10. As residential broadband grows, traffic should approach the 100% per year growth rate.
- Internal corporate (intranet) traffic is growing at about 30% per year, but corporate traffic to the public Internet is growing at 100% per year.
- According to Peter Davidson, of Davidson Consulting, about 300 billion seconds of fax transmissions take place annually. At 45 seconds per page, this means 400 billion pages per year are faxed. At 15 Kb per page, this comes to around 6,000 terabytes of fax information per year.

References

- **Internet growth: Is there a "Moore's Law" for data traffic?,** K. G. Coffman and A. M. Odlyzko, 2000. [Abstract] [PostScript] [PDF] [LaTeX]

- **The size and growth rate of the Internet,** K. G. Coffman and A. M. Odlyzko, First Monday 3(10) (October 1998), http://firstmonday.org/. [Abstract] [PostScript] [PDF] [LaTeX] [First Monday version]

- **The Worldwide Fax Traffic Statistical Market Review and Forecast, 1997-2002,** Peter Davidson, [Link to Report]

How Much Information?

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Mail - Summary

This table of facts about US mail is from Odlyzko (2000). About half of all mail is currently first class and about half is junk mail. If we assume 5 pages per piece of mail, and digitize it at 15 Kbytes per page, 1998 US mail is about **150 petabytes** per year.

Table 1: Statistics about US mail service, from Odlyzko (2000).

Year	Cost (millions)	Cost/GDP (percent)	Pieces (millions)	Mail per Person
1790	0.032	0.02	0.8	0.20
1800	0.214	0.05	3.9	0.73
1810	0.496	0.09	7.7	1.07
1820	1.161	0.18	14.9	1.55
1830	1.933	0.21	29.8	2.32
1840	4.718	0.28	79.9	4.68
1850	5.213	0.20	155.1	6.66
1860	14.87	0.39		
1870	24.00	0.33		
1880	36.54	0.35		
1890	66.26	0.51	4,005	63.7
1900	107.7	0.58	7,130	93.8
1910	230.0	0.65	14,850	161
1920	454.3	0.50		
1930	803.7	0.89	27,887	227
1940	807.6	0.81	27,749	211
1950	2,223	0.78	45,064	299
1960	3,874	0.77	63,675	355
1970	7,876	0.81	84,882	418
1980	19,412	0.70	106,311	469
1990	40,490	0.70	166,301	669
1998	57,778	0.68	197,943	733

- **The history of communications and its implications for the Internet,** A. M. Odlyzko. [Abstract] [PostScript] [PDF] [LaTeX]

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Acknowledgements

Overall

- Rita Gildea, EMC
- Jim Gray, Microsoft
- Andrew Odlyzko, AT&T Labs
- Dave Patterson, UC Berkeley
- Gil Press, EMC
- Robert Wilensky, UC Berkeley

Print

- Eric Chaskas, US National Archives and Records Administration
- Merilyn Dunn at CAPventure, via Kathy Jarvis at XeroxParc
- Rebecca Green and Lee Leighton, UC Berkeley Library
- Richard J. Hill, Newspapers Librarian, State Library of Pennsylvania
- Aimee Pyle and Amy Kirschoff, JSTOR
- Ginger Ogle, UC Berkeley Digital Library Project
- Rebecca Green and Lee Leighton, UC Berkeley Library
- Clara R. Williams, Information Consultant at the Hasleton Library, Institute of Paper Science and Technology

Optical

- Brewster Kahle, Internet Archive
- Brian Lewin, International Recording Media Association
- Jim Taylor, DVD Association

Magnetic

- Fred Moore, Horizon Information Strategies

Telephone

- Peter Davidson, Davidson Consulting

Internet

- John McCredie and Jerry Berkman, UC Berkeley
- Steve Lawrence, NTT
- Scott Kirkpatrick, Archive.org

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Site Map

Sound Bytes

Terror of terabytes.
One terabyte, the smallest practical measure for our project, is a million megabytes, which is equivalent to the textual content of a million books. An exabyte, which is what we use to report the final results, is a billion gigabtyes.

Capabilities of compression.
Conversion to ASCII, MP3, MP4 and other compression technologies dramatically reduces storage requirements by one to two orders of magnitude.

Democratization of data.
Individuals produce significant amounts of non-digital information. As photos and videos move to digital formats, households will have to manage terabytes of data.

Dominance of digital.
Ninety-three percent of the information produced each year is stored in digital form. Hard drives in stand-alone PCs account for 55% of total storage shipped each year.

Magnetic migration.
Print and film content is rapidly moving to magnetic and optical storage. This is true for professional use now, and will become increasingly true at the level of individual users.

Tape in transition.
Magnetic tape is about 10 times as large as disk storage, but is used almost exclusively for archives. Disk storage is much more attractive, even for archives, due to its rapidly declining cost and the fact that it is much easier to access data stored on disk.

Paucity of print.
If all printed material published in the world each year were expressed in ASCII, it could be stored in less than 5 terabytes.

Immensity of images.
Over 80 billion photographs are taken every year, which would take over 400 petabytes to store, more than 80 million times the storage requirements for text.

Convenience of copies.
There is a lot of redundancy both across and within media. A newspaper, for example, is composited using digital technology, printed on paper, then archived on microfilm. Estimates of ``unique'' information can only be taken as approximate.

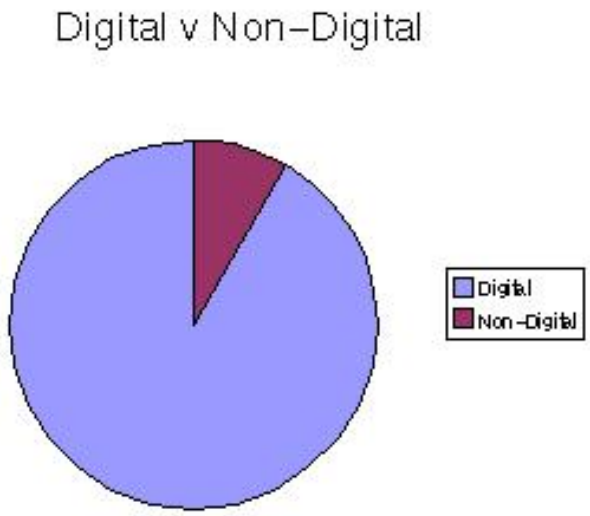
Ubiquity of the US.
The US produces 35% of all print material, 40% of the images and well over 50% of the digitally stored content produced in the world each year.

How Much Information?

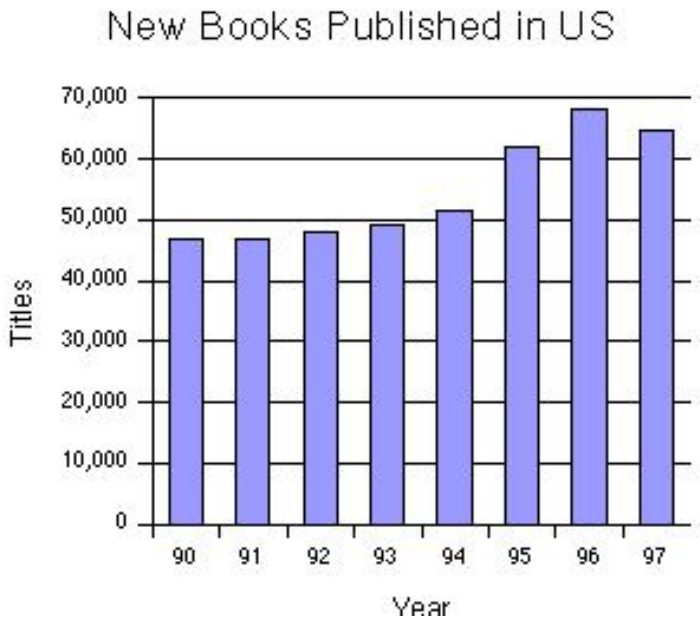
About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Charts

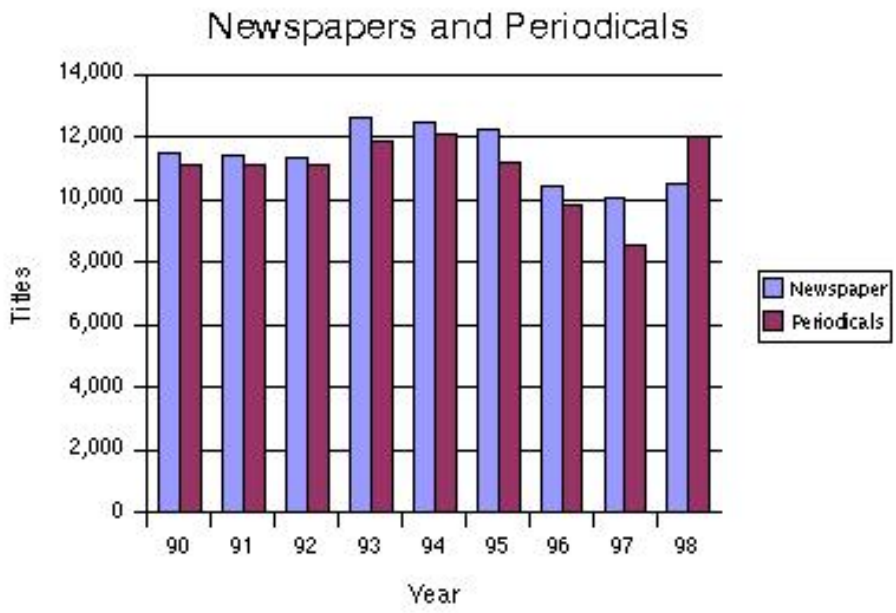
Here are some charts that show the relative sizes of various interesting magnitudes.



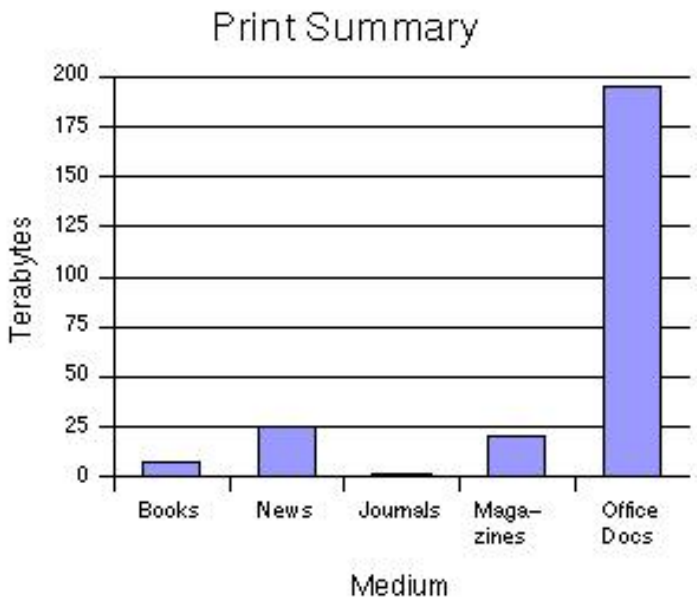
Over 93 percent of the information produced in 1999 was in digital format. (Authors' calculations.)



New books published in US. (*US Statistical Abstract 1999*, Table 938. Some years interpolated.)



Newspapers and periodicals published in US. (*US Statistical Abstract 1999*, Table 941.)



Print summary. (Calculations by authors.)

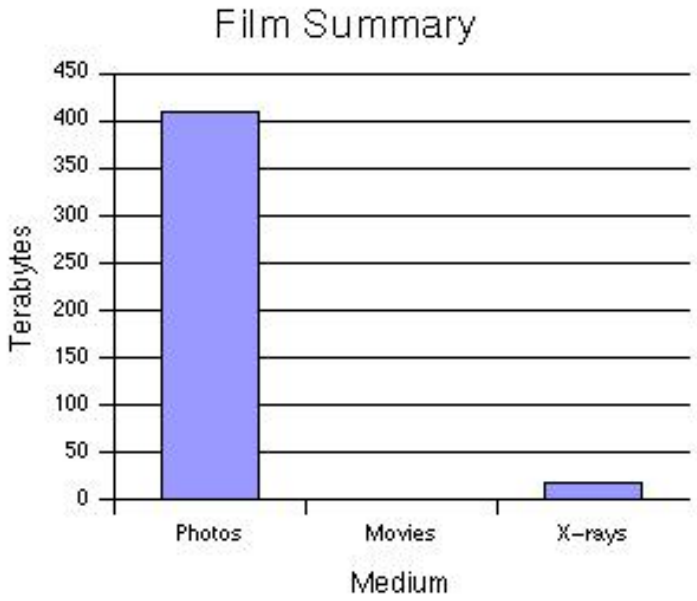
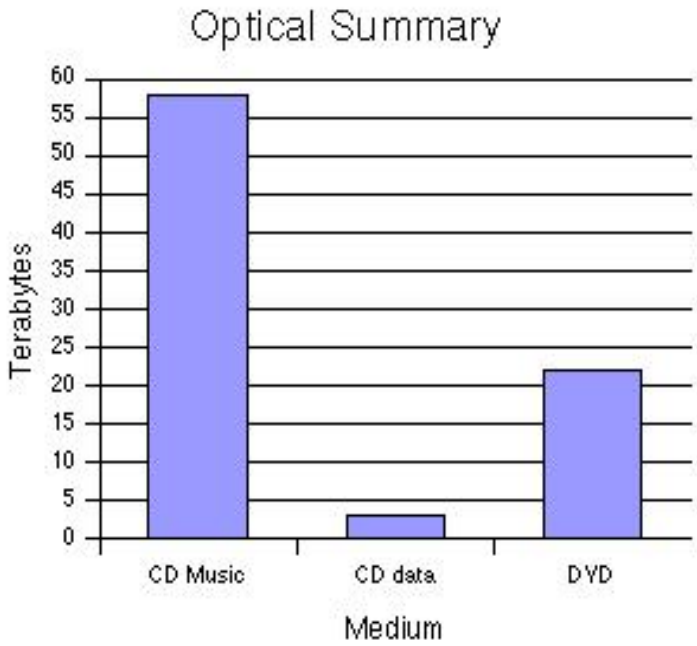
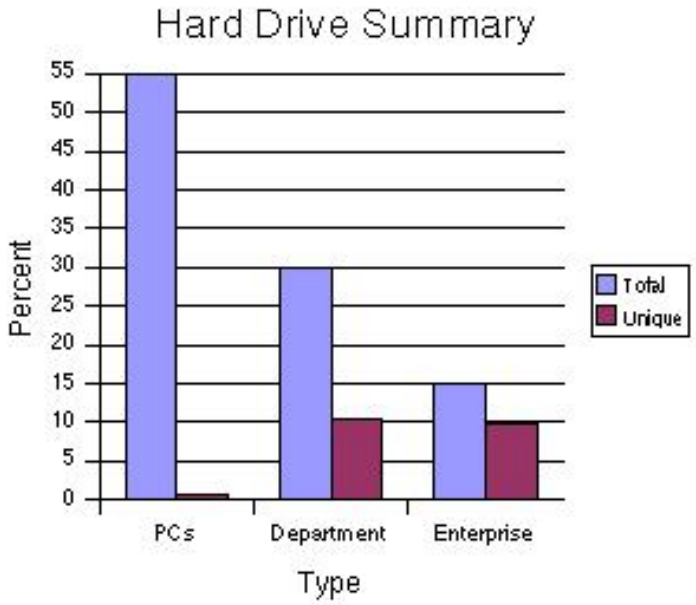


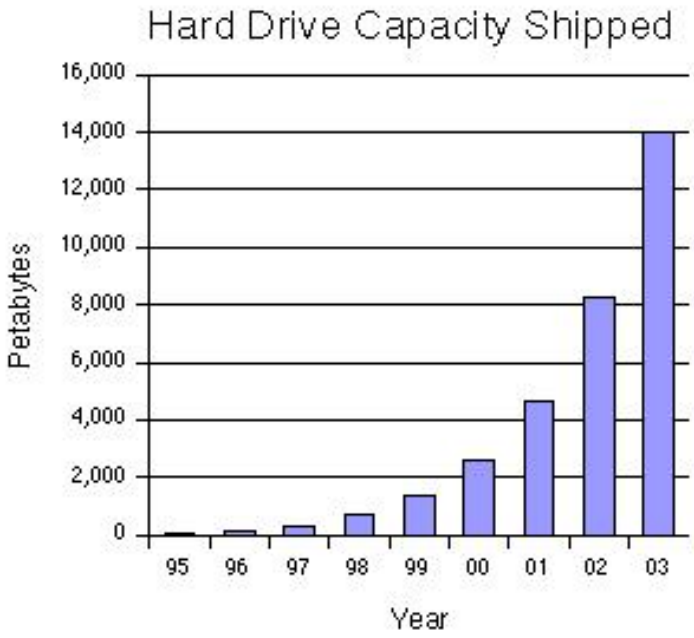
Photo summary. (Calculations by authors.)



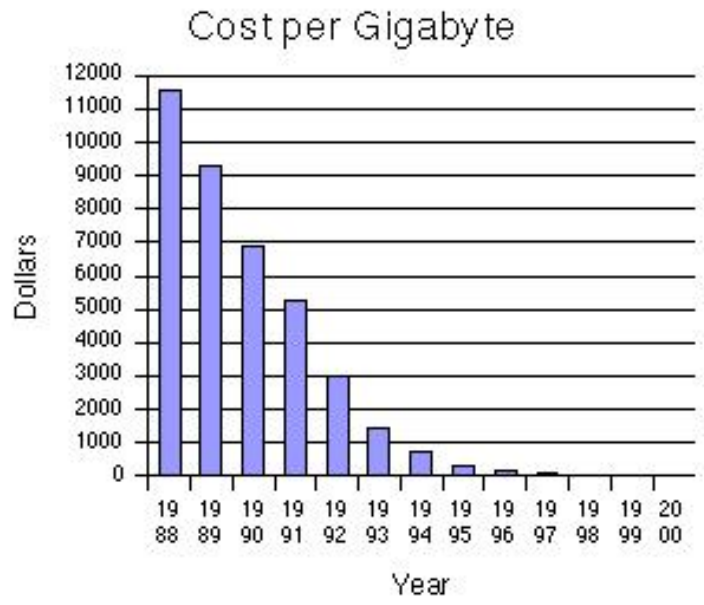
Optical summary. (calculations by authors)



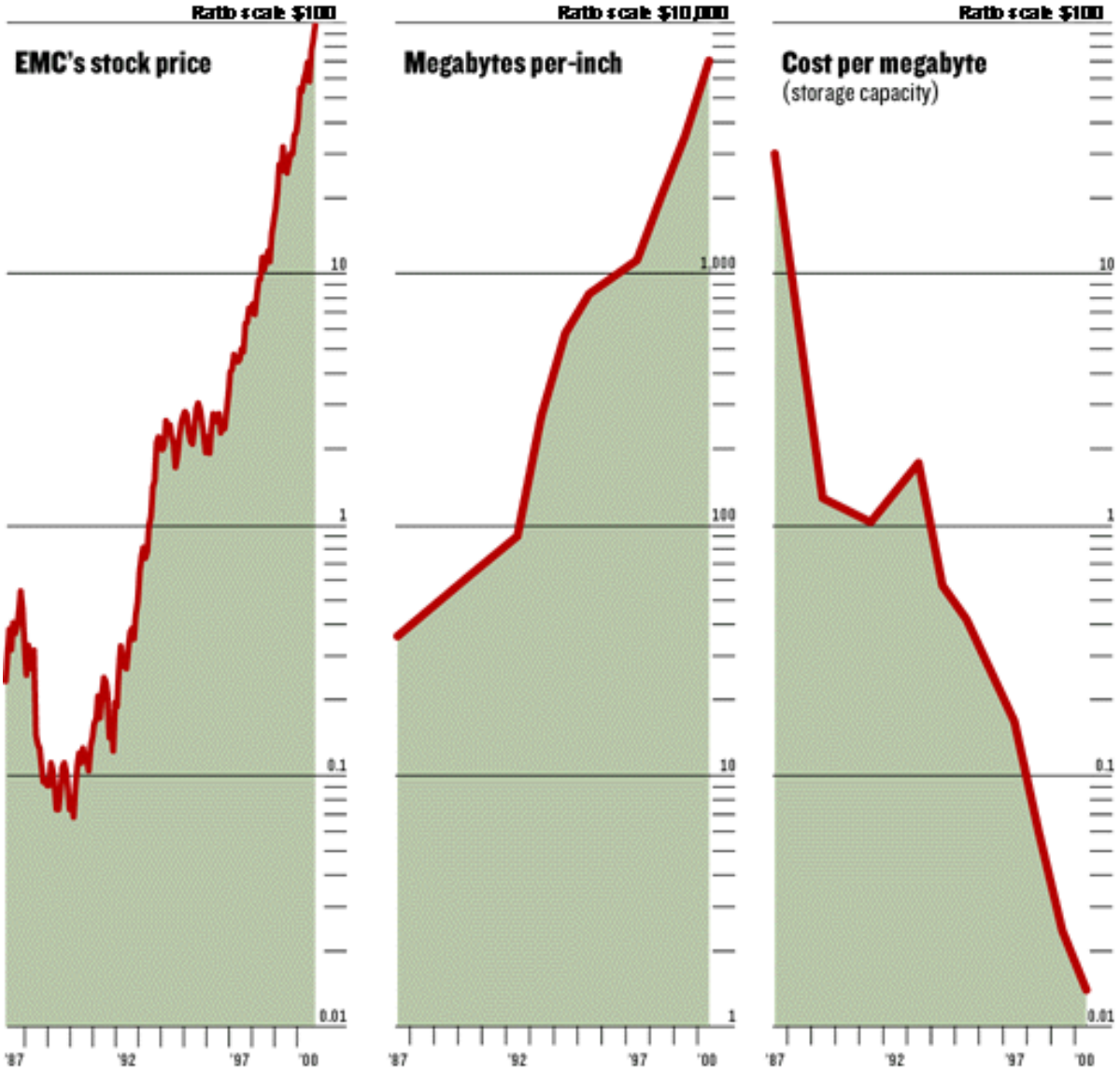
Total and unique information on various types of hard drives. (Calculations by authors)



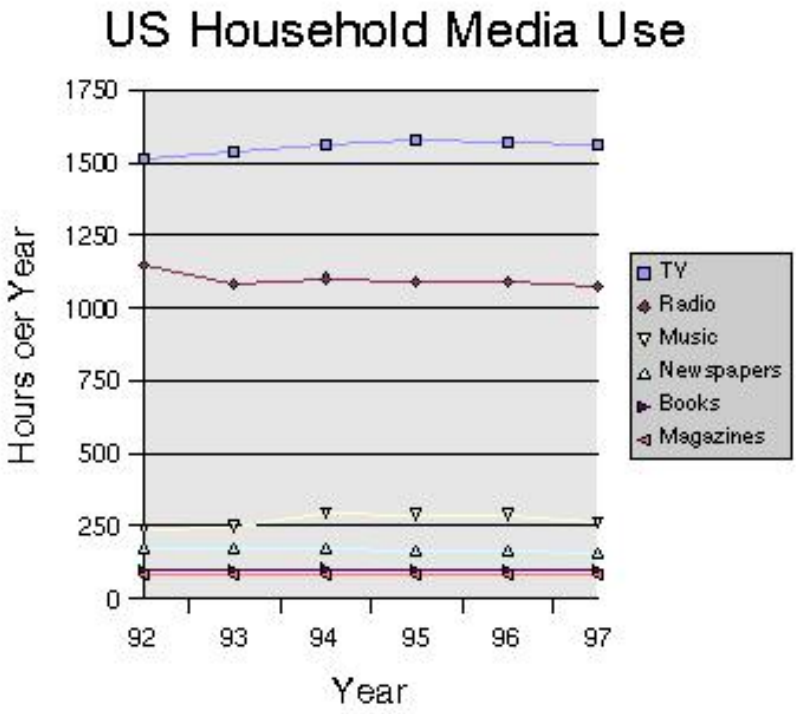
Worldwide PC hard drive capacity shipped. (*1999 Winchester Disk Drive Market Forecast and Review*, Table C5, International Data Corporation report. Some years forecast.)



Hard drive cost per gigabyte. (IDC reports.)

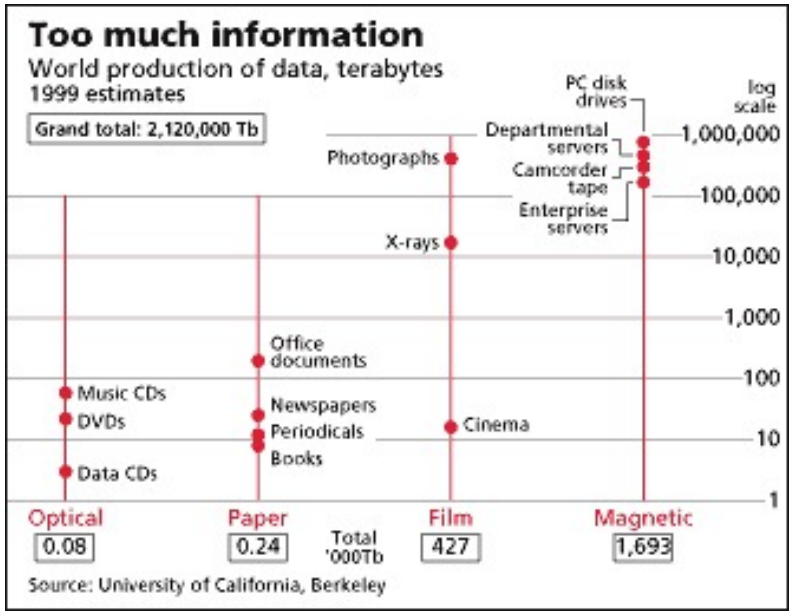


Trends in EMC's stock price, megabytes per inch, and cost per megabyte. From Forbes, October 2, 2000.

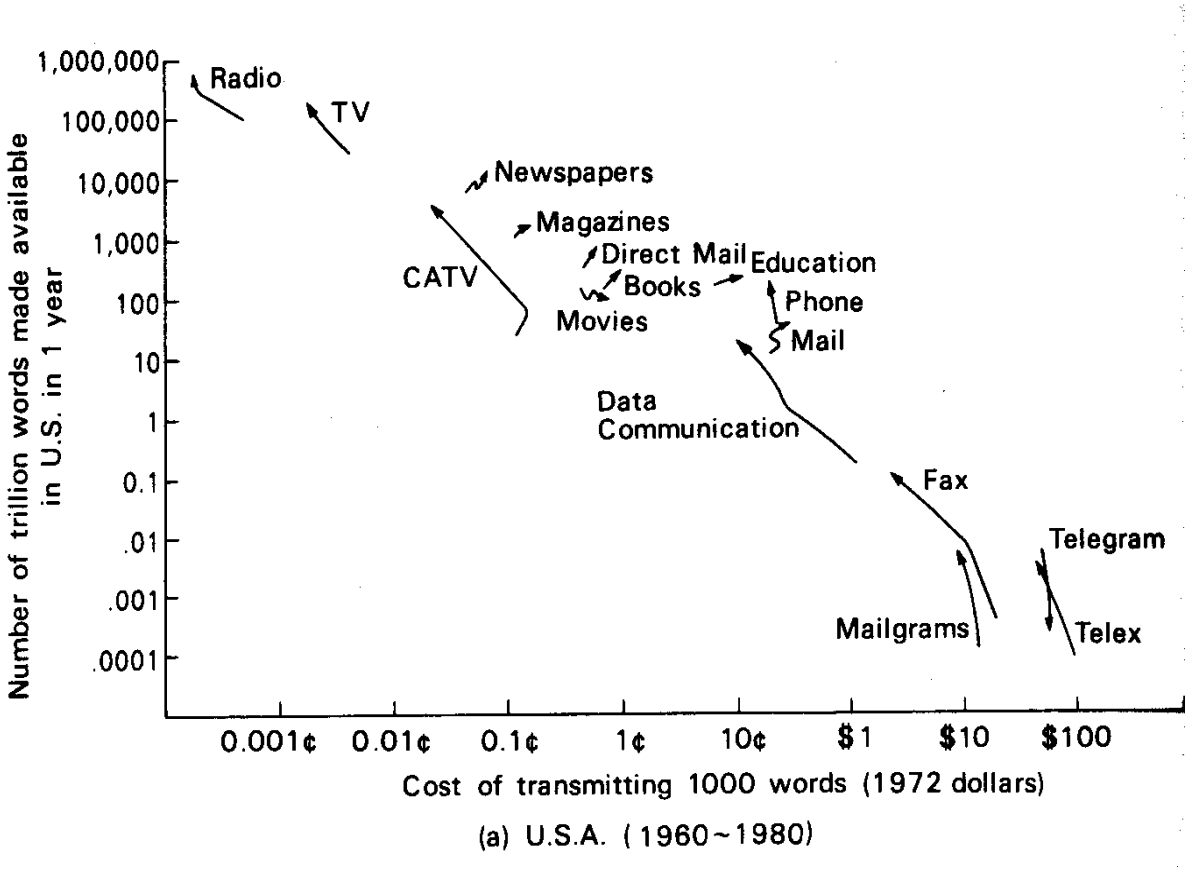


US household media use. (*US Statistical Abstract 1999*, Table 920.)

Activities performed by online households in the US during an average week, Q2, 2000. (*Economist Magazine*, October 7, 2000, E-Entertainment Survey, page 11.)



Summary of upper estimate of worldwide information production. (*Economist Magazine*, October, 2000)



Cost of transmitting 1000 words v Number of words available in USA 1960-1980. (Ithiel de Sola Pool, Hiroshi Inose, Nozomu Takasaki, Roger Hurwitz, *Communication Flows: A Census in the United States and Japan,* Elsevier Science Publishers, New York, 1984.)

About the Project

Senior Researchers: Peter Lyman and Hal R. Varian
Research Assistants: James Dunn, Aleksey Strygin, Kirsten Swearingen

This study is an attempt to measure how much information is produced in the world each year. We look at several media and estimate yearly production, accumulated stock, rates of growth, and other variables of interest.

If you want to understand what we've done, we offer different recommendations, depending on the degree to which you suffer from *information overload:*

Heavy information overload: *the world's total yearly production of print, film, optical, and magnetic content would require roughly 1.5 billion gigabytes of storage. This is the equivalent of 250 megabytes per person for each man, woman, and child on earth.*

Moderate information overload: read the *Sound Bytes* and look at the *Charts* illustrating our findings.

Normal information overload: read the *Executive Summary.*

Information deprived: read the detailed reports by clicking on the contents to your left. Or download the entire Web site as a PDF file. (It is about 100 pages.)

This study was produced by faculty and students at the School of Information Management and Systems at the University of California at Berkeley. We gratefully acknowledge financial support from EMC. We have put "[???]" in the text where we had to make "questionable" assumptions. If you have suggestions, corrections, or comments, please send email to how-much-info@sims.berkeley.edu. We view this as a "living document" and intend to update it based on such contributions.

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Executive Summary

Abstract

The world produces between 1 and 2 exabytes of unique information per year, which is roughly 250 megabytes for every man, woman, and child on earth. An exabyte is a billion gigabytes, or 10^{18} bytes. Printed documents of all kinds comprise only .003% of the total. Magnetic storage is by far the largest medium for storing information and is the most rapidly growing, with shipped hard drive capacity doubling every year. Magnetic storage is rapidly becoming the universal medium for information storage.

Introduction

The cost of magnetic storage is dropping rapidly; as of Fall 2000 a gigabyte of storage costs less than $10 and it is predicted that this cost will drop to $1 by 2005. Soon it will be technologically possible for an average person to access virtually all recorded information. The natural question then becomes: how much information is there to store? If we wanted to store "everything," how much storage would it take?

We have conducted a study to answer this question. In particular, we have estimated yearly US and world production of originals and copies for the most common forms of information media. We have also attempted to estimate the cumulated stock of information in various formats. Finally, we have described the magnitudes of some communication flows that are currently not stored but may well be

in the future.

Information produced by medium

Most information is stored in four physical media: paper, film, optical (CDs and DVDs), and magnetic. There are very good data for the worldwide production of each storage medium, and there are reasonably good estimates of how much original content is produced in each of these different formats.

We have identified production of content by media type, translated the volume of original content into a common standard (terabytes), determined how much storage each type takes under certain assumptions about compression, attempted to adjust for duplication of content, and added up to get total estimates.

Table 1 depicts yearly worldwide production of original stored content as of 1999. In general, the upper estimate is based on the raw data, while the lower estimate reflects an attempt to adjust for duplication and compression. We discuss these adjustments below and in the medium-specific documents. Note that the growth rate estimates are very rough. See the "Qualifications" section and Appendix A for further discussion; the details of the calculations are presented in the accompanying documents.

Table 1: Worldwide production of original content, stored digitally using standard compression methods, in terabytes circa 1999.

Storage Medium	Type of Content	Terabytes/Year, Upper Estimate	Terabytes/Year, Lower Estimate	Growth Rate, %
Paper	Books	8	1	2
	Newspapers	25	2	-2
	Periodicals	12	1	2
	Office documents	195	19	2
	Subtotal:	**240**	**23**	**2**
Film	Photographs	410,000	41,000	5
	Cinema	16	16	3
	X-Rays	17,200	17,200	2
	Subtotal:	**427,216**	**58,216**	**4**
Optical	Music CDs	58	6	3
	Data CDs	3	3	2
	DVDs	22	22	100
	Subtotal:	**83**	**31**	**70**
Magnetic	Camcorder Tape	300,000	300,000	5
	PC Disk Drives	766,000	7,660	100
	Departmental Servers	460,000	161,000	100
	Enterprise Servers	167,000	109,000	100
	Subtotal:	**1,693,000**	**635,660**	**55**
TOTAL:		**2,120,539**	**693,930**	**50**

Three striking facts emerge from these estimates. The first is the "paucity of print." Printed material of all kinds makes up less than .003 percent of the total storage of information. This doesn't imply that print is insignificant. Quite the contrary: it simply means that the written word is an extremely efficient way to convey information.

The second striking fact is the ``democratization of data.'' A vast amount of unique information is created and stored by individuals. Original documents created by office workers are more than 80% of all original paper documents, while photographs and X-rays together are 99% of all original film documents. Camcorder tapes are also a significant fraction of total magnetic tape storage of unique content, with digital tapes being used primarily for backup copies of material on magnetic drives.

As for hard drives, roughly 55% of the total are installed in single-user desktop computers. Of course, much of the content on individual user's hard drives is not unique, which accounts for the large difference between the upper and lower bounds for magnetic storage. However, as more and more image data moves onto hard drives, we expect to see the amount of digital content produced by individuals stored on hard drives increase dramatically.

This democratization of data is quite remarkable. A century ago the average person could only create and access a small amount of information. Now, ordinary people not only have access to huge amounts of data, but are also able to create gigabytes of data themselves and, potentially, publish it to the world via the Internet, if they choose to do so.

The third interesting finding is the "dominance of digital" content. Not only is digital information production the largest in total, it is also the most rapidly growing. While unique content on print and film are hardly growing at all, optical and digital magnetic storage shipments are doubling each year. Even today, most textual information is "born digital,'' and within a few years this will be true for images as well. Digital information is inexpensive to copy and distribute, is searchable, and is malleable. Thus the trend towards democratization of data---especially in digital form---is likely to continue.

Qualifications

It goes without saying that the numbers in Table 1 can only be taken as rough estimates. We have had to make various assumptions in order to construct our these figures, and some data sources are contradictory or simply not available. Here we list some of the most serious methodological qualifications, each of which offers interesting challenges for those who would seek to refine these estimates.

Duplication.

It is very difficult to distinguish "copies" from "original" information. A newspaper, for example, is published on paper, often published on the Web as well, and is generally archived on microfilm. In fact, most printed materials are produced and/or archived magnetically. There is also lot of duplication within each medium: many newspapers reproduce stock prices, wire stories, advertisements and so on. Ideally, we would like to measure the storage required for the *unique* content in the newspaper, but it is very hard to measure that number. As indicated above, the duplication issue is particularly serious for digital storage, since little of what is stored on individual hard drives is unique. We've tried to adjust for this the best we can, and documented our assumptions in the detailed treatment of each medium.

Compression.

Unlike print or film, there is no unambiguous way to measure the size of digital information. A 600 dot per inch scanned digital image of text can be compressed to about one hundredth of its original size. A DVD version of a movie can be 1000 times smaller than the original digital image. We've made what we thought were sensible choices with respect to compression, steering a middle course between the high estimate (based on ``reasonable'' compression) and the low estimate (based on highly compressed content). It is worth noting that the fact that digital storage can be compressed to different degrees depending on needs is a

significant advantage for digital over analog storage.

Archival Media.

Should information stored as "backup" be included in the total? This question arises for microfilm, rewritable CD ROMS, and even with print, but digital magnetic tape is the most difficult case. Tape's most common use is to archive material on hard drives and therefore should not count towards the stock of ``original information'' produced each year. Industry rules of thumb suggest that there is about 10 times as much storage on tape as on hard drives. This fraction has been falling as more and more data is stored on arrays of hard drives, which are much more convenient to use. We've omitted most tape storage for this reason. However, we should also note that vast quantities of original scientific data are stored in tape libraries; we describe a few such repositories in the detailed treatment of magnetic storage.

World and US production.

The US produces about 25% of all textual information and about 30% of the photographic information, a significant fraction of the world's total. We don't have good data on magnetic storage, but it seems plausible that the US produces at least half of the content stored on magnetic media. We've used numbers for world production when available, but in some cases have had to extrapolate from US production. Little data is available about information production in the Third World.

Growth rates.

The production of unique content in books, photos, and CDs is barely growing. DVD content is growing rapidly, but that's because it is a new medium and a significant amount of legacy content is being converted. By contrast, shipments of digital magnetic storage are essentially doubling every year.

TV and Radio.

Original TV content produced each year is generally stored on magnetic camcorder tapes, and so is counted in that category of storage media. Much radio content is simply broadcast music, which we have already captured with the CD statistics. See Table 3 for information on how much storage it would take to back up all TV and radio broadcasts, with minimal adjustment for duplication.

Digital Communication

Our project is primarily concerned with content that is stored, either by institutions or by individuals. But there is a lot of material that is communicated, without being systematically stored. Some of this material is born digital, such as email, Usenet, and the Web. Some of it is non-digital, such as telephone calls and letters.

We expect that digital communications will be systematically archived in the near future, and thus will contribute to the demand for storage. Table 2 shows how much storage would be required to archive the major forms of digital communication.

Table 2: Summary of yearly unique computer-mediated information flows.

Content	Terabytes
Email	11,285
Usenet	73

In 2000 the World Wide Web consisted of about 21 terabytes of static HTML pages, and is growing at a rate of 100% per year. Many Web pages are generated on-the-fly from data in databases, so the

total size of the "deep Web" is considerably larger.

Although the social impact of the Web has been phenomenal, about 500 times as much email is being produced per year as the stock of Web pages. It appears that about 610 billion emails are sent per year, compared to 2.1 billion static Web pages. Even the yearly flow of Usenet news is more than 3 times the stock of Web pages. As Odlyzko (2000) puts it, "communication, not content, is the killer app."

Non-digital Communication

We also estimated the storage requirements if one attempted to archive all the non-digital communication flows in the United States. We consider only the US since we didn't have very good data for worldwide communication. The results are shown in Table 3.

Table 3: Summary of yearly non-digital communication flows in the United States 1999.

Content	Terabytes
Radio	788
TV	14,150
Telephone	576,000
Postal	150,000

The striking thing here is the volume of voice telephone traffic, most of which is presumably unique content. Radio and TV, by contrast, have a huge amount of duplication from station to station, since many of the broadcasts are reusing the same content.

Consumption of Information

Though the main focus of our report is on the supply of information, it is interesting to look at data measuring the consumption of information as well. Table 4 depicts hours per year of time spent on various media in US households in 1992 and in 2000. We do not have good data on information use in the workplace.

Table 4: Summary of yearly media use by US households in hours per year, with estimated megabyte equivalent. (Hours from Statistical Abstract of the United States, 1999, Table 920, (projected)).

Item	1992 Hours	2000 Hours	2000 MBytes	% Change
TV	1510	1571	3,142,000	4
Radio	1150	1056	57,800	-8
Recorded Music	233	269	13,450	15
Newspaper	172	154	11	-10
Books	100	96	7	-4
Magazines	85	80	6	-6
Home video	42	55	110,000	30
Video games	19	43	21,500	126

Internet	2	43	9	2,050
Total:	**3,324**	**3,380**	**3,344,783**	**1.7**

The notable features of this table are 1) the hours spent on TV and radio consumption and their consistency over time; 2) the reduction in time spent on printed information; and, 3) the dramatic increase in home video, video games, and Internet usage. However, it is important to note that the latter three categories are still very small in terms of total hours.

It is also noteworthy that total time spent in media access has hardly changed in eight years. Even while information supply is growing dramatically (especially in electronic media) the actual consumption of information is barely changing: a smaller and smaller fraction of what is produced is actually consumed, on average, a trend noted by Pool (1984). Census data indicate that over 40% of the US population has access to the Internet, so this trend is likely to increase.

Individual and Published Information

We remarked above that technological advances have allowed for a "democratization of data:" individuals can now generate a huge amount of information on their own. Table 5 summarizes the yearly production of information by and about individuals.

Table 5: Yearly production of individual information

Item	Amount	Terabytes
Photographs	80 billion images	410,000
Home videos	1.4 billion tapes	300,000
X-rays	2 billion images	17,200
Hard disks	200 million installed drives	13,760
Total:		**740,960**

The production of individual information can be compared to the amount of "published" information in Table 6. Note that the amount of "individual" information is over 600 times larger as the amount of published information.

Although the Web, Usenet, and email include a great deal of individual information, they have been omitted from both of these tables, since it is difficult to know whether to classify this material as "individual" or "public." In the future we expect the distinction between "individual" and "public" to become increasingly blurred.

Table 6: Yearly production of published information

Item	Titles	Terabytes
Books	968,735	8
Newspapers	22,643	25
Journals	40,000	2
Magazines	80,000	10
Newsletters	40,000	.2
Office Documents	7,500,000,000	195
Cinema	4,000	16

Music CDs	90,000	6
Data CDs	1,000	3
DVD-video	5,000	22
Total:		**285**

Conclusion

The world's total production of information amounts to about 250 megabytes for each man, woman, and child on earth. It is clear that we are all drowing in a sea of information. The challenge is to learn to swim in that sea, rather than drown in it. Better understanding and better tools are desperately needed if we are to take full advantage of the ever-increasing supply of information described in this report.

About this Report

Financial support for this study was provided by EMC. We view this report as a "living document" and intend to revise it based on comments, corrections, and suggestions. Please send such materials to how-much-info@sims.berkeley.edu.

About the School of Information Management and Systems

UC Berkeley's School of Information Management and Systems is the first school in the nation to explicitly address the growing need to manage information more effectively.

With respect to education, we are training a new type of professional: "information managers". Our graduates are familiar with the latest and most powerful techniques for locating, organizing, retrieving, manipulating, protecting, and presenting information. They study not only technology, but also the institutional, legal, economic and organizational factors necessary for creating information systems that meet peoples' needs.

With respect to research, we are examining ways to build more effective tools and systems for managing information. This effort is inherently multidisciplinary, involving computer science, information science, social science, cognitive science, and legal studies.

Appendices

- **A. Powers of Ten**
  The Powers of Ten table is helpful in illustrating the relative size of gigabytes, terabytes, petabytes and the like.

- **B. Upper and lower estimates**
  The upper estimate is a reasonably "hard" number; based on published data. The lower estimate is an attempt to adjust for duplication and compression. Here is a quick summary of some of those adjustments.

  - **Paper.**
    There is some duplication with ISBN numbers due to paperback, hardback, different editions, etc. There is duplication with financial papers, ads, and so on in newspapers. We used CPC compression, which captures images; conversion to ASCII eliminates images, but compresses text dramatically.

  - **Film.**
    If we used JPEG compression, rather than PhotoCD, we get a much smaller number

for the storage requirements for images.

- **Music CDs.**
  If we use MP3 compression we get a much smaller number for the storage requirements of audio files.

- **Magnetic.**
  We assume that about 20 percent of magnetic storage is unique.

- **C. Reading the data**
  The left-side navigation and Site Map provide links to summary reports on each medium. The summaries provide links to detailed reports and spreadsheets containing the raw data.

  Within each media type, we have distinguished between originals and copies, and between the yearly flow of production and the accumulated stock. We've also described growth rates and compression issues for each medium.

- **D. Acknowledgements**
  Gray and Shenoy (2000) provides useful information on trends in magnetic storage. Lesk (1997) conducted an earlier study that attempted to estimate the total stock of information. Pool (1984) examined the flow of information in the US circa 1980. See the individual acknowledgements for the names of people who helped us.

Bibliography

- Jim Gray and Prashant Shenoy.
  Rules of thumb in data engineering.
  in *Proceedings of 16th International Conference on Data Engineering, pages 3-12. IEEE, 2000.*
  http://www.research.microsoft.com/~Gray/.

- Michael Lesk.
  How much information is there in the world?
  Technical report, lesk.com, 1997.
  http://www.lesk.com/mlesk/ksg97/ksg.html.

- Andrew Odlyzko.
  Content is not king.
  Technical report, AT&T Labs, 2000.
  http://www.research.att.com/~amo/doc/networks.html.

- Ithiel De Sola Pool, Hiroshi Inose, Nozomu Takasaki, Roger Hurwitz.
  Communications flows : a census in the United States and Japan.
  Elsevier Science, New York, 1984.

- Ithiel De Sola Pool. "Tracking the Flow of Information".
  Science (12 August), 1983, 221:4611, 609-613.

- U.S. Census Bureau.
  Statistical Abstract of the United States, 1999
  Washington, D.C., 1999.
  http://www.census.gov/prod/www/statistical-abstract-us.html

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Print Media - Summary

Originals

Flow

Approximately **240 terabytes** (compressed) of unique data are recorded on printed media worldwide each year, as shown in the following table:

Table 1: World Flow

Media Type (Sources and Year Cited)*	Unique Items per Year	Conversion Factor	Total Terabytes (Annual Worldwide)
Books (UNESCO 1996)	**968,735**	Scanned image (600 dpi): 40 MB/book	39
		Digital compression: 8 MB/book	8
		Plain text: 1 MB/book	1

Newspapers (ISSN 1999)	**22,643**	Scanned image (600 dpi): 5,475 MB/year	124
		Digital compression: 1095 MB/year	25
		Plain text: 110 MB/year	2.5
Scholarly journals (Ulrich's 2000)	**40,000**	Scanned image (600 dpi): 225 MB/year	9
		Digital compression: 45 MB/year	2
		Plain text: 4 MB/year	.2
Mass-market periodicals (Ulrich's 2000)	**80,000**	Scanned image (600 dpi): 650 MB/year	52
		Digital compression: 130 MB/year	10
		Plain text: 13 MB	1
Newsletters (Oxbridge Directory 1997)	**40,000**	Scanned image (600 dpi): 20 MB/item	.8
		Digital compression: 4 MB/item	.2
		Plain text: .4 MB/item	.02
Archivable, original office documents (National Archives 1998)	**7.5 X 10^9 pages**	Scanned image (600 dpi): 130 KB/page	975
		Digital compression: 26 KB/page	195
		Plain text: 2.5 KB/page	19
		Totals:	**Scanned: 1200 TB**
			Compressed: 240 TB
			Text: 24 TB

* Detailed source information listed at end of this document.

Table 2: United States Flow

Media Type (Sources and Year Cited)*	**Unique Items per Year**	**Conversion Factor**	**Total Terabytes** (Annual Worldwide)
Books (US Statistical Abstract 1999)	**64,711**	Scanned image (600 dpi): 40 MB/book	3
		Digital compression: 8 MB/book	.5
		Plain text: 1 MB/book	.05
Newspapers (Newspaper Association of America)	**2,386**	Scanned image (600 dpi): 5,475 MB/year	13
		Digital compression: 1095 MB/year	3
		Plain text: 110 MB/year	.3
Scholarly journals	**10,500**	Scanned image (600 dpi): 225 MB/year	2

(Tenopir and King)		Digital compression: 45 MB/year	.5
		Plain text: 4 MB/year	.04
Mass-market periodicals (Ulrich's 2000)	**20,000**	Scanned image (600 dpi): 650 MB/year	13
		Digital compression: 130 MB/year	2.6
		Plain text: 13 MB	.26
Newsletters (NEPA)	**10,000**	Scanned image (600 dpi): 20 MB/item	.2
		Digital compression: 4 MB/item	.04
		Plain text: .4 MB/item	.004
Archivable, original office documents (National Archives 1998)	**3 X 10^9 pages**	Scanned image (600 dpi): 130 KB/page	390
		Digital compression: 26 KB/page	78
		Plain text: 2.5 KB/page	7.5
		Totals:	**Scanned: 421 TB**
			Compressed: 84 TB
			Text: 8.2 TB
* Detailed source information listed at end of this document.			

Notes on Conversion Assumptions

Books. Estimate 300 pages per book. (*Source: Robert M. Hayes, UCLA, "The Economics of Digital Libraries" www.usp.br/sibi/economics.html*)

Newspapers. Estimate 30 pages per newspaper, then multiply by 365 days per year. (*The page number is low, to reflect the number of small and non-daily newspapers published around the world.*)

Scholarly Journals. Estimate 1,700 pages per periodical per year. (*Source: Donald W. KLing and Carol Tenopir. "Economic Cost Models of Scientific Scholarly Journals," 1998. www.bodley.ox.ac.uk/icsu/kingppr.htm*)

Mass Market Periodicals. Estimate 5,000 pages per periodical per year. (*Source: Robert M. Hayes, UCLA, "The Economics of Digital Libraries" www.usp.br/sibi/economics.html*)

Newsletters. Estimate 150 pages per newsletter per year. (*Source: Oxbridge Directory of Newsletters - 1997*)

Office documents. The estimate above is limited to documents that an organization might retain permanently such as documents comparable to those retained by the National Archives in Washington D.C., which estimates that they retain 2% of US government documents produced each year.

More detail on the conversion factors used for the above estimates appears in the Print Detail Report.

Stock

United States

According to a press release from January 2000, booksinprint.com 2000 includes **3.2 million titles** - about **26 TB** total. This figure is supported by online booksellers such as Amazon.com and Barnes&Noble.com who claim to offer access to 3 to 4 million titles.

If one wished to more fully address the universe of book titles in the United States, including those that are no longer in print, one could look to the holdings of the larger national libraries and copyright repositories - for example, the Library of Congress print media collection includes almost **26 million books (208 terabytes).**

World

To estimate the international stock of books currently available for purchase, we extrapolate from the United States production figures. The US engages in the world's largest trade in printed products, producing about 40% of the world's printed material, according to the US Industry and Trade Outlook 2000. The world stock of original titles might be about **8 million titles** - equivalent to **64 TB.**

Using the same 40% rule of thumb, we can also estimate the worldwide stock of books (including those out of print). The national library and copyright repository of the United States - the Library of Congress - contains about 26 million books. Therefore, the world stock of books might be approximately **65 million titles.**

Rate of change

The number of titles within most print media forms have increased each year worldwide - between 2 and 10%. Within the US, the number of book titles increased every year until 1996, when there was a 5% downturn.

Copies

Flow

If all of the writing paper and newsprint produced each year were used to store printed information, this would be equivalent to about **980,000 terabytes** worldwide.

World

As of 1997, the world was producing 90 million metric tons of printing and writing paper and 36 million metric tons of newspaper. In equivalent bytes, this translates to **540,000 TB** (world) for printing and writing paper, and **432,000 TB** for newsprint.

The number of books sold worldwide may be estimated using the 40% rule of thumb (cited above) and the US book sales statistics (cited below): about **2.75 billion books**, equivalent to **22,000 TB**.

United States

The US produces about 30% of the world's paper and paperboard output (*Source: US Industry & Trade Outlook 2000*). In 1999, the US produced 23.8 million metric tons of printing and writing paper and 6.4 million metric tons of newsprint. In bytes, this translates to **142,800 TB** for printing and writing paper and **76,800 TB** for newsprint. These figures provide an upper bound on the total number of bytes required to digitally store all the information produced in printed format each year.

About **1.1 billion books** were sold in the United States in 1999. Using the 8 MB/book estimate, this is equivalent to **8,800 TB**. (*Source: Wall Street Journal, July 17, 2000, "A New Chapter: Independent Booksellers Hope to Find Strength in Numbers" by Scott Eden.*)

In the United States **55,979,332 daily newspapers** and **59,894,381 Sunday newspapers** circulate each year. (Source: *Newspaper Association of America, citing Editor and Publisher.*)

The total number of US magazines circulated annually exceeds **500 million**. (*Source: US Industry and Trade Outlook.*)

Each year, almost **500 billion copies** are produced on copiers in the US; nearly **15 trillion copies** are produced on copiers, printers, and multi-function machines. (*Source: XeroxParc*). For specific information on fax printing, see the Telecommunications Summary.

Stock

According to the 1999 US Industry and Trade Outlook, the United States produces more printed products than any other country in the world. NAFTA countries have a 50% world market share. Estimating that the US produces 40% of the world's printed materials, we can estimate that each year the world produces **7.5 billion archiveable pages**, which would be equivalent to **195 terabytes** (compressed). (*Source: National Archives and Records Administration.*)

Rate of Change

As with the other print industries, growth in paper production is expected to be incremental but fairly consistent, both within the United States and internationally. Globally, paper and paperboard production capacity is forecast to grow from 333.6 million metric tons in 1998 to 348.1 million metric tons in 2001, an increase of 14.5 million metric tons (about 4%) over those three years. (*Source: U.S. Industry and Trade Outlook, 2000*).

According to the American Forest and Paper Association, US capacity to produce paper will increase by an average of only 0.7% annually over the next three years (2000-2002).

References

- Bogart, Dave, ed. *The Bowker Annual Library and Book Trade Almanac*, 44th edition. New Jersey: R.R. Bowker, 1999.
- Cummings, Anthony M., Marcia L. Witte, William G. Bowen, and Laura O. Lazarus. *University Libraries and Scholarly Communication: A Study Prepared for the Andrew W. Mellon Foundation*. The Association of Research Libraries, 1992
- Hayes, Robert M., UCLA School of Information Science, "The Economics of Digital Libraries"
- King, Donald W. and Carol Tenopir. "Economic Cost Models of Scientific Scholarly Journals," 1998.
- American Forest and Paper Association, *1999 Statistics for Paper, Paperboard and Wood Pulp*. Washington, DC, 1999. To order a copy, see www.afandpa.org/about/about.html or call (800) 244-3090.
- *Annual Report of the Librarian of Congress*. Washington, DC: Library of Congress, 1999.
- ArchiveBuilders.com White Paper, "Computer Storage Requirements for Various Digitized Document Types"
- Association of Research Libraries Statistics
- *Books in Print, 1999-2000.* New Jersey: R.R. Bowker, 1999.
- International Standard Serial Number Register
- International Standard Book Number Register
- JSTOR digital library
- Magazine Publishers of America, Information Center, (212) 872-3745.
- National Directory of Magazines.

- Newsletter and Electronic Publisher's Association (NEPA)
- *Newspaper Association of America, Facts About Newspapers*.
- Newspaper Project, National Endowment for the Humanities.
- *Oxbridge Directory of Newsletters 1997*. New York: Oxbridge Communications, Inc., 1997.
- *Ulrich's International Periodical Directory*. New York: Bowker Publishing, 2000.
- *UNESCO Statistical Yearbook 1999.* Paris, UNESCO, 1999.
- U.S. Census Department, 1999 Statistical Abstract of the United States
- *U.S. Department of Labor, Bureau of Labor Statistics*
- U.S. Government Printing Office, *Prepared Statement before the Committee on Rules and Administration, U.S. Senate on Public Access to Government Information in the 21st Century* July 1996.
- *U.S. Industry and Trade Outlook*. Available in print from McGraw Hill/U.S Department of Commerce Washington, D.C. or download from www.ntis.gov/products
- Chaskas, Eric. US National Archives and Records Administration.
- *Walden's Paper Report*. Twice-monthly newsletter, published by Walden-Mott Corporation, Ramsey, NJ. Available by subscription only. See www.walden-mott.com/PaperReport/PAP_RPT.HTM or call (201) 818-8630.

Charts

Click here to see charts supporting the above estimates, with time-series data.

More Discussion

Click here to read additional discussion of the conversion factors and related issues and to obtain detailed bibliographical information.

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Film - Summary

Originals

Flow

Photographs

There are over 2700 photographs taken every second around the world, adding up to well over 80 billion new images a year taken on over 3 billion rolls of film, according to estimates published by the United States Department of Commerce.

Photo CD, a format for digitized photography introduced by Kodak in 1992, has been widely adopted both by professional and amateur photographers. Kodak reports that the typical photograph can be digitized in this format in 5 megabytes without loss of picture quality. Utilizing this conversion factor, then, the world's 82 billion photos store 410 petabytes of data every year in photographs.

Motion Pictures

Apart from still photography, film is also used to store moving pictures. In the years from 1990 to 1995, UNESCO reports that there were 4,250 films produced annually throughout the world. The Motion Picture Association of America reports that for the year 1998, its members released 221 movies (compared to 219 in 1997), while releases by all U.S. companies, including independent film companies, rose from 461 in 1997 to 490 in 1998.

It takes approximately 2 gigabytes to store an hour of motion picture images in digital form using the MPEG-2 compression standard. If the images in 4500 full length movies were converted into bits, the world's annual original cinematic production would, therefore, consume about 16 terabytes.

X-Ray Film

The other major use for film is the storage of x-ray images for medical, dental and industrial purposes. Approximately 2 billion radiographs are taken around the world each year, including chest x-rays, mammograms, CT scans, and so on. (Traditionally, 8% of x-ray film is used in dentistry and industrial applications.) When x-ray films are converted to digital format, it is important that there is no important clinical information lost. The University of Pittsburgh Clinical Multimedia Laboratory suggests that an average conversion of a chest x-ray to digital storage with lossless compression will require 8 megabytes. To store all the world's x-rays to a computer file of this size would, therefore, require 17 petabytes each year.

Table 1: Original Data On Film Annually Worldwide

	Units	Digital Conversion	Total Petabytes
Photography	82,000,000,000	5 mb per photo	410
Motion Pictures	4,000	4 gb per movie	0.016
X-Rays	2,160,000,000	8 mb per radiograph	17.2
		All Film Total:	427.216

Stock

Photographs

The number of original photographs stored around the world is not a widely reported topic. There are commercial firms that collect professional photographs for resale and the largest of these report collections in the tens of millions. For example, Getty Images reports holding 70 million images mostly on silver halide film and its principal rival, Corbis Images, claims 65 million images.

As for amateur photographs, in 1997 it was estimated that there were 150 billion photographs stored in the United States. This is approximately 8 to 10 years production of photographs as of that time. Assuming that similar storage rates apply worldwide, it is likely that on the order of 750 billion photographs now exist. Using the same PhotoCD estimate of 5 megabytes per photo, there are 3.5 exabytes of original photographic data.

Motion Pictures

The number of motion pictures made around the world from 1895 to 1990 was approximately 242,000. *The International Film Index, 1895-1990*, lists entries of that many titles of films. (There is no guarantee that some of these movies still exist, however.) The overall total is broken down into these categories of film types:

Table 2: Film Breakdown by Type

Type	Number

FEATURE	110,586
SHORT	50,056
ANIMATED	9,787
DOCUMENTARY	7,277
TELEVISION	4,197
SERIAL	553
SILENT	59,680
TOTAL	242,136

The "Television" category refers to films that were originally made for television broadcast rather than theatrical release. "Serials" are films made primarily in the United States between 1914-1936 and were made to be shown in weekly installments.

As discussed in the precedng section, in the decade since 1990, there have been around 4500 movies made each year, adding another 45,000 to the total world stock of original motion pictures. Accordingly, a good estimate of the world's original pictures is approximately 300,000.

X-Rays

The clinical and legal uses of medical x-rays continue for an indefinite time and, therefore, prudent practice is to preserve x-rays and medical records generally for as long as possible. The same principle applies to dental x-rays. The only use of x-ray that may result in regular destruction of the resulting images is industrial testing, but even there it is likely that images are retained for a substantial period of time. Therefore, it is believed that there is little systematic destruction of the flow of new x-rays and virtually all of them are added to the stock. For the sake of calculation, it is assumed that a full ten years of x-ray images will constitute the stock. This is equivalent to approximately 21.6 billion images or 172.8 petabytes.

Copies

Flow

Photographs

There has been very little history of the large mass of photographs being copied. Kodak estimates that only about 2 percent of photographs are ever copied or modified in any way after they are originally developed. Of course, some photographic images are widely distributed in newspapers and magazines. But these represent a miniscule fraction even of the professional photographer's work, itself a small minority of all photographs.

One copy of 2 percent of the annual new photographs would be 1.6 billion photographic copies. With PhotoCD compression, this would represent 8 petabytes per year of photographic copies.

Motion Pictures

The Wolfman Report on the Photographic and Imaging Industry in the United States states that the average number of prints per original motion picture is 700. The Silver Institute, however, reports that 6000 release prints are made for each feature movie. A figure of 1000 copies for motion pictures will be used on the assumption that many of the world's motion pictures have more limited releases than the typical Hollywood blockbuster. American studios only account for 400 to 500 movies per year. 45,000 copies of motion pictures per year at 4 gigabytes per copy is 18 petabytes of copies.

X-Rays

The clinical requirements for medical x-rays demand that originals be used in almost any situation. There is no siginificant use of copies of x-rays at all.

Stock

Photographs

The stock of copies of photographs can be calculated by reference to the assumptions made for the storage of the originals. Ten years of copies of photographs would be 16 billion. If these were all digitized at the rate of 5 megabytes per picture, there are 80 petabytes of photographic copies on hand.

Motion Picture

The copies on film of motion pictures made for distribution are short-lived. Most of these copies are deliberately destroyed when the general theatrical release of the movie is over. If even ten of these survive for the approximately 4,000 motion pictures made annually for the last twenty-five years, this would be equivalent to about 1,000,000 copies. If each copy is equivalent to 4 gigabytes of data, this stock of copies of motion picture film is 4 petabytes.

X-Rays

There is no significant stock of copies of x-rays.

References

- United States Industrial and Trade Outlook 2000
- UNESCO Statistical Yearbook
- 1999 Photo Marketing Association
- Silver Institute

More Details About Film

How Much Information?

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Optical - Summary

Originals

For optical media, we focused on the three major industry categories: CD audio, CD-ROM, and DVD.

Flow

Annual world title production of the 3 media types appears in the following chart. (All figures are rounded to the nearest hundred or terabyte.)

Table 1: Annual world title production of the 3 media types.

Media Type (Source* & Year Cited)	Unique Items per Year (US)	Unique Items per Year (World)	Conversion Factor	Total Terabytes (Annual US)	Total Terabytes (Annual Worldwide)
CD - Music (1998)	33,100	90,000	Uncompressed: .650 GB/item	22	58
			Compressed (to MP3): .065 GB/item	2	6
CD- ROM (1999)	500	1,000	Uncompressed (to MP3): .650 GB/item	1	3
			Compressed: .065 GB/item	.3	.6
DVD-video (1999)	3,000	5,000	4.38 GB/item	13	22

Totals:	**Uncompressed**	**36**	**83**
	Compressed	**15.3**	**29**
*Detailed source information listed at end of this document.			

World

To estimate how many CD-audio originals are created each year worldwide, we used RIAA statistics regarding the US market share and US record releases (see below). The United States holds a 37% share of the world music market and releases about 33,100 items per year. Therefore, the world produces roughly **90,000 originals** per year, equivalent to **58 TB** (uncompressed).

Between 1998 and 1999, 1,000 new CD-ROM data titles were added to *CD-ROMs in Print*, an international directory published by Gale Research. This equals about **3 TB** of new information in one year (uncompressed).

The United States currently produces about 60% of the DVD titles available worldwide. Using US title production statistics, we estimate that about **5,000 new DVD titles** are produced internationally each year - this is about **22 TB**. (Source: Jim Taylor's *DVD FAQ).*

United States

The US produces approximately 37% of the world's CD-audio titles, 50% of the world's CD-ROM titles and 60% of the world's DVD titles. (Sources: Recording Industry Association of America, International Recording Media Association, and Jim Taylor's *DVD FAQ*).

The Recording Industry Association of America (RIAA) reports the number of new releases and album re-releases each year. In 1998, **33,100 titles** were released, roughly equivalent to **22 TB.**

The US share of the CD-ROM replication market is 52%, according to the International Recording Media Association. If we assume that the US holds a similar share of CD-ROM title production, then about **500 titles** are produced by the United States each year, equivalent to about **1 TB** per year.

For the past three years, new DVD titles have been added at a rate of about **3,000 per year** (**13 TB per year)** --a tremendous rate of content growth, but that's because it is a new medium and a significant amount of legacy content is being converted. This rate should decrease as the medium becomes more well-established. (Source: *DVD Entertainment Group*)

Stock

The All Music Guide (a comprehensive database tool used by industry leaders) reports a total of **523,363 titles** (445,735 popular music and 77,628 classical music albums). If each work were stored on a 650 MB CD, this would be equivalent to **340 TB.** We can assume that this figure represents the US portion of original works, then extrapolate the world stock to be about **1,400,000 titles**. [???]

According to the 1999 edition of *CD-ROMs in Print,* internationally there are about 16,200 unique CD-ROM titles (**about 11 TB**)--business applications (such as word processing and spreadsheet packages), games, reference tools, and instructional programs. This figure is consistent with other CD-ROM directories, such as the *Multimedia & CD-ROM Directory*, which lists 17,000 titles.

The United States produces about two-thirds of the total DVD titles. As of June 2000, there are about 8,500 titles available in the United States, 13,000 worldwide. This is equivalent to **37 TB** (US) and **57 TB** (world).

Rate of Change

Production of CD-Audio and CD-ROM originals is not increasing dramatically. However, the production of DVD originals is growing at a tremendous rate (about 100% per year), due to the fact that material previously available in another format is being reissued on DVD.

Copies

Flow

Replication

In 1999, there were about 4.7 million audio CDs and 3.6 million CD-ROMs replicated worldwide, according to the International Recording Media Association (IRMA). In addition, in 1999 194 million DVD-Video units, 12 million DVD-ROM units, and 2 million DVD-Audio units were replicated.

A total of 1.583 billion recordable CDs (CD-Rs) were sold around the world in 1999, according to Santa Clara Consulting Group. (Source: News Release, *"Record sales drive Memorex up to third in world CD market" www.lewisvpr.com/Releases/UKReleases/000620_memorex.html*) Estimates of growth in demand for CD-R vary--one consulting firm projects global demand to reach 4.74 billion in the year 2000, while others project demand to be between 2.5 and 3 billion units. (Photonics Industry & Technology Development Association, cited in www.taiwanheadlines.gov)

Retail

During 1999, according to the Recording Industry Association of America (RIAA), **938.9 million CDs** were shipped to retail by U.S. producers. The US has a 37% share of the world's sales, according to the International Federation of the Phonographic Industry (IFPI). Therefore, one can extrapolate that CD shipments worldwide are about **2,500 million units**. This is equivalent to **1,625 TB**.

According to the DVD Entertainment Group, more than 130 million DVD video movies and music video titles were shipped to retail between spring 1997 (when the format launched) and January 2000. **100 million discs** were shipped during 1999 alone.

Stock

The stock of audio CDs in the United States can be estimated by summing the CD unit sales since the format became popular. The RIAA only provides statistics going back to 1990 - these 10 years of shipments add up to **6,200 million units**, equivalent to about **4,030 TB**. Ten years of worldwide shipments are approximately **15.2 billion units**, equivalent to **10,937 TB**.

Since the format was launched in 1997, more than **1.5 billion DVDs** have been replicated worldwide, almost **1 billion** in North America alone.

Optical Media Bibliography

- All Music Guide
- The CD Information Center
- *CD-ROMs in Print, 13th Edition*: An International Guide to CD-ROM, CD-I, 3DO, MMCD, CD32, Multimedia, Laserdisc, and Electronic Products. New York: The Gale Group, 1999.
- *CD-ROM Finder: The World of CD-ROM Products for Information Seekers*. Medford, NJ: Learned Information, 1993
- The International Recording Media Association
- DVD Channel News
- DVD Entertainment Group
- DVD FAQ
- DVD Insider
- International Federation of the Phonographic Industry

- Medialine
- Optical Storage Technology Association
- SUN CD-ROM FAQ

Charts

Click here to see charts supporting the above estimates, with time-series data.

More Discussion

Click here to read additional discussion of the conversion factors and related issues and to obtain detailed bibliographical information.

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Magnetic - Summary

Originals

Flow

Tape

Analog

Video

In the year 2000, 1.4 billion blank VHS video tapes will be produced for the entire world. If all of these tapes were filled to their 120 minute capacity and then converted to digital using MPEG-2 compression, there would be approximately 4 gigabytes of data per tape. One year's production of blank videotape, therefore, provides storage space adequate for 5600 petabytes of data.

Assuming twenty percent [???] of this tape is used for the storage of original data, the flow of new data stored on analog VHS videotape per year would be 1120 petabytes.

Video camcorder tapes (all formats except VHS) are produced at the rate of 150 million per year according to the Japan Recording Media Industry Association. Almost all of this tape is used for the storage of original data. Assuming one hour per tape in MPEG-2 format yields 300 petabytes.

Total original analog video production worldwide runs at about 1420 petabytes annually. [???]

Audio

In the year 2000, 921 million blank audio tape cassettes will be produced for the entire world according to British research firm, Understanding & Solutions. If all of these tapes were filled to their 120 minute capacity and then converted to digital using the common CD audio format, there would be approximately 1 gigabyte of data per tape. One year's production of blank audiotape, therefore, provides storage space adequate for 921 petabytes of data.

Assuming twenty percent [???] of this tape were used for the storage of original data, the flow of new data stored on analog audiotape per year would be 184.2 petabytes.

Digital

There are 25 million computer tape drives installed in the world at present. These drives provide storage capacity for all range of computers - from desktop personal computers to the most mammoth supercomputers. Fred Moore estimates that the amount of data stored on tape is between 4 and 15 times the amount of enterprise data on disks and that there is about $1 billion per year of computer tape media sold worldwide.

In order to estimate the amount of data stored on tape, we will first estimate the number of enterprise storage systems and then multiply by 10. IDC estimates that 250 petabytes of RAID storage capacity will be shipped worldwide in 2000. RAID storage systems are taken as most closely approximating the storage capacity deployed to data-intensive corporate, government and scientific uses, the largest consumers of magnetic tape backup.

A midrange estimate of the total amount of data stored on magnetic tape would, therefore, be 2.5 exabytes.

In all but the largest computer applications, however, tape is generally used solely for backup of data already stored on hard disk drives. Quantum, the manufacturer of DLT tape, the most popular format for enterprise storage, estimates that 90 percent of the tape capacity in that format is used for backup. Fred Moore also points out that it is more and more common for multiple copies of data to now be stored on tape.

If it is assumed that ten percent of the total amount of data stored on tape is original data of the sort generated by scientific experiments in high-energy physics or by observational earth satellites or archival storage on tape where the data is no longer stored on disk, original magnetic tape data is roughly the same as all the RAID capacity shipped annually - 250 petabytes in the year 2000.

250 petabytes is also generally consistent with estimates derived by use of forecasts of producer revenue of around $1 billion for tape media and an average cost of around $5 per gigabyte of tape storage.

Disks

Floppy Disks

In the year 2000, 1 billion 3.5 inch floppy disks, each capable of storing 1.44 megabytes, will be produced for the world. This is an aggregate storage capacity of 1.4 petabytes. If [???] five percent of this is original data, new data per year on floppy disk would be 0.07 petabytes.

Removable Disks

In the year 2000, 88 million removable 100 megabyte disks and 25 million removable 1 gigabyte disks will be produced for the world. Together, these two varieties of removable disks provide 33.8 petabytes of storage capacity. If five [???] percent of this is original data, new data per year stored on removable disks would be 1.69 petabytes.

Hard Disks

In the year 2000, hard disk drives capable of storing 2500 petabytes of data will be produced for the world.

The amount of original data stored on hard disks is most likely to vary according to the computing environment in which the disks are deployed. It is possible to divide all hard disk storage into three categories:

1. Personal computer, laptop, or workstation. This type of computer is responsible for approximately 55% of the computer disk storage capacity currently shipped.
2. Departmental server. This class of computing environment is responsible for about 30% of the overall storage market.
3. Enterprise server. These big computers account for about 15% of the hard disk storage.

As with all such broad categorizations, there may be quite a bit of blurring around the edges. However,

consideration of these three different environments leads to the conclusion that the amount of original data stored on the computers in each is probably substantially different.

Single user computers and the applications software usually found on them is not suited for the production of large amounts of original data. A recent study (McKenzie, "Microsoft's 'Applications Barrier to Entry': The Missing 70,000 Programs") found that most people used only a few applications other than those found in the Microsoft Office application suite. These applications are usually text-based, such as word processing or spreadsheets, and so require minimal storage space. Most personal computers now sold come with hard disk storage capacity in the range of 10 gigabytes. 100 megabytes of original data constitutes 1 percent of disk capacity, which is the estimate for this category of computer disk. [???]

Departmental servers would be commonly found in business, government, educational or other organizational settings. These servers provide disk space for a group of users, who all contribute to the production of organizational data. Aside from the databases and spreadsheets, there may be product catalogs and other graphic intensive marketing material, PowerPoint presentations, and so on. An estimate of the original data stored in these hard disks is 35%. [???]

Enterprise servers are the large-scale computing environments where "big iron" traditionally has reigned. The applications here are corporate or governmental transaction processing on a large scale or the generation of huge data sets from scientific research missions. The amount of original data stored on these computers is estimated at around 65%. [???]

Table 1: Original Data Stored On Hard Disk By Computing Environment (1995-1999) [In Petabytes]

	Original Data	1995		1996		1997		1998		1999	
		Total	Original	Total	Original	Total	Original	Total	Original	Total	Original
Personal	1%	58	0.6	101	1	189	2	399	4	766	8
Departmental	35%	32	15	55	19	114	40	239	84	460	161
Enterprise	65%	16	8	27	17	41	27	86	56	167	109
Total:		105	23.6	183	37	344	69	724	144	1,393	278

Table 2: Original Data Stored On Hard Disk By Computing Environment (2000-2004) [In Petabytes]

	Original Data	2000		2001		2002		2003		2004	
		Total	Original	Total	Original	Total	Original	Total	Original	Total	Original
Personal	1%	1,405	14	2,553	26	4,466	45	7,165	72		
Departmental	35%	843	295	1,532	536	2,680	938	4,299	1,505		
Enterprise	65%	306	199	557	362	974	633	1,563	1,016		
Total:		2,554	508	4,642	924	8,120	1,616	13,027	2,593		

The amount of original data compared to the total amount of data stored on magnetic hard disks works out to twenty percent over all the years under consideration. Accordingly, in the following table summarizing the annual flow of original data in various media, 20% of all hard disks are assumed to store original data, the actual amount will vary depending on the capacity of the hard drive itself.

Table 3: Original Data Flow Storage Estimates

Media Type	Unique Items per Year	Conversion Factor	Total Annual Petabytes (World)
Blank Audio Tape (2000)	184,200,000	1 gb per tape (CD audio format - no compression)	184.2
Blank Video Tape (2000)	355,000,000	4 gb per tape (MPEG-2 compression)	1420
Computer Tape Drives (2000)	5,000,000	Varies	250
Floppy Disk (2000)	5,000,000	1.44 mb per disk	0.07
Removable Disks (2000)	4,400,000 1,250,000	100 mb (low capacity) 1 gb (high capacity)	1.69

Hard Disks (2000)	37,400,000	Varies	500

Stock of Originals

Tape

Analog

Video

[???] 10 billion video cassettes, both prerecorded and recorded at home, have been accumulated. This is equivalent to about five years production of these tapes. One billion camcorder format tapes have also been added over the same time. This would be equivalent to 3 billion hours of stored original video, which if digitally encoded would produce a stock of about 6000 petabytes of original videotaped data. [???]

Audio

The total stock of original audio content stored on tape may be estimated by assuming twenty [???] percent of five years production of blank cassette tapes contains original content. This equals 1 billion cassette tapes. The digital equivalent of this audio information is 1000 petabytes.

Digital

The stock of original data on magnetic tape may be approximated by adding the yearly flow of original data over the course of the expected lifetime of the medium. Some unfortunate experiences with the loss of computer data stored on magnetic tape has led to the practice of continuous migration of this data to new media every five to ten years. This process leads to the reduction in the number of tape cartridges that need to be managed as tape capacity inexorably rises as well as insuring modernization of the tape format.

Therefore, the stock of original data on magnetic tape may be taken as five year's worth of original data flow. [???]

Disks

Floppy Disks

Floppy disks useful life is estimated to be three years. The amount of original data stored on the 4.5 billion floppies produced over the course of the past three years is around 5 percent of the total data on those disks, or 0.32 petabytes.

Removable Disks

The amount of original data stored on removable disks over the past three years is approximately 5 petabytes.

Hard Disks

Over the past three years, hard disk capacity of 4625 petabytes has been produced. If twenty [???] percent of that capacity has been used to store original content, the stock in that format is now 925 petabytes.

Copies

Flow

Tape

Analog

Video

In the year 2000 1.8 billion prerecorded video tapes will be distributed worldwide according to the International Recording Media Association. This entire production will be copies, principally of feature films. If converted to digital using MPEG-2 compression, prerecorded videotape would consume 7,200 petabytes per year.

If [???] 80 percent of the blank videotape distributed per year were used for copies, this would constitute a digital equivalent of 4,480 petabytes.

The total yearly world production of copied data on analog videotape, therefore, is 11,680 petabytes. [???]

Audio

In the year 1999, 125 million prerecorded audio cassette tapes were distributed in the United States according to the Recording Industry Association of America. This entire production was copies, principally of music. If converted to digital using audio CD format, and assuming about one hour of music per tape, prerecorded audio tape would consume 62.5 petabytes per year.

If 80 percent [???] of the 921 million blank audiotapes distributed per year worldwide were used for copies, this would constitute a digital equivalent of 737 petabytes.

The total yearly production of copied data on analog audiotape, therefore, is 800 petabytes.

Digital

If ninety percent of the computer tape distributed annually were used for copies, this would constitute 2.25 exabytes of copied data for the year 2000. The amount of data on tape, however, is calculated by reference to the RAID storage capacity shipped, which has been rising rapidly. Similar trends will most likely be seen in the amount of data stored on tape, particularly as it becomes more common to make multiple tape copies of data.

Disks

Floppy Disks

If 95 [???] percent of the 1.4 petabytes of annual floppy disk storage were used for copies of data, this would add 1.33 petabytes to the stock of digital data stored on floppy disks.

Removable Disks

If 95 [???] percent of the 33.8 petabytes of annual removable disk storage were used for copies of data, this would add 32.1 petabytes to the stock of digital data stored on removable disks.

Hard Disks

If 80 [???] percent of the 2500 petabytes of annual hard disk storage were used for copies of data, this would add 2000 petabytes to the stock of digital data stored on hard disks.

Stock of Copies

Tape

Analog

Video

Five years production of blank T-120 VHS videotapes is approximately 7 billion units. If 80 [???] percent of this were used for storage of copied data, the total stock of such data in MPEG-2 conversion would be 2240 petabytes. Five years production of prerecorded videotapes is approximately 8 billion units. This is the digital equivalent of 3200 petabytes. The total world stock of copied data on videotape is 5440 petabytes.

Audio

The world production of blank audiotape for the past five years is approximately 5 billion units. There has also been sales of prerecorded audiotapes during that period of 6 billion units. Assuming that 80 [???] percent of the blank audiotape has been used for the storage of copied data, the overall stock of copies of audio data on tape is 4000 petabytes on blank tape and 6000 petabytes on prerecorded tape. Therefore, there is a total stock of 10,000 petabytes of copies of audio data on magnetic tape.

Digital

Ninety percent of the data stored on tape is copies and tape is estimated to store 4 to 15 the amount of enterprise data on hard disk. IDC estimates that approximately 400 petabytes of RAID have been shipped in the past three years. If all data on tape is calculated as ten times that amount, or 4 exabytes, and copies as 90 percent of that overall amount, copies of data on magnetic tape worldwide would be 3.6 exabytes.

Disks

Floppy Disks

There have been 4.5 billion floppy disks produced in the last three years, creating a total storage capacity of 6.5 petabytes. If 95 percent of that storage were used for copies, the total stock of copies on floppy disk would be 6.1 petabytes.

Removable Disks

Over the past three years approximately 100 petabytes have been stored on removable disk. Assuming that 95 [???] percent of that is backups and copies, the total stock of copied data in this media format is 95 petabytes.

Hard Disks

Over the past three years, hard disk capacity of 4625 petabytes has been produced. If eighty [???] percent of that capacity has been used to store copies of data, the stock of copied data on hard disk is 3700 petabytes.

More Details

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Internet - Summary

The Internet is one of the youngest and fastest growing media in today's world. Internet growth is still accelerating, which indicates that the Internet has not yet reached its highest expansion period [1]. It should be noted, however, that while the Internet is a completely new kind of medium, by separating it into a distinct category, we are allowing for a certain amount of double counting, because all the Internet-based stock of information is already accounted for under "magnetic" or "tape" categories. Furthermore, we should make clear the distinction between the stock and the flow of information. While web sites and some portion of email messages are being stored and accounted for under different storage categories, there are other "components" of what we know as "Internet," such as Internet Relay Chat (IRC) or Telnet, which exist only as a flow of communication. What makes the Internet extremely successful is that it is one of a handful of media (such as radio and TV), where one unit of storage might generate terabytes of flow, as opposed to books and newspapers, where one exemplar is usually read by one or two people, and the flow of information is relatively low.

World Wide Web

There are two groups of Web content. One, which we would call the "surface" Web is what everybody knows as the "Web," a group that consists of static, publicly available web pages, and which is a relatively small portion of the entire Web. Another group is called the "deep" Web, and it consists of specialized Web-accessible databases and dynamic web sites, which are not widely known by "average" surfers, even though the information available on the "deep" Web is 400 to 550 times larger than the information on the "surface." [2]

The "surface" Web consists of approximately 2.5 billion documents [1 and 5], up from 1 billion pages at the beginning of the year [3], with a rate of growth of 7.3 million pages per day [1]. Estimates of the average "surface" page size vary in the range from 10 kbytes [1] per page to 20 kbytes per page [4]. So, the total amount of information on the "surface" Web varies somewhere from **25 to 50 terabytes** of information [HTML-included basis]. If we want to obtain a figure for textual information, we would use a factor of 0.4 [4], which leads to an estimate of **10 to 20 terabytes** of textual content. At 7.3 billion new pages added every day, the rate of growth is [taking an average estimate] 0.1 terabytes of new information [HTML-included] per day.

If we take into account all web-accessible information, such as web-connected databases, dynamic pages, intranet sites, etc., collectively known as "deep" Web, there are **550 billion web-connected documents**, with an average page size of 14 kbytes, and 95% of this information is publicly accessible [2]. If we were to store this information in one place, we would need **7,500 terabytes** of

storage, which is 150 times more storage than we would need for the entire "surface" Web, even taking the highest estimate of 50 terabytes. 56% of this information is the actual content [HTML excluded], which gives us an estimate of **4,200 terabytes** of high-quality data. Two of the largest "deep" web sites - National Climatic Data Center and NASA databases - contain **585 terabytes** of information, which is 7.8% of the "deep" web. And 60 of the largest web sites contain **750 terabytes** of information, which is 10% of the "deep" web.

When we look at the distribution of the web sites, the most apparent trend is that English loses its dominant position. Currently, only 50% of all Internet users are native English speakers, though English web sites continue to dominate with approximately 78% of all web sites and 96% of e-commerce web sites being in English [6]. It's hard to estimate what percentage of web sites have their origins in the United States, because .com domains can be registered in virtually any country, English-language web sites are often created in countries like Japan, and many international web sites are hosted in the United States. 17 million out of 27.5 million domains registered worldwide are .com, and 2 million are .uk, making Great Britain's domain the biggest country domain in the world [7].

More Details

Email & Mailing Lists

Email has become one of the most widespread ways of communication in today's society. A white-collar worker receives about 40 email messages in his office every day [8]. Aggregately, based on different estimates, there will be from 610 billion [9] to 1100 billion [10] messages sent this year alone. With the average size of an email message 18,500 bytes [11] and growing, the amount of flow becomes surprisingly gigantic, somewhere between **11,285 and 20,350 terabytes**. Of course, not all of this email gets stored. Mail.com has 14.5 million email boxes and uses 27 terabytes of storage; with approximately 500 million mailboxes worldwide, the required storage space is more than **900 terabytes**, which means that only one in 17 messages is kept for some period of time.

Mailing lists can be viewed as a subcategory in email. It is hard to determine the number of mailing lists in existence, but we can approximate it based on some available statistics. One of the most frequently used mailing list managers - LISTSERV - is used to send 30 million messages per day in approximately 150,000 mailing lists [12]. A sample of mailing lists has shown that 30% of them are managed using LISTSERV. Using this information, we would estimate the total number of mailing list messages at 36.5 billion per year with aggregate volume of **675 terabytes**.

Distribution of mailboxes has the same pattern as the distribution of web sites. While in 1984, 90% of the world's e-mailboxes were located in the U.S., at the end of 1999 this number dropped to 59%, and is expected to decrease even further. [13].

More Details

Usenet

Most of the statistics in this category are vague, so the numbers we have should be regarded with a certain skepticism. Cidera, which is the 14th biggest news provider on the Internet [14], gets approximately 0.150 terabytes of Usenet feeds per day. We would estimate the total amount of original news feeds at 0.2 terabytes per day, which leads to **73 terabytes** of original Usenet postings per year, which are redistributed by local ISPs and news servers an endless number of times.

FTP

We are missing any significant data on this sector, but we know that Walnut Creek CD-ROM archive contains a total of **0.412 terabytes** of data on two servers [ftp.cdrom.com and ftp.freesoftware.com] and the amount of storage was expanding at 100% every year over the past 6 years [15]. It should be noticed that the distinction between FTP and HTTP becomes more blurred, as more and more file archives become available through HTTP.

IRC, Messaging Services, Telnet...

These categories mostly represent a flow of information as opposed to the stock. Liszt.com has one of the biggest directories of IRC channels - 37750 channels on 27 networks, with 150,000 users, all of them typing text as fast as they can. [16]

References

[1] "Sizing the Internet," *Cyveillance*, http://www.cyveillance.com/resources/library.asp
[2] "The Deep Web: Surfacing Hidden Value," *BrightPlanet LLC*, http://www.completeplanet.com/Tutorials/DeepWeb/index.asp
[3] "Web Surpasses One Billion Documents," *Inktomi Corp.*, http://www.inktomi.com/new/press/billion.html
[4] "Accessibility of Information on the Web," *Nature Magazine, Volume 400, Number 6740, Page 107*
[5] "Size of the Web: A Dynamic Essay for a Dynamic Medium," *The Censorware Project*, http://censorware.org/web_size/
[6] "State of the Internet 2000," *United States Internet Council & ITTA Inc.,*http://usic.wslogic.com/intro.html
[7] "Domain Statistics," *DomainStats.com*, http://www.domainstats.com
[8] "Sending AOL a Message," *Newsweek, Aug 9, 1999, p.51*
[9] "Email Facts," *24/7 Media*, http://www.247media.com/research/trends/email.html
[10] "Like It Or Not, You've Got Mail," *BusinessWeek*, http://businessweek.com/1999/99_40/b3649026.htm
[11] UC Berkeley Email Stats
[12] "LISTSERV Statistics," *L-Soft*, http://www.lsoft.com/news/default.asp?item=statistics
[13] "Year-End 1999 Mailbox Report," *Messaging Online*, http://www.messagingonline.com/
[14] "Top 1000 Usenet Sites" *Freenix*, http://www.freenix.org/reseau/top1000/
[15] David Greenman, Walnut Creek CD-ROM Archive
[16] *Liszt.Com*, http://www.liszt.com/

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Broadcast - Summary

Originals

World

Table 1: World					
Media Type	Number of Stations	Unique Items per Year	Conversion Factor	Total Terabytes (Annual)	
				Lower Bound	Upper Bound
Radio (CIA Factbook 2000)	43,973	65.5 million hours of original programming	.05 GB/hour	3,274	3,274
Television (broadcast only) (CIA Factbook 2000)	33,071	48 million hours of original programming	1.3 GB - 2.25 GB hour	62,769	108,638
			Total:	**66,043**	**111,912**

Table 1: United States					
Media Type	Number of Stations	Unique Items per Year	Conversion Factor	Total Terabytes (Annual)	
				Lower Bound	Upper Bound

Radio (FCC, 1999)	12,600	15.8 million hours	.05 GB/hour	788 TB	788 TB
Television (broadcast and cable networks) (FCC, NCTA, 1999)	1,884	3.4 million hours of original programming	1.3 GB- 2.25 GB/hour	4,470 TB	7,736 TB
			Total:	**5,258**	**8,524 TB**

Radio

Conversion Factor

Each hour of audio requires about **50 MB**, if stored at MP3 quality. (Different sources cite different figures, depending upon assumptions made about compression and sound quality.)

World

There are **43,773** active radio stations in the world, according to the CIA World Factbook 2000: about 16,500 AM stations, 26,000 FM stations, and 1,500 shortwave stations.

We estimate that FM radio stations broadcast 20 hours per day, AM stations 16 hours per day, and shortwave stations 12 hours per day. Therefore, there are approximately **290 million** hours (188 million FM, 98 million AM, and 6 million shortwave) of radio programming per year. Applying the 50 MB/hour rule of thumb, one may estimate an annual storage requirement of about **14,500** TB if one were to record everything broadcast on the radio.

United States

As of 1999, there are **12,615** radio stations in the United States, according the Federal Communications Commission: 4,783 AM, 5,766 FM and 2,066 FM Educational stations. As noted above, the two formats broadcast different numbers of hours each day: 20 hours for FM stations, 16 hours for AM. Total US broadcasting hours would therefore be roughly **85 million hours** per year. Again, each hour of broadcasting would require 50 MB of storage, using the MP3 format. Total storage required for all US radio broadcasts is about **4,300 TB.**

About 84% of US radio stations have music as their primary focus, and provide little original content. (Source: Radio Marketing Guide & Fact Book, Radio Advertising Bureau 2000-2001) The remaining 16% are news, talk and religious stations, providing, presumably, almost 100% "original" information each day. Regardless of format, a percentage of most stations' broadcast time includes some commentary, weather reports, news updates, and traffic reports - perhaps an average of 5 minutes per hour. In addition, radio stations average between 12 and 16 minutes of commercials per hour. We can use this information to estimate how much "original" programming appears on the United States radio airwaves:**15.8 million hours**. (This estimate excludes advertisements and music.) The equivalent in bytes is **788 TB**.

Television

Conversion Factor

Satellite TV transmits between 3 and 5 mb per second. Therefore, one hour of broadcasting will require 10.8-18 Gbits of storage (compressed to MPEG-2). We can estimate about **1.3 - 2.25 GB** per hour of TV broadcasting. Source: Internet Archive.

World

There are **33,071 television stations** in the world, according to the CIA World Factbook 2000. If these stations broadcast about 16 hours per day, this would equal about 193 million hours total programming. We estimate about 1/4 of the programs are "original," - this is **48 million hours** each year. Estimating that one hour of video requires 1.3 GB of storage, then worldwide, program storage would be about **63,000 TB**; using the 2.25 GB estimate, it would be about **109,000 TB**.

United States

As of 1999, there are 1,616 broadcast television stations in the United States, according the US Federal Communications Commission. This figure includes the major networks (ABC, CBS, NBC, FOX, PBS and newcomers WB, UPN and PAX), the networks' affiliates, as well as local and public broadcasting stations. In addition, the National Cable Television Association reports that there are 210 national cable networks and 54 regional networks, as of August 2000. (54 more cable networks are planned but not yet operational.)

If all 1,884 of these stations broadcast 20 hours per day, that would equal just under **14 million hours** per year. We estimate that about 1/4 of the television programs broadcast are "original" - this is **3.5 million hours** each year, equivalent to **between 4,400 and 7,800 TB**.

Stock

In 55 years of programming, the networks have accumulated the following stock of material (*Source: Library of Congress Report, Television/Video Preservation Study: Volume 1: Report, October 1997*).

Table 3: Stock of Material Accumulated by the Major Networks.	
ABC	1,037,000 films/tapes
CBS	1,045,000 tapes and more than 150,000,000 feet of film
NBC	600,000 film reels (currently estimated at 100,000,000 feet) and 1,600,000 videotapes

Meanwhile, some of the major studios have accumulated original materials as well:

Table 4: Materials Accumulated by the Major Studios.	
Disney	6,500 television programs on 80,000 reels and tapes
Fox	54,000 television programs on 780,000 reels and tapes
MCA/Universal	18,000 (through 1994) television programs on 217,000 reels and tapes
Paramount (Viacom)	8,000 television programs on 1,200,000 reels and tapes
Sony/Columbia	35,000 television programs on 600,000 reels and tapes
Turner Entertainment	20,000 television programs on 337,000 reels and tapes
Warner Brothers	28,000 television programs on 1,000,000 reels and tapes

These figures overlap, of course, with those we have compiled for magnetic tape.

As of 1998, there are well over **18,000 hours** of programs in syndication available to be aired. This is equivalent to **18 TB** of information. (*Source: Television and Video Almanac 1998*)

Rate of Change

The number of radio stations and broadcast television stations in the United States has increased slightly - between 1 and 2% - every year since 1990.

While the number of cable television systems has decreased since 1993, the number of cable networks (i.e. channels) has increased annually. 54 new cable networks are scheduled to be launched in the coming year.

Copies

Radio

World

We estimate that FM radio stations broadcast 20 hours per day, AM stations 16 hours per day, and shortwave stations 12 hours per day. Therefore, there are approximately **290 million hours** (188 million FM, 98 million AM, and 6 million shortwave) of radio programming per year. Applying the 50 MB/hour rule of thumb, one may estimate an annual storage requirement of about **14,500 TB** if one were to record everything broadcast on the radio.

United States

As of 1999, there are **12,615 radio stations** in the United States, according the Federal Communications Commission: 4,783 AM, 5,766 FM and 2,066 FM Educational stations. As noted above, the two formats broadcast different numbers of hours each day: 20 hours for FM stations and 16 hours for AM. Total US broadcasting hours would therefore be roughly **85 million hours** per year. Again, each hour of broadcasting would require 50 MB of storage, using the MP3 format. Total storage required for all US radio broadcasts is about **4,300 TB**.

Television

World

There are about 33,000 television stations in the world (including some but not all cable stations), according to the CIA World Factbook 2000. This means that there are approximately **193 million hours** of television programming per year (assuming each station broadcasts 16 hours per day). Estimating that one hour of video requires 1.3 - 2.25 GB of storage, then worldwide, television would require between **250,000 and 435,000 TB**.

United States

As of 1999, there are 1,616 broadcast television stations in the United States as well as about 260 cable networks operating on nearly 10,500 cable systems. Each station broadcasts an average of 7,300 hours per year (again estimating 20 hours of broadcasting per day). If one wished to capture all of the programming generated by every station and cable system, regardless of duplication, it would be about **88 million hours**, requiring between **114,000 and 198,000 TB** of storage.

Fun Facts about Broadcast Media.

Television

- For many years, most large TV stations and the major networks subscribed to the Code of Good Practices of the National Association of Broadcasting, which established limits on the number of commercial minutes that could be telecast each hour. The limits were voluntary but widely followed: 9 1/2 minutes of commercials during primetime; higher amounts during other times of night and day. In 1992, however, the guidelines were ruled a violation of Federal antitrust law. Throughout the industry, most pledged to continue the limits - but gradually that eroded, as networks added more ad time. Prime time today has an average of 15 minutes of ads per hour. The FCC regulates advertising only during children's programming: 10.5 minutes/hour on weekends, 12 minutes/hour on weekdays. (*Source: Gerald Baldasty, University of Washington* *http://faculty.washington.edu/baldasty/Feb3.htm*)
- Approximately 7 in 10 television households, more than 65 million households, subscribe to cable. (*Source: National Cable Television Association*)

Radio

- There are currently more than 4,500 streaming radio stations on the Internet, distributed as follows: Africa (16); Asia (109); Europe (970); North America (2,786); Oceania (235); South America (166) and Internet Only (313). (*Source: RadioDirectory,* *http://www.radiodirectory.com/Stations/*)

- "Talk radio," in which celebrities and experts from various fields answer listener "call-in" questions and offer their advice on various topics, has grown spectacularly in recent years. The "call-in" format is the fastest-growing in radio, accounting for nearly 1,000 of the 10,000 commercial radio stations in the United States. (*Source: U.S. MEDIA IN THE 1990s: THE BROADCAST MEDIA By Fredric A. Emmert, US Information Agency*)
- 6,108 new radio stations have been started in United States since 1970 (*Source: Arbitron Ratings, http://www.arbitronratings.com*)
- The average US listener 12 and older hears approximately 1,100 hours of radio each year. (*Source: Arbitron Ratings, http://www.arbitronratings.com*)
- Since 1996, commercial spot loads have averaged 12-16 minutes per hour. These averages can be even higher during morning and afternoon drive hours. (*Source: "An Analysis of the Effects of Consolidation on the Radio Industry" http://mmstudio.gannon.edu/~gabriel/rapela.html.*)

Bibliography

- Arbitron Ratings http://www.arbitronratings.com/
- Castleman, Harry. Podrazik, Walter J. *The TV schedule book : four decades of network programming from sign-on to sign-off*. New York: McGraw-Hill, 1984.
- Central Intelligence Agency World Factbook 2000 http://www.odci.gov/cia/publications/factbook/
- Federal Communications Commission Audio Services Division, *Broadcast Station Totals*, 1990 - 1999 http://www.fcc.gov/mmb/asd/totals/index.html
- Fletcher, William H. Making Instructional Digital Video and Audio Work, http://miniappolis.com/mpeg/mpeg.html. April 1998.
- Internet Archive. Phone 415-561-6767 or see http://www.archive.org/
- National Cable Television Association http://www.ncta.com/directory.html
- *International Television and Video Almanac*. New York: Quigley Publishing, 1998.
- UNESCO, *Latest statistics on radio and television broadcasting*. Paris: UNESCO [United Nations Organization for Education, Science and Culture] Division of Statistics on Culture and Communication, 1987.
- U.S. Census Department, 1999 Statistical Abstract of the United States, www.census.gov/prod/www/statistical-abstract-us.html

See chart for more details

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Phone - Summary

The material in this section is drawn from Coffman and Odlyzko (1998, 2000).

The International Telecommunications Union (ITU) database provides estimates of telephone traffic for 207 countries for 1997-98, which total 2.5 x 10^{12} minutes per year. Adding in an estimate for the missing countries brings us to 7.5 x 10^{12} minutes per year, or roughly 600,000 terabytes per month. Compression would reduce storage requirements by a factor of 6 to 8.

The US accounts for about 250,000 terabytes per month, of which roughly a third is modem calls.

We have somewhat better estimates for long-distance traffic in the US, including voice, Internet, public data networks and private lines.

Table 1: Traffic on U.S. long distance networks in terabytes, year-end 1999.

Network	Traffic (terabytes/month)
US voice	48,000
Internet	10,000 - 16,000
Other Public Data Networks	2,000
Private Line	5,000 - 8,500

Notes

- Internet traffic has, on average, been doubling every year for 30 years, though the growth rates varied significantly during that period.
- The current rate of growth of Internet traffic is roughly 100% per year.
- By 2002 data traffic will surpass phone traffic.
- Residential Internet use in the US is growing at about 30% a year. When residential users switch to broadband, the total volume of data accessed increases by a factor of 5-10. As residential broadband grows, traffic should approach the 100% per year growth rate.
- Internal corporate (intranet) traffic is growing at about 30% per year, but corporate traffic to the public Internet is growing at 100% per year.
- According to Peter Davidson, of Davidson Consulting, about 300 billion seconds of fax transmissions take place annually. At 45 seconds per page, this means 400 billion pages per year are faxed. At 15 Kb per page, this comes to around 6,000 terabytes of fax information per year.

References

- **Internet growth: Is there a "Moore's Law" for data traffic?,** K. G. Coffman and A. M. Odlyzko, 2000. [Abstract] [PostScript] [PDF] [LaTeX]

- **The size and growth rate of the Internet,** K. G. Coffman and A. M. Odlyzko, First Monday 3(10) (October 1998), http://firstmonday.org/. [Abstract] [PostScript] [PDF] [LaTeX] [First Monday version]

- **The Worldwide Fax Traffic Statistical Market Review and Forecast, 1997-2002,** Peter Davidson, [Link to Report]

How Much Information?

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Mail - Summary

This table of facts about US mail is from Odlyzko (2000). About half of all mail is currently first class and about half is junk mail. If we assume 5 pages per piece of mail, and digitize it at 15 Kbytes per page, 1998 US mail is about **150 petabytes** per year.

Table 1: Statistics about US mail service, from Odlyzko (2000).

Year	Cost (millions)	Cost/GDP (percent)	Pieces (millions)	Mail per Person
1790	0.032	0.02	0.8	0.20
1800	0.214	0.05	3.9	0.73
1810	0.496	0.09	7.7	1.07
1820	1.161	0.18	14.9	1.55
1830	1.933	0.21	29.8	2.32
1840	4.718	0.28	79.9	4.68
1850	5.213	0.20	155.1	6.66
1860	14.87	0.39		
1870	24.00	0.33		
1880	36.54	0.35		
1890	66.26	0.51	4,005	63.7
1900	107.7	0.58	7,130	93.8
1910	230.0	0.65	14,850	161
1920	454.3	0.50		
1930	803.7	0.89	27,887	227
1940	807.6	0.81	27,749	211
1950	2,223	0.78	45,064	299
1960	3,874	0.77	63,675	355
1970	7,876	0.81	84,882	418
1980	19,412	0.70	106,311	469
1990	40,490	0.70	166,301	669
1998	57,778	0.68	197,943	733

- **The history of communications and its implications for the Internet,** A. M. Odlyzko. [Abstract] [PostScript] [PDF] [LaTeX]

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Acknowledgements

Overall

- Rita Gildea, EMC
- Jim Gray, Microsoft
- Andrew Odlyzko, AT&T Labs
- Dave Patterson, UC Berkeley
- Gil Press, EMC
- Robert Wilensky, UC Berkeley

Print

- Eric Chaskas, US National Archives and Records Administration
- Merilyn Dunn at CAPventure, via Kathy Jarvis at XeroxParc
- Rebecca Green and Lee Leighton, UC Berkeley Library
- Richard J. Hill, Newspapers Librarian, State Library of Pennsylvania
- Aimee Pyle and Amy Kirschoff, JSTOR
- Ginger Ogle, UC Berkeley Digital Library Project
- Rebecca Green and Lee Leighton, UC Berkeley Library
- Clara R. Williams, Information Consultant at the Hasleton Library, Institute of Paper Science and Technology

Optical

- Brewster Kahle, Internet Archive
- Brian Lewin, International Recording Media Association
- Jim Taylor, DVD Association

Magnetic

- Fred Moore, Horizon Information Strategies

Telephone

- Peter Davidson, Davidson Consulting

Internet

- John McCredie and Jerry Berkman, UC Berkeley
- Steve Lawrence, NTT
- Scott Kirkpatrick, Archive.org

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

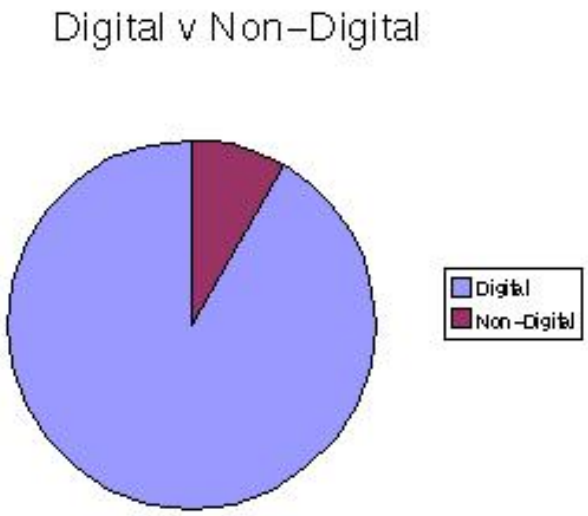
Site Map

How Much Information?

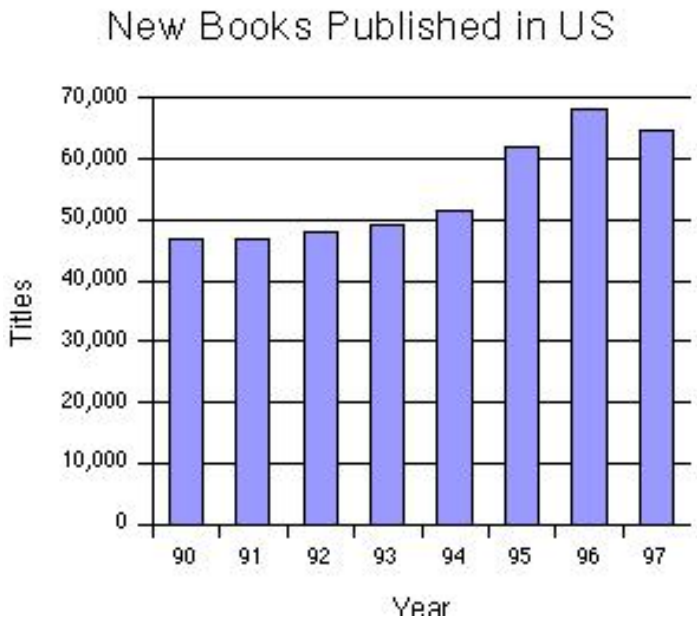
About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Charts

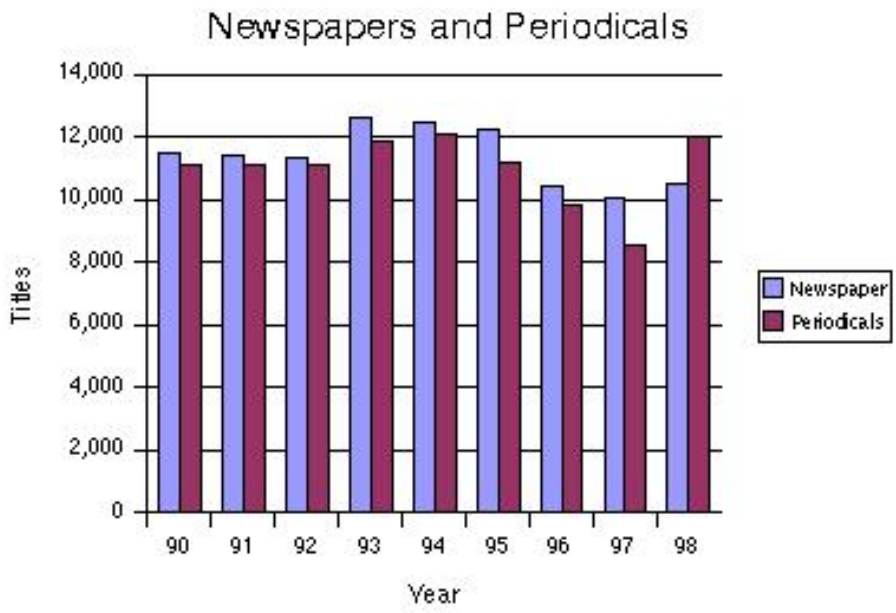
Here are some charts that show the relative sizes of various interesting magnitudes.



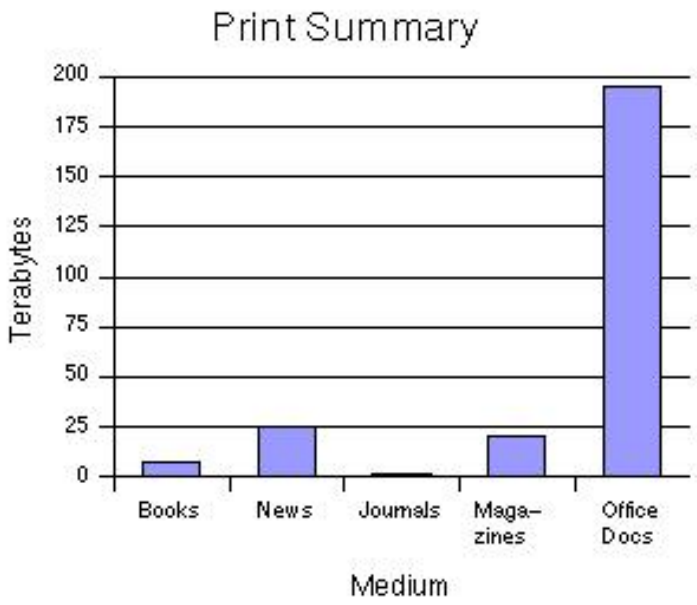
Over 93 percent of the information produced in 1999 was in digital format. (Authors' calculations.)



New books published in US. (*US Statistical Abstract 1999*, Table 938. Some years interpolated.)



Newspapers and periodicals published in US. (*US Statistical Abstract 1999*, Table 941.)



Print summary. (Calculations by authors.)

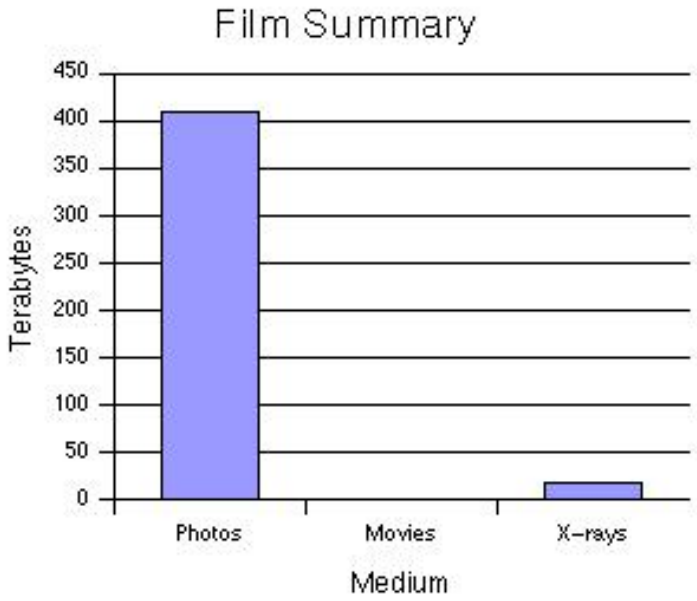
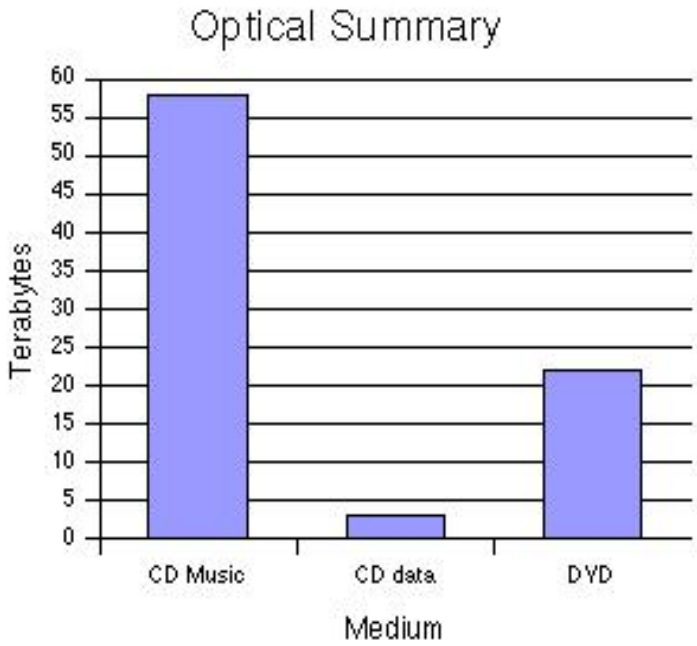
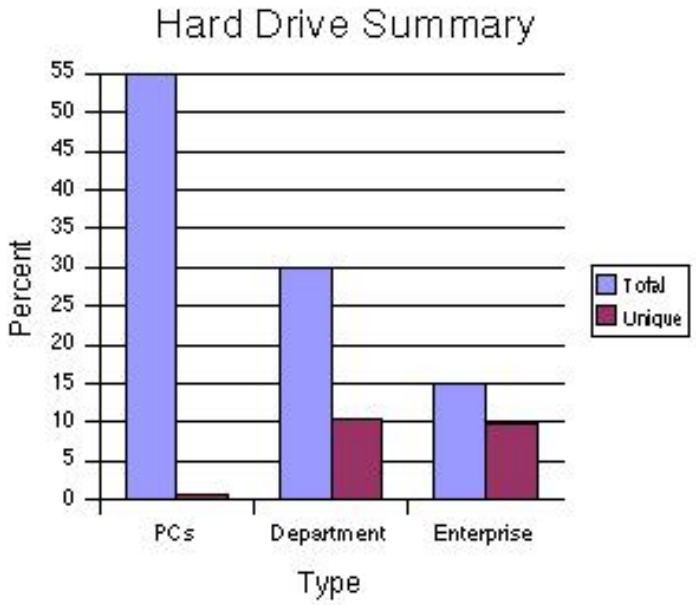


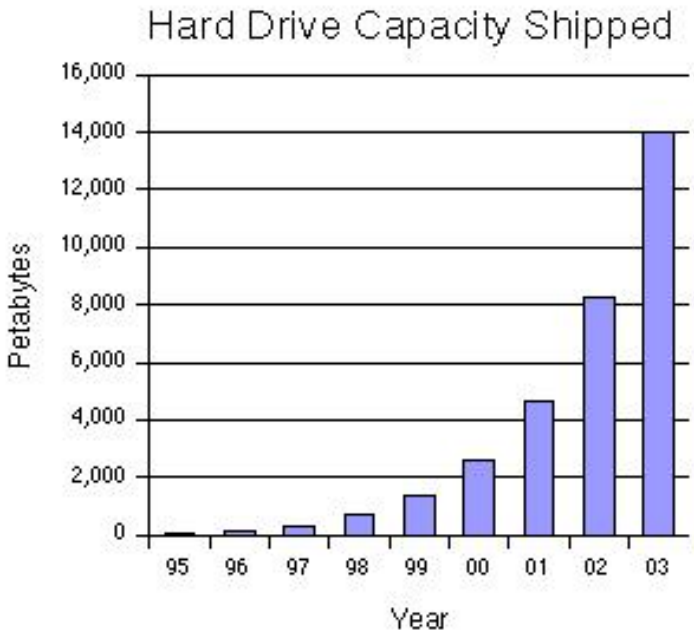
Photo summary. (Calculations by authors.)



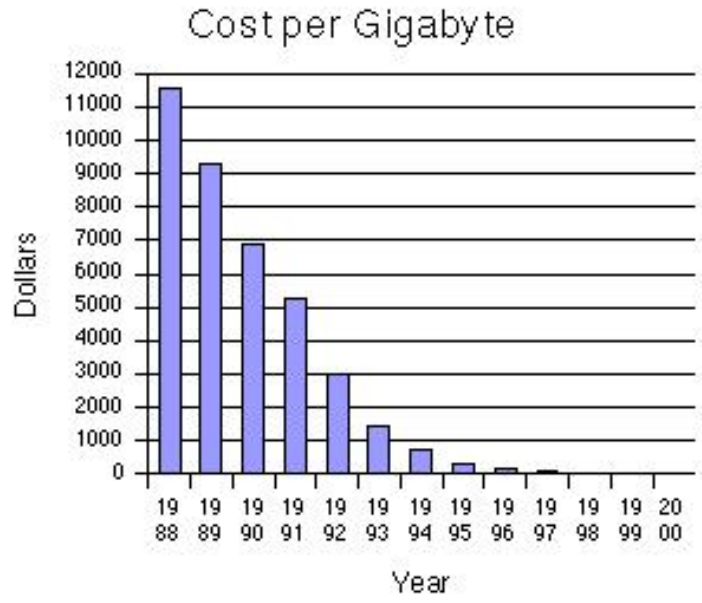
Optical summary. (calculations by authors)



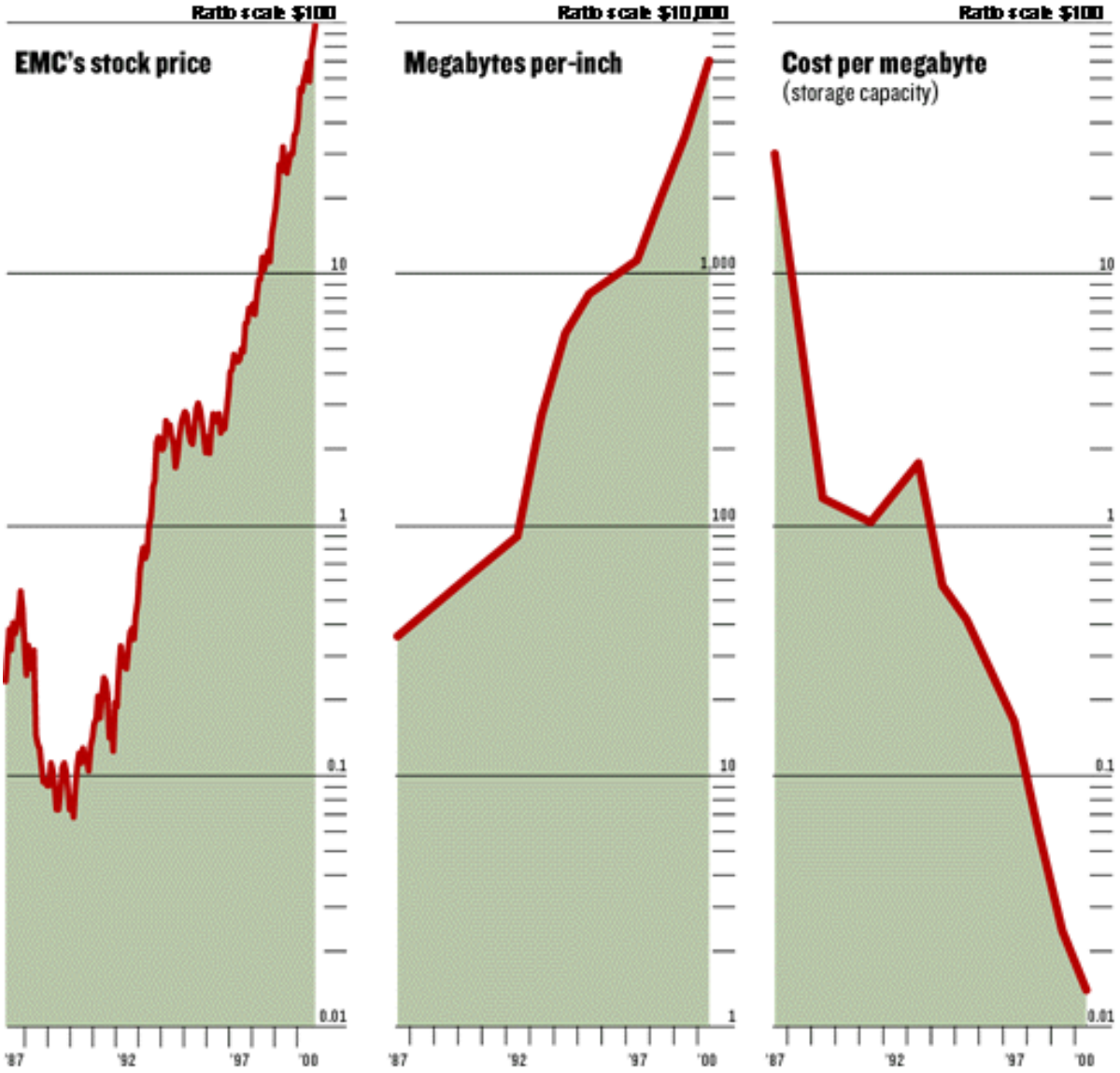
Total and unique information on various types of hard drives. (Calculations by authors)



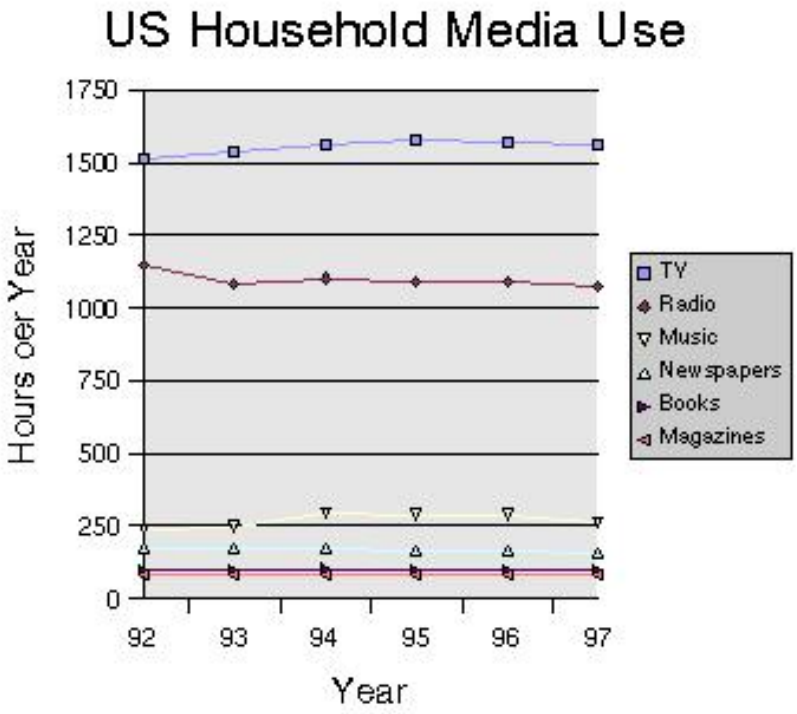
Worldwide PC hard drive capacity shipped. (*1999 Winchester Disk Drive Market Forecast and Review*, Table C5, International Data Corporation report. Some years forecast.)



Hard drive cost per gigabyte. (IDC reports.)

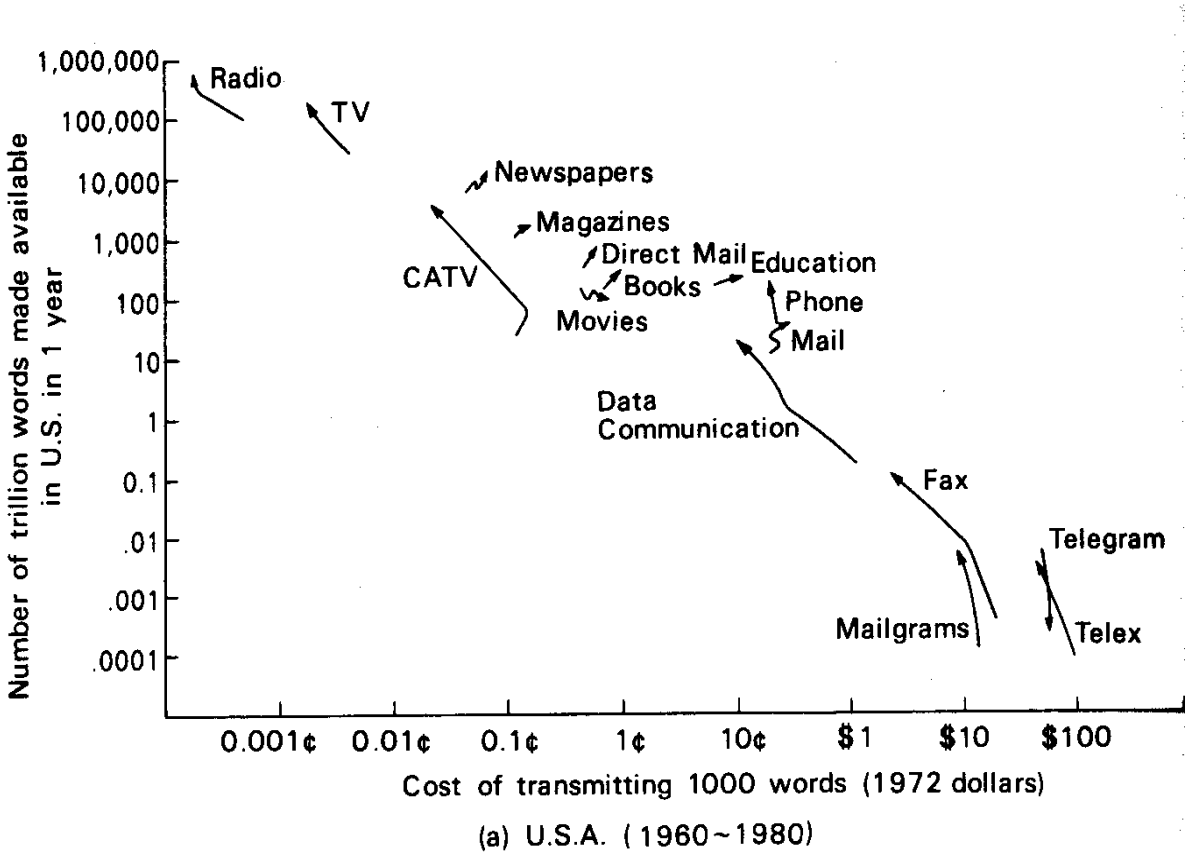


Trends in EMC's stock price, megabytes per inch, and cost per megabyte. From Forbes, October 2, 2000.



US household media use. (*US Statistical Abstract 1999*, Table 920.)

Activities performed by online households in the US during an average week, Q2, 2000. (*Economist Magazine*, October 7, 2000, E-Entertainment Survey, page 11.)



Cost of transmitting 1000 words v Number of words available in USA 1960-1980. (Ithiel de Sola Pool, Hiroshi Inose, Nozomu Takasaki, Roger Hurwitz, *Communication Flows: A Census in the United States and Japan,* Elsevier Science Publishers, New York, 1984.)

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Sound Bytes

Terror of terabytes.
One terabyte, the smallest practical measure for our project, is a million megabytes, which is equivalent to the textual content of a million books. An exabyte, which is what we use to report the final results, is a billion gigabtyes.

Capabilities of compression.
Conversion to ASCII, MP3, MP4 and other compression technologies dramatically reduces storage requirements by one to two orders of magnitude.

Democratization of data.
Individuals produce significant amounts of non-digital information. As photos and videos move to digital formats, households will have to manage terabytes of data.

Dominance of digital.
Ninety-three percent of the information produced each year is stored in digital form. Hard drives in stand-alone PCs account for 55% of total storage shipped each year.

Magnetic migration.
Print and film content is rapidly moving to magnetic and optical storage. This is true for professional use now, and will become increasingly true at the level of individual users.

Tape in transition.
Magnetic tape is about 10 times as large as disk storage, but is used almost exclusively for archives. Disk storage is much more attractive, even for archives, due to its rapidly declining cost and the fact that it is much easier to access data stored on disk.

Paucity of print.
If all printed material published in the world each year were expressed in ASCII, it could be stored in less than 5 terabytes.

Immensity of images.
Over 80 billion photographs are taken every year, which would take over 400 petabytes to store, more than 80 million times the storage requirements for text.

Convenience of copies.
There is a lot of redundancy both across and within media. A newspaper, for example, is composited using digital technology, printed on paper, then archived on microfilm. Estimates of ``unique'' information can only be taken as approximate.

Ubiquity of the US.
The US produces 35% of all print material, 40% of the images and well over 50% of the digitally stored content produced in the world each year.

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Data Powers of Ten

Many of the following facts were taken from Roy Williams "Data Powers of Ten" page at Caltech.

- **Byte [8 bits]**
  - 0.1 bytes: a binary decision;
  - 1 byte: a single character;
  - 10 bytes: a single word;
  - 100 bytes: a telegram ORr a punched card;
- **Kilobyte [1,000 bytes OR 10^3 bytes]**
  - 1 Kilobyte: A very short story;
  - 2 Kilobytes: A tTypewritten page;
  - 10 Kilobytes: An encyclopaedic page OR a deck of punched cards;
  - 10 Kilobytes: static web page;
  - 50 Kilobytes: A compressed document image page;
  - 100 Kilobytes: A low-resolution photograph;
  - 200 Kilobytes: A box of punched cards;
  - 500 Kilobytes: A very heavy box of punched cards;
- **Megabyte [1,000,000 bytes OR 10^6 bytes]**
  - 1 Megabyte: A small novel OR a 3.5 inch floppy disk;
  - 2 Megabytes: A high resolution photograph;
  - 5 Megabytes: The complete works of Shakespeare OR 30 seconds of TV-quality video;
  - 10 Megabytes: A minute of high-fidelity sound OR a digital chest X-ray;
  - 20 Megabytes: A box of floppy disks;
  - 50 Megabytes: A digital mammogram;
  - 100 Megabytes: 1 meter of shelved books OR a two-volume encyclopaedic book;
  - 200 Megabytes: A reel of 9-track tape OR an IBM 3480 cartridge tape;
  - 500 Megabytes: A CD-ROM OR the hard disk of a PC;
- **Gigabyte [1,000,000,000 bytes OR 10^9 bytes]**
  - 1 Gigabyte: a pickup truck filled with paper OR a symphony in high-fidelity sound OR a movie at TV quality;
  - 2 Gigabytes: 20 meters of shelved books OR a stack of 9-track tapes;
  - 5 Gigabytes: 8mm Exabyte tape;
  - 20 Gigabytes: A good collection of the works of Beethoven OR 5 Exabyte tapes OR a VHS tape used for digital data;
  - 50 Gigabytes: A floor of books OR hundreds of 9-track tapes;
  - 100 Gigabytes: A floor of academic journals OR a large ID-1 digital tape;
  - 200 Gigabytes: 50 Exabyte tapes;

- 500 Gigabytes: The biggest FTP site.

- **Terabyte [1,000,000,000,000 bytes OR 10^{12} bytes]**
  - 1 Terabyte: An automated tape robot OR all the X-ray films in a large technological hospital OR 50000 trees made into paper and printed OR daily rate of EOS data (1998);
  - 2 Terabytes: An academic research library OR a cabinet full of Exabyte tapes;
  - 10 Terabytes: The printed collection of the US Library of Congress;
  - 50 Terabytes: The contents of a large Mass Storage System;
  - 400 Terabytes: National Climactic Data Center (NOAA) database;
- **Petabyte [1,000,000,000,000,000 bytes OR 10^{15} bytes]**
  - 1 Petabyte: 3 years of EOS data (2001);
  - 2 Petabytes: All US academic research libraries;
  - 8 Petabytes: All information available on the Web;
  - 20 Petabytes: Production of hard-disk drives in 1995;
  - 200 Petabytes: All printed material OR production of digital magnetic tape in 1995;
- **Exabyte [1,000,000,000,000,000,000 bytes OR 10^{18} bytes]**
  - 2 Exabytes: Total volume of information generated worldwide annually.
  - 5 Exabytes: All words ever spoken by human beings.
- **Zettabyte [1,000,000,000,000,000,000,000 bytes OR 10^{21} bytes]**
- **Yottabyte [1,000,000,000,000,000,000,000,000 bytes OR 10^{24} bytes]**

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Printed Media - Details

Conversion Factors

To estimate the storage requirements for these non-digital information sources, one must make

some assumptions about the form in which they will be stored and the degree to which they will be compressed. There is considerable variation in the "bytes per page" proposed by various sources, partly because of differing assumptions about the content (e.g. ratio of text to pictures), formatting (e.g. number of characters per page, amount of white space), and file storage method (e.g. whether converted to a text or compressed document image file). Here is a sample of the range:

- According to Webopoedia: "A disk that can hold 1.44 megabytes, for example, is capable of storing approximately 1.4 million characters, or about 3,000 pages of information." 1.44 MB = 1,509,949 bytes = 3,000 pages = **503 bytes per page**
- According to Horison Information Strategies: 2,400 bytes **(2.3 KB)** for one page of paper. 300 - 500 MB for large encyclopedia
- According to **Michael Lesk**: 5,000 bytes (about **5 KB**) per page (Source: "How Much Information is There in the World?", 1997)
- 30 - **50 KB** for a page of mathematical text in bitmap form (Source: Andrew Odlyzko, "Tragic loss or good riddance? The impending demise of traditional scholarly journals," 1994)
- According to Roy Williams of CalTech, "Data Powers of Ten": **10 KB** per encyclopedic (small font, dense formatting) page, **50 KB** for a compressed document image page
- According to ArchiveBuilders.com **50 KB** for one scanned page but **800 KB** for one scanned engineering drawing
- According to JSTOR, a digital library of scholarly journals, one scanned document page is **130 KB**; compressed using Cartesian Perceptual Compression (CPC) a page is only **26 KB**.
- According to Internet Archive: **1 MB** for one mystery novel. 1 copy of Encyclopedia Brittanica = 1 gigabyte (2,619 pages per copy so 409,981 bytes or about **400 KB** per page)

To demonstrate this range, for our estimates, we will use three different numbers: one for a scanned image (uncompressed, at archival quality - 600 dpi), one for a compressed image file, and one for and ASCII (or plain text) file. For the summary statements, will use the mid-range number, as most documents are scanned with some measure of compression. We opted to use Cartesian Perceptual Compression (CPC) as our compression standard because although it is a lossy algorithm, it is "non-degrading" - the human eye can't determine any difference. And it keeps the full 600 dpi resolution.

Originals

Books

Conversion Factors

To estimate the size (in bytes) of all books published in the world, use the following formula:

Average number of pages per book * Average file size per page * Number of books published per year.

Table 1: Books Conversion Factors

Average # of Pages per Item	Storage Format	Average File Size	Conversion Factor (rounded)
300	Scanned TIFF (600 dpi)	130 KB	40 MB/Item
	Compressed	26 KB	8 MB/item
	Plain text	2.5 KB	1 MB/Item

Flow

World

Worldwide, in 1996 there were 808,066 titles published, according to *UNESCO's Statistical Yearbook 1998*, as cited in *The Bowker Library and Trade Almanac*. This is roughly equivalent to **7 TB/year**. The editor of the *Almanac* notes problems with UNESCO's data-gathering--inconsistencies in reporting, data collection lags, and missing data; nonetheless, the editor acknowledges that UNESCO provides researchers the best and most usable overview of international book title output. If one includes the most recently available title production figures for countries that did not show 1996 data (such as China, France, and the Netherlands), the annual world total increases to **968,735** - about **8 TB/year**. This number is still not quite accurate because there are still about 70 countries of the world not documented by UNESCO. However, it is the best available estimate we were able to locate.

United States

In 1997, about **64,711 books** were published in the United States, according to the *U.S. Census Bureau's 1999 Statistical Abstract of the United States*. *The Bowker Annual Library and Book Trade Almanac* provides a similar figure for 1997 (65,796 titles) and notes that this is a decrease from 1996's all-time high figure of 68,175 titles. The preliminary estimate for 1998 suggested another decrease, to 56,129 books, roughly equivalent to **.5 TB** per year.

Stock

World

To estimate the international stock of books currently available for purchase, we note that the United States produces about 40% of the world's printed material. (Source: US Industry and Trade Outlook 2000) The national library and copyright repository of the United States--the Library of Congress--contains about 26 million books. Therefore, using the 40% rule of thumb, the world stock of books might be approximately **65 million titles**.

United States

Books In Print is an authoritative source for books published and/or distributed in the United States. The 1999-2000 edition of Books in Print (with supplement) includes **1,663,815 books** - about **13 TB** in all. (This number does not include: books not published and/or not distributed in the U.S.; books not available to the trade or general public; free books not included with a title for sale; unbound material, pamphlets and booklets; periodicals and serials; books only available to members of an organization; subscription-only material, and music manuscripts, sheet music, and librettos.) According to a press release from January 2000, booksinprint.com 2000 includes **3.2 million titles** - about **26 TB** total. (Of these, about 500,000 titles are out-of-print.) This figure is consistent with online booksellers: Barnes & Noble.com asserts that it has 3 million titles available while Amazon claims to have 4.5 million titles. *Source: Wall Street Journal, July 17, 2000, "A New Chapter: Independent Booksellers Hope to Find Strength in Numbers" by Scott Eden.*

Rate of change

The number of titles published each year in the United States and internationally has risen gradually and fairly consistently over the last 10 years, apart from a slight downturn in US title production in 1997.

Newspapers

Conversion Factors

A large metropolitan daily US newspaper will run approximately 60 pages each day, and may vary from 48 to 72 pages. Therefore one year of a large US newspaper would consist of about 21,900 pages per year. To account for the smaller regional papers and those which are not published daily, we opted to use 30 for the average number of pages, which results in 10,950 pages per year.

A double truck (center fold) full broadsheet is 24 in X 36 in. Because a newspaper page would be scanned at higher resolution and contains detailed graphics, a double - truck would require about 1 Megabyte (uncompressed) and a single full broadsheet page (18 X 24 inches) would require about .5 MB.

Table 2: Newspapers Conversion Factors

Average Number of Pages per Year	Storage Format	Average File Size (per page)	Conversion Factor (rounded)
10,950	Scanned TIFF (600 dpi)	500 KB	5475 MB/item
	Compressed	100 KB	1095 MB/item
	Plain text	10 KB	110 MB/item

Flow

World

The number of newspapers in the world as of 1999 is **22,643**, according to the International Standard Serial Number Register. UNESCO's Statistical yearbook from 1996 offers a different, much smaller figure: **8,391**. The discrepancy is due largely to the fact that multiple ISSNs may be assigned to what is essentially the same information product if it exists in two different formats (paper, online, floppy disk, CD-ROM, microform). For example, the New York Times print version and NYT online have two different ISSNs. Other differences may be attributed to the fact that UNESCO's excludes non-daily newspapers from its statistics or that the UNESCO data set is less complete than that of ISSN.

Using the ISSN figure and the compressed format:

.1MB/page * 10,950 pages/year * 22,643 newspapers/world = **124 TB/year**

United States

According to the Newspaper Association of America, there were **1,489** daily newspapers and **897** Sunday - only newspapers published in the United States in 1999.

Stock

The Library of Congress provides useful statistics on the nationwide stock of printed periodicals and newspapers. Most library print periodical holdings have been transferred to microfilm, microfiche, or simply made accessible online, through efforts such as the National Endowment for the Humanities' Newspaper Project. However, bound periodicals do still exist at the LOC: about 30,570 newspapers and 557,738 serials.

Rate of Change

The number of newspaper titles and the circulation figures have both declined slowly over the past 10 years.

Periodicals

Conversion Factors

Table 3: Periodicals Conversion Factors				
Periodical Type	**Average # of Pages per Item (annual total)**	**Storage Format**	**Average File Size (per page)**	**Conversion Factor (rounded)**
Scholarly Journal	**1,723**	Scanned TIFF (600 dpi)	130 KB	225 MB/item
		Compressed	26 KB	45 MB/item
		Plain text	2.5 KB	4 MB/item
Mass-Market Periodical	**5,000**	Scanned TIFF (600 dpi)	130 KB	650 MB/item
		Compressed	26 KB	130 MB/item
		Plain text	2.5 KB	13 MB/item
Newsletter	**150**	Scanned TIFF (600 dpi)	130 KB	20 MB/item
		Compressed	26 KB	4 MB/item
		Plain text	2.5 KB	.4 MB/item

World

According to the 2000 Ulrich International Periodical Directory, there are **158,000** unique periodical titles in the world. If one wished to include all serial publications, for example, yearbooks, annual publications of all kinds, not just journals, one could cite the International Standard Serial Number statistics: **617,801** for all serials around the world.

It is useful to differentiate between scholarly journals and mass market periodicals because of the differences in number of pages per year. According to a recent study on Economic Cost Models of Scientific Scholarly Journals, "A typical 1995 scientific scholarly journal has 8.3 issues... and 1,723 pages (or 208 pages per issue)." (*Source:* http://www.bodley.ox.ac.uk/icsu/kingppr.htm) A typical mass-market periodical, (such as Time, Newsweek, or People) that is published weekly, will be about 5,000 pages per year. (*Source: Robert M. Hayes, "The Economics of Digital Libraries",* http://www.usp.br/sibi/economics.html). There are also newsletters, smaller than both the scholarly journal and mass-market publications, at about 100 pages per year.

Using the ratios within the US periodical statistics as a reference point (see below), we can estimate that about half of the world's periodical publications are mass - market magazines/tabloids, one-fourth are scholarly journals, and one-fourth are newsletters.

United States

According to the *1999 US Statistical Abstract*, there are **12,036 periodicals** published in the United States. Given the world total of 158,000, this seems low. Another estimate on United States periodical production comes from the National Directory of Magazines: **20,613 titles** as of August 2000 (this figure includes Canada).

The estimated number of scientific journals published in the United States increased 62 percent, from 4,175 in 1975 to **6,771** in 1995 (*Source: Designing Electronic Journals With 30 Years of Lessons from Print, Tenopir and King, 1998* http://www.press.umich.edu/jep/04-02/king.html) Data on periodicals in the field of modern languages and literature may serve as a rough index of the trends in the general availability of scholarly journals outside the sciences. According to the MLA

Directory of Periodicals, there are currently **3,700 periodicals** in the areas of literature, language, linguistics, and folklore covered regularly in the MLA International Bibliography. This means that there are about **10,500 scholarly journals** in the United States.

According to the Newsletter and Electronic Publisher's Association (NEPA), there is no way of really knowing exactly how many newsletters exist, because they are not required to be registered anywhere. There are two directories of newsletters: the Hudson Subscription Newsletter Directory lists over 5,000 subscription newsletters while the Oxbridge Directory of Newsletters includes over 20,000 newsletter titles, of which they say 9,000 are subscription newsletters. NEPA's best guess is that there are approximately **10,000 subscription newsletters** in print today. (*Source: http://www.newsletters.org/faqs.htm*)

Office Documents

Conversion Factors

Table 4: Office Documents Conversion Factors			
Average # of Document Pages per Year	**Storage Format**	**Average File Size**	**Total Annual Production**
7,500,000,000	Scanned TIFF (600 dpi)	130 KB	975 TB/year
	Compressed	26 KB	195 TB/year
	Plain text	2.5 KB	19 TB/year

Flow

United States

To estimate the amount of information generated by offices, one might begin by looking at the statistics of the Federal Government, which is the single largest employer in the United States, with 1.8 million civilian workers as of 1998 and 1.2 million individuals in the armed services. The Federal Government, in total, employs about 3% of the nation's workforce.

The National Archives in Washington D.C. retains 2% of what the government produces, across a range of media. According to Archives representative Eric Chaskas, the Archives retains only what is deemed to be of some permanent historical value. Document types include correspondence, registers, reports, forms, treaties, case files, and log books. The perceived value determines how long a record will be retained--some will be kept indefinitely, while others are retained for no more than 6 months. An effort is made to prevent duplicating records but there is still some degree of overlap. Although the NARA representative was unable to quantify how much new information is accessioned each year, he did state that the current archival holdings, as of July 2000, include 4 billion pieces of paper, occupying a total of 21.5 million cubic feet. That is about 200 pieces of paper per cubic foot.

If one divides 4 billion by the number of years the United States government has existed (224 years), one could obtain a rough number of pages collected per year by dividing 4 billion by 224 - the result is about 18,000,000 pages per year.

The current accession rate, however, appears to be much higher. Each year, Federal agencies submit about 4,000 items and about 75% of these (3,000) are processed for archival. Although the Archives does not publish statistics on the average size of these items, it is known that NARA adds a total of 500,000 cubic feet of mostly paper-based records each year. As previously noted, in

archives, 1 cubic foot can hold 200 pieces of paper, so the total annual accession rate is therefore about **100,000,000 pages** per year.

If this represents 3% of the nation's workforce, then one could estimate that United States companies produce a total of more than **3 billion archiveable pages** each year, equivalent to **78 terabytes.**

World

Using our rule of thumb that the US produces about 40% of the world's printed materials, we can estimate that each year the world produces **7.5 billion archiveable pages**, which would be equivalent to **195 terabytes.**

Stock

Again, consider the US National Archives with 4 billion pieces of paper. If this is 3% of the United States total, then the total stock of paper held by US companies is somewhere on the order of 130 billion sheets of paper - about **3,380 terabytes**.

Rate of Change

Despite the increasing use of computers for communication and storage, the production and retention of paper office records is on the rise. The materials within the National Archives has increased in volume by more than one-third, just in the past decade.

Visual Material (Maps, Posters, Prints and Drawings)

This is a difficult category to quantify because, unlike the book, newspaper, and periodical industries, there is no central agency monitoring the production of originals or serving as a repository for the items. The best method for approximating the production statistics is to consider the holdings of the National Archives and the Library of Congress.

Conversion Factors

According to the Library of Congress, maps scanned in at 300 dpi tend to be large, approximately 100 to 300 MB. (*Source: http://lcweb2.loc.gov/ammem/formats.html#V*)

Flow

For photographs, see the Film Summary portion of this paper.

During 1998, the Library of Congress added 60,150 maps, 1,881 posters, and 4,723 prints and drawings to its collections.

Stock

World

[???]

United States

The Library of Congress has 4,523,049 maps in its collections. 85,216 posters and 405,708 prints and drawings. The National Archives has another 5 million maps and drawings, as well as 13 million still photographs and 9 million aerial photographs.

Copies

For the past 10 years, the US has produced about 30% of the world's paper and paperboard output (*US Industry & Trade Outlook 2000*). Other major producers are Asia, Latin America, Canada, and Europe. As of 1998, total global capacity was 333.6 million metric tons, with growth predicted so that by 2001 the annual total is expected to be 348.1 metric tons. A recent article in the Economist supports this forecast stating that, during the period 1993 - 1998, "the production of printing and writing paper in North America has grown by over 13%. Worldwide it has doubled since 1982."

There are four different commodity segments in the paper/paperboard industry, only two of which are relevant to this study: printing and writing paper and newsprint. In 1999, the US produced 23.8 million metric tons of printing and writing paper and 6.4 million metric tons of newsprint. The most recent global production figures (available from UNESCO's Statistical Yearbook) are for 1997, at which time the world was producing 90 million metric tons of printing and writing paper and 36 million metric tons of newspaper.

For the estimates of **printing/writing paper**, we used the midrange figure of **26 KB per 8 1/2 X 11 page**. We estimated 500 sheets of standard grade 8 1/2 X 11 paper would weigh 5 pounds. Therefore, each metric ton (2204 pounds) equals about 440 500-sheet-reams or 220,000 sheets. Multiplying by 26 KB per page results in **6 GB per metric ton**, for a total annual storage capacity of **142,800 possible terabytes** (US) and **540,000 terabytes** (world).

For **newspapers**, we estimate that storage requirements would be be about 2 times that for writing/printing paper (**12 GB per metric ton**). Newspapers tend to contain more words and graphics per page, requiring, on average 1 MB per page. Furthermore, newsprint is thinner and lighter, so each metric ton contains more individual sheets. US newsprint production in 1997 equalled about **80,000 terabytes** and international production was about **432,000 terabytes**.

These figures provide an extreme upper bound on the total number of bytes required to digitally store information currently produced in printed format each year.

Books

About **1.1 billion books** were sold in the United States in 1999. (*Source: Wall Street Journal, July 17, 2000, "A New Chapter: Independent Booksellers Hope to Find Strength in Numbers" by Scott Eden.*) This is an increase of more than 3% from 1998 when about 1,037,000 books were sold. (*Source: The 1999 Consumer Research Study on Book Purchasing, (Section II) Detailed Findings: Consumer Adult Book Purchasing*).

Newspapers

In the United States **55,979,332 daily newspapers** and **59,894,381 Sunday newspapers** circulate each year. (*Source: Newspaper Association of America.*)

Periodicals

The total number of US magazines circulated annually exceeds **500 million.** (*Source: US Industry and Trade Outlook 2000.*)

Office Documents

Each year, almost **500 billion copies** are produced on copiers in the US (**13,000 TB**); nearly **15 trillion copies** are produced on copiers, printers, and multi - function machines. 15 trillion copies is roughly equal to **390,000 TB**. This accounts for the majority of the printing/writing paper used each year. (*Source: XeroxParc.*) For specific information on fax printing, see the Telecommunications Summary of this report.

A 1998 Coopers & Lybrand study showed that the average office makes 19 copies of each document. The average office loses 1 out of 20 office documents. It then costs: $120 to search for the document; $250 to recreate it, if lost (1 lost document = $370). (*Source: NAGARA, Records Management Technical Bulletin www.nagara.org/rmbulletins/bulletin_2.htm*)

The Government Printing Office, which handles the printing needs of the entire Federal community,

uses or sells more than 55 million pounds of paper each year. Approximately 10,000 titles are available for sale to the public at any given time.

Fun Facts About Print Media

Paper Size Equivalencies

According to ArchiveBuilders.com: 1 pulp tree (loblolly pine) = 1/10th cord of wood = 10,000 pages = 1 file cabinet = 4 boxes = 1/2 Gigabyte = 1 CD

Scope of the Library of Congress Collections

"LC21: A Digital Strategy for the Library of Congress," (a report from an advisory group at the National Academy of Sciences about strategic directions for the applications of information technology in the Library) reports the following statistics about the collections of the Library of Congress:

- 26 million volumes (pages 2-3)
- 1 million serials (pages 2-3)
- 4.5 million maps (pages 2-5)
- 1,000,000 audio recordings (pages 2-6)
- 200,000 cans of film (pages 2-6)
- 13.5 million images (pages 2-7)
- 70,000 periodicals (pages 2-7)
- 1,400 newspapers (pages 2-7)

The Library receives some 22,000 items each working day and adds approximately 10,000 items to the collections daily.

Paperless Society?

With the spread of the personal computer, many predicted the advent of the paperless office. So far, this prediction has not come to pass. One one hand, according to BIFMA, a business and institutional furniture manufacturer's association, as of Dec. 16, 1999, the percentage of file cabinets in relation to other types of office furniture, like seating, desks, and tables, has steadily decreased, from 16% in 1988 to 12.9% in 1998. This would lead one to believe that less paper is being generated to require storage space. On the other hand, the installation of home computers and expansion of home offices has led some analysts to predict that by 2003 the consumption of standard office paper will double from the 1996 level (*Source: 1999 U.S. Industry and Trade Outlook, 10-3*).

Growth of Online Periodicals

Of the more than 157,000 serial titles reported in the latest edition of the international periodicals directory, 10,332 were available exclusively online or in addition to a paper counterpart (*Source: Ulrich's International Periodicals Directory, 1999, p. vii*).

Bulk Mail

As of 1990, the United States Postal Service handled almost 90 billion pieces of first-class mail per year and another 75 billion pieces of second- , third-, and fourth-class mail.

Americans receive almost 4 million tons of junk mail a year. About 44% of the junk mail is never opened. Every person in the United States receives junk mail that represents the equivalent of one and a half trees a year. (*Source: The Consumer Research Institute's Stop Junk Mail Page*)

For more information on flows of information through the mail, please see the Mail Summary of this report.

Print Media Bibliography

- Bogart, Dave, ed. *The Bowker Annual Library and Book Trade Almanac*, 44th edition. New Jersey: R.R. Bowker, 1999.
- Cummings, Anthony M., Marcia L. Witte, William G. Bowen, and Laura O. Lazarus. *University Libraries and Scholarly Communication: A Study Prepared for the Andrew W. Mellon Foundation*. The Association of Research Libraries, 1992
- Hayes, Robert M., UCLA School of Information Science, "The Economics of Digital Libraries" www.usp.br/sibi/economics.html
- King, Donald W. and Carol Tenopir. "Economic Cost Models of Scientific Scholarly Journals," 1998. www.bodley.ox.ac.uk/icsu/kingppr.htm
- American Forest and Paper Association, *1999 Statistics for Paper, Paperboard and Wood Pulp*. Washington, DC, 1999. To order a copy, see www.afandpa.org/about/about.html or call (800) 244-3090.
- *Annual Report of the Librarian of Congress*. Washington, DC: Library of Congress, 1999.
- ArchiveBuilders.com White Paper, "Computer Storage Requirements for Various Digitized Document Types"
- Association of Research Libraries Statistics www.arl.org/stats/index.html
- *Books in Print, 1999-2000.* New Jersey: R.R. Bowker, 1999.
- International Standard Serial Number Register www.issn.org
- International Standard Book Number Register www.isbn.org
- JSTOR digital library. www.jstor.org/
- Magazine Publishers of America, Information Center, (212) 872-3745. www.magazine.org/resources/fact_sheets/ed1_8_99.html
- National Directory of Magazines, www.mediafinder.com/mag_home.cfm
- Newsletter and Electronic Publisher's Association (NEPA) www.newsletters.org/faqs.htm
- *Newspaper Association of America, Facts About Newspapers*.
- Newspaper Project, National Endowment for the Humanities, www.neh.gov/preservation/usnp.html
- *Oxbridge Directory of Newsletters 1997*. New York: Oxbridge Communications, Inc., 1997.
- *Ulrich's International Periodical Directory*. New York: Bowker Publishing, 2000.
- *UNESCO Statistical Yearbook 1999.* Paris, UNESCO, 1999.
- U.S. Census Department, 1999 Statistical Abstract of the United States, www.census.gov/prod/www/statistical-abstract-us.html
- *U.S. Department of Labor, Bureau of Labor Statistics*
- U.S. Government Printing Office, *Prepared Statement before the Committee on Rules and Administration, U.S. Senate on Public Access to Government Information in the 21st Century* July 1996.
- *U.S. Industry and Trade Outlook*. Available in print from McGraw Hill/U.S Department of Commerce Washington, D.C. or download from www.ntis.gov/products
- Chaskas, Eric. US National Archives and Records Administration.
- *Walden's Paper Report*. Twice-monthly newsletter, published by Walden-Mott Corporation, Ramsey, NJ. Available by subscription only. See www.walden-mott.com/PaperReport/PAP_RPT.HTM or call (201) 818-8630.

Charts

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Film - Details

Impact of Digital Cameras On Rate of Growth of New Photographs

80 billion new photographs are taken around the world every year. Photography generally follows the trends of the overall economy, for example, in the mid-1990's there was a slight slowdown in photography concurrently with the economic recession experienced in many major world economies, particularly those in Asia. However, it is expected that in the next five years there will be continuing growth in the overall number of photographs as the enormous potential of China, India and other developing regions is realized. At the present 25% of film sales occur outside of North America, Western Europe and Japan. In June 2000 Kodak announced that China was its second largest market after the United States. To give some perspective to the growth potential for photography in China, film usage there is currently less than one roll per capita as opposed to 3.6 rolls per capita in the United States. Similarly, only 15% of Chinese households own cameras as opposed to over 80 percent in the United States.

A countervailing trend to the growth of silver halide film based photography is the increasing popularity of digital photography. Very rapid growth is predicted over the next five years for this method of taking pictures. In 1999, in the United States almost 2 million digital cameras were sold, double the number from the year before and representing about 12 percent of all cameras sold. In a survey by NPD Intellect, more than 70% of consumers planning to buy a new camera say they will choose a digital one. Kodak projects that digital photography will account for forty-five percent of its revenue by 2005 from a current 17%.

Photofinishing News and Lyra Research predict that there will be an 80 percent growth in the number of digital cameras worldwide by the year 2002. There were 8.3 million digital cameras in use worldwide in 1999 according to this study. (compared to 200 million conventional film cameras in the United States alone). Salomon Smith Barney predicts that the growth of photographic exposures worldwide will grow from 71.4 billion in 1997 to 96.8 billion by 2002. At the same time, they expect digital camera exposures to grow 1500 percent. Many of these photographers will use web photo services. For example, Photo Works already hosts over 100 million images on its web site and there

are many more companies competing in this same market. These companies indicate that at this time 60 to 80 percent of the photographs on their web sites are scanned from silver halide film.

Kodak predicts that the growth trend will predominate until at least 2004, at which point traditional film will gradually decline in use as the digital camera becomes ascendant.

The Internet may actually be increasing demand at the present for traditional photographs points out Morgan Stanley Dean Witter. eBay has 4 million items for sale every day and the inventory turnover averages about 7 days. About 80 percent of these items are accompanied by a photograph. About one-half of one percent of current U.S. photographs are pictures of merchandise offered for sale on eBay.

Table 1: Photographic Exposures (Billions)

Year:	1992	1997	1998	1999	2000	2001	2002	2005
World	56	84.4	83.3	82			89	
US	21	24.9	26.9		34.4			41.2

Source: 1992 statistics are from U.S. Industrial Outlook 1994; 1997 and 1998 U.S. Statistics are from U.S. Industry and Trade Outlook 2000; 1997 and 1998 world statistics are from the Silver Institute. The 1999 world figure is from Kodak's corporate web page. The 2000 U.S. figure is from Kodak and printed in May 8, 2000 Wall Street Journal. The 2002 projection for the world is from the Silver Institute. The 2005 projection is from the May 8, 2000 Wall Street Journal article and attributed to Kodak.

Chart of the annual production of film used in conventional photography in the United States.

Conversion and Compression

Photographs

To assess the amount of data in terms of bits in a photograph, certain assumptions must be made about the size of the film used and the technical way in which the photograph is digitized and stored. Professional photographers generally use physically larger film formats with more chemical grains on the film to store information than do amateur photographers. Conversion of these professional photographs to a digital medium, therefore, generates a much bigger computer file. For example, high quality professional photographs may require 40 megabytes or more of storage space.

In the early 1990's Kodak introduced the PhotoCD format for the storage of photographs on compact disks. Since then, the format has been widely used by consumers, professional photographers and archivists. Kodak estimates that most photographs can be converted to its PhotoCD format with little to no loss of image data in 5 megabytes. Kodak's rule of thumb is that it can place 100 photographs on one CD, which holds approximately 650 megabytes. For the sake of carrying out our estimates of total photographic data, therefore, we use this same 5 megabyte figure.

Of course, actual digital storage of photographs might consume far less space if any of the very popular compression schemes are used. For example, the JPEG standard is commonly used to reduce photographic file sizes to one-tenth of their original size. However, this is a "lossy" compression, i.e., one where data is irrecoverably discarded in the compression.

The 5 megabyte conversion factor used is also supported empirically by the extensive image digitizing experience of the University of California, Berkeley Digital Library. The photographic collection there is reported as containing 164,702 images which require 888 gigabytes of storage space, or just over 5 megabytes per photograph. (http://elib.cs.berkeley.edu/admin/rpts/TestSuite/00q2.html)

Motion Pictures

The conversion of motion pictures to digital media can result in very large file sizes. It is estimated that each frame of film can be digitized in about 12 megabytes. However, to create the illusion of seamless motion, 24 images are used per second of film. Thus, one second of film requires 288 megabytes of digital storage, one minute requires 17.28 gigabytes, and one hour requires 1 terabyte. At this rate a feature length film of 100 minutes would fill about 400 typical DVD's. Fox Animation Studios, for example, reports that its animated feature "Anastasia" is 1.49 terabytes in size. This particular film was digitally inked and painted and only then transferred to film for theatrical distribution.

In cases where film will be distributed or stored on digital media such as DVD's some form of compression must be used to make the file sizes tractable. The most common compression format now used for motion pictures is MPEG-2. Using this format, a feature lenth film can be reduced to several gigabytes in size and placed on a single DVD. This is a "lossy" compression format where data that is present on the film is irrecoverably lost in the compression process.

X-Rays

The conversion of x-ray images to bits requires careful attention to the loss of information because of the attendant risk of harm to patients. See, The Economist, "Why JPEG's Can Be Bad For Your Health", June 2000, citing a study published in the Journal of the American College of Cardiology in which researchers found that using JPEG with a compression ration of 16:1, the error rate was 30% higher than for uncompressed images. In fact, in the United States conversion of x-rays to digital format is regulated by the U.S. Government for the protection of the patients. FDA Regulation 510(k).

1 chest X-ray = 1 megabyte (14 x 17 inches), 150 dpi, 12 bits (compressed). 12 bits per pixel, provides 4,096 shades of grey) (wavelet compression, lossless mode, has FDA 510(k) approval) 150 dpi, 12 bit images recommended by American College of Radiology for primary reads). (Source: Steve Gilheany, Archive Builders)

A team of radiologists from the University of Florida estimated the size of a converted conventional radiograph as 10 megabytes assuming a matrix size of 2048 x 2580 pixels with a 16-bit sample of each pixel. This estimate assumes there will be no compression. Honeyman, Huda, Frost, Palmer & Staab, "Picture Archiving and Communications System Bandwidth and Storage Requirements", Journal of Digital Imaging, Vol. 9, no. 2, May 1996.

A slightly smaller estimate of the amount of storage required for a standard chest x-ray comes from the University of Pittsburgh, Clinical Multimedia Lab, which estimates the size for a radiograph (2K * 2K * 16 bits) as approximately 8 megabytes (uncompressed). Clunie, David A., Lossless Compression of Grayscale Medical Images - Effectiveness of Traditional and State of the Art Approaches (Link to this paper (PDF))

This latter figure will be used because it does not assume compression and is generally consistent with the assumptions made for converting traditional photographic film to digital format. Furthermore, Dr. H.K. Huang, Radiologist at UC San Francisco, estimates that "a typical examination generates betwen 10 and 20 MBytes." Huang, "Teleradiology Technologies and Some Service Models" Computerized Medical Imaging and Graphics, Vol.20, No.2 (1996). Applying this assumption of 8 mb per x-ray image results in a total amount of data stored on x-ray film of 16 petabytes.

Flow of New X-Rays

There are three common uses of x-ray film: medical imaging, dental imaging and non-destructive

testing in manufacturing processes. The Silver Institute Reports that the medical uses of x-ray film overwhelm the other categories, accounting for 92% of the world's overall x-ray film usage.

Medical Imaging

The administration of radiologic procedures in the United States and the developed world are fairly well documented. The number of such procedures in the developing world, however, are harder to come by. The United Nations Scientific Committee on the Effects of Atomic Radiation found that no data at all could be found for radiological procedures for half the world's population and that there is only fragmentary data on examination rates for another quarter of the world's population.

Table 2: Medical X-Ray Procedures in United States

1992	1993	1994	1995	1996	1997	1998	1999	2000
275	280	286	291	297	303	309	315	320

Source: Theta Reports

The UN Committee observed that the developed nations of the world use x-rays at a rate generally consistent with that of the United States, i.e., approximately 1000 procedures (including dental uses) for 1000 population. Therefore, on the assumption that the developed world's population is 1.2 billion, then a rate of x-ray procedures of slightly more than one per person per annum would yield approximately 1.2 billion x-ray procedures per year in the developed world. The population of the less developed world is 4.9 billion. If it is assumed that there are another 0.5 billion x-ray procedures performed for this population, a rate one-tenth that of the developed world, then the world total of annual medical and dental x-ray procedures is 1.7 billion.

According to Clinica Reports and Theta Reports, the world market for x-ray film is approximately $3.5 billion per year with the United States market accounting for around $1.4 billion of that total amount, a 40% share of the world market. If the dollar share of the market is reflective of the share of procedures performed than there are about 750 million x-ray procedures annually. On the other hand, it is reasonable to assume that film is sold for less in dollar terms around the world than in the United States. If it is assumed that the price of film in other markets is 50% of that in the United States, then this would work out to about 913 million x-ray procedures outside of the U.S., and around 1.2 billion for the entire world.

Another approach to verifying these estimates is to take into account the use of silver for the production of x-ray film as estimated by the Silver Institute. In 1998, 71.5 million troy ounces of silver were used for medical x-ray film and that was sufficient to produce 387 million square meters, or 4.2 billion square feet, of x-ray film. This implies that one ounce of silver is enough to produce 5.4 square meters, or 58 square feet, of medical x-ray film. Theta Consulting estimated U.S. consumption of 1.7 billion square feet of medical x-ray film to perform 309 million procedures in 1998, indicating usage of an average of 5.5 square feet of film per procedure. If the same amount of film were used per procedure in the rest of the world, 450 million medical x-ray procedures were performed there. This is consistent with a world wide x-ray procedure total of 750 million.

We have chosen to estimate the total number of world medical x-ray procedures at 1 billion annually as a reasonable balance of the available statistics and assumptions discussed.

In order to determine the actual number of x-ray images taken some calculation must be made based upon the average number of films taken per procedure (more than one image may be captured on one film). In fact, the number of films used per procedure varies considerably. If 2 films are shot for each x-ray procedure, there are approximately 2 billion x-ray medical images taken worldwide every year.

Other X-Ray Uses

The other major uses for x-ray film are dental imaging and the non-destructive testing of materials in manufacturing and fabrication processes. These uses are approximately 8 percent of overall x-ray film usage in developed nations according to the Silver Institute. Accordingly, based upon the finding of 2 billion x-rays for medical purposes, then industrial and dental uses would amount to 160 million x-rays in the developed world for all purposes. The total world use of x-ray film is therefore approximately 2.16 billion images annually.

Copies

As more photographs are digitized through inexpensive home scanners or are taken on digital cameras in the first place, it is not yet clear how often paper copies will be made. Furthermore, it is still unknown whether output from digital storage will typically be on photographic paper, or by means of ink jet or thermal printers. Current indications are that ink jet printing will be the most common process. In the year 2000, it is estimated that United States consumers will print out 5.4 billion photographic pictures, mostly at home. By the year 2005, this will quadruple to 26 billion. At the same time, traditional photoprocessing is expected to grow only 20 percent to 41.2 billion prints. (Source May 8, 2000 Wall Street Journal.)

The most common use made by consumers of digital images so far has been sharing them through e-mail. Similarly many new businesses are counting on the consumer's desire to share digital photos by hosting them on web sites. In December 1999 there were over 25 million unique home visitors and 12 million work visitors to the web sites of over one hundred companies providing photographic hosting services.

It is anticipated that the growing number of digital images owned by consumers will have a dramatic impact on their need to consider making backup copies of their personal disk drives. Traditionally, only a small percentage of home computer data is original data that requires backup at all. This need will grow dramatically if many photographs are stored digitally. The web hosting companies are counting on consumers storing the data on their sites and ordering prints from time to time. Currently, consumers are able to get free photo developing and unlimited disk space from these companies.

Copies of Motion Pictures

The Wolfman Report on the Photographic & Imaging Industry in the United States states that the average number of prints per original motion picture is 700. The Silver Institute, however, reports that 6,000 release prints are made for each feature movie. Interestingly, however, these copies are short-lived. 98 percent of all films for theatrical distribution made in the United States are destroyed by FPC, Inc.of Mountain City, Tennessee. After the films are no longer being shown in movie theaters, they are sent to FPC, which destroys10 million pounds of film every year. The film is shredded and sent to FPC's parent company, Kodak, where it is recycled and made into new film or fuel used in power plants. (Source: Associated Press)

Film Factoids

- Kodak describes the photography market as follows: 82 billion pictures processed a year throughout the world with 750 million rolls of film processed annually in the United States and 2.9 billion rolls consumed worldwide. Kodak also estimates that of the photographs that are processed approximately 2 percent are later reprinted or reused in some way.

  SOURCE: http://kodak.com/US/en/corp/georgeFisher/shihPres.shtml, Presentation to Imaging Technology Analysts Group on February 24, 1999. by Willie Shih, President, Digital and Applied Imaging, and Vice President, Eastman Kodak Company.

  More than 82% of U.S. households have cameras and use them to take over 17 billion pictures annually. It is estimated that there are over 150 billion photographs stored in those households. "Instant Images" Fortune, Winter 1997 (Technology Buyer's Guide Supplement) 184-187. This article cites these figures as according to the Photo Marketing Association.

According to Kodak, China has become its second largest market. The per-capita film consumption in China averages one roll a year, as compared with 3.6 rolls in America. Only 15 percent of Chinese own cameras. (XINHUA ECONOMIC NEWS SERVICE, 6/13/00).

- The United States Library of Congress reports that it holds 12 million photographs in its collection.

- Number of UK Feature Films Produced Annually (1912-1998)

- The Department of Commerce cites the journal *Medical Imaging* for the statistic that "U.S. health care systems spend up to $7 billion a year on film alone." U.S. Industry and Trade Outlook 2000, 44-18. Unfortunately for the film manufacturers, this probably overstates the amount of sales by a wide margin. Theta Reports puts the sales of x-ray film at $1.4 billion in 1996 and expects it to reach $1.5 billion by 2000.

- American women undergo approximately 30 million mammograms annually. Medical Industry Today, December 20, 1999.

- Approximately 150 million chest x-rays are done in the United States each year.(Source: University of Pittsburgh, Clinical Multimedia Lab).

- The Silver Institute reports that 92 million ounces of silver was used in 1999 for radiography throughout the world. This association also reports that dental and commercial uses of x-ray film traditionally account for about 8% of the total radiography market.

- In 1996 it was estimated that the average institutional radiology department would generate approximately 15.7 gigabytes of data per day and 3.5 terabytes per year. Honeyman, Huda, Frost, Palmer & Staab, "Picture Archiving and Communications System Bandwidth and Storage Requirements," Journal of Digital Imaging, Vol. 9, no. 2, May 1996.

- Production of x-ray film is essentially controlled by three companies: Kodak, Sterling Diagnostic Imaging/Agfa and Fuji Medical Systems. The U.S. market for x-ray film is about $1 billion annually. Modern HealthCare January 18, 1999.

- "More than 32.5 million mammograms and 4.5 million cardiac catheterization procedures are performed each year in the U.S., and X-rays account for 70 percent of all medical imaging procedures." PR Newswire, Feb. 29, 2000, " New GE Digital Imaging Technology"

- Kodak sells about $800 million year in x-ray film. Associated Press, Nov. 26, 1999. This also seems to suggest that the statistic cited above is incorrect as to the overall U.S. x-ray film market.

- "The reduction in the number of competitors is clearly seen in the medical X-ray film market. In 1995, there were five global players: Du Pont, 3M, Kodak, Fuji and Agfa, sharing approximately 90% of the world market. In 1996, Du Pont sold its X-ray film business to Sterling, which subsequently announced the resale of the business to Agfa in January 1999. The imaging division of 3M was spun off into a new company, Imation, in 1996, and this was subsequently purchased by Kodak in August 1998 for US$520 million. There are now only three major players in the global market: Kodak, Agfa and Fuji, assuming that the Agfa/Sterling deal is approved by the US FTC and the European Commission. This obviously leads to dominant companies with market shares in excess of 40% in some major markets. Further consolidation of these groups, however, would almost certainly run into legal difficulties." Medical Device Technology (May 1999), Vol 10, no. 4

- "Greater volumes of products and hence bargaining power are being concentrated into the hands of a declining number of purchasers. The greater volume and value involved with each negotiation mean that the market-share impact of either winning or losing a contract will

increase. This can lead to a decline in prices, as witnessed following the merger of Columbia and HCA, which was widely considered to have started the rapid decline of film prices in the US market between 1996 and 1998. Prices fell by more than 20%, wiping approximately US$150 million per annum off the value of the US X-ray film market." Id.

Film References and Sources

- Photo Marketing International: http://www.pmai.org/
- Photo Marketing Magazine: http://www.photomarketing.com/
- Berkeley Digital Library Sunsite: Digitizing Imaging and Text: http://sunsite.berkeley.edu/Imaging/

Reference Books

- The Photography Encyclopedia (call number TR9.M39 1999)
- Journal of Electronic Imaging (TK8315.J68)
- NAICS 512110 Motion Picture and video production
- SIC Code 7812 Motion picture and video production
- The Film Encyclopedia
- Guinness Book of Movie Facts and Feats
- International Film Guide
- Screen International
- Screen Finance
- Screen Digest
- The Motion Picture Guide Annual 1999
- The International Film Index, 1895-1990, edited by Alan Goble published by Bowker-Saur (London: 1990).
- The Silver Institute
- Theta Reports, X-Ray Film Markets, Report No. 671, January 1997
- University of Pittsburgh Clinical MultiMedia Lab
- United States Government, U.S. Food and Drug Administration, Center for Devices and Radiological Health
- United Nations Scientific Committee on the Effects of Atomic Radiation, Reports to the General Assembly, (New York 1993, 1994).

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Optical - Details

Originals

Conversion Factors

In its most common format, a compact disc (CD) holds about 650 MB.

A digital video disc or digital versatile disc (DVD) holds about 4.38 GB per disk - about 20 times more data than a compact disc (CD). The DVD specifications describe four disk configurations: single-sided (SS) vs. double-sided (DS) and single-layered (SL) vs. double-layered (DL) disks. Total storage capacities for DVDs range from 1.36 GB to 15.90 GB as seen in the following chart, reproduced from JimTaylor's DVD Demystified, an authoritative set of answers to Frequently Asked Questions about DVD:

Table 1: DVD Storage Capacities

Format	Capacity
DVD-5 (12 cm, SS/SL)	4.38 GB of data, over 2 hours of video
DVD-9 (SS/DL)	7.95 GB, about 4 hours
DVD-10 (12 cm, DS/SL)	8.75 GB, about 4.5 hours
DVD-14 (12 cm, DS/ML)	12.33 GB, about 6.5 hours
DVD-18 (12 cm, DS/DL)	15.90 GB, over 8 hours
DVD-1 (8cm, SS/SL)	1.36 GB, about .5 hour
DVD-2 (8cm, SS/DL)	2.48 GB, about 1.3 hours

DVD-3 (8cm, DS/SL)	2.72 GB, about 1.4 hours
DVD-4 (8cm, DS/DL)	4.95 GB, about 2.5 hours
DVD-R 1.0 (12 cm, SS/SL)	3.68 GB
DVD-R 2.0 (12 cm, SS/SL)	4.38 GB, 8.75 GB for rare DS discs
DVD-RW 2.0 (12 cm, SS/SL)	4.38 GB, 8.75 GB for rare DS discs
DVD-RAM 1.0 (12 cm, SS/SL)	2.40 GB
DVD-RAM 1.0 (12 cm, DS/SL)	4.80 GB
DVD-RAM 2.0 (12 cm, SS/SL)	4.38 GB
DVD-RAM 2.0 (12 cm, DS/SL)	8.75 GB
DVD-RAM 2.0 (8 cm, DS/SL)	1.36 GB
CD-ROM (12 cm, SS/SL)	0.635 GB
CD-ROM (8 cm, SS/SL)	0.18 GB

Flow

World

It is a fairly simple matter to obtain information about the world music market value and units shipped; organizations such as the Recording Industry Association of America and the International Federation of the Phonographic Industry report these statistics each year. More difficult to obtain are international statistics on the number of unique titles released each year. To obtain this figure, we used statistics regarding the US market share and US record releases (see below) to estimate the total releases worldwide. The United States holds a 37% share of the world music market and releases about 33,100 items per year. Therefore, the world produces about **90,000 originals** per year, equivalent to **58 TB** (uncompressed).

United States The Recording Industry Association of America (RIAA) reports annual figures for new releases and album re-releases. The 1998 releases are equivalent to about **21 TB** of data.

Table 2: Audio Releases

1992	1993	1994	1995	1996	1997	1998
18,400	20,300	36,600	30,200	30,200	33,700	33,100

Source: http://www.riaa.com/MD-US-6.cfm

Stock

The All Music Guide (www.allmusic.com), a comprehensive entertainment database for music, videos, DVDs and video games, provides statistics on the number of CD-audio originals in the world. The AMG database is licensed by major music sites such as CDNow, ArtistDirect and Tunes.com. AMG's listings indicate a total of 523,363 albums (445,735 popular music and 77,628 classical music albums). Each CD can hold 650 MB, so the total AMG catalog would equal roughly **340 TB**.

According to the 1999 edition of CD-ROM's in Print, there are about **16,200** unique CD-ROM titles. This figure includes business applications (such as word processing and spreadsheet packages), games, reference tools, and instructional programs.

Copies

Flow

World

Replication

Replication statistics include discs for retail as well as discs used for promotions, training, and rental. Some percentage of the replicated discs also turn out badly and are never used. These figures give us an upper bound on the amount of information that could be stored on compact disc each year.

In 1999, there were **4,654 million audio CDs** and **3,591 million CD-ROMs** replicated worldwide, according to the International Recording Media Association (IRMA). In addition, **194 million DVD-Video** units, **12 million DVD-ROM** units, and **2 million DVD-Audio** units were replicated in 1999.

United States

Replication

For audio CDs, data CD-ROMs, DVD-ROM, and DVD-Audio disks, North America produced roughly **50%** of all the disks replicated. In addition, North America replicated about 75% of the DVD videos.

Retail

During 1999, according to the Recording Industry Association of America, 938.9 million CDs were shipped to retail by U.S. producers. The U.S. has a 37% share of the world's sales. The nine next largest markets are Japan, which follows our lead with 16.7%, followed by the United Kingdom (7.6%), Germany (7.4%), France (5.2%), Canada (2.3%), Brazil, Australia and Spain (1.7% each), and Mexico (1.6%). From the US figures and market share, one can extrapolate that CD shipments worldwide are about 2,537 million units. This is equivalent to **1.6 billion TB**.

Rate of Change

According to the London-based International Federation of the Phonographic Industry (IFPI), the global music market was worth US $38.5 billion in 1999 - up by 1% in constant dollar terms with total unit sales of US$3.8 billion in 1999. Overall units remained level, with a continued growth of 3% in the CD market offset by a 10% decline in cassette sales and an 11% decline in singles.

Bibliography

- The All Music Guide, www.allmusic.com
- The CD Information Center www.cd-info.com/CDIC/index.html
- CD-ROM Finder: The World of CD-ROM Products for Information Seekers. Medford, NJ: Learned Information, 1993
- CD-ROMs in Print, 13th Edition: An International Guide to CD-ROM, CD-I, 3DO, MMCD, CD32, Multimedia, Laserdisc, and Electronic Products. New York: The Gale Group, 1999.
- DVD Entertainment Group www.dvdinformation.com
- DVD Insider www.dvdinsider.com
- Taylor, Jim. DVD FAQ www.dvddemystified.com/dvdfaq.html
- DVD Channel News www.dvdchannelnews.com
- International Federation of the Phonographic Industry www.ifpi.org

- International Recording Media Association, Statistics page www.recordingmedia.org/statistics/statistics_idx2.html
- Medialine www.medialinenews.com
- Optical Storage Technology Association www.osta.org/
- SUN CD-ROM FAQ saturn.tlug.org/suncdfaq/

Charts

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Magnetic - Details

Magnetic Storage Media

The first magnetic storage mechanism was the Telegraphone invented in 1898 by Danish scientist Valdemar Poulsen.

The digital revolution has so far been magnetic. The vast majority of the world's data is now created, transported and stored in electro-magnetic systems.

Hard Disk Drives (World)

No storage medium has ever had the explosive growth demonstrated by the hard disk.

Year Disks	Sold (Thousands)	Storage Capacity (PetaBytes)
1995	89,054	104.8
1996	105,686	183.9
1997	129,281	343.63
1998	143,649	724.36
1999	165,857	1394.60
2000	187,835	2553.7
2001	212,800	4641
2002	239,138	8119
2003	268,227	13027

Source: IDC (1999) "1999 Winchester Disk Drive Market Forecast and Review"

The incredible growth in hard disk shipments has been accompanied by a relentless decrease in the cost

per gigabyte of storage capacity:

Year	Cost per GB	GB's PER $200
1988	$11,540	0.02
1989	9,300	0.02
1990	6,860	0.03
1991	5,230	0.04
1992	3,000	0.07
1993	1,460	0.14
1994	705	0.28
1995	330	0.61
1996	179	1.12
1997	94	2.13
1998	43	4.65
1999	23	8.70
2000	13	15.38
2001	6	33.33
2002	3	66.67

Source: Wall Street Journal, June 26, 2000

Details regarding hard disk shipments.

Rates of Growth

The projections for the growth over the next few years in hard disk sales are for units shipped to increase at an annual rate of 15 to 20% but for the actual capacity shipped to grow much faster - 70 - 80% annually.

According to Disk/Trend, 75% of the disk drives sold are for desktop computers, followed by 13% for servers and 12% mobile drives for laptop computers. Therefore, although enterprise level disk storage systems each may have huge capacity, the vast number of disks deployed to individual workstations really accounts for the enormous scale of the world's current digital storage capability.

The lifespan of a hard disk is approximately 3 years. The storage capacity of the hard disks shipped in 1998, 1999 and 2000 is 4672 petabytes, or roughly 5 exabytes. In order to appreciate the scale of this statistic, consider that Roy Williams of CalTech advises that 5 exabytes is equivalent to the number of words ever spoken by all human beings.

The hard disk typically shipped with a desktop personal computer in 2000 holds 10 gigabytes.

Disk drives are expected to be deployed in applications other than personal computers, such as TV set-top boxes. These increasingly popular devices allow users to store TV shows on disk rather than tape and to stop and rewind during a broadcast while still recording. In all, by 2003 the IDEMA predicts that 8-10 percent of the disk drive market will be such devices.Similarly, starting in the year 2000 audio jukeboxes and computer game consoles will also include hard disk drives. Already the higher resolutions of digital cameras are creating such large file sizes that small hard disks will be incorporated into cameras as well.

Floppy Disks (World)

The number of floppy disk drives sold every year has remained relatively constant at around 100 million units for the past several years. Little change is expected. (Source: Computer Tech Review, April 1, 1999). The number of floppy disks being sold is diminishing rapidly as their storage capacity is too small to be useful in light of the much larger file sizes now common.

Year	3.5" Disks (billions)	Total Capacity (terabytes)	5.25" Disks (millions)
1996	1.823	2625	32
1997	1.179	1698	11.7

Source: International Recording Media Association

The Japanese Recording Media Industries Association, however, claim somewhat higher figures for floppy disk production, although confirming that the trend is still dramatically downward. This organization, for example, anticipated sales of 2.21 billion floppies in 1998. It would appear that an estimate of 2 billion for 1998, 1.5 billion for 1999 and 1 billion for 2000 would be a reasonable compromise around these figures. This would indicate that there would be a total stock of floppy disks of 4.5 billion, assuming 3 years production as stock.

One of the world's largest producers of 3.5 inch floppy disks is CMC Magnetics. They claim to have produced over 700 million disks in 1998. They also are purported to produce 56% of the world's floppy disks in 2000, implying a market of somewhere north of 1 billion disks.

Floppy disks are used primarily for backup and are little used now for original content creation.

Removable magnetic disk drives (World)

Removable drives are primarily used for backup, transfer of files, e.g,, desktop publishing files to service bureaus, or video or image editing. The general trend for low-end disk (capacity of around 100 to 250 megabytes, e.g. Iomega Zip Drives) sales is upward with a strong possibility that, if manufacturer incompatibility issues are ever resolved, this format could replace the 1.44 mb floppy. However, high capacity removable drives, with capacity in the gigabyte or better range, are being replaced by recordable CD's, which in turn may be replaced by recordable DVD's.

Year	Low-End Disk Drives	High-End Disk Drives
1996	3723	992
1997	7724	1334
1998	12035	1164
1999	17039	701
2000	21775	623
2001	26087	578
2002	30182	554
2003	34287	541

Source: IDC (1999), "1999 Optical/Removable Storage Market Forecast and Review"

The high capacity drives (Iomega Jaz) come with a free cartridge and it is assumed that an additional three cartridges are sold with each drive. Source: San Francisco Chronicle, Jan. 23, 1998 "Does Bad News for Iomega Mean Horrible News for HMT?"

The amount of original content created directly to this medium is, therefore, probably quite low. Furthermore, the disks are regularly reused and not generally viewed as archival solutions.

Magnetic Tape

Tape was the primary storage medium for the first generation of electronic computers in the 1950's. Reel-to-reel half-inch tape was used for data storage on mainframe computers from the earliest days of computing into the 1970's. Since that time, numerous tape formats have been developed. The worldwide installed base for tape drive units is 25.2 million.Michael Lesk estimated that in 1995, the magnetic tape industry would ship 200 petabytes of blank tape. (Lesk, M., "Preserving Digital Objects: Recurrent Needs and Challenges").

Current estimates are that approximately $1 billion of tape media will be sold every year. Source, Infostore July 1, 1999 (Tape: The Media Is the Message).

Worldwide Tape Drive Shipments - 1996-2003 (in thousands)								
	1996	1997	1998	1999	2000	2001	2002	2003
DC2000	3231.5	2365	1695.3	1867.3	1814.8	1749.4	1711.0	1688.7
SLR	350.4	327.1	311.7	290.0	278.3	272.0	267.5	266.8
4 mm	1370.0	1607.5	1634.8	1689.4	1650.5	1592.8	1525.9	1426.7
8 mm	201.8	213.6	180.9	138.2	131.0	127.7	135.1	149.3
DLT	160.0	356.4	366.1	483.5	550.3	612.5	677.4	746.5
LTO Ultrium	0	0	0	0	28.5	59.5	112.0	192.1
0.5" Cartridge	45.2	46.2	46.6	46.5	45.4	44.1	42.6	40.9
Total:	5358.8	4915.7	4235.3	4514.9	4498.9	4458.0	4471.5	4511.0

In a filing with the United States Securities and Exchange Commission in July 1999, Storage Tek, a large tape drive manufacturer, wrote that the cost of data storage on computer tape media was less than $.005/megabyte ($5.00/gb). Therefore, if predictions are correct that approximately $1 billion of computer tape media is sold every year, that implies worldwide annual tape storage capability of 200 petabytes. There may be some incongruity in these figures because the $1 billion may reflect manufacturer revenue for the product, rather than the retail cost of the product to end users, which would presumably be much higher. In fact, the Department of Commerce figures for the mid-1990's showed factory revenue for computer tape manufacturer's of around $600 to 700 million. Substantial amounts of tape media are manufactured in other countries so it is likely that $1 billion is a producer revenue figure.

If it is assumed that the retail price of the tape media is twice that of the manufacturer's then $2 billion of retail sales of tape would work out to around 400 petabytes of storage capacity. The markup at retail over manufacturer's prices is likely limited due to competition and the common practice of users purchasing bulk quantities of tape. Further, some moderation is due to the lower price of DAT format.

Low-End Formats

The Imation DC2000, or Travan, quarter-inch tape drive is a low-end product used primarily for the backup of desktop PC's. Their general capacity is in the range of 500 megabytes to 4 gigabytes. In August 2000, a Sony Travan Formatted MiniCartridge capable of holding 4 gb uncompressed was advertised for sale on the Internet for $29.49 each. A comparable 4gb tape cartridge manufactured by Maxell was available for $30.79.

Tandberg SLR (Scalar Linear Recording) is also a backup format for desktops and workstations and typically store 350 megabytes to 4 gigabytes.

4 mm tape drives are the largest segment of the market and use digital audio tape (DAT) format. They are commonly deployed as backups for PC servers. These drives generally provide backup in the range of 5 to 40 gigabytes (uncompressed). This format has an installed base of 7.6 million users.

Mid-Level Formats

8 mm tape drives provide storage in the 14 to 50 gigabyte range. Vendors include Exabyte (Mammoth), Sony (AIT), IBM (Magstar 3570).

DLT: (Digital Linear Tape) produced by Quantum Corporation. mid-range computer backup with 15 to 40 gigabytes of native capacity. There are more than 1.4 million DL Tape drives deployed and there have been approximately 40 million tape cartridges in this format sold. Quantum estimates that by the end of 2000, there will be 1.9 million DLT drives shipped to customers.

LTO Ultrium. A new format from consortium of IBM, Seagate and Hewlett Packard. The specification for the Ultrium format is for 100 gigabytes of native storage.

Enterprise Level Formats

1/2-inch cartridge: The dominant format in the mainframe, enterprise level storage market.

Automated tape libraries - which provide completely automated hands-off storage management, including random tape access, sophisticated robotics, unattended backup, and reduced labor costs - are expected to grow from less than 18,000 units shipped in 1996 to close to 120,000 units by 2002. (Source: Freeman Reports)

There is an industry rule of thumb that suggests a three-to-one ratio of disk capacity over tape be maintained.

Format	3490E	3480	Reel-to-Reel
1996	9.3 million	11.7 million	2.1 million
1997	10.7 million	9.5 million	1.9 million

Source: International Recording Media Association

The retail price in August 2000 of 3590 tape cartridges with 10 gb native capacity was $53.21. Fuji Film DLT Tape cartridges were also available at retail for $51.10 for 10 gb native capacity. Sony DLT tapes were being sold for $49.72 for 10 gb. uncompressed. (http://www.cleansweepsupply.com/pages/skugroup2599.html)

From low-end formats such as Travan, through most popular format DLT, to high-end 3590 format, retail price of roughly $5.00 per gigabyte of native storage capacity on tape seems reasonable estimate. (DAT tapes are the only exception and are a lot cheaper.) Of course, if larger purchases result in substantial discounts, then the revenue assumptions would commensurately be pushed more toward the $1 billion wholesale estimate; therefore, not really affecting the calculation of the price per gigabyte of storage capacity.

According to Computer Technology Review (March 1998) the total storage at a typical Fortune 1000 site is projected to escalate from 10 TB in 1997 to 1 PB by the year 2000. In the next five years, a typical large database system for US Government agencies is expected to accept 5TB per day and archive from 15 to 100 PB.

In 1995 Freeman Associates predicted that the total number of tape libraries would increase from 6,454 in 1994 to about 90,000 by the year 2000.

The estimate of the amount of original data stored on tape will, therefore, focus only on mass storage applications from large-scale scientific applications to heavily transaction oriented business applications. The installed base of IBM mainframe OS390 class computers is estimated by IDC to be around 16,500 in 2000.

The number of tape cartridges required to backup small sized computer disks is relatively few and will never substantially exceed the capacity necessary to backup the entire hard disk or disk array. Typical

backup storage strategy is to store the entire file system once and then do incremental updates of any changes made, reducing the amount of storage necessary to keep a current copy of the entire file system at hand.

In large scale tape libraries, there may be thousands or even tens of thousands of magnetic tapes providing primary storage of the application data. The scale of storage requirements is growing rapidly as new facilities, such as the Large Scale Hadron Collider, are built and start performing experiments. Large scale databases are also becoming more common as corporations make increasing efforts to comprehensively track consumer transactions.

The number of households conducting banking transactions may reach 32 million by 2003. The cost of an Internet banking transaction is an estimated 1 cent, compared with $1.14 per transaction by teller, 55 cents by phone, 29 cents by ATM and 2 cents by proprietary computer system. (Source: "Banking on the Internet")

The importance of massive scale databases in general commercial arenas is exemplified by the experience of Wal-Mart, a leader in so-called "data mining" technology and the owner of one of the largest privately held data sets. The U.S. Department of Commerce in its July 2000 report "Digital Economy 2000" points out that "over a three-year period, Wal-Mart achieved a 47 percent increase in sales on only a 7 percent increase in inventories by using a relational database system running on massively parallel computers. The system allows vendors to access almost realtime information on sales and customer transactions and handles 120,000 queries each week from 7,000 suppliers."

Digital Data Creation

Computers for the most part may not greatly contribute to the production of new and original data, but the great exception is in scientific explorations where huge data sets are commonplace and where new discoveries rely on computing and storage.

High Energy Physics

The Large Hadron Collider is being built at CERN in Switzerland, it is expected to be conducting production experiments in around 2005. It is expected to generate approximately 20 petabytes of data per experiment at rates of 100-1500 megabytes per second. Currently experiments in high energy physics generate data at the rate of 35 megabytes per second and many hundred terabytes per experiment. Obviously, this is all original data.

Source: Shiers, Jamie, "Massive-Scale Data Management using Standards-Based Solutions" IEEE 16th Symposium on Mass Storage Systems.

The BaBar experiment at SLAC will generate approximately 200TB/year of data at a rate of 10MB/sec for 10 years.

Los Alamos National Laboratory estimated total storage capacity in its open storage system at 243 terabytes and in its secure system at 2.31 petabytes as of 1998. It is also anticipated that storage capacity will grow to 5 petabytes in 2001.

GeoScience

The majority of data held and administered by the National Oceanic and Atmospheric Administration are held at three national data centers: the National Climatic Data Center at Asheville, NC; the National Oceanographic Data Center at Silver Spring, MD; and the National Geophysical Data Center at Boulder, CO. The climatic data is by far the largest of the three collections, holding approximately 640 terabytes on 350,000 magnetic tapes. The geophysical and oceanographic data total a combined 12 terabytes on 14,500 tapes.

The NASA Center for Computational Sciences in Greenbelt MD has 27,692 tapes holding data as of August 2000. This Center is using 3590 and 9840 tapes which hold 20 gb per tape uncompressed. This Center also automatically makes duplicate tapes for all new data generated. As of August 2000, this storage facility holds 92.5 terabytes of unique data and over 162 terabytes counting the duplicate data. New data is received at the rate of approximately 200-300 gigabytes per month.

The University of Tokyo stores satellite images in an environmental digital library of approximately 6

terabytes, approximately 60,000 images that average 100 megabytes in size.

The San Diego SuperComputer Center stored as of August, 1998 approximately 65 terabytes. The SDSC as of that time held the data on approximately 11,000 tapes. As of mid-2000, the SDSC is storing 1.5e+15 bytes.

The National Center for Atmospheric Research (NCAR) in 1996 had about 68 TB and was growing at the rate of 1.5 TB per month.

Analog Storage Tape

The first uses of magnetic media for data storage occurred about fifty years ago with the development of magnetic tape. A number of formats have evolved over the decades but today by far the most prevalent are the cassette tape used for the mass market distribution of prerecorded music. The other important use of magnetic tape for the storage of analog signals is videotape recordings.

Analog Audio Tape

The distribtution of prerecorded music is one of the most common uses of magnetic tape. The sales of music in this format, however, are now much smaller than they have been historically and are generally expected to continue to decline as digital media become more prevalent and convenient.

United States										
U.S. Sales	**1982**	**1985**	**1990**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**
Cassettes(mil.)	182.3	339.1	442.2	339.5	345.4	272.6	225.3	172.6	158.5	123.6
Cassette Singles	x	x	9.2	11	11.2	12.6	16.9	18.6	27.2	14.2

Source: Recording Industry Association of America

Although there has been a dramatic decline in the overall sales of cassette tapes due to the availability of music on alternative formats, there has been a booming market for books-on-tape. According to the Audio Publishers Association, this market is estimated to have grown 100% since 1990. The reasons for this are that cassettes can run 40 minutes longer than CDs, they have a built-in "bookmark", and they are frequently listened to in the car, and 75% of cars are manufactured with only an AM/FM cassette radio.

The IFPI reports that for the entire world prerecorded music sales on cassette tape were down by 11% in 1998 to 1.2 billion due to depressed sales in Asia.

Blank Audio Tape

The U.S. shipments of blank audio tape dropped dramatically during the 1990's.

Manufacturers' U.S. Shipments of Blank Audio Cassettes		
Year	Dollars (000s Omitted)	Unit (000s Omitted)

1984	$268,287	243,061
1985	$286,865	295,313
1986	$336,179	368,488
1987	$363,336	387,518
1988	$369,550	396,587
1989	$397,734	429,963
1990	$387,895	437,840
1991	$367,716	436,659
1992	$369,769	436,739
1993	$353,022	437,783
1994	$338,428	438,949
1995	$301,316	415,028
1996	$247,442	330,353
1997	$215,576	296,151

Source: International Recording Media Association

Worldwide shipments of blank audio tape are expected to decline in 2000 to 921 million units from 971 million in 1999, with an anticipated market for 771 million cassette tapes by 2003. Source: Consumer MultiMedia Report, Dec. 27, 1999.

Analog Video Tape

Prerecorded VideoTapes (VHS Format) (World)

1997 - 1.666 billion
1998 - 1.719 billion
1999 - 1.748 billion
2000 - 1.664 billion
2001 - 1,561 billion

Source: International Recording Media Association

The main use of the blank video tape is the consumer's use to record televison programs. It is anticipated that there would be a large drop in the sales of this tape if pay-per-view television shows carried copy protection. It is estimated that a very large share of the users of video cassette recorders do so for time shifting of viewing programs.

Blank VideoTapes (VHS Format T-120 equivalent units) (World)

1997 - 1,485 million
1998 - 1,446 million
1999 - 1,463 million
2000 - 1400 million
2001 - 1,275 million

Source: International Recording Media Association

Another view of the blank video market came from British research firm Understanding & Solutions. Their prediction was for 1.147 billion blank videocassettes in 1999 compared to 1.146 in 1998.

The stock of videotape in 1997 was estimated at about 4.6 billion by Richard Kelly, Cambridge Associates.It is not clear whether this is referring to all videos, including those sold blank, or just prerecorded. Further, it is not specified whether this estimate is solely for the United States or includes the whole world. (Feb. 1997 Newsletter, IRMA) We have taken this estimate into account but not used it directly because it seems that the flow of new videotapes worldwide each year would yield a considerably higher figure, even assuming a lot of videotapes may be viewed as disposable after a few years. We have instead estimated a world stock of videotape of all format at around 10 billion.

Conversion

Audio Conversion Issues

In translating the vast quantity of audio information available on cassette tape into its digital equivalent, we have chosen to use the CD format, linear PCM audio at a 16-bit word length and 44.1kHz sample rate. Although, professional recording studios use a sampling rate of 96kHz, the vast majority of tape recorded audio material is music for consumer use and the CD format is the digital format of choice for this application. The amount of data generated by this format is easily calculated. There are 44,100 16-bit samples taken each second for two tracks. Thus,1.4 million bits per second and 5.08 gigabits per hour are generated. The conversion to bytes yields 605 mBs per hour. (1 mByte = 1,048,576 bytes). This data is not compressed and yields a reasonable representation of music for most people.

Video Conversion Issues

In making assumptions about the size of analog videotape stores we have chosen to make conversions assuming the use of MPEG-2 video compression standard. In the case of videotape, the use of this conversion factor is seen as appropriate because it was designed as a generic format for digital multimedia and includes coding schema for both video and audio.

In the case of video, the massive amount of data generated requires that for any practical purpose some compression scheme must be used. MPEG-2 is now the international standard for video storage. Compression is achieved in two ways: spatial compression and temporal compression. The spatial compression is achieved by reducing the number of bits used to represent a single frame. Temporal compression, where the bulk of the savings come, attempts to encode only the bits that represent the portions of a frame that have changed from the previous frame.

The actual amount of compression that can be achieved with MPEG-2 varies quite a bit, we have assumed that 2 gigabytes is adequate to represent 1 hour of high-fidelity audio and high-definition video data.

References and Resources

Gibson, G.D. (1994): Audio, Film and Video Survey. A report on an international survey of 500 audio, motion picture film and video archives. Library of Congress, Washington DC.

Moving Pictures Expert Group

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Internet - WWW Details

- "A Cyveillance Study: Sizing the Internet" Summarized
- "The Deep Web: Surfacing Hidden Value" Summarized
- Excel Spreadsheet with further information.

"Sizing the Internet: A Cyveillance Study"
Source: Cyveillance, 10-July-2000

- 2.1 billion unique, publicly accessible web pages, and about 4 billion by early 2001 if the current rate of growth continues;
- 7.3 million uniques ages added per day;
- Average page size: 10,060 bytes;
- Average number of images on page: 14.38 (median);
- Percentage of US vs. international pages: 84.7%/15.37%
- Internet still continues to grow at accelerating rate;

"The Deep Web: Surfacing Hidden Value"
Source: BrightPlanet LLC, July 2000

- The deep Web contains 7,500 tarabytes of information, compared to 19 terabytes of information in the surface Web;
- The deep Web contains nearly 550 billion individual documents compared to the 1 billion of the surface Web;
- 60 of the largest web sites contain about 750 terabytes of information;
- More than half of the deep Web content resides in topic specific databases;
- Average page size on the surface Web is 18.7 kbytes, while on the deep Web - 13.7 kbytes, and median for the deep Web is 19.7 kbytes;
- Average deep Web site has a Web-expressed (HTML included basis) database size of 74.4 megabytes, and a median of 169 kbytes;

More Information

Excel Spreadsheet: rawdata.xls
HTML Created from Excel Spreadsheet: click here. [202K]

Internet Growth Data

Source: Hobbes' Internet Timeline
URL:
http://info.isoc.org/guest/zakon/Internet/History/HIT.html

Date	Hosts	Date	Hosts	Networks	Domains
Dec-69	4	Jul-89	130,000	650	3,900
Jun-70	9	Oct-89	159,000	837	
Oct-70	11	Oct-90	313,000	2,063	9,300
Dec-70	13	Jan-91	376,000	2,338	
Apr-71	23	Jul-91	535,000	3,086	16,000
Oct-72	31	Oct-91	617,000	3,556	18,000
Jan-73	35	Jan-92	727,000	4,526	
Jun-74	62	Apr-92	890,000	5,291	20,000
Mar-77	111	Jul-92	992,000	6,569	16,300
Dec-79	188	Oct-92	1,136,000	7,505	18,100
Aug-81	213	Jan-93	1,313,000	8,258	21,000
May-82	235	Apr-93	1,486,000	9,722	22,000
Aug-83	562	Jul-93	1,776,000	13,767	26,000
Oct-84	1,024	Oct-93	2,056,000	16,533	28,000
Oct-85	1,961	Jan-94	2,217,000	20,539	30,000
Feb-86	2,308	Jul-94	3,212,000	25,210	46,000
Nov-86	5,089	Oct-94	3,864,000	37,022	56,000
Dec-87	28,174	Jan-95	4,852,000	39,410	71,000
Jul-88	33,000	Jul-95	6,642,000	61,538	120,000
Oct-88	56,000	Jan-96	9,472,000	93,671	240,000
Jan-89	80,000	Jul-96	12,881,000	134,365	488,000
		Jan-97	16,146,000		828,000
		Jul-97	19,540,000		1,301,000

Notes: Hosts = a computer system with registered IP address;
Domain = registered domain name (with name server record)

A more accurate survey mechanism was produced in 1999, new and corrected numbers are shown below:

Date	Hosts	Date	Hosts	Date	Hosts
Jan-95	5,846,000	Jan-97	21,819,000	Jan-99	43,230,000
Jul-95	8,200,000	Jul-97	26,053,000	Jul-99	56,218,000
Jan-96	14,352,000	Jan-98	29,670,000	Jan-00	72,398,092
Jul-96	16,729,000	Jul-98	36,739,000		

Note: the original source for the data above is: Network Wizards Internet Domain Survey

URL: ftp://ftp.nw.com/pub/zone/WWW-9807/top.html

This statistics is also reproduced at the following locations:
1. Internet Statistics: Growth and Usage of the Web and the Internet
URL: http://www.mit.edu/people/mkgray/net/
2. Internet Software Consortsium: Internet Domain Survey
URL: http://www.isc.org/ds/

Source: The NSFNET Backbone Project,
1987-1995
URL:
ftp://nic.merit.edu/statistics/nsfnet/index.html

Date	Networks	Date	Networks	Date	Networks
Jul-88	217	Nov-90	2125	Mar-93	10498
Aug-88	241	Dec-90	2190	Apr-93	11252
Sep-88	292	Jan-91	2338	May-93	12349
Oct-88	305	Feb-91	2417	Jun-93	13170
Nov-88	334	Mar-91	2501	Jul-93	14121
Dec-88	346	Apr-91	2622	Aug-93	15160
Jan-89	384	May-91	2763	Sep-93	16696
Feb-89	410	Jun-91	2982	Oct-93	17979
Mar-89	467	Jul-91	3086	Nov-93	19664
Apr-89	516	Aug-91	3258	Dec-93	21430
May-89	564	Sep-91	3389	Jan-94	23494
Jun-89	603	Oct-91	3556	Feb-94	25706
Jul-89	650	Nov-91	3751	Mar-94	28578
Aug-89	745	Dec-91	4305	Apr-94	30626
Sep-89	809	Jan-92	4526	May-94	32370
Oct-89	837	Feb-92	4740	Jun-94	34051
Nov-89	897	Mar-92	4976	Jul-94	36153
Dec-89	927	Apr-92	5291	Aug-94	38307
Jan-90	1233	May-92	5515	Sep-94	39977
Feb-90	1290	Jun-92	5739	Oct-94	41520
Mar-90	1356	Jul-92	6031	Nov-94	42883
Apr-90	1525	Aug-92	6385	Dec-94	44689
May-90	1580	Sep-92	6640	Jan-95	46318
Jun-90	1639	Oct-92	7354	Feb-95	48514
Jul-90	1727	Nov-92	7854	Mar-95	50344
Aug-90	1894	Dec-92	8561	Apr-95	50766
Sep-90	1988	Jan-93	9117		
Oct-90	2063	Feb-93	9604		

Date	Hosts	Date	Domains
Aug-81	213	Jul-88	900

Aug-83	562	Jul-89	3,900
Oct-85	1,961	Oct-90	9,300
Dec-87	28,174	Jul-91	16,000
Oct-89	159,000	Oct-92	18,100
Oct-90	313,000	Oct-93	28,000
Oct-91	617,000	Oct-94	56,000
Oct-92	1,136,000	Jul-95	120,000
Oct-93	2,056,000	Jul-96	488,000
Oct-94	3,864,000	Jul-97	1,301,000
Jul-95	6,642,000		
Jul-96	12,881,000		
Jul-97	19,540,000		

Source: NetSizer
URL: http://www.netsizer.com/

Estimate of the number of web hosts (As of Jul 21, 2000): 86.437 millions

Source: Domain Stats
URL: http://www.domainstats.com/

Total domains registered worldwide (As of Jul 21, 2000): 17,804,717

Source: Center for Next Generation Internet
URL:
http://www.ngi.org/trends.htm

Annual hosts growth rate: 63%

Predictions:	100 million hosts by the end of 2000
	1 billion hosts by 2005
Expansion:	69 new hosts per minute
	23 new domains per minute

Contains detailed statistics by country

WWW Growth Data

Source: Hobbes' Internet Timeline
URL:
http://info.isoc.org/guest/zakon/Internet/History/HIT.html

Date	Sites	Date	Sites	Date	Sites
Jun-93	130	May-97	1,044,163	Dec-98	3,689,227
Sep-93	204	Jun-97	1,117,255	Jan-99	4,062,280
Oct-93	228	Jul-97	1,203,096	Feb-99	4,301,512
Dec-93	623	Aug-97	1,269,800	Mar-99	4,389,131
Jun-94	2,738	Sep-97	1,364,714	Apr-99	5,040,663
Dec-94	10,022	Oct-97	1,466,906	May-99	5,414,325
Jun-95	23,500	Nov-97	1,553,998	Jun-99	6,177,453
Jan-96	100,000	Dec-97	1,681,868	Jul-99	6,598,697
Jun-96	252,000	Jan-98	1,834,710	Aug-99	7,078,194
Jul-96	299,403	Feb-98	1,920,933	Sep-99	7,370,929
Aug-96	342,081	Mar-98	2,084,473	Oct-99	8,115,828
Sep-96	397,281	Apr-98	2,215,195	Nov-99	8,844,573
Oct-96	462,047	May-98	2,308,502	Dec-99	9,560,866
Nov-96	525,906	Jun-98	2,410,067	Jan-00	9,950,491
Dec-96	603,367	Jul-98	2,594,622	Feb-00	11,161,811
Jan-97	646,162	Aug-98	2,807,588	Mar-00	13,106,190
Feb-97	739,688	Sep-98	3,156,324	Apr-00	14,322,950
Mar-97	883,149	Oct-98	3,358,969	May-00	15,049,382
Apr-97	1,002,512	Nov-98	3,518,158	Jun-00	17,119,262

Source: article "Accessibility of Information on the Web"
URL: n/a, only a printed copy is available

Note: all numbers below refer to publicly indexable web pages; publicly indexable web pages exclude pages that are not normally considered for indexing by web search engines, such as pages with authorization requirements(including firewalls), pages excluded from indexing using the robots exclusion standard, dynamic pages, etc;

December, 1997:	at least 320 million pages;
February, 1999:	2.8 million servers on the publicly indexable web; 289 average pages per server; 800 million publicly indexable web pages; 18.7 kilobytes is the mean size of a page;

February, 1999, continued:	3.9 kbytes is the median size of a page; 7.3 - kbytes average size of the textual content [after removing HTML tags, comments, and extra white space]; 0.98 kbytes - median size of the textual content; 15 terabytes of pages is the amount of data on the publicly indexable web; 6 terabytes is the amount of text data; 62.8 images per web server; 15.2 kbytes - average image size; 5.5 kbytes - median image size; 180 million images on the publicly indexable web; 3 terabytes - total amount of image data;

Note: the distribution of pages on web servers is extremely skewed, following a universal power law; many sites have few pages, and a few sites have vast numbers of pages, which limits the accuracy of the estimates above; the true value could be higher because of very rare sites that have millions of pages (for example, GeoCities reportedly has 34 million pages), or because some sites could not be crawled completely because of errors;

Source: "Size of the Web: A Dynamic Essay for a Dynamic Medium"
URL:
http://censorware.org/web_size/

As of 7/5/2000, the Web has roughly:	2,170,000,000 pages; 40,800,000,000,000 bytes of text; 489,000,000 images; 8,160,000,000,000 bytes of image data.
In the last 24 hours, the Web has added:	4,420,000 new pages; 82,800,000,000 new bytes of text; 994,000 new images; and 16,600,000,000 new bytes of image data. 49,400,000 pages changed; and 11,100,000 images changed.
Average lifespan of the Web page:	44 days;

Source: "The Truth About the Web Crawling Towards Eternity"
URL:
www.alexa.com/company/internet_stats.html

Note: Published in Web Techniques Magazine, May 1997,
Issue 5

How many Web sites are there?

One million Web-site names are in common usage.
There are about 450,000 unique host machines.
If you request the top page from these 450,000, about 300,000 will return one within reasonable time. The rest appear to be intermittent or archaic.
About 95 percent of the 300,000 servers are "up" at any given time.

How big is the Web?
Note: "We estimate there are 80 million HTML pages on the public Web as of January 1997. The figure is fuzzy because some sites are entirely dynamic (a database generates pages in response to clicks or queries). The typical Web page has 15 links (HREFs) to other pages or objects and five sourced objects (SRC), such as sounds or images." Moreover:

The typical HTML page is 5
kbytes;
The typical image (GIF or JPEG) is 12 kbytes;
The average object served via HTTP is 15
kbytes;
The typical Web site is about 20 percent HTML, 80 percent images, sounds, and executables (by size in bytes);

"The upshot of this data is that it takes about 400 GB to store the text of a snapshot of the public Web and about 2000 GB (2 TB) to store nontext files."

How big are individual Web sites?

The median size for a Web site is about 300 pages; only 50 sites have more than 30,000 pages.

About 5 percent of all servers have a robot.txt file (for governing how crawlers visit).

About 1 percent of all servers have a sitelist.txt file (to aid site mapping and robot revisiting).

How fast is the Web growing?

The size of the Web is doubling yearly, but this statistic is losing its meaning because of the growth of dynamic sites.
The typical Web page is only about two months old.
Dynamic sites are becoming a significant presence; JavaScript is widespread, Java much less so, but growing.

Note: "Facts above are mostly based on data gathered by Internet Archive, but augmented with some stats from Larry Page of Stanford University and public documents from the Web."
Author: Z Smith

Source: Measuring the Web
URL: http://www5conf.inria.fr/fich_html/papers/P9/Overview.html

What is the "average page" like? (as of May, 1996)

Mean size: 6518
Median size: 2021
SD: 31678

Source: "The Web: Growing by 2 Million Pages a Day"
URL: www.thestandard.com/article/display/0,1151,12329,00.html
Author: David Lake
Published in: The Industry Standard Magazine, February 28, 2000

Metric	1999 Total	Estimated Growth per Day 1998-1999
Web pages	1,500,000,000	1,917,808
Hosts	72,398,092	79,913
Domain Names	8,100,000	12,981
Unique Web Sites	3,649,000	4,422

Cumulative Domain Name Registrations in the .com, .net, .org Domains (Millions)

Year	Existing	New
1993	0	0.01
1994	0.01	0.03
1995	0.04	0.16
1996	0.2	0.4
1997	0.6	0.9
1998	1.6	1.8
1999	3.4	4.7

Web Site Growth Trend (Millions)

Year	Online Computer Library Center	Alexa Internet
1997	1.2	n/a
1998	2	2.5
1999	3.6	3.4
2000	n/a	n/a
2001	n/a	10

Number of Hosts in the Domain Name System (Millions)

Year	Number of Hosts
1995	5.8
1996	14.4
1997	21.8
1998	29.7
1999	43.2
2000	72.4

Source: Inktomi Corp: Web Surpasses One Billion Documents
URL:
http://www.inktomi.com/new/press/billion.html

Jan 18, 2000: Inktomi announces that WWW surpassed 1 billion pages; This figure also coincides with Internet Archive's (www.archive.org) estimate. Internet Archive also says that these 1 billion pages correspond to 13.8 terabytes of text-only data, with a rate of growth 2 terabytes per month;

Source: Online Computer Library Center June 1999 Web Statistics
URL:
http://www.oclc.org/oclc/research/projects/webstats/statistics.htm

Number of IP addresses in 32-bit address space:	4,294,967,296
Number of IP addresses in the 0,1% random sample:	4,294,967
Number of Web Sites:	4,882,000 (+/- 3%)
Number of Unique Web Sites:	3,649,000 (+/- 3%)
Number of Public Web Sites:	2,229,000 (+/- 4%)
Number of Private Web Sites:	389,000 (+/- 10%)
Number of Provisional Web Sites:	1,031,000 (+/- 6%)
Number of Web Pages:	288,221,000 (+/- 35%)
Number of Files:	500,491,000

Web Growth

	1997	1998	1999
Web Sites:	1,570,000	2,851,000	4,882,000
Unique Sites:	1,230,000	2,035,000	3,649,000
Unique Public Sites:	800,000	1,457,000	2,229,000
% Change:	97 to 98	98 to 99	97 to 99

Web Sites:	82	71	211
Unique Sites:	65	79	197
Unique Public Sites:	82	53	179

Web Volatility: 44%
(IP addresses identifying a Web site in 1998 that no longer identify a Web site in 1999)

Explanation of the Statistics

The 32-bit Internet Protocol address space consists of 4,294,967,296 unique IP addresses. A 0.1% random sample (without replacement) was taken from this address space, resulting in 4,294,967 unique IP addresses. An attempt was made to connect to each of the sample addresses on Port 80.

Web Site
Identified by an IP address that returns a response code of 200 and a Web page in reply to an HTTP request for the home page.

Unique Web Sites
It is not uncommon for the same Web site to be duplicated at multiple IP addresses (e.g., for server load distribution purposes). To ensure that each unique Web site has the same probability of being selected for the sample, the following rule was followed: if a site is located at multiple IP addresses, the site is retained in the sample only if the numerically lowest IP address is in the sample. Three diagnostic tests were developed to assist in identifying sites with multiple IP addresses. A description of these tests is available in the paper A Methodology for Sampling the World Wide Web.

Public Web Sites
A public Web site offers content that, from a general perspective, is meaningful and non-trivial, and is freely accessible without fee payment or prior authorization.

Private Web Sites
A private Web site requires payment or prior authorization to access its content; typically, a private site will not permit free access beyond its home page.

Provisional Web Sites
A provisional Web site is in a transitory or unfinished state, and/or offers only content that from a general perspective, is meaningless or trivial.

Web Pages
According to the W3C Working Draft "Web Terminology & Definitions Sheet" (May 24, 1999), a Web page is defined as:
A collection of information, consisting of one or more Web resources, intended to be rendered simultaneously, and identified by a single URI. More specifically, a Web page consists of a Web resource with zero, one, or more embedded Web resources intended to be rendered as a single unit, and referred to by the URI of the one Web resource which is not embedded.
The key point in this definition is that a Web page often is a composite object, consisting of multiple Web resources: e.g., text, images, applets, etc. The Web page is the single entity representing the combination of

these resources.

For each unique public Web site, a harvesting agent was used to download and store all text-based files internal to the site. Harvested files had one of the following extensions:

.asc	.asp	.dhtm	.dhtml
.ephtm	.ephtml	.htm	.html
.jsp	.mhtm	.mhtml	.php
.php3	.phtm	.phtml	.shtm
.shtml	.text	.tx8	.txt

In addition, files with no extension and those identified implicitly in the URL - i.e., URLs whose path ends with a directory, and the server automatically loads the correct file when the URL is accessed - were also harvested.
Duplicate pages on the same site were eliminated. Elimination of duplicate pages was conducted strictly on an intra-site basis, and was not extended across sites.
Once the de-duping process was completed, the remaining files were processed by software that counted the number of Web pages present on each Web site. For a given Web site, Web pages fell into one of two categories:
1. the harvested pages
2. references in the harvested pages to other Web pages internal to the Web site, through the HTML <a href="..."> convention.

The counting algorithm was configured so that a Web page with frames is counted only once, rather than multiple times for each constituent part of the frame.

Number of Files
An estimate of the number of files located at public Web sites can also be obtained from the harvest data. For each public Web site, files were identified as one of the following:
1. harvested files
2. references in the harvested files to other files internal to the Web site, through the HTML <a href="..."> convention.
3. references in the harvested files to images internal to the Web site, through the HTML <img src="..."> convention.
4. references in the harvested files to applets internal to the Web site, through the HTML <applet code="..." codebase="..."> convention.

Source: Cyveillance's "Sizing the Internet"
URL: http://www.cyveillance.com/resources/Sizing_the_Internet_whitepaper.pdf

Number of unique pages on Internet:	2.1 billion
Unique pages added per day:	7.3 million

Average size of pages:	10,060 bytes
Average number of images on a page:	14.38

Source: Cyveillance Corp.
URL: http://www.cyveillance.com/resources/facts.asp#5

Web grows by 300,000 pages every 5 days (Original source: Lycos)
The amount of Web pages is estimated to grow seven-fold to 7.7 billion by 2002. (Source: IDC)

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Internet - Email Details

"E-Mail Growth Hogs Enterprise Resources"
Source: Network World Magazine, 31-Jan-2000

- Study of corporate email usage, citing David Ferris, president of Ferris Research;
- Average number of messages received by end users is expected to jump 81% to 34 per day by the beginning of 2001;
- Average size of a message is expected to increase 192% to 286 kbytes by the beginning of 2001 [with growth attributable to attachments];
- There are nearly 170 million corporate email boxes worldwide, more than three times the number of boxes five years ago, according to Eric Arnum, editor of "Messaging Online";
- There are approximately 440 million corporate and personal mailboxes worldwide

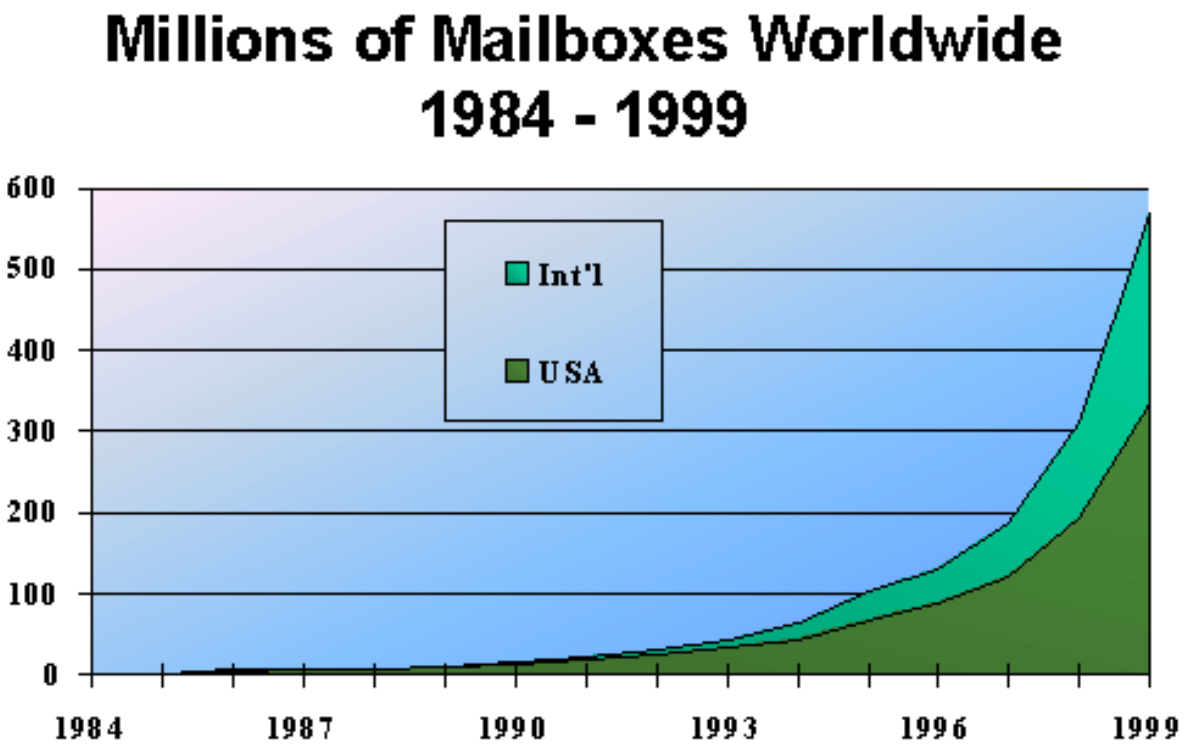
---

"AOL Per-User Email Figures Climb 60 Percent in 1999"
Source: "Messaging Online," 4-Feb-2000

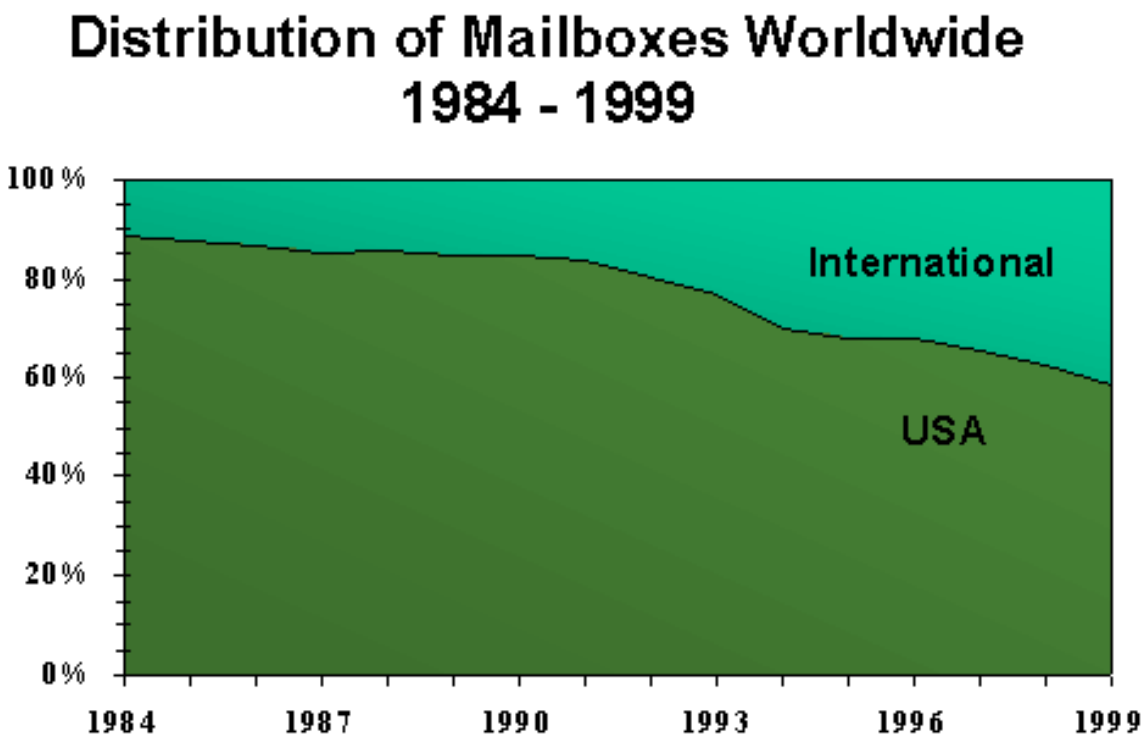
- 3.5 messages per AOL user per day in 1998, and 5.6 in 1999;
- 110 messages sent in 1999, up from 50 million in 1998;
- 20.5 million users at the end of 1999
- Email usage per person increased 60 percent in 1999;
- "If you believe every person in the U.S. has an email account (and it's beginnign to seem that way), then you are talking 1.54 billion messages per day, or 560 billion messages per year. If you believe half the population has email, then your numbers are 770 million messages per day or 280 billion messages per year. Adjust those numbers to reflect heavier usage by the workforce email users and lighter usage by Webmail and ISP users, and you possibly could come up with a trillion messages per year."
- In 1999, the U.S. Postal Service delivered over 200 billion pieces of mail, so email volume now outpaces postal mail volumes;

"Messaging Today: Worldwide Trends"
Source: "Messaging Online," 14-Mar-2000

- The total number of electronic mailboxes in the world has soared 83.5 percent in the past year to 569,171,660 mailboxes;
- In the U.S., the number of mailboxes has jumped 73 percent to 333.5 million mailboxes since the end of 1998. In the rest of the world, the total number of mailboxes has grown 101 percent to 235.6 million mailboxes in the past year;
- In the U.S., the average corporate email user has around 1.5 mailboxes, and the average household using email has about 4 mailboxes.
- There are about 89 million Americans using email at work and roughly 50 million households using email
- There are probably 110 million Americans using email at home or at work, 40 percent of the population.
- There are fewer than one billion televisions in the world, fewer than 800 million phone lines, and 569.2 million mailboxes;
- This paper contains other interesting information and graphs. Here are two samples::



Source: Messaging Online



Source: Messaging Online

24/7 Media: Email Facts
Source: 24/7 Media

- Opt-in email volume will jump 52.3% to 61.1 billion by year-end 2000, and reach 240 billion messages by 2003.
- There were 78 million active e-mail users, aged 13 and older, in the U.S. at year-end 1999, accounting for 35% of the total U.S. population of adults and teens (13+)
- By year-end 2002, there will be 135 million e-mail users, representing 59% of the overall U.S. population of adults and teens. (Source: Emarketer)
- There were 409 million e-mail boxes worldwide in 1999, up from 234 million a year earlier. In 1999, an estimated 20% of all e-mail received in the U.S. was commercial, split almost evenly between spam and permission e-mail (Source: Emarketer)

Nov. 92 - Nov. 94 - Messages Per Month
Source: Internet Society

Month	# of Messages
Nov. 92	279,060,000
Dec. 92	262,320,000
Jan. 93	286,920,000
Feb. 93	317,616,667
Mar. 93	363,180,000
Apr. 93	352,410,000
May 93	410,700,000

Jun. 93	371,700,000
Jul. 93	383,340,000
Aug. 93	397,860,000
Sep. 93	421,320,000
Oct. 93	479,853,333
Nov. 93	508,980,000
Dec. 93	469,370,000
Jan. 94	516,540,000
Feb. 94	527,170,000
Mar. 94	636,140,000
Apr. 94	695,220,000
May 94	731,220,000
Jun. 94	747,960,000
Jul. 94	765,840,000
Aug. 94	803,100,000
Sep. 94	899,400,000
Oct. 94	1,033,920,000
Nov. 94	1,007,590,000

Junk Email

- "Is the Junk Mail Problem Disappearing?" contains different numbers and time-series analysis of the junk email.

Other Information

- Mail.com has 14.4 million e-mail boxes, which require 27 terabytes of storage.
  Source: Network World, Network Storage White Paper
- 300 million e-mails are sent per day in the US.
  Source: Thomas Staffing of California
- Today's users send about 15 messages a day and receive about 20 a day, on average. This colume is expected to grow by some 60% and 80% respectively over the next year.
  Source: Ferris Research
- Electronic mailboxes will more than double in the next two years to an estimated 112 million in 1999.
  Source: Electronic Mail & Messaging Systems
- UC Berkeley's average message size in August, 2000 was 18,517 bytes, with median size of 1863 bytes.

Final Thoughts

Based on the information above, **500 to 600 billion email messages in 2000** seems to be a reasonable non-inflated estimate [though it might be lower than the actual number]. While the average size of email message is hard to tell, we can predict the volume of storage needed from the data above. For example, Mail.com uses 27 terabytes of storage space for its 14.4 million email

boxes, and the number of email boxes can be estimated to be between 450 and 500 million. With the average of **475 million email boxes** in 2000, the volume of storage needed can be estimated at **900 terabytes**, which is lower than the number obtained by multiplying the number of messages by their [probable] average size. The explanation may be that this is the amount of messages stored, not the measure of flow.

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Data Powers of Ten

Many of the following facts were taken from Roy Williams "Data Powers of Ten" page at Caltech.

- **Byte [8 bits]**
  - 0.1 bytes: a binary decision;
  - 1 byte: a single character;
  - 10 bytes: a single word;
  - 100 bytes: a telegram ORr a punched card;
- **Kilobyte [1,000 bytes OR 10^3 bytes]**
  - 1 Kilobyte: A very short story;
  - 2 Kilobytes: A tTypewritten page;
  - 10 Kilobytes: An encyclopaedic page OR a deck of punched cards;
  - 10 Kilobytes: static web page;
  - 50 Kilobytes: A compressed document image page;
  - 100 Kilobytes: A low-resolution photograph;
  - 200 Kilobytes: A box of punched cards;
  - 500 Kilobytes: A very heavy box of punched cards;
- **Megabyte [1,000,000 bytes OR 10^6 bytes]**
  - 1 Megabyte: A small novel OR a 3.5 inch floppy disk;
  - 2 Megabytes: A high resolution photograph;
  - 5 Megabytes: The complete works of Shakespeare OR 30 seconds of TV-quality video;
  - 10 Megabytes: A minute of high-fidelity sound OR a digital chest X-ray;
  - 20 Megabytes: A box of floppy disks;
  - 50 Megabytes: A digital mammogram;
  - 100 Megabytes: 1 meter of shelved books OR a two-volume encyclopaedic book;
  - 200 Megabytes: A reel of 9-track tape OR an IBM 3480 cartridge tape;
  - 500 Megabytes: A CD-ROM OR the hard disk of a PC;
- **Gigabyte [1,000,000,000 bytes OR 10^9 bytes]**
  - 1 Gigabyte: a pickup truck filled with paper OR a symphony in high-fidelity sound OR a movie at TV quality;
  - 2 Gigabytes: 20 meters of shelved books OR a stack of 9-track tapes;
  - 5 Gigabytes: 8mm Exabyte tape;
  - 20 Gigabytes: A good collection of the works of Beethoven OR 5 Exabyte tapes OR a VHS tape used for digital data;
  - 50 Gigabytes: A floor of books OR hundreds of 9-track tapes;
  - 100 Gigabytes: A floor of academic journals OR a large ID-1 digital tape;
  - 200 Gigabytes: 50 Exabyte tapes;

- 500 Gigabytes: The biggest FTP site.

- **Terabyte [1,000,000,000,000 bytes OR 10^{12} bytes]**
  - 1 Terabyte: An automated tape robot OR all the X-ray films in a large technological hospital OR 50000 trees made into paper and printed OR daily rate of EOS data (1998);
  - 2 Terabytes: An academic research library OR a cabinet full of Exabyte tapes;
  - 10 Terabytes: The printed collection of the US Library of Congress;
  - 50 Terabytes: The contents of a large Mass Storage System;
  - 400 Terabytes: National Climactic Data Center (NOAA) database;
- **Petabyte [1,000,000,000,000,000 bytes OR 10^{15} bytes]**
  - 1 Petabyte: 3 years of EOS data (2001);
  - 2 Petabytes: All US academic research libraries;
  - 8 Petabytes: All information available on the Web;
  - 20 Petabytes: Production of hard-disk drives in 1995;
  - 200 Petabytes: All printed material OR production of digital magnetic tape in 1995;
- **Exabyte [1,000,000,000,000,000,000 bytes OR 10^{18} bytes]**
  - 2 Exabytes: Total volume of information generated worldwide annually.
  - 5 Exabytes: All words ever spoken by human beings.
- **Zettabyte [1,000,000,000,000,000,000,000 bytes OR 10^{21} bytes]**
- **Yottabyte [1,000,000,000,000,000,000,000,000 bytes OR 10^{24} bytes]**

How Much Information?

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Printed Media - Details

Conversion Factors

To estimate the storage requirements for these non-digital information sources, one must make

some assumptions about the form in which they will be stored and the degree to which they will be compressed. There is considerable variation in the "bytes per page" proposed by various sources, partly because of differing assumptions about the content (e.g. ratio of text to pictures), formatting (e.g. number of characters per page, amount of white space), and file storage method (e.g. whether converted to a text or compressed document image file). Here is a sample of the range:

- According to Webopoedia: "A disk that can hold 1.44 megabytes, for example, is capable of storing approximately 1.4 million characters, or about 3,000 pages of information." 1.44 MB = 1,509,949 bytes = 3,000 pages = **503 bytes per page**
- According to Horison Information Strategies: 2,400 bytes **(2.3 KB)** for one page of paper. 300 - 500 MB for large encyclopedia
- According to **Michael Lesk**: 5,000 bytes (about **5 KB**) per page (Source: "How Much Information is There in the World?", 1997)
- 30 - **50 KB** for a page of mathematical text in bitmap form (Source: Andrew Odlyzko, "Tragic loss or good riddance? The impending demise of traditional scholarly journals," 1994)
- According to Roy Williams of CalTech, "Data Powers of Ten": **10 KB** per encyclopedic (small font, dense formatting) page, **50 KB** for a compressed document image page
- According to ArchiveBuilders.com **50 KB** for one scanned page but **800 KB** for one scanned engineering drawing
- According to JSTOR, a digital library of scholarly journals, one scanned document page is **130 KB**; compressed using Cartesian Perceptual Compression (CPC) a page is only **26 KB**.
- According to Internet Archive: **1 MB** for one mystery novel. 1 copy of Encyclopedia Brittanica = 1 gigabyte (2,619 pages per copy so 409,981 bytes or about **400 KB** per page)

To demonstrate this range, for our estimates, we will use three different numbers: one for a scanned image (uncompressed, at archival quality - 600 dpi), one for a compressed image file, and one for and ASCII (or plain text) file. For the summary statements, will use the mid-range number, as most documents are scanned with some measure of compression. We opted to use Cartesian Perceptual Compression (CPC) as our compression standard because although it is a lossy algorithm, it is "non-degrading" - the human eye can't determine any difference. And it keeps the full 600 dpi resolution.

Originals

Books

Conversion Factors

To estimate the size (in bytes) of all books published in the world, use the following formula:

Average number of pages per book * Average file size per page * Number of books published per year.

Table 1: Books Conversion Factors

Average # of Pages per Item	Storage Format	Average File Size	Conversion Factor (rounded)
300	Scanned TIFF (600 dpi)	130 KB	40 MB/Item
	Compressed	26 KB	8 MB/item
	Plain text	2.5 KB	1 MB/Item

Flow

World

Worldwide, in 1996 there were 808,066 titles published, according to *UNESCO's Statistical Yearbook 1998*, as cited in *The Bowker Library and Trade Almanac*. This is roughly equivalent to **7 TB/year**. The editor of the *Almanac* notes problems with UNESCO's data-gathering--inconsistencies in reporting, data collection lags, and missing data; nonetheless, the editor acknowledges that UNESCO provides researchers the best and most usable overview of international book title output. If one includes the most recently available title production figures for countries that did not show 1996 data (such as China, France, and the Netherlands), the annual world total increases to **968,735** - about **8 TB/year**. This number is still not quite accurate because there are still about 70 countries of the world not documented by UNESCO. However, it is the best available estimate we were able to locate.

United States

In 1997, about **64,711 books** were published in the United States, according to the *U.S. Census Bureau's 1999 Statistical Abstract of the United States*. *The Bowker Annual Library and Book Trade Almanac* provides a similar figure for 1997 (65,796 titles) and notes that this is a decrease from 1996's all-time high figure of 68,175 titles. The preliminary estimate for 1998 suggested another decrease, to 56,129 books, roughly equivalent to **.5 TB** per year.

Stock

World

To estimate the international stock of books currently available for purchase, we note that the United States produces about 40% of the world's printed material. (Source: US Industry and Trade Outlook 2000) The national library and copyright repository of the United States--the Library of Congress--contains about 26 million books. Therefore, using the 40% rule of thumb, the world stock of books might be approximately **65 million titles**.

United States

Books In Print is an authoritative source for books published and/or distributed in the United States. The 1999-2000 edition of Books in Print (with supplement) includes **1,663,815 books** - about **13 TB** in all. (This number does not include: books not published and/or not distributed in the U.S.; books not available to the trade or general public; free books not included with a title for sale; unbound material, pamphlets and booklets; periodicals and serials; books only available to members of an organization; subscription-only material, and music manuscripts, sheet music, and librettos.) According to a press release from January 2000, booksinprint.com 2000 includes **3.2 million titles** - about **26 TB** total. (Of these, about 500,000 titles are out-of-print.) This figure is consistent with online booksellers: Barnes & Noble.com asserts that it has 3 million titles available while Amazon claims to have 4.5 million titles. *Source: Wall Street Journal, July 17, 2000, "A New Chapter: Independent Booksellers Hope to Find Strength in Numbers" by Scott Eden.*

Rate of change

The number of titles published each year in the United States and internationally has risen gradually and fairly consistently over the last 10 years, apart from a slight downturn in US title production in 1997.

Newspapers

Conversion Factors

A large metropolitan daily US newspaper will run approximately 60 pages each day, and may vary from 48 to 72 pages. Therefore one year of a large US newspaper would consist of about 21,900 pages per year. To account for the smaller regional papers and those which are not published daily, we opted to use 30 for the average number of pages, which results in 10,950 pages per year.

A double truck (center fold) full broadsheet is 24 in X 36 in. Because a newspaper page would be scanned at higher resolution and contains detailed graphics, a double - truck would require about 1 Megabyte (uncompressed) and a single full broadsheet page (18 X 24 inches) would require about .5 MB.

Table 2: Newspapers Conversion Factors			
Average Number of Pages per Year	**Storage Format**	**Average File Size (per page)**	**Conversion Factor (rounded)**
10,950	Scanned TIFF (600 dpi)	500 KB	5475 MB/item
	Compressed	100 KB	1095 MB/item
	Plain text	10 KB	110 MB/item

Flow

World

The number of newspapers in the world as of 1999 is **22,643**, according to the International Standard Serial Number Register. UNESCO's Statistical yearbook from 1996 offers a different, much smaller figure: **8,391**. The discrepancy is due largely to the fact that multiple ISSNs may be assigned to what is essentially the same information product if it exists in two different formats (paper, online, floppy disk, CD-ROM, microform). For example, the New York Times print version and NYT online have two different ISSNs. Other differences may be attributed to the fact that UNESCO's excludes non-daily newspapers from its statistics or that the UNESCO data set is less complete than that of ISSN.

Using the ISSN figure and the compressed format:

.1MB/page * 10,950 pages/year * 22,643 newspapers/world = **124 TB/year**

United States

According to the Newspaper Association of America, there were **1,489** daily newspapers and **897** Sunday - only newspapers published in the United States in 1999.

Stock

The Library of Congress provides useful statistics on the nationwide stock of printed periodicals and newspapers. Most library print periodical holdings have been transferred to microfilm, microfiche, or simply made accessible online, through efforts such as the National Endowment for the Humanities' Newspaper Project. However, bound periodicals do still exist at the LOC: about 30,570 newspapers and 557,738 serials.

Rate of Change

The number of newspaper titles and the circulation figures have both declined slowly over the past 10 years.

Periodicals

Conversion Factors

Table 3: Periodicals Conversion Factors				
Periodical Type	**Average # of Pages per Item (annual total)**	**Storage Format**	**Average File Size (per page)**	**Conversion Factor (rounded)**
Scholarly Journal	**1,723**	Scanned TIFF (600 dpi)	130 KB	225 MB/item
		Compressed	26 KB	45 MB/item
		Plain text	2.5 KB	4 MB/item
Mass-Market Periodical	**5,000**	Scanned TIFF (600 dpi)	130 KB	650 MB/item
		Compressed	26 KB	130 MB/item
		Plain text	2.5 KB	13 MB/item
Newsletter	**150**	Scanned TIFF (600 dpi)	130 KB	20 MB/item
		Compressed	26 KB	4 MB/item
		Plain text	2.5 KB	.4 MB/item

World

According to the 2000 Ulrich International Periodical Directory, there are **158,000** unique periodical titles in the world. If one wished to include all serial publications, for example, yearbooks, annual publications of all kinds, not just journals, one could cite the International Standard Serial Number statistics: **617,801** for all serials around the world.

It is useful to differentiate between scholarly journals and mass market periodicals because of the differences in number of pages per year. According to a recent study on Economic Cost Models of Scientific Scholarly Journals, "A typical 1995 scientific scholarly journal has 8.3 issues... and 1,723 pages (or 208 pages per issue)." (*Source:* http://www.bodley.ox.ac.uk/icsu/kingppr.htm) A typical mass-market periodical, (such as Time, Newsweek, or People) that is published weekly, will be about 5,000 pages per year. (*Source: Robert M. Hayes, "The Economics of Digital Libraries",* http://www.usp.br/sibi/economics.html). There are also newsletters, smaller than both the scholarly journal and mass-market publications, at about 100 pages per year.

Using the ratios within the US periodical statistics as a reference point (see below), we can estimate that about half of the world's periodical publications are mass - market magazines/tabloids, one-fourth are scholarly journals, and one-fourth are newsletters.

United States

According to the *1999 US Statistical Abstract*, there are **12,036 periodicals** published in the United States. Given the world total of 158,000, this seems low. Another estimate on United States periodical production comes from the National Directory of Magazines: **20,613 titles** as of August 2000 (this figure includes Canada).

The estimated number of scientific journals published in the United States increased 62 percent, from 4,175 in 1975 to **6,771** in 1995 (*Source: Designing Electronic Journals With 30 Years of Lessons from Print, Tenopir and King, 1998* http://www.press.umich.edu/jep/04-02/king.html) Data on periodicals in the field of modern languages and literature may serve as a rough index of the trends in the general availability of scholarly journals outside the sciences. According to the MLA

Directory of Periodicals, there are currently **3,700 periodicals** in the areas of literature, language, linguistics, and folklore covered regularly in the MLA International Bibliography. This means that there are about **10,500 scholarly journals** in the United States.

According to the Newsletter and Electronic Publisher's Association (NEPA), there is no way of really knowing exactly how many newsletters exist, because they are not required to be registered anywhere. There are two directories of newsletters: the Hudson Subscription Newsletter Directory lists over 5,000 subscription newsletters while the Oxbridge Directory of Newsletters includes over 20,000 newsletter titles, of which they say 9,000 are subscription newsletters. NEPA's best guess is that there are approximately **10,000 subscription newsletters** in print today. (*Source: http://www.newsletters.org/faqs.htm*)

Office Documents

Conversion Factors

Table 4: Office Documents Conversion Factors			
Average # of Document Pages per Year	**Storage Format**	**Average File Size**	**Total Annual Production**
7,500,000,000	Scanned TIFF (600 dpi)	130 KB	975 TB/year
	Compressed	26 KB	195 TB/year
	Plain text	2.5 KB	19 TB/year

Flow

United States

To estimate the amount of information generated by offices, one might begin by looking at the statistics of the Federal Government, which is the single largest employer in the United States, with 1.8 million civilian workers as of 1998 and 1.2 million individuals in the armed services. The Federal Government, in total, employs about 3% of the nation's workforce.

The National Archives in Washington D.C. retains 2% of what the government produces, across a range of media. According to Archives representative Eric Chaskas, the Archives retains only what is deemed to be of some permanent historical value. Document types include correspondence, registers, reports, forms, treaties, case files, and log books. The perceived value determines how long a record will be retained--some will be kept indefinitely, while others are retained for no more than 6 months. An effort is made to prevent duplicating records but there is still some degree of overlap. Although the NARA representative was unable to quantify how much new information is accessioned each year, he did state that the current archival holdings, as of July 2000, include 4 billion pieces of paper, occupying a total of 21.5 million cubic feet. That is about 200 pieces of paper per cubic foot.

If one divides 4 billion by the number of years the United States government has existed (224 years), one could obtain a rough number of pages collected per year by dividing 4 billion by 224 - the result is about 18,000,000 pages per year.

The current accession rate, however, appears to be much higher. Each year, Federal agencies submit about 4,000 items and about 75% of these (3,000) are processed for archival. Although the Archives does not publish statistics on the average size of these items, it is known that NARA adds a total of 500,000 cubic feet of mostly paper-based records each year. As previously noted, in

archives, 1 cubic foot can hold 200 pieces of paper, so the total annual accession rate is therefore about **100,000,000 pages** per year.

If this represents 3% of the nation's workforce, then one could estimate that United States companies produce a total of more than **3 billion archiveable pages** each year, equivalent to **78 terabytes.**

World

Using our rule of thumb that the US produces about 40% of the world's printed materials, we can estimate that each year the world produces **7.5 billion archiveable pages**, which would be equivalent to **195 terabytes.**

Stock

Again, consider the US National Archives with 4 billion pieces of paper. If this is 3% of the United States total, then the total stock of paper held by US companies is somewhere on the order of 130 billion sheets of paper - about **3,380 terabytes**.

Rate of Change

Despite the increasing use of computers for communication and storage, the production and retention of paper office records is on the rise. The materials within the National Archives has increased in volume by more than one-third, just in the past decade.

Visual Material (Maps, Posters, Prints and Drawings)

This is a difficult category to quantify because, unlike the book, newspaper, and periodical industries, there is no central agency monitoring the production of originals or serving as a repository for the items. The best method for approximating the production statistics is to consider the holdings of the National Archives and the Library of Congress.

Conversion Factors

According to the Library of Congress, maps scanned in at 300 dpi tend to be large, approximately 100 to 300 MB. (*Source: http://lcweb2.loc.gov/ammem/formats.html#V*)

Flow

For photographs, see the Film Summary portion of this paper.

During 1998, the Library of Congress added 60,150 maps, 1,881 posters, and 4,723 prints and drawings to its collections.

Stock

World

[???]

United States

The Library of Congress has 4,523,049 maps in its collections. 85,216 posters and 405,708 prints and drawings. The National Archives has another 5 million maps and drawings, as well as 13 million still photographs and 9 million aerial photographs.

Copies

For the past 10 years, the US has produced about 30% of the world's paper and paperboard output (*US Industry & Trade Outlook 2000*). Other major producers are Asia, Latin America, Canada, and Europe. As of 1998, total global capacity was 333.6 million metric tons, with growth predicted so that by 2001 the annual total is expected to be 348.1 metric tons. A recent article in the Economist supports this forecast stating that, during the period 1993 - 1998, "the production of printing and writing paper in North America has grown by over 13%. Worldwide it has doubled since 1982."

There are four different commodity segments in the paper/paperboard industry, only two of which are relevant to this study: printing and writing paper and newsprint. In 1999, the US produced 23.8 million metric tons of printing and writing paper and 6.4 million metric tons of newsprint. The most recent global production figures (available from UNESCO's Statistical Yearbook) are for 1997, at which time the world was producing 90 million metric tons of printing and writing paper and 36 million metric tons of newspaper.

For the estimates of **printing/writing paper**, we used the midrange figure of **26 KB per 8 1/2 X 11 page**. We estimated 500 sheets of standard grade 8 1/2 X 11 paper would weigh 5 pounds. Therefore, each metric ton (2204 pounds) equals about 440 500-sheet-reams or 220,000 sheets. Multiplying by 26 KB per page results in **6 GB per metric ton**, for a total annual storage capacity of **142,800 possible terabytes** (US) and **540,000 terabytes** (world).

For **newspapers**, we estimate that storage requirements would be be about 2 times that for writing/printing paper (**12 GB per metric ton**). Newspapers tend to contain more words and graphics per page, requiring, on average 1 MB per page. Furthermore, newsprint is thinner and lighter, so each metric ton contains more individual sheets. US newsprint production in 1997 equalled about **80,000 terabytes** and international production was about **432,000 terabytes**.

These figures provide an extreme upper bound on the total number of bytes required to digitally store information currently produced in printed format each year.

Books

About **1.1 billion books** were sold in the United States in 1999. (*Source: Wall Street Journal, July 17, 2000, "A New Chapter: Independent Booksellers Hope to Find Strength in Numbers" by Scott Eden.*) This is an increase of more than 3% from 1998 when about 1,037,000 books were sold. (*Source: The 1999 Consumer Research Study on Book Purchasing, (Section II) Detailed Findings: Consumer Adult Book Purchasing*).

Newspapers

In the United States **55,979,332 daily newspapers** and **59,894,381 Sunday newspapers** circulate each year. (*Source: Newspaper Association of America.*)

Periodicals

The total number of US magazines circulated annually exceeds **500 million.** (*Source: US Industry and Trade Outlook 2000.*)

Office Documents

Each year, almost **500 billion copies** are produced on copiers in the US (**13,000 TB**); nearly **15 trillion copies** are produced on copiers, printers, and multi - function machines. 15 trillion copies is roughly equal to **390,000 TB**. This accounts for the majority of the printing/writing paper used each year. (*Source: XeroxParc.*) For specific information on fax printing, see the Telecommunications Summary of this report.

A 1998 Coopers & Lybrand study showed that the average office makes 19 copies of each document. The average office loses 1 out of 20 office documents. It then costs: $120 to search for the document; $250 to recreate it, if lost (1 lost document = $370). (*Source: NAGARA, Records Management Technical Bulletin www.nagara.org/rmbulletins/bulletin_2.htm*)

The Government Printing Office, which handles the printing needs of the entire Federal community,

uses or sells more than 55 million pounds of paper each year. Approximately 10,000 titles are available for sale to the public at any given time.

Fun Facts About Print Media

Paper Size Equivalencies

According to ArchiveBuilders.com: 1 pulp tree (loblolly pine) = 1/10th cord of wood = 10,000 pages = 1 file cabinet = 4 boxes = 1/2 Gigabyte = 1 CD

Scope of the Library of Congress Collections

"LC21: A Digital Strategy for the Library of Congress," (a report from an advisory group at the National Academy of Sciences about strategic directions for the applications of information technology in the Library) reports the following statistics about the collections of the Library of Congress:

- 26 million volumes (pages 2-3)
- 1 million serials (pages 2-3)
- 4.5 million maps (pages 2-5)
- 1,000,000 audio recordings (pages 2-6)
- 200,000 cans of film (pages 2-6)
- 13.5 million images (pages 2-7)
- 70,000 periodicals (pages 2-7)
- 1,400 newspapers (pages 2-7)

The Library receives some 22,000 items each working day and adds approximately 10,000 items to the collections daily.

Paperless Society?

With the spread of the personal computer, many predicted the advent of the paperless office. So far, this prediction has not come to pass. One one hand, according to BIFMA, a business and institutional furniture manufacturer's association, as of Dec. 16, 1999, the percentage of file cabinets in relation to other types of office furniture, like seating, desks, and tables, has steadily decreased, from 16% in 1988 to 12.9% in 1998. This would lead one to believe that less paper is being generated to require storage space. On the other hand, the installation of home computers and expansion of home offices has led some analysts to predict that by 2003 the consumption of standard office paper will double from the 1996 level (*Source: 1999 U.S. Industry and Trade Outlook, 10-3*).

Growth of Online Periodicals

Of the more than 157,000 serial titles reported in the latest edition of the international periodicals directory, 10,332 were available exclusively online or in addition to a paper counterpart (*Source: Ulrich's International Periodicals Directory, 1999, p. vii*).

Bulk Mail

As of 1990, the United States Postal Service handled almost 90 billion pieces of first-class mail per year and another 75 billion pieces of second- , third-, and fourth-class mail.

Americans receive almost 4 million tons of junk mail a year. About 44% of the junk mail is never opened. Every person in the United States receives junk mail that represents the equivalent of one and a half trees a year. (*Source: The Consumer Research Institute's Stop Junk Mail Page*)

For more information on flows of information through the mail, please see the Mail Summary of this report.

Print Media Bibliography

- Bogart, Dave, ed. *The Bowker Annual Library and Book Trade Almanac*, 44th edition. New Jersey: R.R. Bowker, 1999.
- Cummings, Anthony M., Marcia L. Witte, William G. Bowen, and Laura O. Lazarus. *University Libraries and Scholarly Communication: A Study Prepared for the Andrew W. Mellon Foundation*. The Association of Research Libraries, 1992
- Hayes, Robert M., UCLA School of Information Science, "The Economics of Digital Libraries" www.usp.br/sibi/economics.html
- King, Donald W. and Carol Tenopir. "Economic Cost Models of Scientific Scholarly Journals," 1998. www.bodley.ox.ac.uk/icsu/kingppr.htm
- American Forest and Paper Association, *1999 Statistics for Paper, Paperboard and Wood Pulp*. Washington, DC, 1999. To order a copy, see www.afandpa.org/about/about.html or call (800) 244-3090.
- *Annual Report of the Librarian of Congress*. Washington, DC: Library of Congress, 1999.
- ArchiveBuilders.com White Paper, "Computer Storage Requirements for Various Digitized Document Types"
- Association of Research Libraries Statistics www.arl.org/stats/index.html
- *Books in Print, 1999-2000.* New Jersey: R.R. Bowker, 1999.
- International Standard Serial Number Register www.issn.org
- International Standard Book Number Register www.isbn.org
- JSTOR digital library. www.jstor.org/
- Magazine Publishers of America, Information Center, (212) 872-3745. www.magazine.org/resources/fact_sheets/ed1_8_99.html
- National Directory of Magazines, www.mediafinder.com/mag_home.cfm
- Newsletter and Electronic Publisher's Association (NEPA) www.newsletters.org/faqs.htm
- *Newspaper Association of America, Facts About Newspapers*.
- Newspaper Project, National Endowment for the Humanities, www.neh.gov/preservation/usnp.html
- *Oxbridge Directory of Newsletters 1997*. New York: Oxbridge Communications, Inc., 1997.
- *Ulrich's International Periodical Directory*. New York: Bowker Publishing, 2000.
- *UNESCO Statistical Yearbook 1999.* Paris, UNESCO, 1999.
- U.S. Census Department, 1999 Statistical Abstract of the United States, www.census.gov/prod/www/statistical-abstract-us.html
- *U.S. Department of Labor, Bureau of Labor Statistics*
- U.S. Government Printing Office, *Prepared Statement before the Committee on Rules and Administration, U.S. Senate on Public Access to Government Information in the 21st Century* July 1996.
- *U.S. Industry and Trade Outlook*. Available in print from McGraw Hill/U.S Department of Commerce Washington, D.C. or download from www.ntis.gov/products
- Chaskas, Eric. US National Archives and Records Administration.
- *Walden's Paper Report*. Twice-monthly newsletter, published by Walden-Mott Corporation, Ramsey, NJ. Available by subscription only. See www.walden-mott.com/PaperReport/PAP_RPT.HTM or call (201) 818-8630.

Charts

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Film - Details

Impact of Digital Cameras On Rate of Growth of New Photographs

80 billion new photographs are taken around the world every year. Photography generally follows the trends of the overall economy, for example, in the mid-1990's there was a slight slowdown in photography concurrently with the economic recession experienced in many major world economies, particularly those in Asia. However, it is expected that in the next five years there will be continuing growth in the overall number of photographs as the enormous potential of China, India and other developing regions is realized. At the present 25% of film sales occur outside of North America, Western Europe and Japan. In June 2000 Kodak announced that China was its second largest market after the United States. To give some perspective to the growth potential for photography in China, film usage there is currently less than one roll per capita as opposed to 3.6 rolls per capita in the United States. Similarly, only 15% of Chinese households own cameras as opposed to over 80 percent in the United States.

A countervailing trend to the growth of silver halide film based photography is the increasing popularity of digital photography. Very rapid growth is predicted over the next five years for this method of taking pictures. In 1999, in the United States almost 2 million digital cameras were sold, double the number from the year before and representing about 12 percent of all cameras sold. In a survey by NPD Intellect, more than 70% of consumers planning to buy a new camera say they will choose a digital one. Kodak projects that digital photography will account for forty-five percent of its revenue by 2005 from a current 17%.

Photofinishing News and Lyra Research predict that there will be an 80 percent growth in the number of digital cameras worldwide by the year 2002. There were 8.3 million digital cameras in use worldwide in 1999 according to this study. (compared to 200 million conventional film cameras in the United States alone). Salomon Smith Barney predicts that the growth of photographic exposures worldwide will grow from 71.4 billion in 1997 to 96.8 billion by 2002. At the same time, they expect digital camera exposures to grow 1500 percent. Many of these photographers will use web photo services. For example, Photo Works already hosts over 100 million images on its web site and there

are many more companies competing in this same market. These companies indicate that at this time 60 to 80 percent of the photographs on their web sites are scanned from silver halide film.

Kodak predicts that the growth trend will predominate until at least 2004, at which point traditional film will gradually decline in use as the digital camera becomes ascendant.

The Internet may actually be increasing demand at the present for traditional photographs points out Morgan Stanley Dean Witter. eBay has 4 million items for sale every day and the inventory turnover averages about 7 days. About 80 percent of these items are accompanied by a photograph. About one-half of one percent of current U.S. photographs are pictures of merchandise offered for sale on eBay.

Table 1: Photographic Exposures (Billions)

Year:	1992	1997	1998	1999	2000	2001	2002	2005
World	56	84.4	83.3	82			89	
US	21	24.9	26.9		34.4			41.2

Source: 1992 statistics are from U.S. Industrial Outlook 1994; 1997 and 1998 U.S. Statistics are from U.S. Industry and Trade Outlook 2000; 1997 and 1998 world statistics are from the Silver Institute. The 1999 world figure is from Kodak's corporate web page. The 2000 U.S. figure is from Kodak and printed in May 8, 2000 Wall Street Journal. The 2002 projection for the world is from the Silver Institute. The 2005 projection is from the May 8, 2000 Wall Street Journal article and attributed to Kodak.

Chart of the annual production of film used in conventional photography in the United States.

Conversion and Compression

Photographs

To assess the amount of data in terms of bits in a photograph, certain assumptions must be made about the size of the film used and the technical way in which the photograph is digitized and stored. Professional photographers generally use physically larger film formats with more chemical grains on the film to store information than do amateur photographers. Conversion of these professional photographs to a digital medium, therefore, generates a much bigger computer file. For example, high quality professional photographs may require 40 megabytes or more of storage space.

In the early 1990's Kodak introduced the PhotoCD format for the storage of photographs on compact disks. Since then, the format has been widely used by consumers, professional photographers and archivists. Kodak estimates that most photographs can be converted to its PhotoCD format with little to no loss of image data in 5 megabytes. Kodak's rule of thumb is that it can place 100 photographs on one CD, which holds approximately 650 megabytes. For the sake of carrying out our estimates of total photographic data, therefore, we use this same 5 megabyte figure.

Of course, actual digital storage of photographs might consume far less space if any of the very popular compression schemes are used. For example, the JPEG standard is commonly used to reduce photographic file sizes to one-tenth of their original size. However, this is a "lossy" compression, i.e., one where data is irrecoverably discarded in the compression.

The 5 megabyte conversion factor used is also supported empirically by the extensive image digitizing experience of the University of California, Berkeley Digital Library. The photographic collection there is reported as containing 164,702 images which require 888 gigabytes of storage space, or just over 5 megabytes per photograph. (http://elib.cs.berkeley.edu/admin/rpts/TestSuite/00q2.html)

Motion Pictures

The conversion of motion pictures to digital media can result in very large file sizes. It is estimated that each frame of film can be digitized in about 12 megabytes. However, to create the illusion of seamless motion, 24 images are used per second of film. Thus, one second of film requires 288 megabytes of digital storage, one minute requires 17.28 gigabytes, and one hour requires 1 terabyte. At this rate a feature length film of 100 minutes would fill about 400 typical DVD's. Fox Animation Studios, for example, reports that its animated feature "Anastasia" is 1.49 terabytes in size. This particular film was digitally inked and painted and only then transferred to film for theatrical distribution.

In cases where film will be distributed or stored on digital media such as DVD's some form of compression must be used to make the file sizes tractable. The most common compression format now used for motion pictures is MPEG-2. Using this format, a feature lenth film can be reduced to several gigabytes in size and placed on a single DVD. This is a "lossy" compression format where data that is present on the film is irrecoverably lost in the compression process.

X-Rays

The conversion of x-ray images to bits requires careful attention to the loss of information because of the attendant risk of harm to patients. See, The Economist, "Why JPEG's Can Be Bad For Your Health", June 2000, citing a study published in the Journal of the American College of Cardiology in which researchers found that using JPEG with a compression ration of 16:1, the error rate was 30% higher than for uncompressed images. In fact, in the United States conversion of x-rays to digital format is regulated by the U.S. Government for the protection of the patients. FDA Regulation 510(k).

1 chest X-ray = 1 megabyte (14 x 17 inches), 150 dpi, 12 bits (compressed). 12 bits per pixel, provides 4,096 shades of grey) (wavelet compression, lossless mode, has FDA 510(k) approval) 150 dpi, 12 bit images recommended by American College of Radiology for primary reads). (Source: Steve Gilheany, Archive Builders)

A team of radiologists from the University of Florida estimated the size of a converted conventional radiograph as 10 megabytes assuming a matrix size of 2048 x 2580 pixels with a 16-bit sample of each pixel. This estimate assumes there will be no compression. Honeyman, Huda, Frost, Palmer & Staab, "Picture Archiving and Communications System Bandwidth and Storage Requirements", Journal of Digital Imaging, Vol. 9, no. 2, May 1996.

A slightly smaller estimate of the amount of storage required for a standard chest x-ray comes from the University of Pittsburgh, Clinical Multimedia Lab, which estimates the size for a radiograph (2K * 2K * 16 bits) as approximately 8 megabytes (uncompressed). Clunie, David A., Lossless Compression of Grayscale Medical Images - Effectiveness of Traditional and State of the Art Approaches (Link to this paper (PDF))

This latter figure will be used because it does not assume compression and is generally consistent with the assumptions made for converting traditional photographic film to digital format. Furthermore, Dr. H.K. Huang, Radiologist at UC San Francisco, estimates that "a typical examination generates betwen 10 and 20 MBytes." Huang, "Teleradiology Technologies and Some Service Models" Computerized Medical Imaging and Graphics, Vol.20, No.2 (1996). Applying this assumption of 8 mb per x-ray image results in a total amount of data stored on x-ray film of 16 petabytes.

Flow of New X-Rays

There are three common uses of x-ray film: medical imaging, dental imaging and non-destructive

testing in manufacturing processes. The Silver Institute Reports that the medical uses of x-ray film overwhelm the other categories, accounting for 92% of the world's overall x-ray film usage.

Medical Imaging

The administration of radiologic procedures in the United States and the developed world are fairly well documented. The number of such procedures in the developing world, however, are harder to come by. The United Nations Scientific Committee on the Effects of Atomic Radiation found that no data at all could be found for radiological procedures for half the world's population and that there is only fragmentary data on examination rates for another quarter of the world's population.

Table 2: Medical X-Ray Procedures in United States

1992	1993	1994	1995	1996	1997	1998	1999	2000
275	280	286	291	297	303	309	315	320

Source: Theta Reports

The UN Committee observed that the developed nations of the world use x-rays at a rate generally consistent with that of the United States, i.e., approximately 1000 procedures (including dental uses) for 1000 population. Therefore, on the assumption that the developed world's population is 1.2 billion, then a rate of x-ray procedures of slightly more than one per person per annum would yield approximately 1.2 billion x-ray procedures per year in the developed world. The population of the less developed world is 4.9 billion. If it is assumed that there are another 0.5 billion x-ray procedures performed for this population, a rate one-tenth that of the developed world, then the world total of annual medical and dental x-ray procedures is 1.7 billion.

According to Clinica Reports and Theta Reports, the world market for x-ray film is approximately $3.5 billion per year with the United States market accounting for around $1.4 billion of that total amount, a 40% share of the world market. If the dollar share of the market is reflective of the share of procedures performed than there are about 750 million x-ray procedures annually. On the other hand, it is reasonable to assume that film is sold for less in dollar terms around the world than in the United States. If it is assumed that the price of film in other markets is 50% of that in the United States, then this would work out to about 913 million x-ray procedures outside of the U.S., and around 1.2 billion for the entire world.

Another approach to verifying these estimates is to take into account the use of silver for the production of x-ray film as estimated by the Silver Institute. In 1998, 71.5 million troy ounces of silver were used for medical x-ray film and that was sufficient to produce 387 million square meters, or 4.2 billion square feet, of x-ray film. This implies that one ounce of silver is enough to produce 5.4 square meters, or 58 square feet, of medical x-ray film. Theta Consulting estimated U.S. consumption of 1.7 billion square feet of medical x-ray film to perform 309 million procedures in 1998, indicating usage of an average of 5.5 square feet of film per procedure. If the same amount of film were used per procedure in the rest of the world, 450 million medical x-ray procedures were performed there. This is consistent with a world wide x-ray procedure total of 750 million.

We have chosen to estimate the total number of world medical x-ray procedures at 1 billion annually as a reasonable balance of the available statistics and assumptions discussed.

In order to determine the actual number of x-ray images taken some calculation must be made based upon the average number of films taken per procedure (more than one image may be captured on one film). In fact, the number of films used per procedure varies considerably. If 2 films are shot for each x-ray procedure, there are approximately 2 billion x-ray medical images taken worldwide every year.

Other X-Ray Uses

The other major uses for x-ray film are dental imaging and the non-destructive testing of materials in manufacturing and fabrication processes. These uses are approximately 8 percent of overall x-ray film usage in developed nations according to the Silver Institute. Accordingly, based upon the finding of 2 billion x-rays for medical purposes, then industrial and dental uses would amount to 160 million x-rays in the developed world for all purposes. The total world use of x-ray film is therefore approximately 2.16 billion images annually.

Copies

As more photographs are digitized through inexpensive home scanners or are taken on digital cameras in the first place, it is not yet clear how often paper copies will be made. Furthermore, it is still unknown whether output from digital storage will typically be on photographic paper, or by means of ink jet or thermal printers. Current indications are that ink jet printing will be the most common process. In the year 2000, it is estimated that United States consumers will print out 5.4 billion photographic pictures, mostly at home. By the year 2005, this will quadruple to 26 billion. At the same time, traditional photoprocessing is expected to grow only 20 percent to 41.2 billion prints. (Source May 8, 2000 Wall Street Journal.)

The most common use made by consumers of digital images so far has been sharing them through e-mail. Similarly many new businesses are counting on the consumer's desire to share digital photos by hosting them on web sites. In December 1999 there were over 25 million unique home visitors and 12 million work visitors to the web sites of over one hundred companies providing photographic hosting services.

It is anticipated that the growing number of digital images owned by consumers will have a dramatic impact on their need to consider making backup copies of their personal disk drives. Traditionally, only a small percentage of home computer data is original data that requires backup at all. This need will grow dramatically if many photographs are stored digitally. The web hosting companies are counting on consumers storing the data on their sites and ordering prints from time to time. Currently, consumers are able to get free photo developing and unlimited disk space from these companies.

Copies of Motion Pictures

The Wolfman Report on the Photographic & Imaging Industry in the United States states that the average number of prints per original motion picture is 700. The Silver Institute, however, reports that 6,000 release prints are made for each feature movie. Interestingly, however, these copies are short-lived. 98 percent of all films for theatrical distribution made in the United States are destroyed by FPC, Inc.of Mountain City, Tennessee. After the films are no longer being shown in movie theaters, they are sent to FPC, which destroys10 million pounds of film every year. The film is shredded and sent to FPC's parent company, Kodak, where it is recycled and made into new film or fuel used in power plants. (Source: Associated Press)

Film Factoids

- Kodak describes the photography market as follows: 82 billion pictures processed a year throughout the world with 750 million rolls of film processed annually in the United States and 2.9 billion rolls consumed worldwide. Kodak also estimates that of the photographs that are processed approximately 2 percent are later reprinted or reused in some way.

  SOURCE: http://kodak.com/US/en/corp/georgeFisher/shihPres.shtml, Presentation to Imaging Technology Analysts Group on February 24, 1999. by Willie Shih, President, Digital and Applied Imaging, and Vice President, Eastman Kodak Company.

  More than 82% of U.S. households have cameras and use them to take over 17 billion pictures annually. It is estimated that there are over 150 billion photographs stored in those households. "Instant Images" Fortune, Winter 1997 (Technology Buyer's Guide Supplement) 184-187. This article cites these figures as according to the Photo Marketing Association.

According to Kodak, China has become its second largest market. The per-capita film consumption in China averages one roll a year, as compared with 3.6 rolls in America. Only 15 percent of Chinese own cameras. (XINHUA ECONOMIC NEWS SERVICE, 6/13/00).

- The United States Library of Congress reports that it holds 12 million photographs in its collection.

- Number of UK Feature Films Produced Annually (1912-1998)

- The Department of Commerce cites the journal *Medical Imaging* for the statistic that "U.S. health care systems spend up to $7 billion a year on film alone." U.S. Industry and Trade Outlook 2000, 44-18. Unfortunately for the film manufacturers, this probably overstates the amount of sales by a wide margin. Theta Reports puts the sales of x-ray film at $1.4 billion in 1996 and expects it to reach $1.5 billion by 2000.

- American women undergo approximately 30 million mammograms annually. Medical Industry Today, December 20, 1999.

- Approximately 150 million chest x-rays are done in the United States each year.(Source: University of Pittsburgh, Clinical Multimedia Lab).

- The Silver Institute reports that 92 million ounces of silver was used in 1999 for radiography throughout the world. This association also reports that dental and commercial uses of x-ray film traditionally account for about 8% of the total radiography market.

- In 1996 it was estimated that the average institutional radiology department would generate approximately 15.7 gigabytes of data per day and 3.5 terabytes per year. Honeyman, Huda, Frost, Palmer & Staab, "Picture Archiving and Communications System Bandwidth and Storage Requirements," Journal of Digital Imaging, Vol. 9, no. 2, May 1996.

- Production of x-ray film is essentially controlled by three companies: Kodak, Sterling Diagnostic Imaging/Agfa and Fuji Medical Systems. The U.S. market for x-ray film is about $1 billion annually. Modern HealthCare January 18, 1999.

- "More than 32.5 million mammograms and 4.5 million cardiac catheterization procedures are performed each year in the U.S., and X-rays account for 70 percent of all medical imaging procedures." PR Newswire, Feb. 29, 2000, " New GE Digital Imaging Technology"

- Kodak sells about $800 million year in x-ray film. Associated Press, Nov. 26, 1999. This also seems to suggest that the statistic cited above is incorrect as to the overall U.S. x-ray film market.

- "The reduction in the number of competitors is clearly seen in the medical X-ray film market. In 1995, there were five global players: Du Pont, 3M, Kodak, Fuji and Agfa, sharing approximately 90% of the world market. In 1996, Du Pont sold its X-ray film business to Sterling, which subsequently announced the resale of the business to Agfa in January 1999. The imaging division of 3M was spun off into a new company, Imation, in 1996, and this was subsequently purchased by Kodak in August 1998 for US$520 million. There are now only three major players in the global market: Kodak, Agfa and Fuji, assuming that the Agfa/Sterling deal is approved by the US FTC and the European Commission. This obviously leads to dominant companies with market shares in excess of 40% in some major markets. Further consolidation of these groups, however, would almost certainly run into legal difficulties." Medical Device Technology (May 1999), Vol 10, no. 4

- "Greater volumes of products and hence bargaining power are being concentrated into the hands of a declining number of purchasers. The greater volume and value involved with each negotiation mean that the market-share impact of either winning or losing a contract will

increase. This can lead to a decline in prices, as witnessed following the merger of Columbia and HCA, which was widely considered to have started the rapid decline of film prices in the US market between 1996 and 1998. Prices fell by more than 20%, wiping approximately US$150 million per annum off the value of the US X-ray film market." Id.

Film References and Sources

- Photo Marketing International: http://www.pmai.org/
- Photo Marketing Magazine: http://www.photomarketing.com/
- Berkeley Digital Library Sunsite: Digitizing Imaging and Text: http://sunsite.berkeley.edu/Imaging/

Reference Books

- The Photography Encyclopedia (call number TR9.M39 1999)
- Journal of Electronic Imaging (TK8315.J68)
- NAICS 512110 Motion Picture and video production
- SIC Code 7812 Motion picture and video production
- The Film Encyclopedia
- Guinness Book of Movie Facts and Feats
- International Film Guide
- Screen International
- Screen Finance
- Screen Digest
- The Motion Picture Guide Annual 1999
- The International Film Index, 1895-1990, edited by Alan Goble published by Bowker-Saur (London: 1990).
- The Silver Institute
- Theta Reports, X-Ray Film Markets, Report No. 671, January 1997
- University of Pittsburgh Clinical MultiMedia Lab
- United States Government, U.S. Food and Drug Administration, Center for Devices and Radiological Health
- United Nations Scientific Committee on the Effects of Atomic Radiation, Reports to the General Assembly, (New York 1993, 1994).

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Optical - Details

Originals

Conversion Factors

In its most common format, a compact disc (CD) holds about 650 MB.

A digital video disc or digital versatile disc (DVD) holds about 4.38 GB per disk - about 20 times more data than a compact disc (CD). The DVD specifications describe four disk configurations: single-sided (SS) vs. double-sided (DS) and single-layered (SL) vs. double-layered (DL) disks. Total storage capacities for DVDs range from 1.36 GB to 15.90 GB as seen in the following chart, reproduced from JimTaylor's DVD Demystified, an authoritative set of answers to Frequently Asked Questions about DVD:

Table 1: DVD Storage Capacities

Format	Capacity
DVD-5 (12 cm, SS/SL)	4.38 GB of data, over 2 hours of video
DVD-9 (SS/DL)	7.95 GB, about 4 hours
DVD-10 (12 cm, DS/SL)	8.75 GB, about 4.5 hours
DVD-14 (12 cm, DS/ML)	12.33 GB, about 6.5 hours
DVD-18 (12 cm, DS/DL)	15.90 GB, over 8 hours
DVD-1 (8cm, SS/SL)	1.36 GB, about .5 hour
DVD-2 (8cm, SS/DL)	2.48 GB, about 1.3 hours

DVD-3 (8cm, DS/SL)	2.72 GB, about 1.4 hours
DVD-4 (8cm, DS/DL)	4.95 GB, about 2.5 hours
DVD-R 1.0 (12 cm, SS/SL)	3.68 GB
DVD-R 2.0 (12 cm, SS/SL)	4.38 GB, 8.75 GB for rare DS discs
DVD-RW 2.0 (12 cm, SS/SL)	4.38 GB, 8.75 GB for rare DS discs
DVD-RAM 1.0 (12 cm, SS/SL)	2.40 GB
DVD-RAM 1.0 (12 cm, DS/SL)	4.80 GB
DVD-RAM 2.0 (12 cm, SS/SL)	4.38 GB
DVD-RAM 2.0 (12 cm, DS/SL)	8.75 GB
DVD-RAM 2.0 (8 cm, DS/SL)	1.36 GB
CD-ROM (12 cm, SS/SL)	0.635 GB
CD-ROM (8 cm, SS/SL)	0.18 GB

Flow

World

It is a fairly simple matter to obtain information about the world music market value and units shipped; organizations such as the Recording Industry Association of America and the International Federation of the Phonographic Industry report these statistics each year. More difficult to obtain are international statistics on the number of unique titles released each year. To obtain this figure, we used statistics regarding the US market share and US record releases (see below) to estimate the total releases worldwide. The United States holds a 37% share of the world music market and releases about 33,100 items per year. Therefore, the world produces about **90,000 originals** per year, equivalent to **58 TB** (uncompressed).

United States The Recording Industry Association of America (RIAA) reports annual figures for new releases and album re-releases. The 1998 releases are equivalent to about **21 TB** of data.

Table 2: Audio Releases

1992	1993	1994	1995	1996	1997	1998
18,400	20,300	36,600	30,200	30,200	33,700	33,100

Source: http://www.riaa.com/MD-US-6.cfm

Stock

The All Music Guide (www.allmusic.com), a comprehensive entertainment database for music, videos, DVDs and video games, provides statistics on the number of CD-audio originals in the world. The AMG database is licensed by major music sites such as CDNow, ArtistDirect and Tunes.com. AMG's listings indicate a total of 523,363 albums (445,735 popular music and 77,628 classical music albums). Each CD can hold 650 MB, so the total AMG catalog would equal roughly **340 TB**.

According to the 1999 edition of CD-ROM's in Print, there are about **16,200** unique CD-ROM titles. This figure includes business applications (such as word processing and spreadsheet packages), games, reference tools, and instructional programs.

Copies

Flow

World

Replication

Replication statistics include discs for retail as well as discs used for promotions, training, and rental. Some percentage of the replicated discs also turn out badly and are never used. These figures give us an upper bound on the amount of information that could be stored on compact disc each year.

In 1999, there were **4,654 million audio CDs** and **3,591 million CD-ROMs** replicated worldwide, according to the International Recording Media Association (IRMA). In addition, **194 million DVD-Video** units, **12 million DVD-ROM** units, and **2 million DVD-Audio** units were replicated in 1999.

United States

Replication

For audio CDs, data CD-ROMs, DVD-ROM, and DVD-Audio disks, North America produced roughly **50%** of all the disks replicated. In addition, North America replicated about 75% of the DVD videos.

Retail

During 1999, according to the Recording Industry Association of America, 938.9 million CDs were shipped to retail by U.S. producers. The U.S. has a 37% share of the world's sales. The nine next largest markets are Japan, which follows our lead with 16.7%, followed by the United Kingdom (7.6%), Germany (7.4%), France (5.2%), Canada (2.3%), Brazil, Australia and Spain (1.7% each), and Mexico (1.6%). From the US figures and market share, one can extrapolate that CD shipments worldwide are about 2,537 million units. This is equivalent to **1.6 billion TB**.

Rate of Change

According to the London-based International Federation of the Phonographic Industry (IFPI), the global music market was worth US $38.5 billion in 1999 - up by 1% in constant dollar terms with total unit sales of US$3.8 billion in 1999. Overall units remained level, with a continued growth of 3% in the CD market offset by a 10% decline in cassette sales and an 11% decline in singles.

Bibliography

- The All Music Guide, www.allmusic.com
- The CD Information Center www.cd-info.com/CDIC/index.html
- CD-ROM Finder: The World of CD-ROM Products for Information Seekers. Medford, NJ: Learned Information, 1993
- CD-ROMs in Print, 13th Edition: An International Guide to CD-ROM, CD-I, 3DO, MMCD, CD32, Multimedia, Laserdisc, and Electronic Products. New York: The Gale Group, 1999.
- DVD Entertainment Group www.dvdinformation.com
- DVD Insider www.dvdinsider.com
- Taylor, Jim. DVD FAQ www.dvddemystified.com/dvdfaq.html
- DVD Channel News www.dvdchannelnews.com
- International Federation of the Phonographic Industry www.ifpi.org

- International Recording Media Association, Statistics page www.recordingmedia.org/statistics/statistics_idx2.html
- Medialine www.medialinenews.com
- Optical Storage Technology Association www.osta.org/
- SUN CD-ROM FAQ saturn.tlug.org/suncdfaq/

Charts

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Magnetic - Details

Magnetic Storage Media

The first magnetic storage mechanism was the Telegraphone invented in 1898 by Danish scientist Valdemar Poulsen.

The digital revolution has so far been magnetic. The vast majority of the world's data is now created, transported and stored in electro-magnetic systems.

Hard Disk Drives (World)

No storage medium has ever had the explosive growth demonstrated by the hard disk.

Year Disks	Sold (Thousands)	Storage Capacity (PetaBytes)
1995	89,054	104.8
1996	105,686	183.9
1997	129,281	343.63
1998	143,649	724.36
1999	165,857	1394.60
2000	187,835	2553.7
2001	212,800	4641
2002	239,138	8119
2003	268,227	13027

Source: IDC (1999) "1999 Winchester Disk Drive Market Forecast and Review"

The incredible growth in hard disk shipments has been accompanied by a relentless decrease in the cost

per gigabyte of storage capacity:

Year	Cost per GB	GB's PER $200
1988	$11,540	0.02
1989	9,300	0.02
1990	6,860	0.03
1991	5,230	0.04
1992	3,000	0.07
1993	1,460	0.14
1994	705	0.28
1995	330	0.61
1996	179	1.12
1997	94	2.13
1998	43	4.65
1999	23	8.70
2000	13	15.38
2001	6	33.33
2002	3	66.67

Source: Wall Street Journal, June 26, 2000

Details regarding hard disk shipments.

Rates of Growth

The projections for the growth over the next few years in hard disk sales are for units shipped to increase at an annual rate of 15 to 20% but for the actual capacity shipped to grow much faster - 70 - 80% annually.

According to Disk/Trend, 75% of the disk drives sold are for desktop computers, followed by 13% for servers and 12% mobile drives for laptop computers. Therefore, although enterprise level disk storage systems each may have huge capacity, the vast number of disks deployed to individual workstations really accounts for the enormous scale of the world's current digital storage capability.

The lifespan of a hard disk is approximately 3 years. The storage capacity of the hard disks shipped in 1998, 1999 and 2000 is 4672 petabytes, or roughly 5 exabytes. In order to appreciate the scale of this statistic, consider that Roy Williams of CalTech advises that 5 exabytes is equivalent to the number of words ever spoken by all human beings.

The hard disk typically shipped with a desktop personal computer in 2000 holds 10 gigabytes.

Disk drives are expected to be deployed in applications other than personal computers, such as TV set-top boxes. These increasingly popular devices allow users to store TV shows on disk rather than tape and to stop and rewind during a broadcast while still recording. In all, by 2003 the IDEMA predicts that 8-10 percent of the disk drive market will be such devices.Similarly, starting in the year 2000 audio jukeboxes and computer game consoles will also include hard disk drives. Already the higher resolutions of digital cameras are creating such large file sizes that small hard disks will be incorporated into cameras as well.

Floppy Disks (World)

The number of floppy disk drives sold every year has remained relatively constant at around 100 million units for the past several years. Little change is expected. (Source: Computer Tech Review, April 1, 1999). The number of floppy disks being sold is diminishing rapidly as their storage capacity is too small to be useful in light of the much larger file sizes now common.

Year	3.5" Disks (billions)	Total Capacity (terabytes)	5.25" Disks (millions)
1996	1.823	2625	32
1997	1.179	1698	11.7

Source: International Recording Media Association

The Japanese Recording Media Industries Association, however, claim somewhat higher figures for floppy disk production, although confirming that the trend is still dramatically downward. This organization, for example, anticipated sales of 2.21 billion floppies in 1998. It would appear that an estimate of 2 billion for 1998, 1.5 billion for 1999 and 1 billion for 2000 would be a reasonable compromise around these figures. This would indicate that there would be a total stock of floppy disks of 4.5 billion, assuming 3 years production as stock.

One of the world's largest producers of 3.5 inch floppy disks is CMC Magnetics. They claim to have produced over 700 million disks in 1998. They also are purported to produce 56% of the world's floppy disks in 2000, implying a market of somewhere north of 1 billion disks.

Floppy disks are used primarily for backup and are little used now for original content creation.

Removable magnetic disk drives (World)

Removable drives are primarily used for backup, transfer of files, e.g,, desktop publishing files to service bureaus, or video or image editing. The general trend for low-end disk (capacity of around 100 to 250 megabytes, e.g. Iomega Zip Drives) sales is upward with a strong possibility that, if manufacturer incompatibility issues are ever resolved, this format could replace the 1.44 mb floppy. However, high capacity removable drives, with capacity in the gigabyte or better range, are being replaced by recordable CD's, which in turn may be replaced by recordable DVD's.

Year	Low-End Disk Drives	High-End Disk Drives
1996	3723	992
1997	7724	1334
1998	12035	1164
1999	17039	701
2000	21775	623
2001	26087	578
2002	30182	554
2003	34287	541

Source: IDC (1999), "1999 Optical/Removable Storage Market Forecast and Review"

The high capacity drives (Iomega Jaz) come with a free cartridge and it is assumed that an additional three cartridges are sold with each drive. Source: San Francisco Chronicle, Jan. 23, 1998 "Does Bad News for Iomega Mean Horrible News for HMT?"

The amount of original content created directly to this medium is, therefore, probably quite low. Furthermore, the disks are regularly reused and not generally viewed as archival solutions.

Magnetic Tape

Tape was the primary storage medium for the first generation of electronic computers in the 1950's. Reel-to-reel half-inch tape was used for data storage on mainframe computers from the earliest days of computing into the 1970's. Since that time, numerous tape formats have been developed. The worldwide installed base for tape drive units is 25.2 million.Michael Lesk estimated that in 1995, the magnetic tape industry would ship 200 petabytes of blank tape. (Lesk, M., "Preserving Digital Objects: Recurrent Needs and Challenges").

Current estimates are that approximately $1 billion of tape media will be sold every year. Source, Infostore July 1, 1999 (Tape: The Media Is the Message).

Worldwide Tape Drive Shipments - 1996-2003 (in thousands)								
	1996	1997	1998	1999	2000	2001	2002	2003
DC2000	3231.5	2365	1695.3	1867.3	1814.8	1749.4	1711.0	1688.7
SLR	350.4	327.1	311.7	290.0	278.3	272.0	267.5	266.8
4 mm	1370.0	1607.5	1634.8	1689.4	1650.5	1592.8	1525.9	1426.7
8 mm	201.8	213.6	180.9	138.2	131.0	127.7	135.1	149.3
DLT	160.0	356.4	366.1	483.5	550.3	612.5	677.4	746.5
LTO Ultrium	0	0	0	0	28.5	59.5	112.0	192.1
0.5" Cartridge	45.2	46.2	46.6	46.5	45.4	44.1	42.6	40.9
Total:	5358.8	4915.7	4235.3	4514.9	4498.9	4458.0	4471.5	4511.0

In a filing with the United States Securities and Exchange Commission in July 1999, Storage Tek, a large tape drive manufacturer, wrote that the cost of data storage on computer tape media was less than $.005/megabyte ($5.00/gb). Therefore, if predictions are correct that approximately $1 billion of computer tape media is sold every year, that implies worldwide annual tape storage capability of 200 petabytes. There may be some incongruity in these figures because the $1 billion may reflect manufacturer revenue for the product, rather than the retail cost of the product to end users, which would presumably be much higher. In fact, the Department of Commerce figures for the mid-1990's showed factory revenue for computer tape manufacturer's of around $600 to 700 million. Substantial amounts of tape media are manufactured in other countries so it is likely that $1 billion is a producer revenue figure.

If it is assumed that the retail price of the tape media is twice that of the manufacturer's then $2 billion of retail sales of tape would work out to around 400 petabytes of storage capacity. The markup at retail over manufacturer's prices is likely limited due to competition and the common practice of users purchasing bulk quantities of tape. Further, some moderation is due to the lower price of DAT format.

Low-End Formats

The Imation DC2000, or Travan, quarter-inch tape drive is a low-end product used primarily for the backup of desktop PC's. Their general capacity is in the range of 500 megabytes to 4 gigabytes. In August 2000, a Sony Travan Formatted MiniCartridge capable of holding 4 gb uncompressed was advertised for sale on the Internet for $29.49 each. A comparable 4gb tape cartridge manufactured by Maxell was available for $30.79.

Tandberg SLR (Scalar Linear Recording) is also a backup format for desktops and workstations and typically store 350 megabytes to 4 gigabytes.

4 mm tape drives are the largest segment of the market and use digital audio tape (DAT) format. They are commonly deployed as backups for PC servers. These drives generally provide backup in the range of 5 to 40 gigabytes (uncompressed). This format has an installed base of 7.6 million users.

Mid-Level Formats

8 mm tape drives provide storage in the 14 to 50 gigabyte range. Vendors include Exabyte (Mammoth), Sony (AIT), IBM (Magstar 3570).

DLT: (Digital Linear Tape) produced by Quantum Corporation. mid-range computer backup with 15 to 40 gigabytes of native capacity. There are more than 1.4 million DL Tape drives deployed and there have been approximately 40 million tape cartridges in this format sold. Quantum estimates that by the end of 2000, there will be 1.9 million DLT drives shipped to customers.

LTO Ultrium. A new format from consortium of IBM, Seagate and Hewlett Packard. The specification for the Ultrium format is for 100 gigabytes of native storage.

Enterprise Level Formats

1/2-inch cartridge: The dominant format in the mainframe, enterprise level storage market.

Automated tape libraries - which provide completely automated hands-off storage management, including random tape access, sophisticated robotics, unattended backup, and reduced labor costs - are expected to grow from less than 18,000 units shipped in 1996 to close to 120,000 units by 2002. (Source: Freeman Reports)

There is an industry rule of thumb that suggests a three-to-one ratio of disk capacity over tape be maintained.

Format	3490E	3480	Reel-to-Reel
1996	9.3 million	11.7 million	2.1 million
1997	10.7 million	9.5 million	1.9 million

Source: International Recording Media Association

The retail price in August 2000 of 3590 tape cartridges with 10 gb native capacity was $53.21. Fuji Film DLT Tape cartridges were also available at retail for $51.10 for 10 gb native capacity. Sony DLT tapes were being sold for $49.72 for 10 gb. uncompressed. (http://www.cleansweepsupply.com/pages/skugroup2599.html)

From low-end formats such as Travan, through most popular format DLT, to high-end 3590 format, retail price of roughly $5.00 per gigabyte of native storage capacity on tape seems reasonable estimate. (DAT tapes are the only exception and are a lot cheaper.) Of course, if larger purchases result in substantial discounts, then the revenue assumptions would commensurately be pushed more toward the $1 billion wholesale estimate; therefore, not really affecting the calculation of the price per gigabyte of storage capacity.

According to Computer Technology Review (March 1998) the total storage at a typical Fortune 1000 site is projected to escalate from 10 TB in 1997 to 1 PB by the year 2000. In the next five years, a typical large database system for US Government agencies is expected to accept 5TB per day and archive from 15 to 100 PB.

In 1995 Freeman Associates predicted that the total number of tape libraries would increase from 6,454 in 1994 to about 90,000 by the year 2000.

The estimate of the amount of original data stored on tape will, therefore, focus only on mass storage applications from large-scale scientific applications to heavily transaction oriented business applications. The installed base of IBM mainframe OS390 class computers is estimated by IDC to be around 16,500 in 2000.

The number of tape cartridges required to backup small sized computer disks is relatively few and will never substantially exceed the capacity necessary to backup the entire hard disk or disk array. Typical

backup storage strategy is to store the entire file system once and then do incremental updates of any changes made, reducing the amount of storage necessary to keep a current copy of the entire file system at hand.

In large scale tape libraries, there may be thousands or even tens of thousands of magnetic tapes providing primary storage of the application data. The scale of storage requirements is growing rapidly as new facilities, such as the Large Scale Hadron Collider, are built and start performing experiments. Large scale databases are also becoming more common as corporations make increasing efforts to comprehensively track consumer transactions.

The number of households conducting banking transactions may reach 32 million by 2003. The cost of an Internet banking transaction is an estimated 1 cent, compared with $1.14 per transaction by teller, 55 cents by phone, 29 cents by ATM and 2 cents by proprietary computer system. (Source: "Banking on the Internet")

The importance of massive scale databases in general commercial arenas is exemplified by the experience of Wal-Mart, a leader in so-called "data mining" technology and the owner of one of the largest privately held data sets. The U.S. Department of Commerce in its July 2000 report "Digital Economy 2000" points out that "over a three-year period, Wal-Mart achieved a 47 percent increase in sales on only a 7 percent increase in inventories by using a relational database system running on massively parallel computers. The system allows vendors to access almost realtime information on sales and customer transactions and handles 120,000 queries each week from 7,000 suppliers."

Digital Data Creation

Computers for the most part may not greatly contribute to the production of new and original data, but the great exception is in scientific explorations where huge data sets are commonplace and where new discoveries rely on computing and storage.

High Energy Physics

The Large Hadron Collider is being built at CERN in Switzerland, it is expected to be conducting production experiments in around 2005. It is expected to generate approximately 20 petabytes of data per experiment at rates of 100-1500 megabytes per second. Currently experiments in high energy physics generate data at the rate of 35 megabytes per second and many hundred terabytes per experiment. Obviously, this is all original data.

Source: Shiers, Jamie, "Massive-Scale Data Management using Standards-Based Solutions" IEEE 16th Symposium on Mass Storage Systems.

The BaBar experiment at SLAC will generate approximately 200TB/year of data at a rate of 10MB/sec for 10 years.

Los Alamos National Laboratory estimated total storage capacity in its open storage system at 243 terabytes and in its secure system at 2.31 petabytes as of 1998. It is also anticipated that storage capacity will grow to 5 petabytes in 2001.

GeoScience

The majority of data held and administered by the National Oceanic and Atmospheric Administration are held at three national data centers: the National Climatic Data Center at Asheville, NC; the National Oceanographic Data Center at Silver Spring, MD; and the National Geophysical Data Center at Boulder, CO. The climatic data is by far the largest of the three collections, holding approximately 640 terabytes on 350,000 magnetic tapes. The geophysical and oceanographic data total a combined 12 terabytes on 14,500 tapes.

The NASA Center for Computational Sciences in Greenbelt MD has 27,692 tapes holding data as of August 2000. This Center is using 3590 and 9840 tapes which hold 20 gb per tape uncompressed. This Center also automatically makes duplicate tapes for all new data generated. As of August 2000, this storage facility holds 92.5 terabytes of unique data and over 162 terabytes counting the duplicate data. New data is received at the rate of approximately 200-300 gigabytes per month.

The University of Tokyo stores satellite images in an environmental digital library of approximately 6

terabytes, approximately 60,000 images that average 100 megabytes in size.

The San Diego SuperComputer Center stored as of August, 1998 approximately 65 terabytes. The SDSC as of that time held the data on approximately 11,000 tapes. As of mid-2000, the SDSC is storing 1.5e+15 bytes.

The National Center for Atmospheric Research (NCAR) in 1996 had about 68 TB and was growing at the rate of 1.5 TB per month.

Analog Storage Tape

The first uses of magnetic media for data storage occurred about fifty years ago with the development of magnetic tape. A number of formats have evolved over the decades but today by far the most prevalent are the cassette tape used for the mass market distribution of prerecorded music. The other important use of magnetic tape for the storage of analog signals is videotape recordings.

Analog Audio Tape

The distribtution of prerecorded music is one of the most common uses of magnetic tape. The sales of music in this format, however, are now much smaller than they have been historically and are generally expected to continue to decline as digital media become more prevalent and convenient.

United States										
U.S. Sales	**1982**	**1985**	**1990**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**
Cassettes(mil.)	182.3	339.1	442.2	339.5	345.4	272.6	225.3	172.6	158.5	123.6
Cassette Singles	x	x	9.2	11	11.2	12.6	16.9	18.6	27.2	14.2

Source: Recording Industry Association of America

Although there has been a dramatic decline in the overall sales of cassette tapes due to the availability of music on alternative formats, there has been a booming market for books-on-tape. According to the Audio Publishers Association, this market is estimated to have grown 100% since 1990. The reasons for this are that cassettes can run 40 minutes longer than CDs, they have a built-in "bookmark", and they are frequently listened to in the car, and 75% of cars are manufactured with only an AM/FM cassette radio.

The IFPI reports that for the entire world prerecorded music sales on cassette tape were down by 11% in 1998 to 1.2 billion due to depressed sales in Asia.

Blank Audio Tape

The U.S. shipments of blank audio tape dropped dramatically during the 1990's.

Manufacturers' U.S. Shipments of Blank Audio Cassettes		
Year	Dollars (000s Omitted)	Unit (000s Omitted)

1984	$268,287	243,061
1985	$286,865	295,313
1986	$336,179	368,488
1987	$363,336	387,518
1988	$369,550	396,587
1989	$397,734	429,963
1990	$387,895	437,840
1991	$367,716	436,659
1992	$369,769	436,739
1993	$353,022	437,783
1994	$338,428	438,949
1995	$301,316	415,028
1996	$247,442	330,353
1997	$215,576	296,151

Source: International Recording Media Association

Worldwide shipments of blank audio tape are expected to decline in 2000 to 921 million units from 971 million in 1999, with an anticipated market for 771 million cassette tapes by 2003. Source: Consumer MultiMedia Report, Dec. 27, 1999.

Analog Video Tape

Prerecorded VideoTapes (VHS Format) (World)

1997 - 1.666 billion
1998 - 1.719 billion
1999 - 1.748 billion
2000 - 1.664 billion
2001 - 1,561 billion

Source: International Recording Media Association

The main use of the blank video tape is the consumer's use to record televison programs. It is anticipated that there would be a large drop in the sales of this tape if pay-per-view television shows carried copy protection. It is estimated that a very large share of the users of video cassette recorders do so for time shifting of viewing programs.

Blank VideoTapes (VHS Format T-120 equivalent units) (World)

1997 - 1,485 million
1998 - 1,446 million
1999 - 1,463 million
2000 - 1400 million
2001 - 1,275 million

Source: International Recording Media Association

Another view of the blank video market came from British research firm Understanding & Solutions. Their prediction was for 1.147 billion blank videocassettes in 1999 compared to 1.146 in 1998.

The stock of videotape in 1997 was estimated at about 4.6 billion by Richard Kelly, Cambridge Associates.It is not clear whether this is referring to all videos, including those sold blank, or just prerecorded. Further, it is not specified whether this estimate is solely for the United States or includes the whole world. (Feb. 1997 Newsletter, IRMA) We have taken this estimate into account but not used it directly because it seems that the flow of new videotapes worldwide each year would yield a considerably higher figure, even assuming a lot of videotapes may be viewed as disposable after a few years. We have instead estimated a world stock of videotape of all format at around 10 billion.

Conversion

Audio Conversion Issues

In translating the vast quantity of audio information available on cassette tape into its digital equivalent, we have chosen to use the CD format, linear PCM audio at a 16-bit word length and 44.1kHz sample rate. Although, professional recording studios use a sampling rate of 96kHz, the vast majority of tape recorded audio material is music for consumer use and the CD format is the digital format of choice for this application. The amount of data generated by this format is easily calculated. There are 44,100 16-bit samples taken each second for two tracks. Thus,1.4 million bits per second and 5.08 gigabits per hour are generated. The conversion to bytes yields 605 mBs per hour. (1 mByte = 1,048,576 bytes). This data is not compressed and yields a reasonable representation of music for most people.

Video Conversion Issues

In making assumptions about the size of analog videotape stores we have chosen to make conversions assuming the use of MPEG-2 video compression standard. In the case of videotape, the use of this conversion factor is seen as appropriate because it was designed as a generic format for digital multimedia and includes coding schema for both video and audio.

In the case of video, the massive amount of data generated requires that for any practical purpose some compression scheme must be used. MPEG-2 is now the international standard for video storage. Compression is achieved in two ways: spatial compression and temporal compression. The spatial compression is achieved by reducing the number of bits used to represent a single frame. Temporal compression, where the bulk of the savings come, attempts to encode only the bits that represent the portions of a frame that have changed from the previous frame.

The actual amount of compression that can be achieved with MPEG-2 varies quite a bit, we have assumed that 2 gigabytes is adequate to represent 1 hour of high-fidelity audio and high-definition video data.

References and Resources

Gibson, G.D. (1994): Audio, Film and Video Survey. A report on an international survey of 500 audio, motion picture film and video archives. Library of Congress, Washington DC.

Moving Pictures Expert Group

How Much Information?

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Internet - WWW Details

- "A Cyveillance Study: Sizing the Internet" Summarized
- "The Deep Web: Surfacing Hidden Value" Summarized
- Excel Spreadsheet with further information.

"Sizing the Internet: A Cyveillance Study"
Source: Cyveillance, 10-July-2000

- 2.1 billion unique, publicly accessible web pages, and about 4 billion by early 2001 if the current rate of growth continues;
- 7.3 million uniques ages added per day;
- Average page size: 10,060 bytes;
- Average number of images on page: 14.38 (median);
- Percentage of US vs. international pages: 84.7%/15.37%
- Internet still continues to grow at accelerating rate;

"The Deep Web: Surfacing Hidden Value"
Source: BrightPlanet LLC, July 2000

- The deep Web contains 7,500 tarabytes of information, compared to 19 terabytes of information in the surface Web;
- The deep Web contains nearly 550 billion individual documents compared to the 1 billion of the surface Web;
- 60 of the largest web sites contain about 750 terabytes of information;
- More than half of the deep Web content resides in topic specific databases;
- Average page size on the surface Web is 18.7 kbytes, while on the deep Web - 13.7 kbytes, and median for the deep Web is 19.7 kbytes;
- Average deep Web site has a Web-expressed (HTML included basis) database size of 74.4 megabytes, and a median of 169 kbytes;

More Information

Excel Spreadsheet: rawdata.xls
HTML Created from Excel Spreadsheet: click here. [202K]

About the Project
Executive Summary
Print
Film
Optical
Magnetic
Internet
Broadcast
Phone
Mail
Acknowledgments
Site Map

Internet - Email Details

"E-Mail Growth Hogs Enterprise Resources"
Source: Network World Magazine, 31-Jan-2000

- Study of corporate email usage, citing David Ferris, president of Ferris Research;
- Average number of messages received by end users is expected to jump 81% to 34 per day by the beginning of 2001;
- Average size of a message is expected to increase 192% to 286 kbytes by the beginning of 2001 [with growth attributable to attachments];
- There are nearly 170 million corporate email boxes worldwide, more than three times the number of boxes five years ago, according to Eric Arnum, editor of "Messaging Online";
- There are approximately 440 million corporate and personal mailboxes worldwide

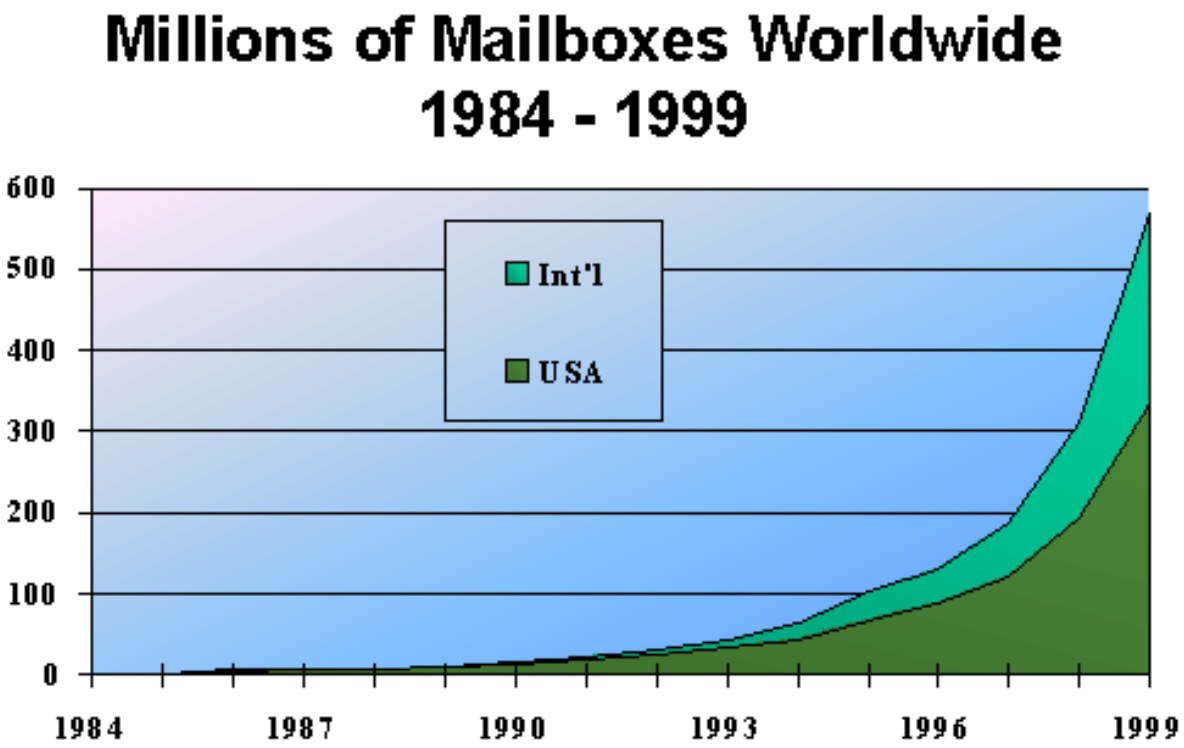
---

"AOL Per-User Email Figures Climb 60 Percent in 1999"
Source: "Messaging Online," 4-Feb-2000

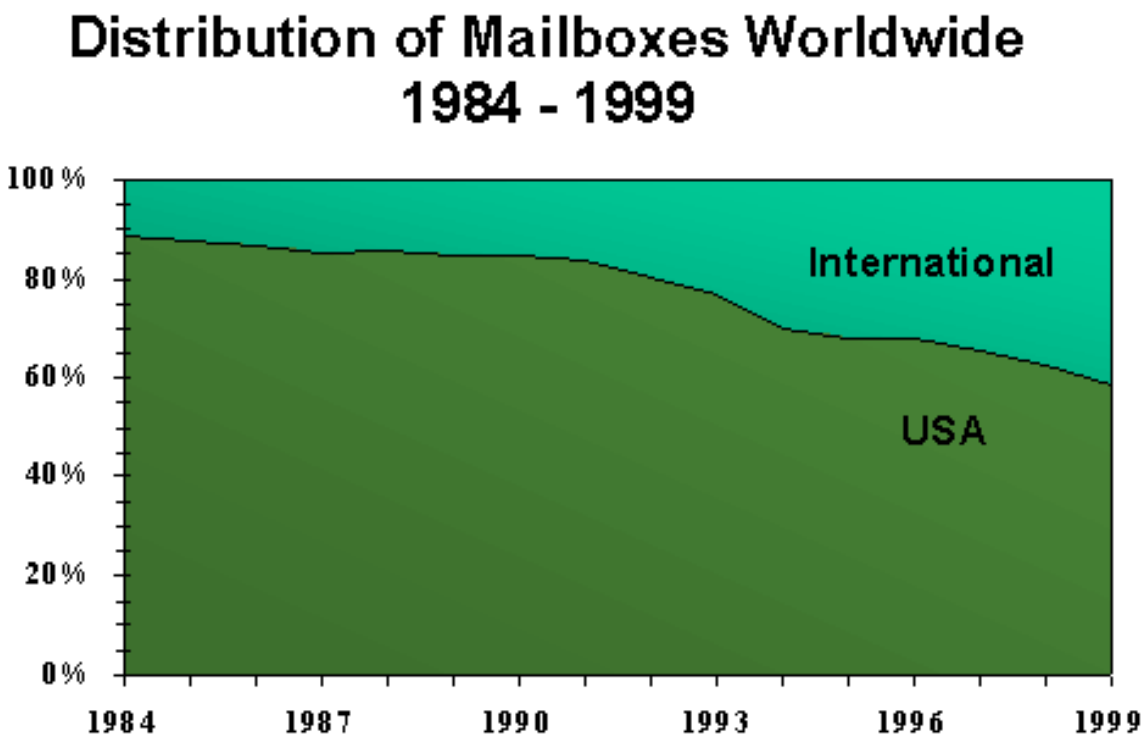
- 3.5 messages per AOL user per day in 1998, and 5.6 in 1999;
- 110 messages sent in 1999, up from 50 million in 1998;
- 20.5 million users at the end of 1999
- Email usage per person increased 60 percent in 1999;
- "If you believe every person in the U.S. has an email account (and it's beginnign to seem that way), then you are talking 1.54 billion messages per day, or 560 billion messages per year. If you believe half the population has email, then your numbers are 770 million messages per day or 280 billion messages per year. Adjust those numbers to reflect heavier usage by the workforce email users and lighter usage by Webmail and ISP users, and you possibly could come up with a trillion messages per year."
- In 1999, the U.S. Postal Service delivered over 200 billion pieces of mail, so email volume now outpaces postal mail volumes;

"Messaging Today: Worldwide Trends"
Source: "Messaging Online," 14-Mar-2000

- The total number of electronic mailboxes in the world has soared 83.5 percent in the past year to 569,171,660 mailboxes;
- In the U.S., the number of mailboxes has jumped 73 percent to 333.5 million mailboxes since the end of 1998. In the rest of the world, the total number of mailboxes has grown 101 percent to 235.6 million mailboxes in the past year;
- In the U.S., the average corporate email user has around 1.5 mailboxes, and the average household using email has about 4 mailboxes.
- There are about 89 million Americans using email at work and roughly 50 million households using email
- There are probably 110 million Americans using email at home or at work, 40 percent of the population.
- There are fewer than one billion televisions in the world, fewer than 800 million phone lines, and 569.2 million mailboxes;
- This paper contains other interesting information and graphs. Here are two samples::



Source: Messaging Online



Source: Messaging Online

24/7 Media: Email Facts
Source: 24/7 Media

- Opt-in email volume will jump 52.3% to 61.1 billion by year-end 2000, and reach 240 billion messages by 2003.
- There were 78 million active e-mail users, aged 13 and older, in the U.S. at year-end 1999, accounting for 35% of the total U.S. population of adults and teens (13+)
- By year-end 2002, there will be 135 million e-mail users, representing 59% of the overall U.S. population of adults and teens. (Source: Emarketer)
- There were 409 million e-mail boxes worldwide in 1999, up from 234 million a year earlier. In 1999, an estimated 20% of all e-mail received in the U.S. was commercial, split almost evenly between spam and permission e-mail (Source: Emarketer)

Nov. 92 - Nov. 94 - Messages Per Month
Source: Internet Society

Month	# of Messages
Nov. 92	279,060,000
Dec. 92	262,320,000
Jan. 93	286,920,000
Feb. 93	317,616,667
Mar. 93	363,180,000
Apr. 93	352,410,000
May 93	410,700,000

Jun. 93	371,700,000
Jul. 93	383,340,000
Aug. 93	397,860,000
Sep. 93	421,320,000
Oct. 93	479,853,333
Nov. 93	508,980,000
Dec. 93	469,370,000
Jan. 94	516,540,000
Feb. 94	527,170,000
Mar. 94	636,140,000
Apr. 94	695,220,000
May 94	731,220,000
Jun. 94	747,960,000
Jul. 94	765,840,000
Aug. 94	803,100,000
Sep. 94	899,400,000
Oct. 94	1,033,920,000
Nov. 94	1,007,590,000

Junk Email

- "Is the Junk Mail Problem Disappearing?" contains different numbers and time-series analysis of the junk email.

Other Information

- Mail.com has 14.4 million e-mail boxes, which require 27 terabytes of storage.
  Source: Network World, Network Storage White Paper
- 300 million e-mails are sent per day in the US.
  Source: Thomas Staffing of California
- Today's users send about 15 messages a day and receive about 20 a day, on average. This colume is expected to grow by some 60% and 80% respectively over the next year.
  Source: Ferris Research
- Electronic mailboxes will more than double in the next two years to an estimated 112 million in 1999.
  Source: Electronic Mail & Messaging Systems
- UC Berkeley's average message size in August, 2000 was 18,517 bytes, with median size of 1863 bytes.

Final Thoughts

Based on the information above, **500 to 600 billion email messages in 2000** seems to be a reasonable non-inflated estimate [though it might be lower than the actual number]. While the average size of email message is hard to tell, we can predict the volume of storage needed from the data above. For example, Mail.com uses 27 terabytes of storage space for its 14.4 million email

boxes, and the number of email boxes can be estimated to be between 450 and 500 million. With the average of **475 million email boxes** in 2000, the volume of storage needed can be estimated at **900 terabytes**, which is lower than the number obtained by multiplying the number of messages by their [probable] average size. The explanation may be that this is the amount of messages stored, not the measure of flow.

Internet Growth Data

Source: Hobbes' Internet Timeline
URL:
http://info.isoc.org/guest/zakon/Internet/History/HIT.html

Date	Hosts	Date	Hosts	Networks	Domains
Dec-69	4	Jul-89	130,000	650	3,900
Jun-70	9	Oct-89	159,000	837	
Oct-70	11	Oct-90	313,000	2,063	9,300
Dec-70	13	Jan-91	376,000	2,338	
Apr-71	23	Jul-91	535,000	3,086	16,000
Oct-72	31	Oct-91	617,000	3,556	18,000
Jan-73	35	Jan-92	727,000	4,526	
Jun-74	62	Apr-92	890,000	5,291	20,000
Mar-77	111	Jul-92	992,000	6,569	16,300
Dec-79	188	Oct-92	1,136,000	7,505	18,100
Aug-81	213	Jan-93	1,313,000	8,258	21,000
May-82	235	Apr-93	1,486,000	9,722	22,000
Aug-83	562	Jul-93	1,776,000	13,767	26,000
Oct-84	1,024	Oct-93	2,056,000	16,533	28,000
Oct-85	1,961	Jan-94	2,217,000	20,539	30,000
Feb-86	2,308	Jul-94	3,212,000	25,210	46,000
Nov-86	5,089	Oct-94	3,864,000	37,022	56,000
Dec-87	28,174	Jan-95	4,852,000	39,410	71,000
Jul-88	33,000	Jul-95	6,642,000	61,538	120,000
Oct-88	56,000	Jan-96	9,472,000	93,671	240,000
Jan-89	80,000	Jul-96	12,881,000	134,365	488,000
		Jan-97	16,146,000		828,000
		Jul-97	19,540,000		1,301,000

Notes: Hosts = a computer system with registered IP address;
Domain = registered domain name (with name server record)

A more accurate survey mechanism was produced in 1999, new and corrected numbers are shown below:

Date	Hosts	Date	Hosts	Date	Hosts
Jan-95	5,846,000	Jan-97	21,819,000	Jan-99	43,230,000
Jul-95	8,200,000	Jul-97	26,053,000	Jul-99	56,218,000
Jan-96	14,352,000	Jan-98	29,670,000	Jan-00	72,398,092
Jul-96	16,729,000	Jul-98	36,739,000		

Note: the original source for the data above is: Network Wizards Internet Domain Survey

URL: ftp://ftp.nw.com/pub/zone/WWW-9807/top.html

This statistics is also reproduced at the following locations:
1. Internet Statistics: Growth and Usage of the Web and the Internet
URL: http://www.mit.edu/people/mkgray/net/
2. Internet Software Consortsium: Internet Domain Survey
URL: http://www.isc.org/ds/

Source: The NSFNET Backbone Project,
1987-1995
URL:
ftp://nic.merit.edu/statistics/nsfnet/index.html

Date	Networks	Date	Networks	Date	Networks
Jul-88	217	Nov-90	2125	Mar-93	10498
Aug-88	241	Dec-90	2190	Apr-93	11252
Sep-88	292	Jan-91	2338	May-93	12349
Oct-88	305	Feb-91	2417	Jun-93	13170
Nov-88	334	Mar-91	2501	Jul-93	14121
Dec-88	346	Apr-91	2622	Aug-93	15160
Jan-89	384	May-91	2763	Sep-93	16696
Feb-89	410	Jun-91	2982	Oct-93	17979
Mar-89	467	Jul-91	3086	Nov-93	19664
Apr-89	516	Aug-91	3258	Dec-93	21430
May-89	564	Sep-91	3389	Jan-94	23494
Jun-89	603	Oct-91	3556	Feb-94	25706
Jul-89	650	Nov-91	3751	Mar-94	28578
Aug-89	745	Dec-91	4305	Apr-94	30626
Sep-89	809	Jan-92	4526	May-94	32370
Oct-89	837	Feb-92	4740	Jun-94	34051
Nov-89	897	Mar-92	4976	Jul-94	36153
Dec-89	927	Apr-92	5291	Aug-94	38307
Jan-90	1233	May-92	5515	Sep-94	39977
Feb-90	1290	Jun-92	5739	Oct-94	41520
Mar-90	1356	Jul-92	6031	Nov-94	42883
Apr-90	1525	Aug-92	6385	Dec-94	44689
May-90	1580	Sep-92	6640	Jan-95	46318
Jun-90	1639	Oct-92	7354	Feb-95	48514
Jul-90	1727	Nov-92	7854	Mar-95	50344
Aug-90	1894	Dec-92	8561	Apr-95	50766
Sep-90	1988	Jan-93	9117		
Oct-90	2063	Feb-93	9604		

Date	Hosts	Date	Domains
Aug-81	213	Jul-88	900

Aug-83	562	Jul-89	3,900
Oct-85	1,961	Oct-90	9,300
Dec-87	28,174	Jul-91	16,000
Oct-89	159,000	Oct-92	18,100
Oct-90	313,000	Oct-93	28,000
Oct-91	617,000	Oct-94	56,000
Oct-92	1,136,000	Jul-95	120,000
Oct-93	2,056,000	Jul-96	488,000
Oct-94	3,864,000	Jul-97	1,301,000
Jul-95	6,642,000		
Jul-96	12,881,000		
Jul-97	19,540,000		

Source: NetSizer
URL: http://www.netsizer.com/

Estimate of the number of web hosts (As of Jul 21, 2000): 86.437 millions

Source: Domain Stats
URL: http://www.domainstats.com/

Total domains registered worldwide (As of Jul 21, 2000): 17,804,717

Source: Center for Next Generation Internet
URL:
http://www.ngi.org/trends.htm

Annual hosts growth rate: 63%

Predictions:	100 million hosts by the end of 2000
	1 billion hosts by 2005
Expansion:	69 new hosts per minute
	23 new domains per minute

Contains detailed statistics by country

WWW Growth Data

Source: Hobbes' Internet Timeline
URL:
http://info.isoc.org/guest/zakon/Internet/History/HIT.html

Date	Sites	Date	Sites	Date	Sites
Jun-93	130	May-97	1,044,163	Dec-98	3,689,227
Sep-93	204	Jun-97	1,117,255	Jan-99	4,062,280
Oct-93	228	Jul-97	1,203,096	Feb-99	4,301,512
Dec-93	623	Aug-97	1,269,800	Mar-99	4,389,131
Jun-94	2,738	Sep-97	1,364,714	Apr-99	5,040,663
Dec-94	10,022	Oct-97	1,466,906	May-99	5,414,325
Jun-95	23,500	Nov-97	1,553,998	Jun-99	6,177,453
Jan-96	100,000	Dec-97	1,681,868	Jul-99	6,598,697
Jun-96	252,000	Jan-98	1,834,710	Aug-99	7,078,194
Jul-96	299,403	Feb-98	1,920,933	Sep-99	7,370,929
Aug-96	342,081	Mar-98	2,084,473	Oct-99	8,115,828
Sep-96	397,281	Apr-98	2,215,195	Nov-99	8,844,573
Oct-96	462,047	May-98	2,308,502	Dec-99	9,560,866
Nov-96	525,906	Jun-98	2,410,067	Jan-00	9,950,491
Dec-96	603,367	Jul-98	2,594,622	Feb-00	11,161,811
Jan-97	646,162	Aug-98	2,807,588	Mar-00	13,106,190
Feb-97	739,688	Sep-98	3,156,324	Apr-00	14,322,950
Mar-97	883,149	Oct-98	3,358,969	May-00	15,049,382
Apr-97	1,002,512	Nov-98	3,518,158	Jun-00	17,119,262

Source: article "Accessibility of Information on the Web"
URL: n/a, only a printed copy is available

Note: all numbers below refer to publicly indexable web pages; publicly indexable web pages exclude pages that are not normally considered for indexing by web search engines, such as pages with authorization requirements(including firewalls), pages excluded from indexing using the robots exclusion standard, dynamic pages, etc;

December, 1997:	at least 320 million pages;
February, 1999:	2.8 million servers on the publicly indexable web; 289 average pages per server; 800 million publicly indexable web pages; 18.7 kilobytes is the mean size of a page;

	3.9 kbytes is the median size of a page;
February, 1999, continued:	7.3 - kbytes average size of the textual content [after removing HTML tags, comments, and extra white space]; 0.98 kbytes - median size of the textual content; 15 terabytes of pages is the amount of data on the publicly indexable web; 6 terabytes is the amount of text data; 62.8 images per web server; 15.2 kbytes - average image size; 5.5 kbytes - median image size; 180 million images on the publicly indexable web; 3 terabytes - total amount of image data;

Note: the distribution of pages on web servers is extremely skewed, following a universal power law; many sites have few pages, and a few sites have vast numbers of pages, which limits the accuracy of the estimates above; the true value could be higher because of very rare sites that have millions of pages (for example, GeoCities reportedly has 34 million pages), or because some sites could not be crawled completely because of errors;

Source: "Size of the Web: A Dynamic Essay for a Dynamic Medium"
URL:
http://censorware.org/web_size/

As of 7/5/2000, the Web has roughly:	2,170,000,000 pages; 40,800,000,000,000 bytes of text; 489,000,000 images; 8,160,000,000,000 bytes of image data.
In the last 24 hours, the Web has added:	4,420,000 new pages; 82,800,000,000 new bytes of text; 994,000 new images; and 16,600,000,000 new bytes of image data. 49,400,000 pages changed; and 11,100,000 images changed.
Average lifespan of the Web page:	44 days;

Source: "The Truth About the Web Crawling Towards Eternity"
URL:
www.alexa.com/company/internet_stats.html

Note: Published in Web Techniques Magazine, May 1997,
Issue 5

How many Web sites are there?

One million Web-site names are in common usage.
There are about 450,000 unique host machines.
If you request the top page from these 450,000, about 300,000 will return one within reasonable time. The rest appear to be intermittent or archaic.
About 95 percent of the 300,000 servers are "up" at any given time.

How big is the Web?
Note: "We estimate there are 80 million HTML pages on the public Web as of January 1997. The figure is fuzzy because some sites are entirely dynamic (a database generates pages in response to clicks or queries). The typical Web page has 15 links (HREFs) to other pages or objects and five sourced objects (SRC), such as sounds or images." Moreover:

The typical HTML page is 5
kbytes;
The typical image (GIF or JPEG) is 12 kbytes;
The average object served via HTTP is 15
kbytes;
The typical Web site is about 20 percent HTML, 80 percent images, sounds, and executables (by size in bytes);

"The upshot of this data is that it takes about 400 GB to store the text of a snapshot of the public Web and about 2000 GB (2 TB) to store nontext files."

How big are individual Web sites?

The median size for a Web site is about 300 pages; only 50 sites have more than 30,000 pages.

About 5 percent of all servers have a robot.txt file (for governing how crawlers visit).

About 1 percent of all servers have a sitelist.txt file (to aid site mapping and robot revisiting).

How fast is the Web growing?

The size of the Web is doubling yearly, but this statistic is losing its meaning because of the growth of dynamic sites.
The typical Web page is only about two months old.
Dynamic sites are becoming a significant presence; JavaScript is widespread, Java much less so, but growing.

Note: "Facts above are mostly based on data gathered by Internet Archive, but augmented with some stats from Larry Page of Stanford University and public documents from the Web."
Author: Z Smith

Source: Measuring the Web
URL: http://www5conf.inria.fr/fich_html/papers/P9/Overview.html

What is the "average page" like? (as of May, 1996)

Mean size: 6518
Median size: 2021
SD: 31678

Source: "The Web: Growing by 2 Million Pages a Day"
URL: www.thestandard.com/article/display/0,1151,12329,00.html
Author: David Lake
Published in: The Industry Standard Magazine, February 28, 2000

Metric	1999 Total	Estimated Growth per Day 1998-1999
Web pages	1,500,000,000	1,917,808
Hosts	72,398,092	79,913
Domain Names	8,100,000	12,981
Unique Web Sites	3,649,000	4,422

Cumulative Domain Name Registrations in the .com, .net, .org Domains (Millions)

Year	Existing	New
1993	0	0.01
1994	0.01	0.03
1995	0.04	0.16
1996	0.2	0.4
1997	0.6	0.9
1998	1.6	1.8
1999	3.4	4.7

Web Site Growth Trend (Millions)

Year	Online Computer Library Center	Alexa Internet
1997	1.2	n/a
1998	2	2.5
1999	3.6	3.4
2000	n/a	n/a
2001	n/a	10

Number of Hosts in the Domain Name System (Millions)

Year	Number of Hosts
1995	5.8
1996	14.4
1997	21.8
1998	29.7
1999	43.2
2000	72.4

Source: Inktomi Corp: Web Surpasses One Billion Documents
URL:
http://www.inktomi.com/new/press/billion.html

Jan 18, 2000: Inktomi announces that WWW surpassed 1 billion pages; This figure also coincides with Internet Archive's (www.archive.org) estimate. Internet Archive also says that these 1 billion pages correspond to 13.8 terabytes of text-only data, with a rate of growth 2 terabytes per month;

Source: Online Computer Library Center June 1999 Web Statistics
URL:
http://www.oclc.org/oclc/research/projects/webstats/statistics.htm

Number of IP addresses in 32-bit address space:	4,294,967,296
Number of IP addresses in the 0,1% random sample:	4,294,967
Number of Web Sites:	4,882,000 (+/- 3%)
Number of Unique Web Sites:	3,649,000 (+/- 3%)
Number of Public Web Sites:	2,229,000 (+/- 4%)
Number of Private Web Sites:	389,000 (+/- 10%)
Number of Provisional Web Sites:	1,031,000 (+/- 6%)
Number of Web Pages:	288,221,000 (+/- 35%)
Number of Files:	500,491,000

Web Growth

	1997	1998	1999
Web Sites:	1,570,000	2,851,000	4,882,000
Unique Sites:	1,230,000	2,035,000	3,649,000
Unique Public Sites:	800,000	1,457,000	2,229,000
% Change:	97 to 98	98 to 99	97 to 99

Web Sites:	82	71	211
Unique Sites:	65	79	197
Unique Public Sites:	82	53	179

Web Volatility: 44%
(IP addresses identifying a Web site in 1998 that no longer identify a Web site in 1999)

Explanation of the Statistics

The 32-bit Internet Protocol address space consists of 4,294,967,296 unique IP addresses. A 0.1% random sample (without replacement) was taken from this address space, resulting in 4,294,967 unique IP addresses. An attempt was made to connect to each of the sample addresses on Port 80.

Web Site
Identified by an IP address that returns a response code of 200 and a Web page in reply to an HTTP request for the home page.

Unique Web Sites
It is not uncommon for the same Web site to be duplicated at multiple IP addresses (e.g., for server load distribution purposes). To ensure that each unique Web site has the same probability of being selected for the sample, the following rule was followed: if a site is located at multiple IP addresses, the site is retained in the sample only if the numerically lowest IP address is in the sample. Three diagnostic tests were developed to assist in identifying sites with multiple IP addresses. A description of these tests is available in the paper A Methodology for Sampling the World Wide Web.

Public Web Sites
A public Web site offers content that, from a general perspective, is meaningful and non-trivial, and is freely accessible without fee payment or prior authorization.

Private Web Sites
A private Web site requires payment or prior authorization to access its content; typically, a private site will not permit free access beyond its home page.

Provisional Web Sites
A provisional Web site is in a transitory or unfinished state, and/or offers only content that from a general perspective, is meaningless or trivial.

Web Pages
According to the W3C Working Draft "Web Terminology & Definitions Sheet" (May 24, 1999), a Web page is defined as:
A collection of information, consisting of one or more Web resources, intended to be rendered simultaneously, and identified by a single URI. More specifically, a Web page consists of a Web resource with zero, one, or more embedded Web resources intended to be rendered as a single unit, and referred to by the URI of the one Web resource which is not embedded.
The key point in this definition is that a Web page often is a composite object, consisting of multiple Web resources: e.g., text, images, applets, etc. The Web page is the single entity representing the combination of

these resources.

For each unique public Web site, a harvesting agent was used to download and store all text-based files internal to the site. Harvested files had one of the following extensions:

.asc	.asp	.dhtm	.dhtml
.ephtm	.ephtml	.htm	.html
.jsp	.mhtm	.mhtml	.php
.php3	.phtm	.phtml	.shtm
.shtml	.text	.tx8	.txt

In addition, files with no extension and those identified implicitly in the URL - i.e., URLs whose path ends with a directory, and the server automatically loads the correct file when the URL is accessed - were also harvested.
Duplicate pages on the same site were eliminated. Elimination of duplicate pages was conducted strictly on an intra-site basis, and was not extended across sites.
Once the de-duping process was completed, the remaining files were processed by software that counted the number of Web pages present on each Web site. For a given Web site, Web pages fell into one of two categories:
1. the harvested pages
2. references in the harvested pages to other Web pages internal to the Web site, through the HTML <a href="..."> convention.

The counting algorithm was configured so that a Web page with frames is counted only once, rather than multiple times for each constituent part of the frame.

Number of Files
An estimate of the number of files located at public Web sites can also be obtained from the harvest data. For each public Web site, files were identified as one of the following:
1. harvested files
2. references in the harvested files to other files internal to the Web site, through the HTML <a href="..."> convention.
3. references in the harvested files to images internal to the Web site, through the HTML <img src="..."> convention.
4. references in the harvested files to applets internal to the Web site, through the HTML <applet code="..." codebase="..."> convention.

Source: Cyveillance's "Sizing the Internet"
URL: http://www.cyveillance.com/resources/Sizing_the_Internet_whitepaper.pdf

Number of unique pages on Internet:	2.1 billion
Unique pages added per day:	7.3 million

Average size of pages:	10,060 bytes
Average number of images on a page:	14.38

Source: Cyveillance Corp.
URL: http://www.cyveillance.com/resources/facts.asp#5

Web grows by 300,000 pages every 5 days (Original source: Lycos)
The amount of Web pages is estimated to grow seven-fold to 7.7 billion by 2002. (Source: IDC)