

How Much Information? 2003

Summary

Exec Summary

Stored Information

Paper | Film | Magnetic | Optical

Information Flows

Broadcast | Telephony | Internet

Wrap-up

Thanks | Printable (PDF)

Executive Summary

[I. Summary of Findings](#)

[II. Method](#)

[III. Stored Information](#)

[A. Paper](#)

[B. Film](#)

[C. Magnetic](#)

[D. Optical](#)

[IV. Information Flows](#)

[A. Broadcasting](#)

[B. Telephony](#)

[C. Internet](#)

[V. Qualifications](#)

[About this Report](#)

[About SIMS](#)

[Printable full report \(PDF\)](#)

878 KB, 103 pages

[Printable Exec Summary
\(PDF\)](#)

108 KB, 14 pages

1. EXECUTIVE SUMMARY

I. Summary of Findings

How much new information is created each year? Newly created information is stored in four *physical media* – print, film, magnetic and optical – and seen or heard in four *information flows through electronic channels* – telephone, radio and TV, and the Internet. This study of information storage and flows analyzes the year 2002 in order to estimate the annual size of the stock of new information recorded in storage media, and heard or seen each year in information flows. Where reliable data was available we have compared the 2002 findings to those of our 2000 study (which used 1999 data) in order to describe a few trends in the growth rate of information.

1. **Print, film, magnetic, and optical storage media produced about 5 exabytes of new information in 2002. Ninety-two percent of the new information was stored on magnetic media, mostly in hard disks.**
 - *How big is five exabytes?* If digitized, the nineteen million books and other print collections in the Library of Congress would contain about ten terabytes of information; five exabytes of information is equivalent in size to the information contained in half a million new libraries the size of the Library of Congress print collections.
 - *Hard disks store most new information.* Ninety-two percent of new information is stored on magnetic media, primarily hard disks. Film represents 7% of the total, paper 0.01%, and optical media 0.002%.

- *The United States produces about 40% of the world's new stored information*, including 33% of the world's new printed information, 30% of the world's new film titles, 40% of the world's information stored on optical media, and about 50% of the information stored on magnetic media.
 - *How much new information per person?* According to the [Population Reference Bureau](#), the world population is 6.3 billion, thus almost 800 MB of recorded information is produced per person each year. It would take about 30 feet of books to store the equivalent of 800 MB of information on paper.
- 2. We estimate that the amount of new information stored on paper, film, magnetic, and optical media has about doubled in the last three years.**
- *Information explosion?* We estimate that new stored information grew about 30% a year between 1999 and 2002.
 - *Paperless society?* The amount of information printed on paper is still increasing, but the vast majority of original information on paper is produced by individuals in office documents and postal mail, not in formally published titles such as books, newspapers and journals.
- 3. Information flows through electronic channels -- telephone, radio, TV, and the Internet -- contained almost 18 exabytes of new information in 2002, three and a half times more than is recorded in storage media. Ninety eight percent of this total is the information sent and received in telephone calls - including both voice and data on both fixed lines and wireless.**
- *Telephone calls* worldwide – on both landlines and mobile phones – contained 17.3 exabytes of new information if stored in digital form; this represents 98% of the total of all information transmitted in electronic information flows, most of it person to person.
 - *Most radio and TV broadcast content is not new information.* About 70 million hours (3,500 terabytes) of the 320 million hours of radio broadcasting is original programming. TV worldwide produces about 31 million hours of original programming (70,000 terabytes) out of 123 million total hours of broadcasting.
 - *The World Wide Web* contains about 170 terabytes of information on its surface; in volume this is seventeen times the size of the Library of Congress print collections.
 - *Instant messaging* generates five billion messages a day (750GB), or 274 Terabytes a year.
 - *Email* generates about 400,000 terabytes of new information each year worldwide.
 - *P2P file exchange on the Internet* is growing rapidly. Seven percent of users provide files for sharing, while 93% of P2P users only download files. The largest files exchanged are video files larger than 100 MB, but the most frequently exchanged files contain music (MP3 files).
 - *How we use information.* Published studies on media use say that the average American adult uses the telephone 16.17 hours a month, listens to radio 90 hours a month, and watches TV 131 hours a month. About 53% of the U.S. population uses the Internet, averaging 25 hours and 25 minutes a month at home, and 74 hours and 26 minutes a month at work – about 13% of the time.

II. Method

In 2000 we conducted a study to estimate how much information is produced every year (see <http://www.sims.berkeley.edu/research/projects/how-much-info/>). We then estimated that in 1999 the world produced between 1 and 2 exabytes of unique information. In Summer 2003 we repeated the study, using 2002 data, in order to begin to identify trends in the production and consumption of information. Some of the 1999 data has been revised in this study because new information sources were identified; our revised estimate is that in 1999 the world produced between 2 and 3 exabytes of new information.

As in 1999, we have estimated the magnitudes of information flows (TV, Radio, Telephone, Internet) that are currently not systematically archived, but may well be in the future. This year we have added two studies of the Internet. We have sampled the World Wide Web, to determine the size of the surface web and to define the source, functions and content of Web pages. And we have studied desktop disk drives, to determine how people consume information on the Internet.

Because information is created and distributed in different media or formats there is no common standard with which to measure the amount of information created each year, thus we have translated the vast array of information formats and media to a single standard – terabytes. Terabytes used as a common standard of measurement of the amount of new information is particularly useful given that most new information is in digital form, and other formats are increasingly giving way to digital form (i.e., digital images replacing film based photographs), or are archived in digital form (i.e., print newspapers also published on the Web).

However, this methodology measures only the volume of information, not the quality of information in a given format or its utility for different purposes, i.e., the relative value of information in an edited book or peer-reviewed journal articles when compared to digital storage of raw data.

Table 1.1: How Big is an Exabyte?

Kilobyte (KB)	<i>1,000 bytes OR 10^3 bytes</i> 2 Kilobytes: A Typewritten page. 100 Kilobytes: A low-resolution photograph.
Megabyte (MB)	<i>1,000,000 bytes OR 10^6 bytes</i> 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. 10 Megabytes: A minute of high-fidelity sound. 100 Megabytes: 1 meter of shelved books. 500 Megabytes: A CD-ROM.
Gigabyte (GB)	<i>1,000,000,000 bytes OR 10^9 bytes</i> 1 Gigabyte: a pickup truck filled with books. 20 Gigabytes: A good collection of the works of Beethoven. 100 Gigabytes: A library floor of academic journals.
Terabyte (TB)	<i>1,000,000,000,000 bytes OR 10^{12} bytes</i> 1 Terabyte: 50000 trees made into paper and printed. 2 Terabytes: An academic research library. 10 Terabytes: The print collections of the U.S. Library of Congress. 400 Terabytes: National Climactic Data Center (NOAA) database.

Petabyte (PB)	1,000,000,000,000,000 bytes OR 10^{15} bytes 1 Petabyte: 3 years of EOS data (2001). 2 Petabytes: All U.S. academic research libraries. 20 Petabytes: Production of hard-disk drives in 1995. 200 Petabytes: All printed material.
Exabyte (EB)	1,000,000,000,000,000,000 bytes OR 10^{18} bytes 2 Exabytes: Total volume of information generated in 1999. 5 Exabytes: All words ever spoken by human beings.

Source: Many of these examples were taken from [Roy Williams "Data Powers of Ten" web page at Caltech.](#)

After production of original content by media type was estimated, the key problem was to identify a common standard of comparison. We have translated the volume of original content into a common standard by figuring the size of analog content in terabytes if it were to be digitized using industry standard practices ('upper bound' estimates). We have then determined how much storage each type would take using industry standards for compression, and defined working assumptions to adjust for duplication of content ('lower bound' estimates). See the [Qualifications](#) section for some of the problems of this methodology.

III. How much new information is recorded every year?

Information is recorded, stored and distributed in four physical media – paper, film, magnetic, and optical. Good data is available for the worldwide production of each storage medium, providing an upper bound for the potential production of original information and copies. There are often good estimates for how much original content is produced in each of these different storage formats, particularly for the advanced economies that produce the most information. Where those data don't exist we have adopted working assumptions to make our estimates; these assumptions are documented in the full report and, as in 2000, we welcome suggestions for improving them.

Table 1.1 summarizes yearly worldwide production of original stored content "circa 2002," because Paper and Film statistics are largely from 2001 while Magnetic and Optical are largely from 2002. Detailed source information and the inferences that were used to produce these calculations are presented in detail in the web pages on [Paper](#), [Film](#), [Magnetic](#), and [Optical](#) accessible from the links at the top of this page.

Table 1.2: Worldwide production of original information, if stored digitally, in terabytes circa 2002. Upper estimates assume information is digitally scanned, lower estimates assume digital content has been compressed.

Storage Medium	2002 Terabytes Upper Estimate	2002 Terabytes Lower Estimate	1999- 2000 Upper Estimate	1999- 2000 Lower Estimate	% Change Upper Estimates
Paper	1,634	327	1,200	240	36%
Film	420,254	76,69	431,690	58,209	-3%
Magnetic	4,999,230	3,416,230	2,779,760	2,073,760	80%
Optical	103	51	81	29	28%
TOTAL:	5,421,221	3,416,281	3,212,731	2,132,238	69%

Source: *How much information 2003*

Summary estimates show that the storage of new information has been growing at a rate of over 30% a year (upper estimate, uncompressed). There has been dramatic growth in storage of new information over the past two years in every storage medium except film. Film-based content – especially photographs – is migrating to digital media, both optical and magnetic.

A. Paper

A tree can produce about 80,500 sheets of paper, thus it requires about 786 million trees to produce the world's annual paper supply. The UNESCO Statistical Handbook for 1999 estimates that paper production provides 1,510 sheets of paper per inhabitant of the world on average. But paper consumption is not equal; annually each of the inhabitants of North America consumes 11,916 sheets of paper (24 reams), and inhabitants of the European Union consume 7,280 sheets of paper (15 reams). At least half of this paper is used in printers and copiers to produce office documents.

Table 1.3: Worldwide production of printed original content, if stored digitally in terabytes circa 2002. Upper estimate is scanned; lower estimate is compressed.

Storage Medium	Type of Content	Terabytes/Yr		1999		% Change Upper Estimates
		Upper Estimate	Lower Estimate	Upper Estimate	Lower Estimate	
Paper	Books	39	8	39	8	0
	Newspapers	138.4	27.7	124	25	12%
	Office Documents	1,397.5	279.5	975	195	43%
	Mass market periodicals	52	10	52	10	0
	Journals	6	1.3	9	2	-33%
	Newsletters	0.9	0.2	0.8	0.2	0
	Subtotal	1,633.8	326.7	1,199.8	240.2	36%

Source: *How much information 2003*, Table 2.3

The amount of new original information stored on paper increased 36% between 1999 and 2002.

- The vast majority of this increase is from the creation of office documents -- largely the production of computer printers. Office documents are a larger proportion of print in the U.S. than in the E.U.
- Also noteworthy is the increase in simultaneous publication of printed information in digital format, such as online newspapers and journals.
- There appears to be an increase in newspaper production in developing countries, although this may be a reflection of better statistical reporting.

For details on this data, our sources and calculations see [Paper](#).

Table 1.4: United States production of printed original content, if stored digitally in terabytes circa 2002. Upper estimate is scanned; lower estimate is compressed.

Storage Medium	Type of Content	Terabytes/Yr Upper Estimate	1999 Upper Estimate	% Change Upper Estimates
Paper	Books	5.5	3	83%
	Newspapers	13.5	13	4%
	Office Documents	559	390	43%
	Mass market periodicals	3.5	13	-73%
	Journals	1.6	2	-20%
	Newsletters	0.3	0.2	50%
	U.S. Mail	6,230	5,940	4.8%
	Subtotal	6,813	6,361	7.1%

Source: *How much information 2003, Table 2.5*

The U.S. produces 35% of the world's new printed information each year and 40% of the world's card and letter postal volume. About half of all postal mail in the United States is currently first class and about half is junk mail. If we assume 2 pages per piece of mail, digitized at 15 kilobytes per page, 2002 U.S. mail was about 6.23 petabytes per year. This represents an increase of about one-half of a petabyte over 1999 estimates.

B. Film

Film is a storage medium for analog images that is evolving towards digital images stored on magnetic and optical media.

- There has been a decline in the number of film-based photographs since 1999, and a dramatic growth in the creation of images using digital cameras (see [Magnetic](#)). In 2002 there were 27.5 million digital still cameras purchased worldwide, compared to 63 million analog (film-based) cameras.
- Film-based cinema and TV is beginning to evolve into digital video because of lower editing costs (see the discussion of DVD in [Optical](#)).
- Medical digital imaging technologies are developing rapidly, but lower technology costs have led to continued growth in film based X-rays.
- There has been a growth in the production of new movies and TV, particularly in developing countries. Approximately 370,000 motion pictures were made around the world from 1890 to 2002. If the entire universe of original film and video titles were played continuously the show would continue for 2,108 years.

For details on this data, our sources and calculations see [Film](#).

Table 1.5: Worldwide production of filmed original content, if stored digitally, in terabytes circa 2002.

Storage Medium	Type of Content	Terabytes/Yr Upper Estimate	Terabytes/Yr Lower Estimate	1999 Upper Estimate	1999 Lower Estimate	% Change Upper Estimates
Film	Photographs	375,000	37,500	410,000	41,000	-9%
	Cinema	6,078	12	4,490	9	35%
	Made for TV films	2,531	2,530	N/A	N/A	N/A
	TV series	14,155	14,155	N/A	N/A	N/A
	Direct to video	2,490	2,490	N/A	N/A	N/A
	X-Rays	20,000	20,000	17,200	17,200	16%
	Subtotal		420,254	74,202	431,690	58,209

Source: *How much information 2003*, Table 3.2

C. Magnetic

Worldwide production of new information recorded on magnetic storage media has grown 80% since 1999.

- Analog-based magnetic tape (audio and videotape) has decreased as digital storage has grown.
- Digital tape continues to be an archival storage media for data.
- The decreasing cost and increasing variety of form factors has made hard disk technologies the fastest growing segment of all storage media for information, as was true in our 1999 study. MiniDV, AudioMD and Flash were not included in the 1999 study.

For details on this data, our sources and calculations, see [Magnetic](#).

Table 1.6: Worldwide production of magnetic original content, if stored digitally using standard compression methods, in terabytes circa 2002.

Storage Medium	Type of Content	Terabytes/Yr Upper Estimate	Terabytes/Yr Lower Estimate	1999 Report Upper Estimate	1999 Report Lower Estimate	% Change Upper Estimates
Magnetic	Videotape	1,340,000	1,340,000	1,420,000	1,420,000	-6%
	Audiotape	128,800	128,800	182,000	182,000	-30%
	Digital tape	250,000	250,000	250,000	250,000	0
	MiniDV	1,265,000	1,265,000	N/A	N/A	N/A
	Floppy disc	80	80	70	70	14%
	Zip	350	350	1,690	1,690	-79%

Audio MD	17,000	17,000	N/A	N/A	N/A
Flash	12,000	12,000	N/A	N/A	N/A
Hard Disk	1,986,000	403,000	926,000	220,000	114%
TOTAL	4,999,230	3,416,230	2,779,760	2,073,760	80%

Source: *How much information 2003*

D. Optical

Optical storage media are the medium of choice for the distribution of software, data, cinema and music -- although a small proportion of digital information overall.

- Decline in the production and sale of retail audio CD titles have been offset by the growing popularity of writeable CDs (CD-R and CD-RW).
- DVDs have achieved the fastest market penetration of any recent technology innovation, although largely in the advanced economies.

For details on this data, our sources and calculations see [Optical](#).

Table 1.7: Worldwide production of optical original content, if stored digitally using standard compression methods, in terabytes circa 2002.

Storage Medium	Genre	Terabytes/Yr Upper Estimate	Terabytes/Yr Lower Estimate	1999 Upper Estimate	1999 Lower Estimate	% Change Upper Estimates
Optical	Audio CD	58	6	58	6	0
	CD ROM	1.1	1.1	0.7	0.7	57%
	DVD	43.8	43.8	22	22	99%
	Subtotal	102.9	50.9	80.7	28.7	28%

Source: *How much information 2003, Table 5.2*

The U.S. produces 37% of the world's audio CD titles, 50% of the CD ROM titles, and 40% of the DVD titles.

Table 1.8: United States production of optical original content, if stored digitally using standard compression methods, in terabytes circa 2002.

Storage Medium	Genre	Terabytes/Yr Upper Estimate	Terabytes/Yr Lower Estimate	1999 Upper Estimate	1999 Lower Estimate	% Change Upper Estimates
Optical	Audio CD	22	2	22	2	0
	CD ROM	0.55	0.06	1	0.3	-45%
	DVD	18	18	13	13	38%
	Subtotal	40.55	20.06	36	15.3	13%

Source: *How much information 2003, Table 5.2*

IV. How Large are New Information Flows in Electronic Channels?

How we use information flows

Communication flows through four electronic channels: radio and television broadcasting, telephone calls, and the Internet. Each channel requires access to a form of information technology: radios, television sets, telephones, and computers. Thus like storage media, information flows are distributed unequally around the world.

How large are new information flows?

Information stored on paper, film, optical, and magnetic media totals about 5 exabytes of new information each year; this is less than one third of the new information that is communicated through electronic information flows – telephone, radio and TV, and the Internet – which is about **17.7 exabytes**.

Table 1.9: Summary of electronic flows of new information in 2002 in terabytes.

Medium	2002 Terabytes
Radio	3,488
Television	68,955
Telephone	17,300,000
Internet	532,897
TOTAL	17,905,340

Source: *How much information 2003*

The striking finding here is that most of the total volume of new information flows is derived from the volume of voice telephone traffic, most of which is unique content. The second largest component of information flows is the Internet.

A. Broadcasting

World radio stations produce 320 million hours of radio broadcasting, which would require 16,000 terabytes to store; we estimate 70 million hours are original programming, which would require an annual storage requirement of about 3,500 terabytes. World television stations produce about 123 million hours total programming; we estimate about 31 million hours are original programming, requiring about 70,000 terabytes of storage.

Table 1.10: World – annual production of original broadcast media items – 2003 sources

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

Radio	47,776	70 million hours of original programming	0.05 GB/hour	3,488	3,488
Television	21,264	31 million hours of original programming	1.3 GB - 2.25 GB hour	39,841	68,955
Total:				43,329	72,443

Source: *How much information 2003*, Table 6.1

In the United States, there are 13,261 radio stations producing 19.7 million hours of original programming, or about 987 terabytes of original programming. As of 2002 there were 1686 broadcast TV stations in the United States producing about 14.5 million hours of content a year; about 3.6 million hours are original information, equivalent to between 4,700 and 8,200 terabytes (depending upon the compression standard used).

For details on this data, our sources and calculations see [Broadcast](#).

Table 1.11: United States – Comparison of production of original broadcast media items – 2003 sources

Media Type	Number of Stations	Unique Items per Year	Conversion Factor	Total Terabytes (Annual)	
				Lower Bound	Upper Bound
Radio stations (2002, FCC)					
Commercial - AM	4,811	7.2 million hours	0.05 GB/hour	358 TB	358 TB
Commercial - FM	6,147	9.2 million hours	0.05 GB/hour	458 TB	458 TB
Educational	2,303	3.4 million hours	0.05 GB/hour	171 TB	171 TB
Total	13,261	19.8 million hours	0.05 GB/hour	987 TB	987 TB
Television (2002, FCC)					
Broadcast Stations	1,686	3.1 million hours	1.3 GB - 2.25 GB hour	4,000 TB	6,923 TB
Cable Stations	308	.6 million hours	1.3 GB - 2.25 GB hour	731 TB	1,265 TB
Total	1,994	3.6 million hours	1.3 GB - 2.25 GB hour	4,731 TB	8,188 TB

Total (radio + TV):	5,718 TB	9,175 TB
----------------------------	-----------------	-----------------

Source: *How much information 2003*, Table 6.3

B. Telephone

Table 1.12: The size of world telephone calls in terabytes.

Line calls	15,000,000
Wireless calls	2,300,000
TOTAL	17,300,000

Source: *How much information 2003*

There are 1.1 billion main telephone lines in the world as of 2002; it is estimated that each line carries an average of 3,441 minutes a year, or 3,785 billion minutes. At 64 kilobits/second, it would take 15 exabytes to store this much information - most of it original. There are 190 million main telephone lines in the U.S., each of them used over an hour a day for all types of calls (i.e., mostly local, including modems, faxes, etc). It would take 9.25 exabytes of storage to hold all U.S. calls each year. The number of landline phones in the U.S. has dropped by more than 5 million, as mobile phones have grown to 43% of all U.S. phones. Mobile phones used more than 600 billion minutes in 2002, an equivalent of 2.3 exabytes of storage.

For details on this data, our sources and calculations see [Telephony](#).

C. Internet

Although the Internet is the newest medium for information flows, it is the fastest growing new medium of all time, becoming the information medium of first resort for its users. Note that the Web consists of the surface web (fixed web pages) and what Bright Planet calls the deep web (the database driven websites that create web pages on demand).

Table 1.13: The size of the Internet in terabytes.

Medium	2002 Terabytes
Surface Web	167
Deep Web	91,850
Email (originals)	440,606
Instant messaging	274
TOTAL	532,897

Source: *How much information 2003*

Around the world about 600 million people have access to the Internet, about 30% of them in North America.

Table 1.14: World Distribution of Internet Users (in millions)

Africa	6.31
--------	------

Asia Pacific	187.24
Europe	190.91
Middle East	5.12
Canada and USA	182.67
Latin America	33.35

Source: Nielsen / NetRatings via CyberAtlas

According to Nielsen/NetRatings, the average Internet user spends 11 hours and 24 minutes online per month; the average user in the United States spends more than twice that amount of time online: 25 hours and 25 minutes at home and 743 hours and 26 minutes at work. In the United States, Internet access is used to send email (52%), get news (32%), use a search engine to find information (29%), surf the web (23%), do research for work (19%), check the weather (17%) or send an instant message (14%) (Source: Pew Internet and American Life Project).

For details on this data, our sources and calculations see [Internet](#).

The Web

In 2000 we estimated the volume of information on the public Web at 20 to 50 terabytes; in 2003 we measured the volume of information on the Web at 167 terabytes - at least triple the amount of information. The surface web is about 167 terabytes as of Summer 2003; BrightPlanet estimates the deep web to be 400 to 450 times larger, thus between 66,800 and 91,850 terabytes.

- The median size of HTM/HTML pages was 8 KB, but the mean was 605 KB. About 23% included images and 4% contained movies or animations, and about 20% contained Javascript applications.
- There are about 2.9 million active weblogs ('blogs'), containing about 81 GB of information.

Email

About 31 billion emails are sent daily, on the Internet and elsewhere, a figure which is expected to double by 2006 (source: International Data Corporation (IDC). The average email is about 59 kilobytes in size, thus the annual flow of emails worldwide is 667,585 terabytes.

- Email ranks second behind the telephone as the largest information flow. Email users include 35% of the total U.S. population (source: eMarketer), and accounts for over 35% of time spent on the Internet (source: Forrester Research).
- Sixty percent of workers with email access receive 10 or fewer messages on an average day, 23% receive more than 20, and 6% more than 50. 73% of workers spend an hour or less per day on their email.
- Only two thirds of email traffic is personal, and spam (defined as unsolicited email) is about one-third of today's email traffic, which is projected to increase to 50% four years from now (source: IDC). Therefore we estimate the upper bound of original content in emails as 440,606 terabytes (uncompressed), lower bound as 333,792 terabytes.

Instant Messaging

Nearly 40% of U.S. Internet users at home logged onto one of the instant messaging (IM) networks at least

once in May 2002, while 31% of U.S. business Internet users used IM (source: Nielsen/NetRatings).

Peer to Peer (P2P) File Sharing

A significant new source of storing, creating and exchanging media and data on the Internet is through P2P file sharing networks. KaZaA, the most popular of these applications, has recently reached over 230 million downloads worldwide, with an average of 2 million more per week (source: Download.com). Users on KaZaA share almost 5,000 terabytes of information, over 600 million files and have over 3 million users active on average at any given time (source: KaZaA.com). Looking at a sample of 400,000 users over a 24 hour period we found about 9% of users (38,256) sharing files. We found 1,980,426 files consisting of 14754 GB (14.4 TB) of information. Files ranged in size from 1 Byte to 1.97 GB, with a mean size of 7.6 MB. Using this sample, we were able to describe how P2P users consume information.

- In looking at file sizes, users frequently exchange files larger than 100 MB, and have desktop collections of files larger than 100 MB; the largest in the sample was 32 GB. The largest number of files was 10,583.
- The largest file types are .AVI Files. The range of these in our sample was 82 Bytes to 2 GB, with most files being in the 100-200 MB range (mean of 162 MB).
- The most common files shared by P2P users are MP3 files, music files encoded using MP3 technology. Images (jpg, bmp) are also popular (~10% of the total number of files in the sample) but take up much less space. Sixty percent of the files on users' hard disks were MP3 files, taking up about 30% of the space.

V. Qualifications

We have had to make various working assumptions in order to construct these estimates, and some data sources are contradictory or simply not available, thus our estimates are often rough. 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.

Estimates

Our documentary research methodology is to estimate yearly U.S. and world production of originals and copies for the most common forms of information media - paper, film, magnetic, and optical. The data supporting these estimates is often difficult to find, or does not exist at all, and key questions often cannot be answered because no data is collected (e.g., about third world information production). Estimates are marked with three question marks [???] to signal our caution about their reliability. For those reasons we have documented our sources in these reports and defined the working assumptions we have made in producing these estimates, hoping that our readers will help us to identify better sources and to improve our working assumptions.

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 determine 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

The advantage of using a single measurement standard such as terabytes to compare the volume of information in different formats is obvious. However, unlike paper 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.

About this Report

We view this report as a "living document" and intend to revise it based on comments, corrections, and suggestions. Please send comments to how-much-info@sims.berkeley.edu.

Many thanks to our sponsors. Financial support for this study was provided by:

- Microsoft Research at <http://www.research.microsoft.com>
- Intel at <http://www.intel.com/go/storage>
- Hewlett Packard <http://www.hp.com>
- and EMC at <http://www.emc.com>.

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.

How Much Information? 2003

Summary

Exec Summary

Stored Information

Paper | Film | Magnetic | Optical

Information Flows

Broadcast | Telephony | Internet

Wrap-up

Thanks | Printable (PDF)

Paper

[I. World Paper Capacity](#)

[II. Genres of Print](#)

[A. World Flow](#)

[B. United States Flow](#)

[C. Rate of Change](#)

[III. Stock of Print / Copies](#)

[A. Books](#)

[B. Newspapers](#)

[IV. Conversion](#)

[Assumptions](#)

[References](#)

[Printable Section \(PDF\)](#)

2. INFORMATION PRINTED ON PAPER

Paper is the primary medium for the formal publication of information, although it represents only 0.01% of new information recorded in all media. Information is recorded on paper in three distinct genres: books; serials (including newspapers, mass market and trade periodicals, scholarly periodicals and newsletters) and documents printed or copied in offices. There are interesting differences among nations in the proportion of printed information in each of these genres: the United States produces far more office documents than other nations, the European Union produces more books and serials, and the rest of the world produces relatively more newspapers and mass market periodicals.

I. World Paper Printed Information Storage Capacity

A tree can produce about 80,500 sheets of paper, thus it requires about 786 million trees to produce the world's annual paper supply. The UNESCO Statistical Handbook for 1999 estimates that paper production provides 1,510 sheets of paper per inhabitant of the world on average, although in fact the inhabitants of North America consume 11,916 sheets of paper each (24 reams), and inhabitants of the European Union consume 7,280 sheets of paper annually (15 reams), according to the ENST report. At least half of this paper is used in printers and copiers to produce office documents.

Table 2.1: World Paper Production 1997 and 2001

Production Statistics	Metric Tons (1997)	Capacity in Terabytes (1997)	Metric Tons (2001)	Capacity in Terabytes (2001)
Printing and Writing Paper (World)	90.0	540,000	94.8	568,800
Newsprint (World)	36.0	432,000	37.8	453,600

Source: *How much information 2003*

World output for printing and writing paper was 94.8 million tons in 2001, and newsprint production was about 37.8 million tons. Using a conversion of 6 GB of data per metric ton of paper there exists a net storage capacity of 568,800 TB on the total amount of printing and writing paper produced in 2001 (94.8 million tons), were every bit of it be used for information recording. Using a conversion of 12 GB of data per metric ton of newsprint there exists a net storage capacity of 453,600 TB on the total amount of newsprint produced in 2001 (37.8 million tons), were every bit of it be used for information recording. If all of the writing paper and newsprint produced in 2001 were used to store printed information, this would be equivalent to about 1,022,400 TB, assuming each page was converted into one large image of 130 KB. This represents a 2% capacity rise from 1999.

U.S. Paper Printed Information Storage Capacity

Total U.S. paper and paperboard production for 2001 was 80.66 million metric tons, or 25.4% of the world's total. This percentage figure has fallen from previous years. At the time of the last study, U.S. production figures were closer to 30% of the world total. The U.S. produced about 22.166 million metric tons of printing and writing paper and approximately 5.77 million tons of newsprint.

Using the conversion rate of 6 GB of data per metric ton of paper there exists a net storage capacity of 132,996 TB using the total amount of printing and writing paper produced in the U.S. be used for information recording. Using the conversion of 12 GB of data per metric ton of newsprint there exists a net storage capacity of 69,240 TB on the total amount of newsprint produced in the U.S. in 2001, should every bit of it be used for information recording.

In 1999, the U.S. produced 23.8 million metric tons of printing and writing paper and 6.4 million metric tons of newsprint.

Table 2.2: United States Paper Production

Production Statistics	Metric Tons (1999)	Capacity in Terabytes (1999)	Metric Tons (2001)	Capacity in Terabytes (2001)
Printing and Writing Paper (U.S.)	23.8	142,800	21.66	132,996
Newsprint (U.S.)	6.4	80,000	5.77	69,240

Source: *How much information 2003*

If all of the writing paper and newsprint produced in the U.S. in 2001 were used to store printed information, this would be equivalent to about 202,236 TB, assuming each page was converted into one large image of 130 KB. This represents a 8% capacity decrease from 1999.

II. GENRES OF ORIGINAL INFORMATION STORED ON PAPER

A. World Flow

In this section, we examine the global flows and stock of original print media. In "World Flow", we discuss the various forms of original information produced around the world and in the United States. This refers to all forms of written information created worldwide, excluding copies, and stored in some form for posterity. The single largest component of print media flow is office documents. The term broadly refers to all paper storage of printed and written materials in various forms throughout the world. This is not surprising, as much original information is created yearly in offices and other institutions for their own (and sometimes public) consumption. Newspapers are a very significant source of original flow information, comprising the single largest source of the public consumption of information, followed by mass-market magazines and books. There has been a slight increase in the worldwide consumption of newspapers, despite the

phenomenal increase in the use of the Internet in information gathering.

Table 2.3: World Original Print Information Flow (2003)

Media Type (Data Sources and Year Data Cited)	Unique Items per Year	Conversion Factor	Total Terabytes (Annual Worldwide)
Books (UNESCO [???)	950,000 [???) books	Scanned image (600 dpi): 39 MB/book (300 page average)	39 TB
		Digital compression: 7.8 MB/book	8 TB
		Plain text: 0.75 MB/book	0.8 TB
Newspapers (ISSN 2001)	25,276 publications	Scanned image (600 dpi): 500 KB/page (30 page average)	138.4 TB
		Digital compression: 100 KB/page	27.7 TB
		Plain text: 10 KB/page	2.7 TB
Mass Market and Trade Periodicals (Ulrich's 2001)	80,000 publications	Scanned image (600 dpi): 13 MB/issue (96 page average)	52 TB
		Digital compression: 5.4 MB/issue	10 TB
		Plain text: 0.5 MB/year	1 TB
Scholarly Periodicals (Ulrich's 2001)	37,609 publications	Scanned image (600 dpi): 27 MB/issue (208 page average)	6.0 TB
		Digital compression: 5.4 MB/issue	1.3 TB

		Plain text: 0.5 MB/year	0.1 TB
Newsletters (Ulrich's 2001)	40,000 publications	Scanned image (600 dpi): 1.6 MB/item (12 page average)	0.9 TB
		Digital compression: 0.3 MB/item	0.2 TB
		Plain text: 0.03 MB/item	0.02 TB
Archiveable, original office documents (National Archives)	(10 ⁷) * 1075 pages (10.75 Bn pgs)	Scanned image (600 dpi): 130 KB/page	1397.5 TB
		Digital compression: 26 KB/page	279.5 TB
		Plain text: 2.5 KB/page	27.9 TB
Totals:		Scanned: TB	1634 TB
		Compressed: TB	327 TB
		Text: TB	33 TB

Source: *How much information 2003*

These estimates are based upon trade publications. However much higher estimates for serials and office documents have been made by the ENST study.

Serials. The ISSN register (see [International Standard Serial Number \(ISSN\) Statistics](#)), used here only for Newspapers, includes some 671,988 registered serials, 98% of which are paper based; we have used the Ulrich numbers because they eliminate duplication, and our goal is to estimate the number of original titles. The ENST numbers estimate that the U.S. produces 17.6% of world paper serial publications, and the European Union 44.5%. (ENST also estimates that the U.S. produces 17% of world books, the EU 42%, and the rest of the world 41%; they agree with our estimate of book production.)

Office Documents. Our estimates of printed office document production is based upon the practices of the U.S. National Archive and Records Administration (NARA), which archive about 2% of office documents produced by the Federal Government. The ENST study uses a methodology based upon working assumptions from the production of printing and copier paper: if 20% of printer paper contains original information (not copies), and if half of this is archived, then the world produces 2238 TB of office documents. Interestingly, they estimate 662 TB (29%) was produced by the EU and 775TB (35%) by the U.S. Our 2% estimate is based upon a very narrow definition of "archival documents," since NARA essentially archives paper forever only if it is thought to have historical value; the ENST estimate that 10% of all paper is archived original information is very broad. We know of no empirical study that identifies the archival practices of different kinds of institutions (i.e., government, business, education) in different countries, thus this issue remains on the research agenda.

NOTE ON OFFICE DOCUMENTS CONVERSIONS

To estimate the amount of information generated by offices, we looked at the statistics of the Federal Government, which is the single largest employer in the United States, with 1.9 million civilian workers and 1.5 million individuals in the armed services as of 2000. The Federal Government, in total, employs about 2.3% of the nation's workforce.

The National Archives in Washington D.C. retains 2% of what the government produces, across a range of media. NARA 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. The current textual archival holdings, as of October 2001 occupy a total of 21.5 million cubic feet. If we use the rule of thumb that 200 pieces of paper fit in one cubic foot (in an archive), then we arrive at a total of 4.3 billion pages.

If one divides 4.3 billion by the number of years NARA has existed (60 years), one could obtain a rough number of pages collected per year by dividing 4.3 billion by 60 - the result is about 72 million 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 million pages per year.

If this represents 2.3% of the nation's workforce, then one could estimate that United States companies produce a total of more than 4 billion archiveable pages each year, equivalent to 1,400 terabytes.

Table 2.4. World Original Print Information Flow by Percent (2003)

Media Type	Total Terabytes (scanned)	Percent of total
Books	39 TB	2.3%
Newspapers	138.4 TB	8.5%
Mass Market Periodicals	52 TB	3.2%
Scholarly Periodicals	6 TB	0.37%
Newsletters	0.9 TB	0.05%
Office Documents	1,397 TB	85.5%
Total:	1,634 TB	100%

Source: *How much information 2003*

B. United States Flow

The U.S. print information flow is dominated by office documents, but unlike for global data, the second most prevalent source of print information is not mass-market periodicals, but books. This is partly a feature

of the United States being a key player in the worldwide publishing industry, and also a factor of its high educational levels and of local markets for books. Third world nations have proportionally higher information being created in mass distribution channels like magazines and newspapers than in books.

Table 2.5: U.S. Original Print Information Flow (2003)

Media Type (Sources and Year Cited)*	Unique Items per Year	Conversion Factor	Total Terabytes (Annual Worldwide)
Books (ISBN)	141,901 books	Scanned image (600 dpi): 39 MB/book <i>(300 page average)</i>	5.5 TB
		Digital compression: 7.8 MB/book	1.1 TB
		Plain text: 0.75 MB/book	0.1 TB
Newspapers (Newspaper Association of America)	10,170 publications	Scanned image (600 dpi): 500 KB/page <i>(30 page average)</i>	13.5 TB
		Digital compression: 100 KB/page	2.8 TB
		Plain text: 10 KB/page	0.3 TB
Mass Market and Trade Periodicals (Bowker 2001)	16,615 publications	Scanned image (600 dpi): 13 MB/issue <i>(96 page average)</i>	3.5 TB
		Digital compression: 5.4 MB/issue	0.7 TB
		Plain text: 0.5 MB/year	0.07 TB
Scholarly Periodicals	10,632 publications	Scanned image (600 dpi): 27 MB/issue <i>(208 page average)</i>	1.6 TB

(Bowker 2001)		Digital compression: 5.4 MB/issue	0.3 TB
		Plain text: 0.5 MB/year	0.03 TB
Newsletters (Bowker 2001)	11,581 publications	Scanned image (600 dpi): 1.6 MB/item <(12 page average)	0.3 TB
		Digital compression: 0.3 MB/item	0.06 TB
		Plain text: 0.03 MB/item	0.006 TB
Archiveable, original office documents (National Archives 2001)	4.3 Bn pgs	Scanned image (600 dpi): 130 KB/page	559 TB
		Digital compression: 26 KB/page	112 TB
		Plain text: 2.5 KB/page	11.6 TB
Totals:		Scanned: TB	583.4 TB
		Compressed: TB	117 TB
		Text: TB	12 TB

Source: *How much information 2003*

Interestingly, the information flow figures for U.S. consumption represent over 35% of the total original worldwide print information flow. While the U.S. is indeed the global leader in production and storage of information, these flow figures are largely influenced by the office document use in the U.S.. If the office documents component were removed from the equation, the U.S. accounts for slightly over 10% of the world's original information flow in print. This underlines the high use of paper in U.S. offices as compared to that in nations around the world.

There have been some changes in the creation of new information between our previous study and the current one, but most of the changes have been fairly small except in the area of office documents. Contrary to notions of paperless offices floated in the late 80s and early 90s, the consumption of office paper has gone up substantially in the recent years, especially following the move to laser/inkjet printers from dot matrix printers. Paper use in offices has further risen with the increasing speed of laser printing coupled with its decreasing cost. Each year, almost **500 billion copies** are produced on copiers in the United States; **nearly 15 trillion copies** are produced on copiers, printers, and multi-function machines. (Source:

XeroxParc).

An increase in the number of information on newsprint is seen - the proportion of this increase is slightly higher worldwide than in the U.S., possibly an outcome of an expansion of media markets in developing nations.

C. Rate of Change

Table 2.6: World Flow: Comparison of 1999 Data with 2001 Data					
		WORLD		U.S.	
		1999 (TB)	2001 (TB)	1999 (TB)	2001 (TB)
Books	Scanned Images	39	39	3	5.5
	Compressed	8	8	0.5	1.1
	Plain Text	1	1	0.05	0.1
Newspapers	Scanned Images	124	138.4	13	13.5
	Compressed	25	27.7	3	2.8
	Plain Text	2.5	2.7	0.3	0.3
Mass Mkt Periodical	Scanned Images	52	52	4.2	3.5
	Compressed	10	10	2.6	0.7
	Plain Text	1	1	0.26	0.07
Scholarly Periodical	Scanned Images	9	6	2	1.6
	Compressed	2	1.3	0.5	0.3
	Plain Text	0.2	0.1	0.04	0.03
Newsletters	Scanned Images	0.8	0.8	0.2	0.3
	Compressed	0.2	0.2	0.04	0.06
	Plain Text	0.02	0.02	0.004	0.006
Office Documents	Scanned Images	975	1397.5	390	559
	Compressed	195	279.5	78	112
	Plain Text	19	26.9	7.5	11.6
Totals:	Scanned Images	1200	1637	421	583
	Compressed	240	327	84	117
	Plain Text	24	33	8.2	12

Source: *How much information 2003*

However, there are other trends that are not shown on this chart. The rise in the amount of office documents has a component of Internet information printouts. Similarly, the amount of information now available online -

such as news, has not severely affected the distribution of newspapers. What it has possibly done is alter the readership of the average newspaper. Consumption studies may show that the number of pages, and possibly even the types of news sections, read by the average newsreader has changed since the Internet came into use.

Consumption of major journals may have also shifted to online formats, as more are published online, as documented in Ulrich's periodicals directories from 1998, 2000, and 2002.

Increasing Shift to New Media by Scholarly Periodicals

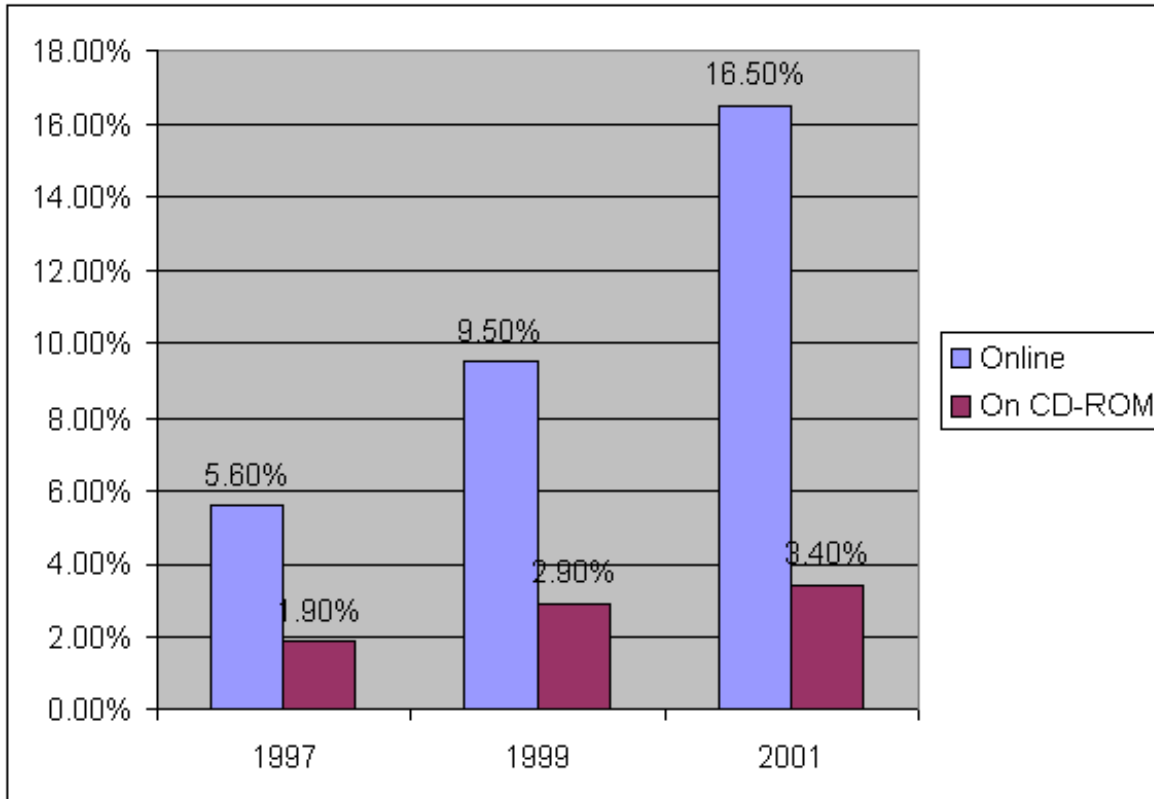


Table 2.7: Increase of Digital Versions of Scholarly Periodicals

Year	Number of Periodicals With Internet Websites	Number of Periodicals With CD-ROM versions
1997	8,672	2,903
1999	14,757	4,625
2001	27,083	5,577

Source: Raw data from Ulrich's international Periodical Directory; Chart and Table, *How much information 2003*

Efforts to place contents online need no explanation. However, the data above is more interesting for the fact that the rate at which publishers have shifted to the use of CD-ROMs for digital storage has been slower than the rate at which they have replicated their print contents online. There was a gap of about 3.7%

between the total number of online and the total number of CD-ROM versions of the print magazines; this has risen to over 13% in 4 years. Publishers clearly see creating and developing an Internet readership strategy as vital to the future of scholarly periodicals.

III. THE STOCK OF PRINTED INFORMATION - COPIES

In this segment we examine the amount of information existing in multiple copies of the original material discussed above. This dimension represents the use of paper as a distribution medium for the consumption of information.

A. Books

1. The flow of books.

World. We calculate that the total world flow of book copies was approximately 4.0 billion books, extrapolating once again from U.S. production figures [???]. This is equivalent to about 32,000 TB of digital data. In our last study, there were 2.75 billion books sold worldwide, which was equivalent to about 22,000 TB of digital data. This represents growth of over 45% since 1999.

United States. About 1.62 billion books were purchased in the United States in 2001, according to Barrie Rappaport at IPSOS-NPD. Using the 8 MB/book estimate, this is equivalent to 129,600 TB. In our last study, there were 1.6 billion books sold worldwide, which was equivalent to about 128,000 TB of digital data. This represents a growth of 1.25% since 1999.

2. The stock of books.

United States. In March 2003, 2,078,051 book titles were available and for sale in some form. In all there were 4,123,094 book titles, including the "out of print" books according to the Books in Print records provided by Andrew Grabois of R.R. Bowker. These numbers include only books made for commercial publication, not books or other materials meant for private circulation. During our last study in 1999, Books in Print had about 1.66 million titles for sale and 3.2 million titles in all listed. This marks a 25% increase in books available for purchase and about a 30% increase in the number of listed books. Increasingly out of print book titles are available from publishers through 'just in time' printing, which may account for some of the increase. This represents a fraction of the total number of books ever written. That there has been an increase in the number of books in Books in Print's universe indicates that even the data on the total number of books being retroactively catalogued is changing constantly. If one wished to more fully address the universe of book titles in the United States, including government publications, books that are no longer in print, and some books that were written and submitted for copyright but never published commercially, one could look to the holdings of the larger national libraries and copyright repositories. For example, the Library of Congress print media collection includes 26 million titles (208 terabytes).

World. To estimate the international stock of books currently available for purchase, we extrapolate from the United States production figures. The U.S. engages in the world's largest trade in printed products, producing about 40% of the world's printed material, according to the U.S. Industry and Trade Outlook 2000. However, in the past, the U.S. has not been as dominant, though production was also significantly less. It could be speculated then that somewhere less than 40% of worldwide stock of books exists in the U.S. - more specifically, it could be speculated that this number is probably in the range of 15% and 35%, more likely on the lower end.

Using the same ballpark estimates, 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 original books might be between 74 million books and 175 million books (the lower figure if the U.S. represents 35% of world's published books, the upper figure if the U.S. is closer to 15%)

NOTE ON eBooks

eBooks can be defined most broadly as books available in digital format. Within this definition, there are several sub-categories: books that are written purely for electronic media, books that are published in both print and electronic media, out of print books that have been digitized for archival purposes, and books that are out of print but are printed or supplied in electronic format. At the time of our last study, eBooks seemed to be a technology to consider closely for the future. In 2001, only 180,000 e-books were sold worldwide (almost a hundred-thousandth of U.S. book sales alone). At the most basic level, eBooks did not do well because of a preference for paper. But a number of other factors also influenced the growth of this technology.

B. Newspapers

Worldwide, there were at least 436.2 million copies of newspapers sold daily in 2001 according to the World Association of Newspapers (WAN). Using this, we calculate at least 159.14 billion copies of newspapers were sold worldwide. **This represents 2,387,100 terabytes of data annually.** The above figure is an "at least" calculation - which excludes data from several print material producing nations including China, Russia, Mexico and Pakistan. Closer examination of the statistics shows that the rise in the worldwide sales of newspapers has almost exclusively been in developing nations worldwide.

Table 2.8: Daily Circulation (Worldwide - Millions), 1997-2001

	1997	1998	1999	2000	2001	% change
USA	58.7	56.1	55.9	55.9	55.5	-2.03
Japan	72.7	72.4	72.2	71.9	71.7	-1.38
EU	82.2	81.4	80.5	80.1	78.7	-4.3
Other	204.5	204.1	212.4	228.3	230.3	12.57
Total	416.2	414.2	421.2	434.2	436.2	4.81

Source: World Association of Newspapers. Note: Figures Exclude Chile Lithuania, Mali, Mexico, Pakistan, Russia, Serbia, South Korea, Taiwan

Table 2.9: Average Circulation Per Title (Worldwide – '000s)

	1997	1998	1999	2000	2001	% change
USA	37.6	37.7	37.7	37.9	37.8	-0.11
Japan	667	670.5	662.8	653.8	676.4	3.48
EU	71.7	71.5	71.1	71	69.5	-2.94
Other	72.2	73.7	69.2	67	68.8	-4.78
Total	74.3	74.7	72.7	71.3	72	-3.08

Source: World Association of Newspapers. Note: Figures Exclude Chile Lithuania, Mali, Mexico, Pakistan, Russia, Serbia, South Korea, Taiwan

It is evident that the growth in newspapers has been away from developed economies, which seem to have a universal stagnation or fall in the sales of newspapers. The downward trend has been slow, yet the fall in circulation has been accompanied by a slight fall in the number of newspapers in existence.

The effects of both Internet technology and of the growing reach of education and access to media are seen in the worldwide newspaper sales trends.

Table 2.10: Selected Annual Copies Sold (Worldwide – Millions)

Country	Sales (Million)	Percent Change (5-years)	Last Year of Available Data
Belgium	484	-3.8	2001
Canada	1753	2.3	2001
Costa Rica	104	15.3	2001
Croatia	173	-3.1	2001
Czech Republic	508	-18.6	2001
Denmark	519	-9.6	2001
Estonia	72	2.1	2001
Finland	87	-11.2	2001
Germany	7368	-5.6	2001
Greece	194	-6.1	2001
Hungary	515	-2	2001
India	10893	19.8	2001
Ireland	183	6.4	2001
Italy	2096	1.4	2001
Latvia	73	-7.6	2001
Luxembourg	36	9.1	2001
Netherlands	1338	-8.9	2001
New Zealand	238	-5.1	2001
Norway	750	-2.3	2001
Slovakia	132	-50.9	2001
South Africa	282	-3	2001
Spain	1556	17	2001
Sri Lanka	164	0.8	2001

Sweden	1233	-3.8	2001
Switzerland	796	-2.5	2001
Turkey	1061	-33.4	2001
United States	17485	-2	1999

Source: World Association of Newspapers.

Thus nations with higher access to the Internet have seen stagnations in newspaper circulation in the 5-year period between 1997 and 2001, while nations like India and China with rapidly expanding economies and consumer bases have seen fairly significant increases in newspaper circulation. Exact percentages are not available but China has seen an increase of roughly 20% in the circulation of dailies since 1997. This trend may also be partly attributed to urbanization and the increase in the number of young professionals (and households) in many developing nations.

According to the WAN, Ireland was the only developed nation with a significant positive increase in annual newspaper sales; growth was dominated by East European, South American and Asian nations. On the contrary, most developed nations saw negative growth, with Turkey heading the list for the most significant fall in newspaper sales. Only two Asian nations, Sri Lanka and Mongolia, were on the negative growth list, otherwise dominated by West European nations.

U.S. Flow: Newspapers

According to the World Association of Newspapers (WAN), there were approximately 55.5 million copies of newspapers sold in the U.S. in 2001. This represents a very slight decrease over a 5-year period. Similarly the total flow of newspapers has decreased since 1997, but remains in the range of 17.4 billion copies sold yearly. **This represents 261,000 terabytes of data annually.** Compared with the change in the number of newspapers in circulation, there is a slight difference, suggesting either that there has been some merger of interest, or simply that some newspapers ceased publishing, and their clientele thereafter moved to other publications.

Table 2.11: Selected Annual United States Newspaper Readership Trends

	1997	1998	1999	2000	2001	5 yr Change
Number of Newspapers	1509	1489	1483	1476	1468	-2.72
Number of Copies Sold (Million)	58.7	56.1	55.9	55.9	55.5	-2.03
Copies Per Publication ('000s)	37.6	37.7	37.7	37.9	37.8	-0.11

Source: World Association of Newspapers (WAN)

C. Mail

This table of facts about postal mail in the United States is from Odlyzko (2000), supplemented by the most recent U.S. Postal Service Annual Report (2002). About half of all mail is currently first class and about half is junk mail. If we assume 2 pages per piece of mail, and digitize it at 15 Kbytes per page, 2002 U.S. mail is about **6.2 petabytes** per year. This represents an increase of about 0.3 petabytes over the 2000 study.

The U.S. Postal Service handles approximately 40 percent of the world's card and letter mail volume. Next largest: Japan with 6 percent. The total volume of mail worldwide equals 15.6 petabytes.

Table 2.12: Statistics about U.S. 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
1999	60,418		201, 644	
2000	62,284		207,882	739
2001	63,425		207,463	717
2002	63,761		202,822	738

Source: *The history of communications and its implications for the Internet*, A. M. Odlyzko. [[Abstract](#)] [[PostScript](#)] [[PDF](#)] [[LaTeX](#)]

IV. CONVERSION FACTORS AND ASSUMPTIONS

Printing / writing paper production. We estimate that one ream (500 sheets) of standard grade 8 1/2 X 11" paper would weigh 5 pounds. Therefore each metric ton (2,204 pounds) equals about 440 reams or 220,000 sheets. Multiplying by 26 KB per page results in 6 GB per metric ton.

Newsprint production. We estimate that storage requirements would be about 2 times that for printing / writing paper: 12 GB per metric ton. Newspapers tend to contain more words and graphics per page, requiring, on average, 1 MB per scanned page. Furthermore, newsprint is thinner and lighter, so each metric ton contains more individual sheets.

Books. We estimate 300 pages per book, for a total of 39 MB apiece. (Source: Robert M. Hayes, UCLA, "The Economics of Digital Libraries" www.usp.br/sibi/economics.html)

Newspapers. We estimate 30 pages per newspaper, then multiply that number by the total units produced every year (The page number is low, to reflect the number of small and non-daily newspapers published around the world.) A double-truck (center fold) full broadsheet is 24 X 36 inches. Because a newspaper 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 0.5 MB.

Mass Market Periodicals. We estimate 96 pages per periodical per issue, averaging larger and smaller sized publications. (Similar calculation by: Robert M. Hayes, UCLA, "The Economics of Digital Libraries" www.usp.br/sibi/economics.html)

Newsletters. On the estimated 150 pages per newsletter per year - on a ballpark 12 pages per issue averaged over all newspapers. (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 U.S. government documents produced each year. This study does not account for the documents generated and disposed of yearly without recording.

Table 2.13: Conversion Assumptions

Average # of Pages per Issue / Item	Storage Format	Average File Size Per Page	Conversion Factor Per Issue/ Item (rounded)
300 (Books)	Scanned TIFF (600 dpi)	130 KB	39 MB
	Compressed	26 KB	7.8 MB
	Plain text	2.5 KB	0.75 MB
208 (Scholarly)	Scanned TIFF (600 dpi)	130 KB	27 MB

Periodicals)	Compressed	26 KB	5.4 MB
	Plain text	2.5 KB	0.5 MB
96 (Mass Market Periodicals)	Scanned TIFF (600 dpi)	130 KB	13 MB
	Compressed	26 KB	2.6 MB
	Plain text	2.5 KB	.25 MB
12 (Newsletters)	Scanned TIFF (600 dpi)	130 KB	1.6 MB/Item
	Compressed	26 KB	0.3 MB/item
	Plain text	2.5 KB	.03 MB/Item

Source: Raw data from Bowker Inc., May 2003. Conversion factors to bytes from the [ArchiveBuilders](#) web page. Compilation by How much information 2003

Table 2.14: Academic and Scholarly Periodicals (Worldwide)			
Frequency	Titles	Multiplier	Total Publications
Annual	4885	(*1)	4,885
Biennial	246	(*0.5)	123
Bimonthly	3990	(*6)	23,940
Daily	27	(*365)	9,855
Irregular	7057	(*4)	28,228
Monthly	7248	(*12)	86,976
Quarterly	9814	(*4)	39,256
Semiannually	4147	(*2)	8,294
Weekly	195	(*52)	10,140
		TOTAL	211,697

Source: Raw data from Bowker Inc., May 2003

Table 2.15: Mass Market and Trade Periodicals (United States)			
Frequency	Titles	Multiplier	Total Publications

Annual	2354	(*1)	2,354
Biennial	130	(*0.5)	65
Bimonthly	2294	(*6)	13,764
Daily	287	(*365)	104,755
Irregular	627	(*4)	2,508
Monthly	6834	(*12)	82,008
Quarterly	2659	(*4)	10,636
Semiannually	484	(*2)	968
Weekly	946	(*52)	49,192
		TOTAL	266,250

Source: Raw data from Bowker Inc., May 2003

Table 2.16: Academic and Scholarly Periodicals (United States)

Frequency	Titles	Multiplier	Total Publications
Annual	1268	(*1)	1,268
Biennial	86	(*0.5)	43
Bimonthly	1089	(*6)	6,534
Daily	11	(*365)	4,015
Irregular	1474	(*4)	5,896
Monthly	1962	(*12)	23,544
Quarterly	3450	(*4)	13,800
Semiannually	1210	(*2)	2,420
Weekly	82	(*52)	4,264
		TOTAL	61,784

Source: Raw data from Bowker Inc., May 2003

Table 2.17: Newsletters (United States)

Frequency	Titles	Multiplier	Total Publications
Annual	653	(*1)	653
Biennial	38	(*0.5)	19
Bimonthly	1210	(*6)	7,260
Daily	220	(*365)	80,300

Irregular	588	(*4)	2,352
Monthly	4670	(*12)	56,040
Quarterly	2714	(*4)	10,856
Semiannually	498	(*2)	996
Weekly	990	(*52)	51,480
		TOTAL	209,956

Source: Raw data from Bowker Inc., May 2003

REFERENCES

- Abromovitz & Mattoon, *Paper Cuts: Recovering the Paper Landscape*. Washington, DC: Worldwatch Institute, 1999
- Bogart, Dave, Ed., *The Bowker Annual Library and Book Trade Almanac*, 46th edition. New Jersey: R.R. Bowker, 2001.
- 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"
- Kenney, Brian, "netLibrary, EBSCO explore new models for E-books" *Library Journal*, April 15, 2003.
- King, Donald W. and Carol Tenopir. "Economic Cost Models of Scientific Scholarly Journals," 1998.
- *Annual Report of the Librarian of Congress*. Washington, DC: Library of Congress, 2001.
- 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
- David Bounie, "The International Production and Dissemination of Information" (Paris: ENST, 2003).
- International Standard Serial Number Register
- International Standard Book Number Register
- [IPSOS-NPD](#)
- 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
- [USPS Postal Facts 2003](#)
- 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.

Special thanks to:

- Shawn Chen and Andrew Grabois, Bowker Inc.
- Gregg Fales, The Pulp & Paper Report
- Barrie Rappaport, IPSOS-NPD Book Trends Manager and Publisher
- Véronique Rat-Morris, ISSN
- Tatiana Repkova, Director of Research, World Association of Newspapers
- Melanie Smuga, UNESCO

How Much Information? 2003

Summary

Exec Summary

Stored Information

Paper | Film | Magnetic | Optical

Information Flows

Broadcast | Telephony | Internet

Wrap-up

Thanks | Printable (PDF)

Film

I. Photographs

[A. Originals](#)

[B. Copies](#)

II. Motion Pictures

[A. Originals](#)

[B. Copies](#)

[C. Trends](#)

[D. Facts/Trivia](#)

III. X-Rays

[References](#)

[Printable Section \(PDF\)](#)

3. FILM

The amount of information produced and stored on film has increased since our last study. This increase has been mainly because of the expansion in the motion picture industry, rather than in film-based photography or in X-rays. Also, this increase may not indicate future expansion in the market for film.

At the time of the last study, digital photography was nascent; by this study it has clearly emerged as the technology that will eventually turn film-based photography into a comparatively smaller market.

The equivalent of the digital camera in moving images is the digital video camera (DV). In the area of motion pictures, the transition to digital technology is likely to be slower than has been in the case with photography. One of the reasons digital camera prints (or computer format images) can take precedence over film in the photography market is that individual users are less likely to be excessively quality-conscious -- and most photographers are amateurs who use cameras for recreation and recording of information rather than for serious photography.

Motion pictures, on the other hand, are made for mass consumption, and thus there is a quality expectation. Although new digital video cameras with capabilities of output closer to motion picture cameras are now available in the market, the quality of film and video are still significantly different, and the film format still enjoys aesthetic preference.

But with the time and material processing costs of digital video editing being exponentially lower than that of film editing, it is very likely that as soon as an equity threshold is reached in terms of visual quality between film and digital video, film will similarly lose market share in motion pictures as it has in still photography.

Digital imaging formats exist in X-ray technology, and these are gaining popularity, especially in the area of dental X-rays. However, in the case of non-dental medical X-rays, the digital format has yet to take off. In this case, the issue is less one of quality than of storage. Digital X-ray images are very storage heavy, and so the traditional film-based X-rays are still predominant. [???

This again means that the route from film to the digital medium in X-rays is related to how the storage technology

develops in the next few years -- if digital storage of visual images decreases in cost and increases in efficiency, X-rays may soon see a shift away from film. Digital storage has seen much more significant adoption in industrialized nations than in developing countries.

Table 3.1: Original Data On Film Annually Worldwide (2002)

	Units	Digital Conversion	Total PB
Photography	75,000,000,000	5 MB per photo	375
Motion Pictures	10,342	(see below)	25
X-Rays	2,000,000,000	10 MB per radiograph	20
Total:			420

Source: *How much information 2003*

Table 3.2: Worldwide production of filmed original content, if stored digitally, in terabytes circa 2002.

Type of Content	Terabytes/Yr Upper Estimate	Terabytes/Yr Lower Estimate	1999 Upper Estimate	1999 Lower Estimate	% Change Upper Estimates
Photographs	375,000	37,500	410,000	41,000	-9%
Cinema	6,078	12	4,490	9	35%
Made for TV films	2,531	2,530	N/A	N/A	N/A
TV series	14,155	14,155	N/A	N/A	N/A
Direct to video	2,490	2,490	N/A	N/A	N/A
X-rays	20,000	20,000	17,200	17,200	16%
Subtotal	420,254	74,202	431,690	58,209	-2.6%

Source: *How much information 2003*

Table 3.3: Total Stock of Film Worldwide (2001)

	Units	Digital Conversion	Total PB
Photography	900,000,000,000	5 MB per photo	4,500
Motion Pictures	368,530	(see charts below)	781
X-Rays	20,000,000,000	10 MB per radiograph	200
Total:			5,481

Source: *How much information 2003*

Table 3.4: Total Copies (Including Originals) of Film Worldwide (2001)

	Units	Digital Conversion	Total PB
Photography	76,500,000,000	5 MB per photo	383
Motion Pictures	5,660	(see above)	2,903
X-Rays	2,000,000,000	10 MB per radiograph	20
Total:			3,306

Source: *How much information 2003*

I. PHOTOGRAPHS

In our previous study, the "How Much Information" project had found that film photographs were increasingly losing market to digital format film. This trend has continued as expected: digital cameras have gained a significant market share and are expected to outsell film cameras by the end of 2003. Twenty-one percent of United States households had a digital camera as of 2002.

Photographs stored in the digital format have been accounted into our calculations of magnetic information. There is however, significant photographic data stored on digital photography devices. In 2002, there were 27.5 million digital still cameras purchased worldwide as compared to 63 million analog still cameras, according to Photo Marketing Association International (PMAI).

Also, the universe of cameras was approximately 53 million cameras. (Figures provided by the InfoTrends Research Group. The figures factor an accumulation of units over time, subtracting those that are retired every 3 years on average.) The storage space on each camera is about 16 MB, on average. Earlier on, there were cameras with smaller memory storage, but currently there are an equally large number of professional and higher-end cameras with greater storage space. The size of utilized storage on these devices is **1.2 petabytes**. (See details in [Magnetic](#).)

A. Original Information Stored on Film Rolls

1. Annual Production of Titles

Among the analog camera prints, there has been a comparative decrease in the total amount of photo output. In 2002, 75 billion prints were made of photographs taken on analog cameras, according to PMAI. The most dramatic rise in the sales of film rolls took place in the mid-1990's. The sharpest fall began in 2000, and is expected to continue in 2003.

Table 3.5: Key Photograph Statistics for 2002

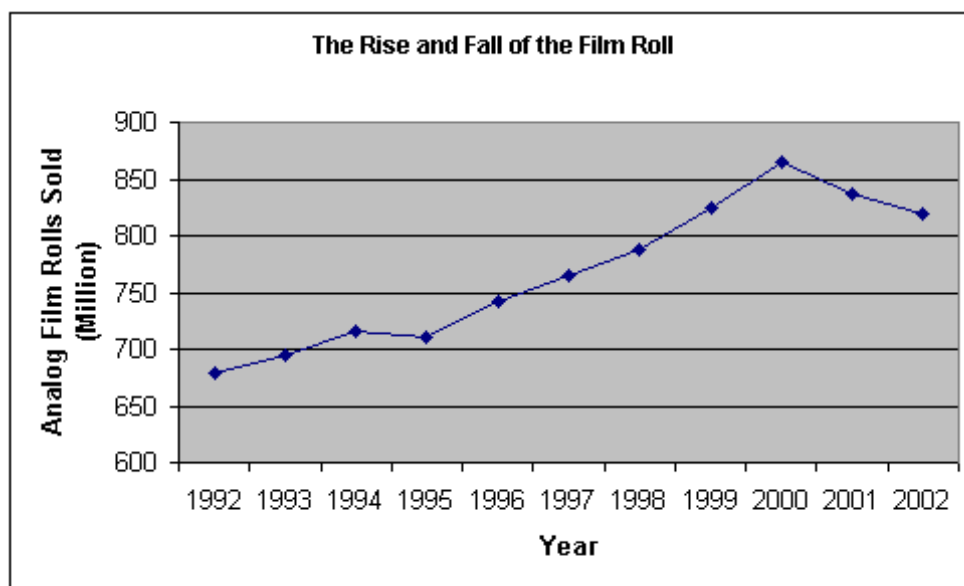
Film Prints Made (World)	75 billion
Film Rolls (World)	2.93 billion (excluding disposable cameras)
Disposable Cameras (World)	0.4 billion
Film Rolls (United States)	1.014 billion

Source: *Photo Marketing Association International*

Table 3.6: Film Rolls Sold in the United States

Year	Rolls Sold (in millions)
1992	679
1993	694
1994	716
1995	710
1996	742
1997	765
1998	787
1999	825
2000	865
2001	837
2002	820

Source: Photo Marketing Association International



Source: Photo Marketing Association International

Each roll potentially produces 30 individual exposure negatives, balancing out 24- and 36-exposure film, which are roughly produced in similar quantities (36 exposure rolls are very popular in many developing nations). This means a net capacity of about 87.9 billion exposures if all the film rolls were to be completely utilized. This figure corresponds well to the worldwide production estimate of 75 billion film prints (2002), accounting for about a 15% wastage of negatives.

Using the above figures, the total digital storage capacity of the film rolls produced in 2002 was about **440 petabytes** (87.9 billion photos at 5 MB apiece).

2. Accumulated Stock

The number of original photographs stored around the world is not a widely reported topic. It is also difficult to calculate, given the differing levels of preservation. There exist firms that stock large databases of photographs for resale; of these, the two largest are Getty Images, with 70 million images, and Corbis Images, with 65 million images.

As per our last study, approximately 750 billion photographs existed worldwide in 1999. There has been an addition of approximately 150 billion photographs in the two following years. Further, it is estimated that 2.1 billion photographs have been printed out from digital cameras. Thus, after accounting for some atrophy of photographs, the universe of photographs currently existing is close to about 900 billion. Using the same calculation of about 5 MB per photograph, this translates to a universe of 4,500 PB or 4.5 exabytes of data in still photographs.

B. Copies of Information Stored/Published on Film

There has been very little history of the large mass of photographs being copied. Kodak estimates that only about 2% 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 are factored into the [Paper](#) section of this study.

Also, there are no statistics to tell what percentage of prints developed have multiple copies at the source. Thus, all original prints from negatives are counted in the first category of original pictures. Given this, the addition to the net figures of yearly photographs through copies is fairly small.

II. MOTION PICTURES

The amount of information stored on film through motion pictures is sizeable, because of the audio-visual data on every frame of motion picture reel. Traditionally, films made for popular release are duplicated with anywhere from a handful to several thousand prints, which are distributed worldwide. This makes the yearly turnover of film a large figure.

A. Original Information Stored on Motion Pictures

1. Annual Production of Titles

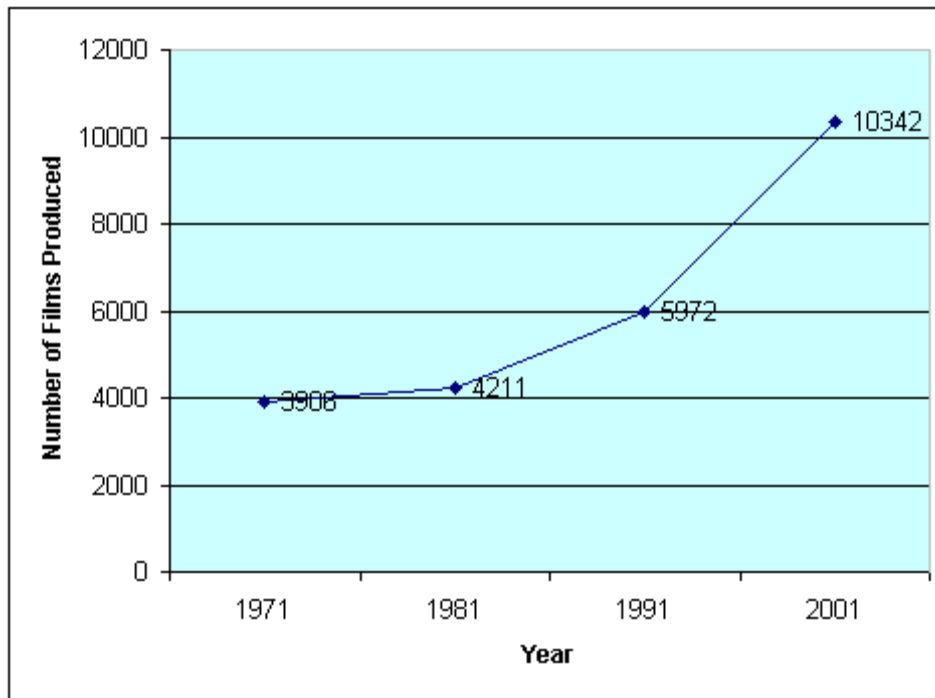
The yearly data for film production is difficult to assess, but 10,342 film titles and video were released worldwide in 2001 according to the Internet Movie Database (IMDB). This figure includes a significant number of television productions as well. Of these, a fair percentage were "short" films (approximately 20%), which is a significant change from previous years, indicating the increasing accessibility of the film medium to small filmmakers. This is evident from the significant increase in total film production during the past ten years, as compared to previous decades (see Table 3.7).

Table 3.7: Worldwide Film Production

Year	1971	1981	1991	2001
Films Produced	3,906	4,211	5,972	10,342

Source: Internet Movie Database (IMDB)

Note: This table represents **all** film production, including those films shot for television. Other parts of this section deal with only those figures of films that are specifically made for screen release. All the films included here are considered "commercial," i.e., for public release in some form. Films shot by individuals and firms, and restricted to their private consumption are not included in these figures.



This factor also points out the growth in independent filmmaking, and a phenomenal increase in digital video usage. The low cost of digital video shooting and editing equipment makes it much more data-storage-friendly than any past filming medium such as 8mm or Super 8. Also, DV tapes are inexpensive and editing equipment is now compatible with personal computers, making the home-movie a growing mode of stored audio-visual information.

It takes approximately 1 terabyte to store an hour of motion picture images in high-quality archival storage. It takes approximately 2 gigabytes to store an hour of motion picture images in digital form using the MPEG-2 compression standard. In Tables 3.8 and 3.9, we have used the larger storage measure instead of the compressed MPEG-2 standard, as this measure offers comparable numbers to archival storage space that we have estimated for photographs and X-rays.

Table 3.8: Annual title production of motion pictures

Type	Number	Multiplier*	Min. / Unit	Total Hours	Total Space (PB)
Feature Films					
Full length feature	3,851	1	90	5,776.5	6.08
Short	1,809	1	10	301.5	
Total:	5,660			6,078.0	
Made for TV	1,687	1	90	2,530.5	2.53
TV Series					
Soaps / Long-running Shows	55	250	30	6,875.0	
Sitcoms / Medium-range	80	30	30	1,200.0	

Mini Series	10	3	60	30.0	
Other TV Series	1,210	10	30	6050.0	
Total:	1,355			14,155.0	14.16
Direct to Video Movies	1,660		90	2,490.0	2.49
GRAND TOTAL	10,432			25,253.5	25.30

Source: Raw data from International Film Index 1895-1990. Chart, How much information 2003

* The Multiplier refers to the "number of units per title" – thus, we estimate that the average soap opera has about 250 shows a year, the average sitcom has 30. The remaining shows are undocumented – but around the world, it can be estimated that several of them run multiple episodes while a few are short-lived. We have therefore used 10 as the average.

2. Accumulated Stock

The number of motion pictures made around the world from 1890 to 2002 was approximately 368,530, according to *The International Film Index, 1895-1990*. The overall total is broken down into these categories of film types:

Table 3.9: Accumulated stock of motion pictures – 2003 sources

Type	Number	Multiplier*	Min. / Unit	Total Hours	Total Space (PB)
Feature Films					
Full length feature	226,771	1	90	340,156.5	
Short	57,825	1	10	9,637.5	
Total:	284,596			349,794.0	349.8
Made for TV	34,540	1	90	51,810.0	51.8
TV Series					
Soaps / Long-running Shows	419	1000	30	209,500.0	
Sitcoms / Medium-range	698	100	30	34,900.0	
Mini Series	4,197	3	60	12,591.0	
Other TV Series	22,148	10	30	110,740.0	
Total:	27,462			367,731	367.7
Direct to Video Movies	7,277	1	90	10,915.5	10.9
Live Action Video Games	553	1	60	553.0	0.5
SUBTOTAL	368,530			780,803.5	780.7

LESS (films lost to nitrate atrophy)	40,000	1	60	40,000.0	40.0
GRAND TOTAL SURVIVING ORIGINAL STOCK:	328,530			740,803.5	740.7

Source: Raw data from International Film Index, 1895 - 1990. Chart, How much information 2003

Also factored (scattered through categories) are the following:

- Animation Films and Series: 15,790
- Documentary Films: 30,475
- Silent Films: 49,417
- Black and White Films: 113,992
- Color: 254,538

UNIVERSE OF FILMS

The total production of 2001 represents 3.3 percent of the entire universe of existing original film and video production.

If the entire universe of available original film and video creations were played continuously, it would continue for 2,108 years.

It is very difficult to estimate the exact atrophy rate of nitrate film. All nitrate film eventually decays; the only way to ensure its survival is to transfer the information to another format. The conversion to acetate film in the 1950s marked a major preservation breakthrough. The rate of preservation of films since the 1950s has been excellent.

The survival of films from the Silent Era is particularly low -- only about 20 to 30% of the features made before the onset of sound still survive.

Table 3.10: Film Survival Among Silent U.S. Films

Year	Films Made	Films Surviving	Percent Surviving
1912	5	3	60.00%
1913	63	6	9.50%
1914	340	51	15.00%
1915	594	93	15.70%

1916	838	152	18.10%
1917	937	186	19.90%
1918	832	98	11.80%
1919	768	124	16.10%
1920	735	154	21.00%
1921	710	164	23.10%
1922	684	137	20.00%
1923	590	135	22.90%
1924	645	184	28.50%
1925	769	271	35.20%
1926	727	269	37.00%
1927	681	234	34.40%
1928	641	201	31.40%
1929	272	103	37.90%
1930	35	11	31.40%
TOTAL	10866	2576	23.71%

Source: Oksana Dykyj, Association of Moving Image Archivists, March 2003

If the question were how much of the film survives in the "film format" this number would be even lower as a lot of the surviving footage survives because it has been transferred into other mediums. Also, information on the archival collection of cinema in developing nations is very limited, but for this project, we have assumed more or less the same rates of survival for films outside of the United States and Europe.

However, it is known that there have been losses of major film collections, including several seminal works, both due to political activities and due to lack of attention to archiving. European and American film industries had established archives by the 1930s but most third world countries joined the bandwagon much later. The first Asian nation with film archives was Iran, in 1949. Thereafter India made its own archives in 1958 and China in 1964. Interestingly, Hong Kong and Japan, with very prolific film collections, did not have organized archives till the 1970s. The Philippines established its national film archives in 1982.

The total loss of film is hard to estimate, partly due to warfare and destruction of (or lack of access to) records in South East Asia, China and Middle East-Central Asia region. The Indian film archives, the largest non-western repository, is also among the most notorious at preservation; less than 10% of all films made before 1931 (some estimate as low as 1%) of Indian films still exist. Among the films lost is India's first talkie. Interestingly, while losses in India have been largely due to fire (most recently in 2002 due to faulty air conditioning in the film archives), there has also been significant loss through producers willfully destroying prints in the past, unmindful of the preservation value. This kind of loss is comparatively smaller in the United States and Europe due to record-keeping by agencies such as the Library of Congress.

B. Copies of Motion Pictures

The number of prints of films has changed drastically as have the sizes of audiences and the sizes of cinema hall facilities. In the 1910's, 40 prints would have been the average number of copies for a major release. By the 1920's, 160 prints would be considered a "big order," according to the Association of Moving Image Archivists (AMIA).

According to the Wolfman Report on the Photographic and Imaging Industry in the United States, the average number of prints per original motion picture was about 700 in the 1980s. In contrast, the Silver Institute reports up to 6,000 release prints may be released for feature movies. The average mass release film in the United States tends to have around 3,000 prints. This phenomenal increase in the number of copies is likely to be a result of the multiplex system of film theaters in the United States.

A number of films made, however, do not secure mass releases and are restricted to limited screenings and film festivals. Such films may have fewer than 50 prints. Indian films, a very large segment of the yearly film copy market, have an average of 300 prints in circulation per film because of the much larger size of cinema viewing halls, and the lack of "multiplexes" for shows on several screens.

A figure of 500 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.

A trend in motion picture information storage is likely to be the use of satellite broadcast for screening technology. This has the potential of doing away with the print copying process. At present, film release prints are largely destroyed after the major screening process is over. Firms exist that make a living by destroying release prints, reclaiming the silver, and turning some of the film into leader they can resell.

Table 3.11: Annual Production of Motion Picture Copies – Worldwide

Type	Number	Avg. Prints	Min. / Unit	Total Hours	Total Space (PB)
Feature films					
Full length feature	3,851	500	90	2,888,250	2888.2
Short	1,809	50	10	15,075	15.1
TOTAL	5,660			2,903,325	2903.3

Source: Wolfman Report on the Photographic and Imaging Industry

The "copies" of television films and series are not calculated here, because they are separately accounted for in the [Broadcast](#) section. The copies on video and DVD format are addressed in the [Magnetic](#) and [Optical](#) sections, respectively.

The prints of feature films are usually destroyed after the films have done their rounds. In earlier days, it was not uncommon for films to reach smaller towns in the United States some two years after initial release. By the time these prints were pulled out of service they were usually practically unwatchable. Then as now, as a film's distribution revenue declines, prints are pulled out of service and destroyed, partly also because the storage is very expensive.

The major studios were and still are very good at tracking and destroying prints. Release prints get junked by the thousands every year in the United States.

C. Trends in Motion Pictures

1. The Growth of Film Information Worldwide

The following chart shows the trends in growth of cinema in some of the top film-producing nations worldwide over the last four decades. The most obvious trends are the phenomenal rise in filmmaking in the United States between 1991 and 2001, contrasted with the significant fall in filmmaking in at least three other film producing nations - Italy, the Soviet Union / Russia, and Mexico. However, since our data sources only include major releases, it is likely that more films were produced in all three nations than reported.

	1971	1981	1991	2001
United States	494	506	762	1740
India	431	737	727	1013
France	165	179	247	453
UK	127	71	118	252
Germany	165	142	141	328
Italy	257	157	149	171
Japan	103	101	137	162
Soviet/ Russia	65	132	164	40
Spain	264	114	88	254
Canada	104	127	160	302
Mexico	124	98	87	43
World Total:	3,128	3,066	3,538	5,717

Source: Raw data compiled from IMDB. Chart by How much information 2003. Note: The world film data here differs from the other net figures of films because this includes only those films specifically made for film release - excluding TV films and series.

2. Comparing U.S. film production to Indian film production

A glance at the figures would indicate that the United States has overtaken India as the major producer of film in the world by a significant margin in the last decade. While this has in fact happened, it has a lot to do with the kind of film information represented in these statistics. Understanding these statistics gives us an idea of the massive growth of independent filmmaking in the United States.

This rise in American film production also has a major short-film factor that figures in the growth of American and some European film growth over the film production in India, for instance. In 2001, the United States produced 443 short films that were shot on 16-mm or 35-mm film (as factored into the statistics in Table 3.12 above), and the total number of short films released for the year in the United States alone was 541. For the same year, India had 2 listed short films.

Documentary cinema is the other area of difference. The statistics in Table 3.12 include 4 documentary films from India in the 1,013 films listed for 2001. In contrast, the figures for the United

States includes 187 documentary films, excluding video and television films. Including television and video, the U.S. documentary films released total 431, and several of these were series of documentaries. The funding for independent documentaries and short films in India is extremely low, and the technology is still very expensive for independent filmmakers to invest in. In contrast, the United States has very well-developed markets and distribution channels for independent cinema.

3. The International Spread of United States Cinema

The decline in film production in Mexico and Italy is accompanied by the massive influx of American cinema both in English language and in dubbed / subtitled versions, a market that has grown more recently. More significantly, the television production has increased very significantly in these nations - especially Mexico, offsetting the demand for indigenous film. Similarly, United States cinema has pervaded several markets where the local film industry has not been developed enough. The only major competition for United States films abroad has been Indian films, prints of which make it in dubbed versions to several Southeast Asian nations, and even Africa. However, even in many of the film-importing nations, American films are seen more favorably than Indian films; common evidence of this is that Indian films usually flood the video markets whereas United States films make it to theatrical releases.

4. Film Information Production and the Government

The most interesting statistics on the production of film information is the contrast between filmmaking in the United States and Western Europe with filmmaking in the former Eastern Bloc countries.

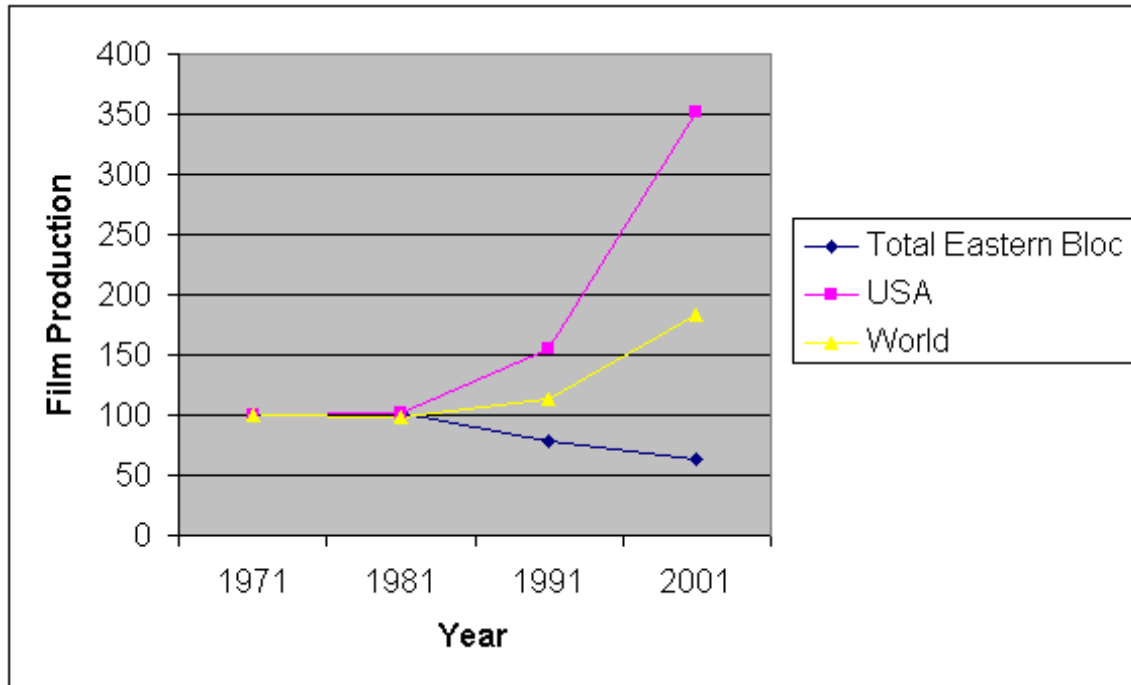
	1971	1981	1991	2001
Soviet U./Russia	66	132	164	63
Czechoslovakia	49	58	27	67
Hungary	30	24	18	48
Poland	44	39	33	29
Romania	45	29	10	10
Former Yugoslavia	153	103	69	46
Bulgaria	21	33	20	12
Albania	29	29	3	2
Total Eastern Bloc	437	447	344	277
United States	494	506	762	1,740
World Total:	3,128	3,066	3,538	5,717

Source: Raw data compiled from IMDB. Chart by How much information 2003. Note: The data excludes East Germany and does not include TV films and series.

Film was predominantly state-funded and sanctioned in most of the communist nations until about 1991. A clear trend exists showing the decline in film production with the end of government monopoly over cinema.

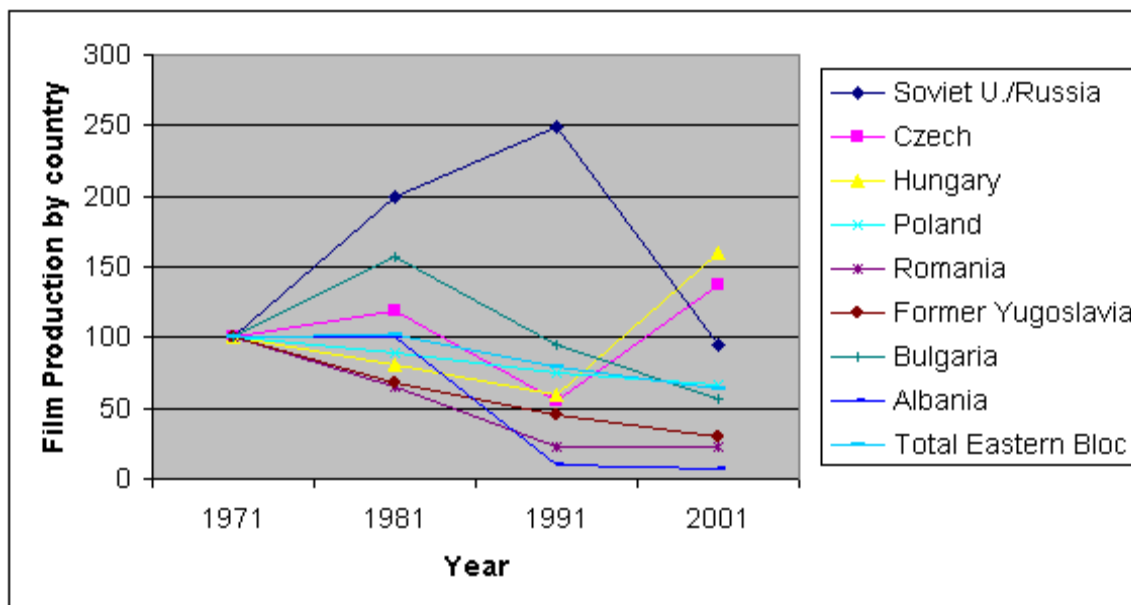
While this does indicate a lesser creation of original film information, the figures do not indicate a diminished flow in film information. On the contrary, the volume of film information in all these nations has increased - the increase mainly being credited to American and West European cinema flooding these markets where they were formerly restricted.

Film Production: Eastern Bloc vs. U.S. (Normalized to 1971 production figures)



Source: Raw data compiled from IMDB. Chart by How much information 2003.

Film in the Eastern Bloc (Normalized to 1971 production figures)



Source: Raw data compiled from IMDB. Chart by How much information 2003.

Only two countries, Hungary and the Czech Republic, were able to increase their film production from the levels attained in 1971 and 1981. Of these, the Czech film industry has traditionally been more independent of the state than in other Eastern Bloc nations. The most precipitous falls have been in Russia, which is estimated to have lost at least 50% of its production between 1991 and 2001. In

smaller countries like Albania and Romania, the film industry has become a fraction of what it was under communist governments. However, the decrease in film-making in many of these nations has been accompanied by a tremendous rise in the local production of TV programming, especially soap operas.

D. Film Trivia

Print Information Destruction

When sound came in (and after all the small-town theaters were wired for sound or closed in the early 1930's) all of the silent prints still in major studio exchanges (distributorships) were junked. Some independent exchanges kept prints of state's rights pictures or of pictures on which they could not trace ownership, but they would often sell these prints to traveling showmen or private collectors (if they could) for a couple of dollars a reel in the 1930's. With no perceived commercial value, the studios did little to maintain their silent libraries except to make periodic inspections and throw out reels that showed signs of nitrate decomposition. Fire was also a problem. The Lubin vault fire in 1914 destroyed all pre-1914 Lubin negatives as well as the negatives on a number of Bosworth and Lasky releases for which Lubin did the printing. Fox had a big fire in its East Coast vaults in 1937, destroying virtually all of the studio's pre-1935 negatives and finegrain masters. Universal destroyed virtually all of its silent negatives, finegrains and studio prints in 1948 to free up vault space. Frances Goldwyn (Goldwyn's wife) ordered all of the silent Samuel Goldwyn Productions destroyed for insurance purposes because she believed they had no value. (Source: Oksana Dykyj, Association of Moving Image Archivists)

Information Censorship!

Interestingly, of all the western nations, the countries to ban the most films are Finland, Sweden and Norway. These three nations are also among the highest producers of adult material, much of which is in turn banned in other nations. Eastern bloc, other communist nations, religious states and nations with media-restrictive policies have been excluded from this calculation. (Based on calculations on unsorted data in the Internet Movie Database.)

Information Chart-Toppers

Top languages of film production, based on calculations from unsorted data on the Internet Movie Database:

1. English
2. French
3. German
4. Spanish
5. Hindi
6. Italian
7. Japanese
8. Tamil
9. Telugu
10. Russian
11. Cantonese
12. Malayalam
13. Mandarin
14. Bengali
15. Dutch

III. X-RAY FILM

A. Original Information Stored on X-Rays

Annual Production of Originals

World. The third 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 American College of Radiology estimates that 10 MB of storage capacity is required for concerting a conventional radiograph to a storable digital format; to convert all radiographs to digital form would require **20 petabytes** of storage each year. Due to the heavy digital storage space requirement for X-rays, they will form a significant storage chunk in the terabyte DVD-RAM libraries market, according to a recent article in Unisys World.

United States. According to an IBM report from 2001, the 2000 American hospitals generate 7 terabytes of data each year from X-rays, mammograms, MRIs, cat-scans, endoscopies and other procedures.

Accumulated Stock

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 20 billion images or **200 petabytes**.

B. Copies of Information Stored/Published on Film

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

REFERENCES

- Ho, Sam, Fragile Heritage and Promising Outlook: Asian Film Archives Look Ahead While Looking Back, *Journal of Film Preservation* N° 62, 2001
- Goble, Alan. *International Film Index 1895-1990*. London: Bowker, 1991.
- [Internet Movie Database](#)
- [Photo Marketing Association International](#)
- Silver Institute
- UNESCO Statistical Yearbook
- United States Industrial and Trade Outlook 2000

Film

- Wilson, Kirk, "Terabyte DVD-RAM libraries play pivotal role in health care for price of DLT tape" in Unisys World
- Wolfman Report on the Photographic and Imaging Industry

Special thanks to:

- Oksana Dykyj, Concordia University (and the Association of Moving Image Archivists) for her extremely valuable assistance on film
- Dimitrios Delis, PMAI, for his cheerful assistance with photography Information
- Michelle Slaughter, InfoTrends Research Group, for her prompt assistance with Digital Camera information

Release date: October 27, 2003. © 2003 Regents of the University of California

How Much Information? 2003

Summary

Exec Summary

Stored Information

Paper | Film | Magnetic | Optical

Information Flows

Broadcast | Telephony | Internet

Wrap-up

Thanks | Printable (PDF)

Magnetic

[I. Video Tape \(Analog\)](#)

[II. Audio Tape \(Analog\)](#)

[III. Digital Tape](#)

[IV. Mini DV](#)

[V. Floppy Disks](#)

[VI. Zip Disks](#)

[VII. Audio MiniDiscs](#)

[VIII. Flash Memory](#)

[IX. Hard Disk Drives](#)

[Fun Facts](#)

[Notes](#)

[References](#)

[Printable Section \(PDF\)](#)

4. MAGNETIC STORAGE MEDIA

Magnetic storage encompasses a wide range of media, from mini cassette tapes to terabyte-sized file servers.

Some new media types have emerged as major storage categories in the last 5 [???] years, including digital videocassettes, audio MiniDiscs, and flash memory. Meanwhile hard disk drives are finding new applications in consumer products such as personal video recorders and video game boxes.

Other magnetic storage media have declined since the 2000 study but remained in demand, despite predictions to the contrary. For example, relatively old-fashioned tape mechanisms still play a major role in safeguarding large amounts of data. PC manufacturers like Dell plan to eliminate floppy disk drives from their personal computers, yet worldwide demand for floppy disks has not disappeared entirely.

In this section we review the production statistics for magnetic media and we estimate the percentages of original information stored on each.

Overall, we estimate that the amount of original data stored annually on magnetic media has increased dramatically—it has nearly doubled since our 2000 study.

Table 4.1: Annual production of original items for the major magnetic media types— 2003 Sources

Media Type		Unique Items per Year	Conversion Factor	Total Annual Petabytes (World)
I	Video tape (VHS and camcorder)	220,000,000	4 GB per tape (MPEG-2 compression)	1,340
II	Audio tape (analog)	128,800,000	1 GB per tape (CD audio format - no compression)	128.8
III	Digital tape	5,000,000 [???	Varies	250
IV	Mini-DV	115,000,000	11 GB per tape	1,265
V	Floppy disks	55,000,000	1.44 MB per disk	0.08
VI	Zip disks	1,400,000	250 MB per disk	0.35
VII	Audio MiniDisks	10,450,000	160 MB per disk	1.7
VIII	Flash memory	43,000,000	Varies	1.2
IX	Hard disk drives	43,928,000	Varies	1,986
Total:				5187.13

Source: *How much information 2003*

Table 4.2: Comparison of production of original information for the major magnetic media types - 2000 sources vs. 2003 sources

Media Type	% change	Year	Unique Items per Year (World)	PB per Year (World)
Video tape (VHS and camcorder)	-6%	2000	355,000,000	1,420
		2003	220,000,000	1,340
Audio tape (analog)	-30%	2000	184,200,000	184.2
		2003	128,800,000	128.8
Digital tape [???	0	2000	5,000,000	250
		2003	5,000,000	250
Floppy disks	-27%	1999	75,000,000	0.11
		2003	55,000,000	0.08
Zip disks	-68%	2000	4,400,000	1.69
		2003	1,400,000	0.35
Flash memory		2000	N/A	N/A
		2003	43,000,000	1.2
Hard disk drives	114%	2000	39,918,000	926
		2003	43,928,000	1,986

I. VIDEO TAPE (ANALOG)

Analog video tapes include the standard VHS tapes using in video cassette recorders (VCRs) as well as the 8 mm video camcorder tapes.

The main use of blank VHS tape is the consumer's use to record television programs. It is estimated that a very large share of the users of VCRs do so for time-shifting of viewing programs. A challenge to this market comes from new time-shifting devices employing hard disk drives—personal video recorders (PVRs) such as TiVo and Replay. The market for such drives is currently small, but growing fast, according to a report released in 2002 by market research company InStat/MDR. The company predicts that the market for PVRs will jump from 1.2 million units in 2001 to over 6 million in 2003; this trend along with the popularity of DVD recorders is likely to cause further declines in the popularity of the VHS tape.

A. Original Information Stored/Published on Video Tape

1. Annual Production of Titles

World. In 2003, 1.1 billion blank VHS video tapes will be produced for the entire world, according to the International Recording Media Association. This statistic is confirmed by the Japanese Recording-Media Industries Association (JRIA). If all 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 video tape, therefore, provides storage space adequate for 4,400 petabytes of data. The vast majority of this is used for recording television programs; a minority is used in video cameras.

Assuming twenty percent [???] of this tape is used for the storage of original data, the annual flow of new data stored on analog VHS video tape would be **880 petabytes**.

As of 2002, video camcorder tapes (all formats except VHS) are consumed at the rate of 230 million per year worldwide, according to JRIA. Almost all of this tape is used for the storage of original data. Assuming one hour per tape in MPEG-2 format yields **460 petabytes** annually.

The production of video tapes has declined by 21 percent since 2000, when 1.4 billion video tapes were produced for the world. In contrast, the production of video camcorder tapes has increased by 50 percent since 2000.

United States. In 2003, 310 million blank VHS video tapes will be produced for the entire world according to IRMA.

2. Accumulated Stock

World. About 8 billion blank video cassettes have been sold since 1998, according to IRMA. If twenty percent [???] of this tape is used for the storage of original data, the total accumulated stock would be **6,400 petabytes**.

One billion camcorder format tapes have also been added over the same period. This would be equivalent to 3 billion hours of stored original video which, if digitally encoded, would produce a stock of about **6,000 petabytes** of original video taped data.

B. Copies of Information Stored/Published on Video Tape

1. Annual Production of Titles

World. In the year 2003, 1.2 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 video tape would consume **4,800 petabytes** per year.

If 80 percent [???] of the blank video tape distributed per year were used for copies, this would constitute a digital equivalent of **3,520 petabytes**.

The total yearly world production of copied data on analog video tape, therefore, is **8,320 petabytes**. This represents a 30% decrease from our 2000 study; significantly fewer copies are being stored on analog video tapes.

2. Accumulated Stock

World. Since 1998, about 9 billion prerecorded videocassettes have been sold worldwide, according to IRMA. If converted to digital using MPEG-2 compression, prerecorded video tapes would consume **36,700 petabytes**.

IRMA also reports that about 8 billion blank video cassettes have been sold since 1998. If 80 percent [???] of this tape is used for the storage of copies, the total accumulated stock of copies would be **25,700 petabytes**.

Therefore the total world stock of copied data on video tape is **62,351 petabytes**.

II. AUDIO TAPE (ANALOG)

The distribution of prerecorded music is one of the most common uses of analog audio tape cassettes. 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.

The International Recording Media Association (IRMA) no longer tracks audio cassette manufacturing because, according to the organization, this medium is a dying business. According to the Japanese Recording-Media Industries Association, "the global blank audio cassette market in 2003 continues to contract significantly by 16% to 644 million units. The decline in demand is largely attributed to the shift to digital processing in the audio recording technology resulting in the growing popularity of alternate media such as the MiniDisk and CD-R for audio use." Worldwide, the compound annual decline rate for duplicated audiocassettes is 10 percent.

In North America, the rate of decline was only 5 percent, due in large part to the growth of "books on tape" (or audiobooks). This segment, with over 726 million cassettes dubbed, accounted for more than 77 percent of all audio cassettes duplicated in North America in 2000, according to IRMA. In 2001, sales of audiobooks in the United States reached 801 million units and increases about 10 – 12% each year, according to the Audio Publishers' Association (APA).

At the same time, the use of CDs for audiobooks increased dramatically from 2000 to 2001. In fact, the APA reports that the average number of hours per week audiobooks are listened to on CD currently almost equals the use of audiobooks on cassettes. This may perhaps be attributed to a shift in car audio systems: audiobook listeners spend most of their time listening to books in the car and 64% of new cars are sold with CD players (as opposed to 42% sold with cassette players).

Even with these declines, in 2002 the analog audio tape cassette was one of only four recording formats with production of over one billion units. According to Magnetic Media Information Services (MMIS), "demand for cassettes has long since peaked, but it remains strong despite the advanced age of the format" (it was first introduced in 1964). The biggest market for music cassettes is Asia, which represents 48% of all sales. Other markets with potential for growth include Latin America, the Middle East, Russia, Eastern Europe, and India, according to Optical Disk Systems magazine.

No one really knows just how large the pirated music cassette business was in 2002 (or any other year), but it is probably close to or even over one billion pieces last year. Much of the demand for these pirated music cassettes is in Asia, and especially in China, Vietnam, Cambodia, Indonesia, and the Philippines, according to MMIS.

A. Original Information Stored/Published on Audio Tape

1. Annual Production of Titles

In 2003, 644 million blank audio tape cassettes will be produced for the entire world, according to the JRIA. If

each tape was filled to its 120 minute capacity and then converted to digital using the common CD audio format, there would be approximately 1 gigabyte of data per tape, for a total of **644 petabytes**. Assuming 20 percent [???] of this tape is used for the storage of original data, the flow of new data stored on analog audio tape would be **128.8 petabytes**.

The flow of audio tape cassettes has declined by 30% since 2000, when 921 million blank cassettes were produced annually. This corresponds to an increase in the use of optical recording media (See [Optical](#) section).

2. Accumulated Stock

The total stock of original audio content stored on tape may be estimated by assuming 20 [???] percent of five years' production of blank cassette tapes contains original content. The digital equivalent of this audio information is **1,000 petabytes**.

B. Copies of Information Stored/Published on Audio Tape

1. Annual Production of Titles

World. According to the Syndicat de l'Edition Phonographique, worldwide sales of prerecorded audio cassettes totalled 800.9 million units in 2000. If converted to digital using audio CD format, and assuming about one hour of music per tape, this would result in **801 petabytes** of data.

If 80 percent [???] of the 644 million blank audio tapes distributed per year worldwide were used for copies, this would constitute a digital equivalent of **515 petabytes**.

The total yearly production of copied data on analog audio tape, therefore, is **1.3 exabytes**. This represents a 30% decrease from our previous study; significantly fewer copies are being stored on analog audio tapes.

United States. In the year 2000, 76 million prerecorded audio cassette tapes were distributed in the United States according to the Recording Industry Association of America. This is less than 10% of the world sales of prerecorded audio cassettes. As of 2002, the figures are even lower: only 31 million tapes sold in the United States, equivalent to **31 petabytes**.

2. Accumulated Stock

The world production of blank audio tape for the past five years is approximately [???] billion units. During that period, there has also been sales of prerecorded audio tapes of 434 million units. Assuming that 80 [???] percent of the blank audio tape has been used for the storage of copied data, the overall stock of copies of audio data on tape is [???] petabytes on blank tape and **434 petabytes** on prerecorded tape. Therefore, there is a total stock of [???] petabytes of copies of audio data on magnetic tape.

III. DIGITAL TAPE

There are 25 million computer tape drives installed in the world at present. These drives provide storage capacity for a range of computers - from desktop personal computers to the most mammoth supercomputers. Storage expert Fred Moore, of Horison Information Strategies, 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 almost US\$1 billion of computer tape media sold each year worldwide. Digital tape types include super tape drives with more than 100 GB capacity such as SuperDLT, LTO (Linear Tape Open), and AIT (Advanced Intelligent Tape), as well as lower-capacity tape formats such as DAT (digital audio tape), QIC (quarter-inch cartridge tape), 8mm, Mammoth, DLT (Digital Linear Tape), ADR, and VXA. The super drive formats are in greater demand than the desktop and entry level drives, as organizations gradually shift from desktop backup to network backup.

There are two sides in the debate over the status and future of digital tape. On one side, there are those who view tape as a throwback technology, advocating instead newer disk-based technologies for backup and archiving. The main advantage of these systems is speed and flexibility: restoring from tape requires a minimum of one hour per terabyte while with disk-to-disk mirroring, one can quickly switch over from failed disk to working disk and keep applications

constantly in service. Relatively inexpensive secondary disk storage is gaining a significant foothold in corporate data centers, according to a 2003 survey of more than 1,000 IT managers conducted by Peripheral Concepts consulting firm. Another indicator of tape's decline relative to disk is noted by StorageSearch, a leading portal for enterprise storage buyers since 1998, with over 0.5 million readers. In April 2003 reader pageviews for disk-to-disk backup on STORAGEsearch overtook pageviews for tape backup for the first time.

On the other side are those who point to the advantages of tape storage, including cost and capacity. Costs associated with tape are generally lower than with other storage formats. The Gartner Group claims that the typical cost of enterprise disk storage amounts to US\$110 per gigabyte, whereas tape costs \$11 per gigabyte. Gartner predicts an 80 percent possibility that this disparity will remain until 2007. Analysts at International Data Corp (IDC) insist that disk prices will never approach the levels of tape. "The cost per gigabyte for tape is so low and getting lower with improving capacity," said Martin Wijaya, senior analyst for storage at IDC Asia-Pacific, quoted at ITWorld.com. "It's also more expensive to manufacture a disk that reaches the density level of tape." The densities of digital tape continue to increase. On March 24, 2003, Sony Electronics announced a new series of PetaSite tape library systems based on SAIT drive technology with a native capacity of up to **1.2 petabytes**. Other advantages noted by analysts and advocates include longer lifecycles, high rates of reliability, less vulnerability to hacking, and portability. Lou Hirsh at the NewsFactor Network writes, "The newest tape technologies will become increasingly important for such storage-hungry corporate functions as data mining and archiving. Also facing increased storage needs are corporate and university researchers in disciplines ranging from geophysics to life sciences. Academic studies, in particular, require systems that allow long-term, repeated access to data, with the ability to make multiple copies and easily transport it to various sites."

Despite major advances in disk technologies, tape continues to be the mainstay of most companies' backup and archival efforts. Secondary disk storage typically involves less than one-fifth of a company's total data, said Farid Neema, president and chief executive officer of Peripheral Concepts, as quoted at ComputerWorld.com. Furthermore, only a small percentage of the data backed up on disks does not get moved to tape devices for archiving, Neema said, "Tape remains by far the most popular medium and does not seem to want to go away."

WORLD'S LARGEST DATABASE

In 1997, the largest database was Knight Ridder's DIALOG, a text database, with 7 terabytes of storage, according to SearchDatabase.com.

As of 2002, the world's largest database is at the Stanford Linear Accelerator Center which stores 500 terabytes of experiment data

A. Original Information Stored/Published on Digital Tape

1. Annual Production of Titles

The digital tape shipped in 2002 has a total storage capacity of about 2,500 petabytes [???].

However, in all but the largest computer applications, digital 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 data on magnetic tape is about **250 petabytes**.

250 petabytes is also generally consistent with estimates derived by use of forecasts of producer revenue of around US\$3.4 billion for tape media and an average cost of around US\$11 per gigabyte of tape storage.

2. Accumulated Stock

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 years' worth of original data flow: **1,250 petabytes**.

B. Copies of Information Stored/Published on Digital Tape

1. Annual Production of Copies

If 90% of the computer tape distributed annually were used for copies, this would constitute 2,250 petabytes of copied data for the year 2002. 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.

2. Accumulated Stock

The stock of copied data on magnetic tape may be taken as five years' worth of copied data flow: 11,250 petabytes.

IV. (MINI) DIGITAL VIDEOCASSETTES

Digital video is recorded on a tape called a Mini DV Cassette (DVC) that occupies less than half the volume of a 8mm cassette and holds one hour of video—11 GB of data on a tape 65 meters long. This small cassette size makes possible much smaller and more portable camcorders. Mini DVC represents 50% of the overall camcorder video cassette market, according to the Japanese Recording-Media Industries Association (JRIA).

A. Original Information Stored/Published on Mini DV

1. Annual Production of Titles

Magnetic Media Information Services estimates that DVC digital camcorder mini-cassettes reached sales of 72 million in 2001 and 88 million in 2002. In 2003, JRIA projects that mini-DV will grow by 19% to 115 million units worldwide. If all the tape produced in 2003 were used to store original information, the total storage would be **1,265 petabytes**.

2. Accumulated Stock

Mini DV was first introduced in 1998. However, we only have sales figures for the last 3 years: 275 million units have been sold. This is equivalent to 3,025 petabytes of storage.

B. Copies of Information Stored/Published on Digital Tape

Mini DV is used primarily for the storage of original information.

V. FLOPPY DISKS

Many industry analysts have declared the micro-floppy disk (MFD) format essentially dead. Dell, the world's largest PC manufacturer, has already eliminated the floppy from its higher-priced desktops and will drop the floppy from most models in 2003, citing two reasons: "computer buyers have become more sophisticated about using their computers, and new technology such as portable USB storage devices and CD-rewritable drives have come down in price over the last few years" (reported at ZDNet, Feb. 2003). Many computer users in the United States, the EU, and Japan haven't used a floppy disk in years, still less purchased new floppy disks. Certainly the statistics show a sharp decline in worldwide demand over the past 7 years, from 4.5 billion diskettes at the peak of the market in 1995 to 1.3 billion in 2002.

However, in many parts of the world such as Latin America, China, and India, the MFD remains a widely-used recording medium, according to Magnetic Media Information Services. The demand for MFDs remained above one billion units in

2002, and it is predicted to stay near that level in 2003 as well. For some companies, the production of MFDs is still a sizable and profitable business. Imation, the world's largest producer, accounts for one-third of the world's total floppy production and claims to produce about two million MFDs a day, both for itself and for other companies still active in selling diskettes.

A. Original Information Stored on Floppy Disks

1. Annual Production of Titles

In 2003, JRIA forecasts that the global market for floppy disks will decrease to 1.1 billion disks, down 15% from 2002 (1.3 billion disks) but up 15% since 1999. If each floppy disk can store 1.44 megabytes, this is an aggregate storage capacity of 1.58 petabytes. If [???] five percent of this is original data, new data per year on floppy disk would be **0.08 petabytes**.

2. Accumulated Stock

The useful life of a floppy disk is estimated to be about three years. The amount of original data stored on roughly 4 billion floppies produced over the course of the past three years is around 5 percent of the total data on those disks, or **0.4 petabytes**.

B. Copies of Information Stored/Published on Floppy Disks

1. Annual Production of Copies

If 95 [???] percent of the 1.6 petabytes of annual floppy disk storage were used for copies of data, this would add 1.52 petabytes to the stock of digital data stored on floppy disks.

2. Accumulated Stock

The useful life of a floppy disk is estimated to be about three years. The amount of copied data stored on roughly 4 billion floppies produced over the course of the past three years is around 95 percent of the total data on those disks, or **7.6 petabytes**.

VI. ZIP DISKS

The general trend for low-end disk (capacity of around 100 to 250 megabytes) is downward, according to lomega president and CEO Werner Heid. At the peak of the format's popularity in 1999, lomega shipped more than 11 million Zip drives and 64 million Zip disks. In 2002 lomega Zip disk unit sales were 28.2 million units, representing a decrease of 6.2 million units from 2001. This 18% decline has been attributed to the advent of recordable CDs.

In order to provide the dedicated Zip user with a format that matches the capacity of a recordable CD, in 2002 lomega introduced the Zip 750MB drive. Initially available as an external USB device, lomega claims the Zip 750 "offers better-than-CD recordable performance, with media designed to be rugged and user friendly." However, Zip disks lack the universal compatibility of a CD.

A. Original Information Stored on Zip Disks

1. Annual Production of Titles

lomega Zip disks are primarily used for backup, transfer of files or video or image editing. If each Zip disk sold in 2002 was filled to capacity, that would equal about 7 petabytes. If five [???] percent of this is original data, new data per year stored on Zip disks would be **0.35 petabytes**.

2. Accumulated Stock

Since the Zip drive was introduced in March 1995, lomega has shipped more than 48 million drives and 300 million disks. This translates to 75 petabytes in all, with about **3.8 petabytes** of original data. (The lifespan of a disk, in heavy use, is about 5 years.)

B. Copies of Information Stored/Published on Zip Disks

1. Annual Production of Copies

If 95 [???] percent of the 7 petabytes of annual removable disk storage were used for copies of data, this would add 6.7 petabytes to the stock of digital data stored on removable disks.

2. Accumulated Stock

The 300 million disks shipped since 1995 have a total storage capacity of 75 petabytes. About **71.2 petabytes** of the storage is used for copies of data.

VII. AUDIO MINIDISCS

The audio MiniDisc, developed by Sony, was first introduced in 1991. This format's main features are random access, good sound quality, easy song editing, and compact size. There are two physically distinct types of discs: Premastered MDs, similar to CDs in operation and manufacture, and Recordable MDs, which can be recorded on repeatedly and employ magneto-optical technology. The disc itself is enclosed in a small (7cm x 7cm) cartridge. According to Sony, each disc can be recorded and erased 1 million times.

Confusion over the MiniDisc's intended purpose plus early technical limitations and high cost made its initial acceptance slow outside of Japan. In 1994, Sony sent 1.1 million MD samplers to Rolling Stone subscribers and in 1997, lowered the prices on players and discs. However, the MD market remains a niche market, a presence established primarily in Japan, accounting for 70% of the world demand. It also has a fairly strong and growing presence in Europe. This is one of the few storage media types where there is not a sizable U.S. market—only 5% of the 274 million discs sold in 2001. In 2002, according to the Magnetic Media Information Services, industry analysts expect some 321 million MiniDiscs to be sold worldwide, of which about a third (100 million) will be sold outside of Japan.

A. Original Information Stored on Audio MiniDiscs

1. Annual Production of Titles

Recordable. Japanese Recording-Media Industries Association (JRIA) estimates that the global market for recordable audio MDs will be 209 million units in 2003; this represents a decline of about 5% from 2002. Each MiniDisc can hold 140MB in data mode or 160MB for 74 minutes in audio mode. This results in a total annual production of about 33 petabytes, assuming that most discs are used for recording audio. If five [???] percent of data stored on recordable MDs is original data, this would equal 1.7 petabytes per year.

Prerecorded. Sony is the only manufacturer currently producing titles on premastered MD; there are currently 531 titles in the Sony MD catalog. There are no titles issued on MD that are not also available on CD; therefore to avoid double-counting, we do not report this format under original information.

2. Accumulated Stock

Estimates of worldwide MD sales vary somewhat and we do not have data for every production year. For the years 1999 – 2002, we estimate that a total of 1 billion blank minidiscs were produced. This results in a total storage capacity of about **160 petabytes**. If 5% of this storage is used for original data, this would equal about **8 petabytes**.

B. Copies of Information Stored/Published on Audio MiniDiscs

1. Annual Production of Copies

Recordable. If 95% of data stored on recordable MDs each year are copies, this would equal approximately **31.3 petabytes**.

Premastered. IFPI first reported on international sales of premastered audio MiniDiscs in 1999. At that time, IFPI stated “in total one million units were sold worldwide, the biggest market being the UK where sales totalled just under 500,000. Minidisc sales were also reported in Austria, Belgium, Denmark, Finland, France, Germany, Norway, Sweden and Japan.”

However the following year, in 2000, only 700,000 premastered MiniDiscs were sold worldwide, a decline of

34% from 1999. We do not have production statistics for 2001 and 2002, but the mastered MiniDiscs sold in 2000 equal **0.11 petabytes**.

2. Accumulated Stock

Recordable. If 95% of data stored on all recordable MDs are copies, this would equal approximately **152 petabytes**.

Premastered. The two years for which we have sales data (1999 and 2000) total **0.29 petabytes**.

VIII. FLASH MEMORY

Flash memory is used for information storage in devices such as mobile phones, digital cameras, MP3 players, personal digital assistants, tablet PCs, home video game consoles and even the Aibo robotic dog from Sony. Here are some examples of flash memory:

- A computer's BIOS chip
- CompactFlash (most often found in digital cameras)
- SmartMedia (most often found in digital cameras)
- Memory Stick (most often found in digital cameras)
- PCMCIA Type I and Type II memory cards (used as solid-state disks in laptops)
- Memory cards for video game consoles

Intel invented flash memory in 1988. In April 2003, Intel said it had shipped over 2 billion flash memory units. It took Intel 12 years to ship its first billion discrete flash units and only three more years to ship the next billion.

In 2002, the largest flash memory manufacturers after Intel were Samsung, Toshiba, AMD (Advanced Micro Devices) and Fujitsu.

According to Techweb (March 11, 2003), flash memory sales in 2002 were driven by continuing demand for digital still cameras and for data storage applications, such as other products, like MP3 players. At the same time, demand for the flash memory used in cell phones and PCs decreased sharply.

Overall, flash unit shipments are expected to increase 21% to 1.8 billion units in 2003, according to IC Insights.

The capacity of flash memory units varies; most memory cards today are 200 megabytes or less, but high-capacity versions just released by SanDisk can hold 2 GB or 4 GB of data.

SMART CARDS

Smart cards currently in deployment have a capacity ranging from a few kilobytes to 224 MB. There is an annual growth rate of about 28% in the number of smart cards produced and sold yearly. There may be anywhere between 4 and 10 billion smart cards in the world by 2006. This figure might get a huge boost if cards that do not need separate readers get mass produced (thus the card plugs directly into the Universal Serial Bus (USB) port on a PC via a passive adapter). The reason this is important is that credit card companies will then shift to smart cards from magnetic cards. Currently, the amount of information on your average magnetic strip credit card is negligible, partly because there is greater reliance on the network (in which case we have already accounted for that in our [Magnetic](#) statistics) - this dependence changes with chip-enabled cards. In the near future the government of China plans to have all its citizens carry smart cards, as do the Philippines, Cambodia and several nations in the Persian Gulf as well. The U.S. government alone plans to buy several million smart cards (the Department of Defense has already implemented the first phase for its own employees). In the future, when biometric information is stored on these

cards, the original data found on this medium will be fairly sizeable.

A. Original Information Stored on Flash Memory

1. Annual Production of Originals

Currently, the majority of original information stored in flash memory is found in digital cameras; digital cameras account for about two-thirds of all shipments of flash memory cards, according to IDC figures reported at CNET.com. In 2002, about 27 million digital cameras were purchased worldwide. The average memory per card as of 2002 is 42 megabytes, but by 2006 the average will increase to about 83 MB per card. If the memory in every camera were filled to capacity, the annual flow of original information would be about **1.1 petabytes**, all of which is original.

High-end cell phones enabled to store data and images also use flash memory for original information. Roughly 16 million such phones were sold worldwide in 2002, the vast majority (96%) in Japan and Korea. If each phone has an average of 128 megabits of flash memory, this adds up to about **0.1 petabyte** of data.

2. Accumulated Stock

As of 2002, there are about 53 million digital cameras in the world. This represents a total possible capacity of about **2 petabytes**, all of which is original.

These results also allow us to estimate the maximum number of photographs stored on digital cameras. If a 42 MB card can store, on average, 80 photographs at a resolution of 2 megapixels, then the total potential stock is **4.2 billion** digital photographs.

B. Copies of Information Stored/Published on Flash Memory

1. Annual Production of Copies

Flash memory is used for storage of copies on MP3 players. The average amount of memory that comes with a player is 64 MB. About 7 million MP3 players were sold in 2002, for a total possible capacity of about **0.4 petabytes**.

2. Accumulated Stock

About 30 million MP3 players have been sold worldwide since the technology was introduced in 1999. With an average of 64 MB of flash memory per player, this equals a total of about 2 petabytes of copied data.

IX. HARD DISK DRIVES (HDD)

Worldwide shipment of hard drives dipped to 196 million in 2001 and rose to 213 million in 2002. Worldwide PC unit shipments grew by 2.7 percent in 2002 from 2001. In spite of two consecutive quarters of growth, Gartner reports the market has yet to show evidence of a significant upturn.

To estimate total capacity, we look at the hard drive sales statistics in combination with the storage capacity per drive. In 2002, **10.85 exabytes** of hard drive storage were sold.

Table 4.3: Annual production of hard disks: units shipped and total storage capacity

Year Disks	Sold (Thousands)	Storage Capacity (Petabytes)
1992	42,000	
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	200,000 (IDEMA)	4,630.5
2001	196,000 (Gartner)	7,279.14
2002	213,000 (Gartner projection)	10,849.56
2003	235,000	15,892.24
TOTAL	1,519,527 (1.5 billion drives)	41,402.73 (41 exabytes)

Source: Raw data, various sources. Chart, How much information 2003

A. Original Information Stored on Hard Drives

1. Annual Production of Titles

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 two major categories:

- **Single-user computers** (personal computer, laptop, or workstation). This type of computer is responsible for approximately 90% of the computer disk storage capacity currently shipped.
- **Servers** (typically used with mainframe or network file servers). This class of computer is responsible for about 10% of the overall storage market.

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 are 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 20 gigabytes. 200 megabytes of original data constitutes 1% of disk capacity, which is the estimate for this category of computer disk.

Servers are 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%.

Table 4.4: Total Storage Capacity and Original Info on Hard Drives, 2002

Drive type	Est. Units Shipped 2002	Total Storage (in TB)	Est. % Original	Total Original (in TB)
Single-user (PC, laptop, and workstation)	200 million	9 million TB	1%	90,000 TB
Server (midrange and enterprise)	20 million	0.9 million TB	35%	313,000 TB

	220 million	9.9 million TB		403,000 TB
--	-------------	----------------	--	------------

Source: Raw data, various sources. Chart, *How much information 2003*

Using these calculations, approximately **403 petabytes** of the total hard disk drive capacity each year is used to store original (user-created) data.

If, instead of assuming 1% original information on PCs and 35% on servers, we estimate 20% original across all hard disk storage, we arrive at an upper bound of **1,986 petabytes**. Clearly, the factor we use in the "% original" estimate has a strong effect on the final result. We hope our investigation of how individuals use storage on their personal computers will provide a more solid foundation for this tentative calculation.

2. Accumulated Stock

Over the past 3 years, hard disk capacity of about 23,000 petabytes has been produced. If 20 [???] percent of that capacity has been used to store original content, the stock in that format is now **4,600 petabytes**.

GENOMIC DATA

One growing source of original information that human genomic sequence information. As of 2002, there are 22 billion sequences stored in the GenBank; this number has grown exponentially since it was started in 1982. If each sequence occupies 10 bytes, this currently represents **220 GB** of information (and growing).

B. Copies of Information Stored/Published on Hard Drives

1. Annual Production of Copies

If 80 [???] percent of the 10,849 petabytes of annual hard disk storage were used for copies of data, this would add **8,679 petabytes** to the stock of digital data stored on hard disks.

2. Accumulated Stock

Over the past 3 years, hard disk capacity of about 23,000 petabytes has been produced. If 80 [???] percent of that capacity has been used to store original content, the stock in that format is now **18,400 petabytes**.

NEW USES FOR DISK DRIVES

Applications for hard disk storage have extended beyond conventional computing uses. As reported in IDEMA Insight Magazine, IDC expects over 50 million drives to ship into consumer applications, representing about 13 percent of all HDDs shipped in 2005. New categories for disk drives include:

- Personal video recorders (IDC expects over 1.5 million units will be sold this year and grow to over 25 million in 2005)
- Video game boxes
- Portable digital music players that integrate hard disks. (Led by models like the iPod and the Rio Riot, worldwide unit shipments are forecasted by In-Stat/MDR to increase from 230,000 in 2001 to over 950,000 in 2003, as reported by MacCentral.)

Interesting Facts about Magnetic Media

- As of February 2003, the cost of disk-drive capacity has dropped below US\$1 per gigabyte. (Reported at

SiliconValley.com)

- Hard drives now come in packages almost as small as a quarter. IBM Corp.'s 1-gigabyte Microdrive holds the equivalent of 700 floppy disks in a half-ounce, one-inch package. Credit-card sized hard drives in laptops can now hold 20 GB of data. (Reported in the San Francisco Examiner, July 29, 2002)
- The typical American consumer now generates some 100 gigabytes of data during his or her lifetime, including medical, educational, insurance, and credit-history data, EMC's Rothnie says. Multiply that by 100 million consumers and you get 10,000 petabytes of data. (Reported in Information Week, Feb. 11, 2002)

Notes on Conversion Decisions

Video tape

In making assumptions about the size of analog video tape stores we have chosen to make conversions assuming the use of MPEG-2 video compression standard. In the case of video tape, 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.

Audio tape

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.

REFERENCES

- "Anatomy of a Mini-DV Tape" <http://www.shortcourses.com/video/chapter04.htm>
- Audio Publisher's Association, Overview of Formats and Storage Media
- Audio Publishers' Association, Consumer Study, May 31, 2001, http://www.audiopub.org/fass_pr.html
- CNET July 29, 2002, "Flash Comes Calling on Cell Phones." <http://news.com.com/2100-1040-946939.html?tag=nl>
- ComputerWorld, Vol. 9 Issue No. 26, 9 - 15 May 2003, "The Persistence of Tape," <http://www.computerworld.com.sg/pcwsg.nsf/unidlookup/55B35C679A57DF1C48256D210015D520?OpenDocument>
- Gartner, Jan. 20, 2003, "PC Shipments Indicate Vendors Should be Cautious in 2003." http://www4.gartner.com/DisplayDocument?doc_cd=112642

- IC Insights, Feb. 27, 2003, "Intel Remains Top Flash Memory Supplier," <http://www.icinsights.com/news/releases/press20030227.html>
- IDEMA Insight Magazine, Sept./Oct. 2001, "How Much Relief Will Emerging Applications Provide the HDD Industry?" http://www.idema.org/public/calendar/industry_market_reinsel.html
- InformationWeek, February 11, 2002. "Tower of Power: IT managers brace for the inevitable -- petabyte-size databases," by Rick Whiting, <http://www.informationweek.com/story/IWK20020208S0009>
- International Federation of the Phonographic Industry (IFPI) <http://www.ifpi.org/site-content/press/20000414.html>
- International Recording Media Association (IRMA):<http://www.recordingmedia.org/news/speech-repna-2001.html>
- Iomega First Quarter 2000 Supplemental Information, http://media.corporate-ir.net/media_files/NYS/iom/reports/q100_supp.pdf
- Iomega press release, 2002:http://www.iomega.com/about/prreleases/2002/750_release.html
- Iomega press releases, 2003
- ITWorld, Dec. 9, 2002, "Don't Rule out Tape, Experts Say." http://www.itworld.com/nl/storage_sol/12092002/
- Japanese Recording-Media Industry Association (JRIA) <http://www.jria.org/english/index.html>
- MacCentral, May 23, 2002, "Hard Drives Booming in Consumer Electronics Products." <http://maccentral.macworld.com/news/0205/23.harddisk.php>
- Magnetic Media Information Services, <http://www.mmislueck.com/WhatsNews.htm>
- Magnetic Media Information Services, <http://www.mmislueck.com/Archives/102001.htm>
- News Factor Network, May 16, 2003, "IBM Puts a Terabyte on Tape" <http://www.newsfactor.com/perl/story/17786.html>
- Optical Disk Systems, Sept-Oct, 2002 <http://www.opticaldisc-systems.com/2002SepOct/TAPEDUPLICATION89.htm>
- San Francisco Examiner, July 29, 2002, "Digital World's Motor." <http://www.examiner.com/headlines/default.jsp?story=n.harddrive.0729w>
- SearchDatabase.com http://searchdatabase.techtarget.com/sDefinition/0,,sid13_gci214145,00.html
- SiliconValley.com, Feb. 16, 2003
- StorageSearch, March 4, 2003, " Think before you invest in disk-to-disk backup." http://searchstorage.techtarget.com/tip/1,289483,sid5_gci882897,00.html
- StorageSearch, May 5, 2003, "The Future of High Performance Tape Technology." <http://www.storagesearch.com/sonyart1.html>
- TechWeb, March 11, 2003, "The New World Order in Flash." <http://www.techweb.com/wire/story/OEG20030311S0026>
- TechTV, "Iomega Zip 750 MB: New Zip disk format offers more storage space but is it for you?"

<http://www.techtv.com/screensavers/print/0,23102,3396122,00.html>

- Wireless News Factor, April 11, 2003, "Intel Scales New Heights with Flash Memory Stacks."
<http://www.wirelessnewsfactor.com/perl/story/21249.html>
- ZDNet, Feb. 6, 2003, "Dell Ready to Flip the Floppy." <http://zdnet.com.com/2100-1103-983596.html>

Release date: October 27, 2003. © 2003 Regents of the University of California

How Much Information? 2003

Summary

Exec Summary

Stored Information

Paper | Film | Magnetic | Optical

Information Flows

Broadcast | Telephony | Internet

Wrap-up

Thanks | Printable (PDF)

Optical

[I. CD-Audio](#)

[A. Originals](#)

[B. Copies](#)

[II. CD-ROM](#)

[A. Originals](#)

[B. Copies](#)

[III. DVD](#)

[A. Originals](#)

[B. Copies](#)

[References](#)

[Printable Section \(PDF\)](#)

5. OPTICAL STORAGE MEDIA

For optical media, we focused on the three major industry categories: compact discs (CD) for audio, CD-ROM, and digital video discs (DVD). These media are relatively young: CDs and CD-ROMS have existed for about 20 years, while DVDs originated about 8 years ago. All 3 optical media types are currently thriving. Decline in the production and sale of retail audio CDs has been offset by the growing popularity of writeable CDs (CD-R and CD-RW). Meanwhile, DVDs have achieved the fastest market penetration of any recent technology.

Table 5.1: Hardware Comparison: Years to Reach 30 Million Players Shipped

VCR	13 years
CD	8 years
DVD	5 years

Source: DVD Entertainment Group press release 1/8/02

Annual title production for the 3 optical media types appears in the Table 5.2. There are significantly more audio CD titles produced each year than CD-ROMs or DVDs. However a DVD can store nearly 10 times as much data as a CD, so the totals in terabytes are roughly equivalent for the two media.

Table 5.2: Annual title production of the 3 optical media types – 2003 sources

Media Type	Unique Titles per Year (US)	Unique Titles Per Year (World)	Conversion Factor	Total Terabytes (annual U.S.)	Total Terabytes (annual worldwide)

CD – Audio (2002)	33,443	90,000	0.650 GB per item	22	58.0
CD ROM (2002)	850	1,700	0.650 GB per item	0.55	1.1
DVD – Video (2001)	4,000	10,000	4.38 GB per item	18	43.8
			Totals:	40.55	102.9

Source: *How much information 2003*

Table 5.3 lists the estimated accumulated stock of titles for each optical storage medium. Note that CD-ROMs and DVDs have roughly the same number of titles available, despite the fact that CD-ROMs have been around for 12 years longer.

Table 5.3: Accumulated stock of the 3 optical media types – 2003 sources					
Media Type	Unique Titles (US)	Unique Titles (World)	Conversion Factor	Total Terabytes (U.S.)	Total Terabytes (worldwide)
CD – Audio (2002)	0.5 million	1.5 million	0.650 GB per item	366	975
CD ROM (2002)	15,000	20,000	0.650 GB per item	10	12
DVD – Video (2001)	14,000	20,000	4.38 GB per item	61	88
			Totals:	437	1,075

Source: *How much information 2003*

Tables 5.4 and 5.5 show the slight decrease in CD-audio title production and the sharp increases in CD-ROM and DVD title production since the last iteration of the “How Much Information” study.

Table 5.4: Comparison of title production for 3 optical media types – 2000 sources vs. 2003 sources				
Media Type	% change	Year	Unique Titles per Year (US)	Unique Titles Per Year (World)
CD - Audio	-14%	1999	38,900	105,135
		2002	33,443	90,000
CD ROM	70%	1999	500	1,000
		2002	850	1,700
DVD – Video	33%	1999	3,000	5,000
		2001	4,000	10,000

Source: *How much information 2003*

Table 5.5: Comparison of total size in terabytes for 3 optical media types – 2000 sources vs. 2003 sources			
Media Type	Year	TB per Year (US)	TB Per Year (World)
CD - Audio	1999	22	58
	2002	22	58
CD ROM	1999	0.33	0.65
	2002	0.55	1.1
DVD – Video	1999	13	22
	2001	18	43.8

Source: *How much information 2003*

I. CD-AUDIO

Original Information Stored on Audio CDs

Annual Production of Titles

World. To estimate how many CD-audio originals are created each year worldwide, we use 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 33,443 items per year, as of 2002.

Therefore, by our best estimate, the world produces roughly 90,000 original titles per year, equivalent to **59 TB** (uncompressed). We acknowledge that using market share to estimate production of originals is questionable; these figures serve as placeholders until a better reference is found.

United States. From 1992 to 1999, there was an upward trend for CD title production in the United States. Between 1999 and 2001, title production of CDs in the United States decreased by 18%, falling from 38,900 to 31,734 releases. In 2002 there were 33,443 new releases, which is an increase over 2001 but still represents 14% fewer titles than the peak in 1999. The 2002 releases are equal to about **22 TB** of new information (uncompressed). (Preceding statistics supplied by the Recording Industry Association of America (RIAA) and include re-releases as well as brand new releases.)

RECORDINGS ON VINYL

The vinyl disc (or LP) is the oldest sound recording medium. According to the statistics of the Syndicat National de l'Edition Phonographique (2001), the stock of vinyl discs since 1984 can be estimated to be more than 4.6 billion units. In recent years, other sound recording formats such as the CD have largely replaced vinyl; the market share of vinyl among all sound recordings represents 0.36% (13.5 million units annually out of 3.7 billion units worldwide). If we suppose that the average duration of a vinyl disc is only 30 minutes, the annual flow of content on vinyl discs is thus equivalent to 34,290 TB.

Source: *ENST Report 2003*

Accumulated Stock

With proper handling, CDs should have a life expectancy of 70 – 200 years, according to the Optical Storage Technology Association. However, many CDs do not receive proper handling and many others (e.g. CDs sent out as promotional material) are discarded immediately upon receipt. CD-Rs are particularly vulnerable, because data may only be written to them once. If a mistake is made during the recording process, the erroneous disk must be trashed. About 1 out of 20 CDs are discarded or become unusable for some reason (scratches, problems during recording). Therefore estimates of CD-audio and CD-ROMS in the following section are adjusted to reflect a decay/destruction rate of 5%.

As with other digital media, there may also be equipment-based preservation issues. What is the life cycle of the equipment to read a CD? Technical standards change rapidly -- we may be seeing a shift from CD to DVD right now that would make CD equipment obsolete, or the density might change.

World. Again, in order to estimate the world's stock of titles, we use the U.S. share of the music market (37%) and extrapolate from our U.S. total (564,406). By this method we arrive at an estimate of 1.5 million titles worldwide, equivalent to **975 TB**.

United States. The All Music Guide (a comprehensive database tool used by industry leaders containing mostly U.S. and some U.K. titles) reports a total of 664,008 albums (575,192 popular and 88,816 classical), as of 1/29/2003. Assuming that 85% of these are available on CD (with the rest only on vinyl or cassette), this means that the stock of original CD titles would be about 564,406 equivalent to **366 TB**.

If one adds up the past 10 years of CD releases posted by the RIAA, one arrives at a total of 306,577 titles (2000 data not available); this gives us a lower bound for the stock of CD originals in the United States as about **200 TB**.

CD burners provide another source of original content. Users may duplicate software and CD titles, but they may also create new musical compilations, by selecting from downloaded MP3 files or CDs. In this instance, the individual songs are exact copies, but the combination of songs is unique. In 2001, 3.7 billion blank CD-Rs were sold worldwide; in April 2002, sales of blank CD-Rs exceeded those of pre-recorded music CDs for the first time, according to a recent article in the Sacramento News and Review. Industry estimates say 6 billion blank CDs will be sold worldwide in 2003 – that's one for every person alive today and represents an additional **3,900 PB** of data annually. 140 million people now own writeable drives and another 44 million people will acquire one this year, according to Wired Magazine. Two-thirds (67%) of CD burner owners report using their burner to copy music in some way; the most common uses for CD burners are making CDs for friends and family (54%) and making CD compilations (51%), according to a study sponsored by the RIAA.

Even with the demise of Napster, peer-to-peer music downloading still thrives. The International Federation of the Phonographic Industry (IFPI) estimates that in May 2002 there were approximately three million users and 500 million files available for copying at any one time on all of the peer-to-peer services worldwide. There are approximately 200,000 Web and FTP sites hosting or linking to about 100 million recorded music files; 99% of these files are unauthorized. An RIAA study reports that of the music consumers with Internet access who have downloaded music, 52% have made copies of that music. See the [Internet](#) section for more details on file-sharing and P2P.

RECORDINGS AT THE LIBRARY OF CONGRESS

The Library of Congress (LOC) has some 2.6 million sound recordings, the largest public collection of sound recordings. The LOC collection includes more than 500,000 LPs; 450,000 78-rpm discs; more than 500,000 unpublished discs, 200,000 compact discs; 175,000 tape reels; 150,000 45-rpm discs; and 100,000 cassettes. These include spoken word and radio broadcasts. All these sound recordings together equal about **1,200 TB**.

Copies of Information Stored/Published on Audio CDs

Annual Production of Copies

World. During 2002, according to the International Federation of the Phonographic Industry (IFPI), 2.2 billion audio CDs were sold worldwide, a decrease of 6% since 2001. IFPI attributes this decline to “mass downloading from unauthorised file sharing on the internet and the massive proliferation of CD burning, combined with competition from other entertainment sectors and economic uncertainty on consumer spending.” The worldwide retail audio market represents about **1,500 PB** of data annually.

Another index of flow is found in the replication market, as reported by the International Recording Media Association (IRMA). Replication is the manufacturing of a CD from the raw elements, where the data is physically embedded in the plastic medium. (Contrast this with duplication, where the data is burned onto an already manufactured, blank CD-R.) 4.7 billion CD-audio units and 3.3 billion units of CD-ROM were replicated worldwide in 2002. Add to this the 5.2 billion units of CD-R and we find a total possible flow of about **8,600 PB** in CD copies.

United States. During 2001, according to the RIAA, 881.9 million CDs were sold in the United States. From IRMA, we find that 1.6 billion units of CD-audio and 1.5 billion units of CD-ROM were replicated in 2002. Add to this the 2 billion units of CD-R and we find a total possible flow of about **3,300 TB**. (Note: these statistics are for North America, not just the United States.)

CD PIRACY

Pirated CDs represent a growing category of duplicated information. According to IFPI, in 2001 worldwide sales of pressed pirate CDs were 500 million units, up from 475 million in 2000, with pirate CD-R discs estimated at around 450 million units, up from 165 million in 2000. IFPI estimates that in 2001, 28% of all CDs sold were pirate – up from 20% the year before. The total is split roughly evenly between CD audio discs made on factory production lines and those made in smaller scale CD-R operations in garages and labs.

Accumulated Stock

World. Between 1983 (the launch of the format) and 2001 (the most recent statistics available from IFPI), 23.8 billion CDs have been sold worldwide.

United States. 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’s year-end statistics for CD shipments only date back to 1990—these 13 years of shipments add up to 8.9 billion units. A lower estimate is provided by the ENST Report estimating the U.S. stock of retail CDs at 4.5 billion units.

II. CD-ROM

Original Information Stored on CD-ROM

Annual Production of Titles

World. Between 2001 and 2002, 1,700 new CD-ROM titles were added to CD-ROMs in Print, an international directory published by Gale Research. This figure includes business applications (such as word processing and spreadsheet packages), games, reference tools, and instructional programs. 1,700 titles equals about 1.1 TB of new information in one year (uncompressed). This represents a 70% percent increase from 1999, when CD-ROM titles were added to this directory at the rate of 1,000 per year (.65 TB).

United States. For CD-ROM title production in the United States, we might extrapolate based upon the number of CD-ROM companies in the United States: roughly 50% of the companies listed in CD-ROM's in Print are located in the U.S. Furthermore, the United States has 44% of the CD-ROM replication market, according to IRMA. However, neither of these statistics can be definitively correlated to title production. If we assume the United States produces 50% of the titles, we find that 850 CD-ROM titles are produced each year, equivalent to **0.55 TB** of new information.

Accumulated Stock

World. According to the 2002 edition of *CD-ROMs in Print*, internationally there are more than 20,000 unique CD-ROM titles, equivalent to **13 TB**.

United States. Using the method for estimating the United States share of the CD-ROM title creation described above, we estimate that there are 10,000 CD-ROM titles published in the U.S. However, this number is probably low, given that Macromedia alone has a catalog of 10,000 software titles. Another data point comes from the All Game Guide, which lists 11,871 titles for a PC (Mac or IBM) platform, as of January 29, 2003. (This index includes video games, reference, educational, and other software.) Finally, the CD-ROM Guide Bargain Finder listed a total of 29,000 results—there may be a fair amount of duplication here. A reasonable estimate for the stock of CD-ROM originals is 15,000, with a size of **10 TB**. [???

Copies of Information Stored/Published on CD-ROM

Annual Production of Copies / Accumulated Stock

World. We do not have statistics for the number of CD-ROM copies distributed worldwide. However, relying once again on IRMA's replication statistics, we find that at least 17 billion CD-ROMS have been replicated since 1997, for a total capacity of 11 exabytes.

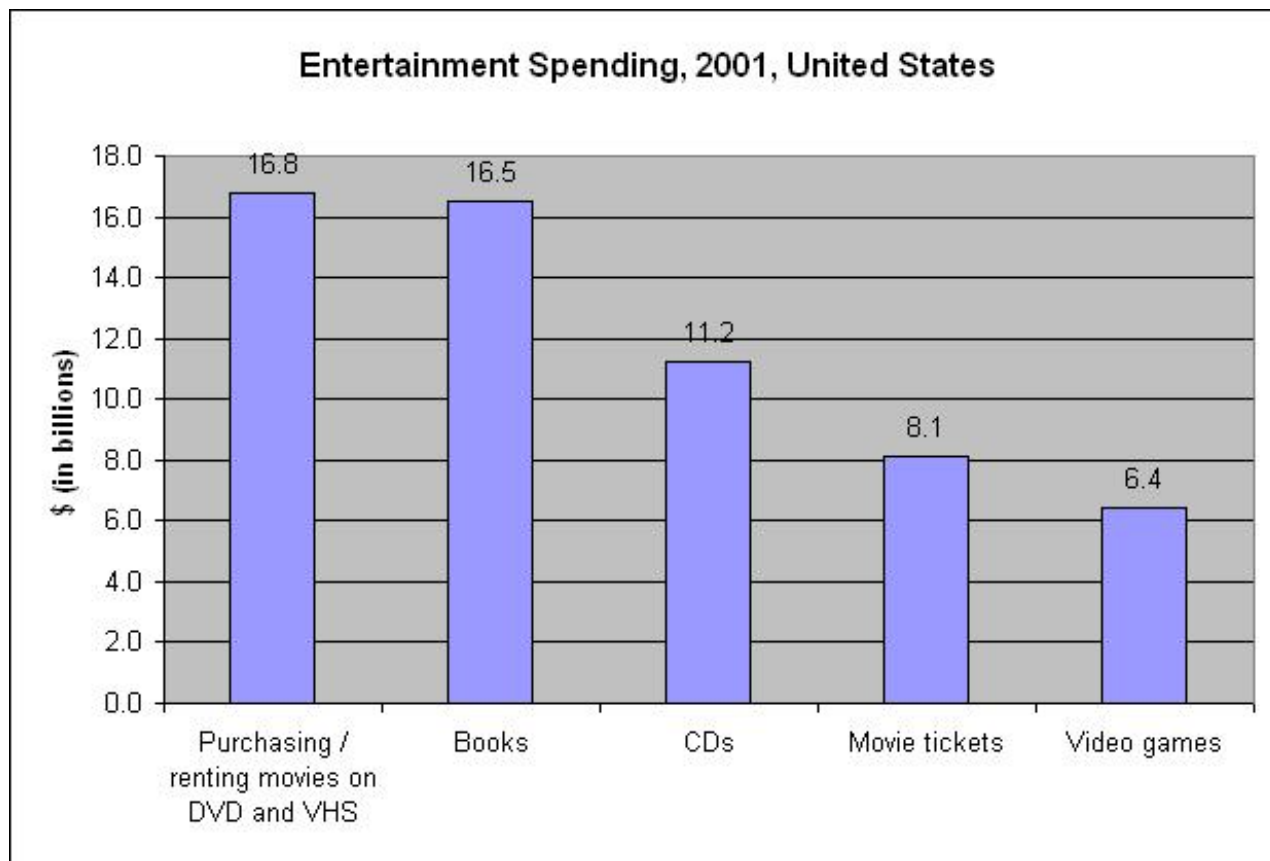
United States. We do not have statistics for the number of CD-ROM copies distributed in the United States. However, relying once again on IRMA's replication statistics, we find that at least 8 billion CD-ROMS have been replicated since 1997.

AOL DISKS

AOL CD-ROMs constitute one significant category of duplicated CD-ROMs. Since the mid-1990s, millions of free disks containing AOL software have been distributed worldwide via magazines, letterboxes, cereal packets, ATMs, service stations and even airline and sports stadium seats. AOL will not release numbers, but it estimated in 1997 that when stacked up, all the CDs it had mailed to date would top the height of Chicago's Sears Tower. Direct Marketing Business Intelligence put the total number at around 300 million. This excess has prompted the creation of protest sites such as No More AOL CDs, which intends to collect 1 million disks and then deliver them to AOL's Virginia headquarters—so far, this group has accumulated 150,000 disks. Other sites display the different disk designs, suggest "101 Uses for AOL disks," or offer a recycling option, along with a chance to set Guinness World Record by contributing unwanted disks.

III. DVD

In 2001, United States consumers spent more money on video than on any other entertainment option, according to the DVD Entertainment Group.



Source: DVD Entertainment Group

Furthermore, in the United States DVD currently outsells VHS two to one and DVD players are in 40% of North American homes, according to IRMA. The average price of players sold in the U.S. in 2002 was \$145. For 2003, the CEA is forecasting an increase in units shipped of 17.5%, to 20.1 million units. The average unit price is expected to fall to about \$125 in 2003, according to Magnetic Media Information Services.

DVD players are in 10% of the world's TV households – nearly 100 million homes. It is expected that nearly half of global TV households will own a player by 2010, according to Mindbranch Global DVD Facts and Forecasts.

As of 2002, the penetration of DVD technology in different countries varies widely:

Table 5.6: International Penetration of DVD Technology		
Continent / Country	Number of players as of 2002	Percentage of homes with a DVD Player
North America	53 million	46%
Australia / New Zealand	2 million	22%
Western Europe	28 million	18%
Japan	6 million	13%
Middle East	1 million	10%
Asia	18 million	5%
Latin America	2 million	2%
Central/Eastern Europe	1.5 million	2%
India	300,000	< 1%
Africa	500,000	< 1%

Source: Screen Digest / Adams Media Research

DVD output also varies by country:

Table 5.7: DVD Replication by Country	
Continent / Country	DVD output (replication)
North America	47%
Western Europe	18%
Asia	13%
Japan	9%
Rest of the world	4%

Source: Singulus Technologies

The DVD standard continues to evolve. Two additional categories of DVD have emerged in the past 3 years: DVD-ROM and DVD-audio. DVD-ROM is intended for computer applications. DVD-Audio is a format specifically designed to provide the highest possible audio fidelity capable on DVD--the audio fidelity of DVD-Audio far exceeds the quality of conventional CDs and audio on DVD-Video.

In 2000, the numbers for DVD-ROM and DVD-audio were negligible, but now they constitute 8% and 0.2% of the DVDs replicated. There are more than 300 DVD-audio titles available in the United States, and 250 more titles will be released in 2003.

DVD recorders, used to record TV programs onto DVD, first became available to consumers in 1999. Four years later, worldwide sales are projected to reach 2.7 million units, and sales of DVD recorders may surpass sales of players by 2005, according to MediaLine. In March, Sony announced the first model of next-generation DVD recorder, a high-definition version that holds more than five times as much data as current DVDs, according to the Wall Street Journal. This product, which uses a blue laser to achieve the higher density recording, has been developed in response to the popularity of high-definition television sets. If the current U.S. catalog of DVDs were converted to HD-DVD, the storage requirements could leap to **61 TB** to **300 TB**.

Original Information Stored on DVD

Annual Production of Titles

World. The rate of worldwide DVD title production doubled between 1999 and 2001. In 1999, 5,000 new titles (22 TB) were added; in 2000, 10,000 new titles (**43.8 TB**) appeared, according to Jim Taylor's DVD FAQ. This tremendous rate of content growth is due to the conversion of legacy film content. For details on the creation of new film content, see [Film](#).

United States. As of 2001, the United States produces about 4,000 DVD titles per year (more than 100 per week). This represents a 33% increase from 1999 when the United States produced about 3,000 new titles per year. 4,000 DVDs equal about **18 TB** of information.

Accumulated Stock

World. The DVD Entertainment Group reports that there are nearly 20,000 DVD-video titles available worldwide, as of January 2003. Remarkably, after only 6 years, this new format has as many titles as the 20-year-old CD-ROM format; as noted earlier, this is due to the relative ease of converting and enhancing legacy content compared to developing brand new content for a CD-ROM title. 20,000 DVDs equal about **88 TB**.

United States. As of 2001, there were 14,000 DVD-video titles available in the United States, according to Jim Taylor's DVD FAQ. This equals about **61 TB**.

Copies of Information Stored/Published on DVD

Annual Production of Copies

World. We do not have statistics on the number of DVD copies sold through retail each year worldwide. However, using IRMA's replication statistics, we find that 1.7 billion units of DVD-video were replicated in 2002, as well as 298 million DVD-ROMs. This adds up to **9 million TB** in DVD copies.

United States. In 2002, 685 million DVD copies were sold (retail) in the United States. For a second data point, we use IRMA's replication statistics, we find that 960 million units of DVD-video were replicated in 2002, as well as 110 million DVD-ROMs. This adds up to about **5 million TB** in DVD copies.

Accumulated Stock

World. We do not have updated statistics for the number of DVD copies distributed worldwide. However, we know that 1.6 billion DVDs have been replicated since the launch of the format in 1997.

United States. A total of 1.36 billion DVDs have been shipped to retail in North America since the launch of the format in 1997, according to the DVD Entertainment Group

REFERENCES

- Adams Media Research, cited in speech by IRMA president <http://www.recordingmedia.org/news/speechdcdexpo2003.html>
- David Bounie, "The International Production and Dissemination of Information" (Paris: ENST, 2003).
- CD-ROM's in Print, 16th edition
- Direct Marketing Business Intelligence, Mar. 1, 2002, http://www.directmag.com/ar/marketing_theyll_mail/
- DVD Entertainment Group press releases 1/8/02, 1/9/2003: <http://www.dvdinformation.com/news/index.html>
- International Federation of the Phonographic Industry (IFPI) World Sales
- International Recording Media Association Statistics
- Jim Taylor's DVD FAQ
- Magnetic Media Information Services, <http://www.mmislueck.com/WhatsNews.htm>
- MediaLine, DVD Recorders to Overtake Players by 2005, April 14, 2003 http://www.medialinenews.com/issues/2003/april/dvd_07.shtml
- Mindbranch Global DVD Facts and Forecasts, November 2002, <http://www.mindbranch.com/catalog/product.jsp?code=R443-0008>
- RIAA MarketData Newsletter, http://www.riaa.org/pdf/public_newsletter_final.pdf
- Sacramento News and Review, May 9, 2002, "Burn Baby Burn." <http://www.newsreview.com/issues/sacto/2002-05-09/editorial.asp>
- Screen Digest / Adams Media Research, <http://www.screendigest.com/NSMH-5CAAWM/Mkt%20report%2002%20-%20HD.pdf>
- Singulus Technologies Presentation, <http://www.singulus.de/pdf/present/20082002.PDF>
- Wall Street Journal, March 10, 2003, "Storing Movies on DVD is So Two Weeks Ago."
- Wired Magazine, Dec. 2002, "Burn Baby Burn." <http://www.wired.com/wired/archive/10.12/view.html?pg=2>

How Much Information? 2003

Summary

Exec Summary

Stored Information

Paper | Film | Magnetic | Optical

Information Flows

Broadcast | Telephony | Internet

Wrap-up

Thanks | Printable (PDF)

Broadcast

[I. Radio](#)

[A. Conversion Factor](#)

[B. World](#)

[C. United States](#)

[D. European Union](#)

[E. Online Radio](#)

[II. Television](#)

[A. Conversion Factor](#)

[B. World](#)

[C. United States](#)

[D. European Union](#)

[E. Consumption](#)

[F. Stock](#)

[III. Interesting Facts](#)

[References](#)

[Printable Section \(PDF\)](#)

[Charts \(Excel\)](#)

6. BROADCAST MEDIA

In this section, we focus on the two broadcast media: radio and television.

Worldwide the number of hours of original programming for both radio and TV has not changed significantly since our 2000 study. Due to difficulties in data collection for this statistic our degree of confidence in this finding is low.

In the United States, there has been a 5% increase in the number of radio stations and a 6% increase in the number of television stations. However, due to increased syndication of content and higher percentages of advertising, these increases may not mean that more original items are being produced.

Satellite radio has emerged as a new category since our 2000 study; this format has yet to achieve wide popularity. Legal challenges have impeded the growth of online radio stations, which had seemed poised to expand in 2000. Personal video recorders such as TiVo and Replay have also not yet received the widespread

consumer acceptance that analysts had predicted.

Table 6.1: World – annual production of original broadcast media items – 2003 sources

Media Type	Number of Stations	Unique Items per Year	Conversion Factor	Total Terabytes (Annual)	
				Lower Bound	Upper Bound
Radio	47,776	70 million hours of original programming	0.05 GB/hour	3,488	3,488
Television	21,264	31 million hours of original programming	1.3 GB - 2.25 GB hour	39,841	68,955
Total:				43,329	72,443

Source: Raw data from CIA World Factbook, 2002. Table, How much information 2003.

Table 6.2: World – Comparison of production of original broadcast media items – 2000 sources vs. 2003 sources

Media Type	Year	Number of Stations	Unique Items per Year
Radio	2000	43,973	65.5 million hours of original programming
	2002	47,776	70 million hours of original programming
Television	2000	21,342*	31 million hours of original programming
	2002	21,264	31 million hours of original programming

Source: Raw data from CIA World Factbook, 2002. Table, How much information 2003.

* Note: We adjusted the 2000 TV figures downward to account for the repeaters and relay stations counted in the previous study.

Table 6.3: United States – Comparison of production of original broadcast media items – 2003 sources

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

Radio stations (2002, FCC)					
Commercial - AM	4,811	7.2 million hours	0.05 GB/hour	358 TB	358 TB
Commercial - FM	6,147	9.2 million hours	0.05 GB/hour	458 TB	458 TB
Educational	2,303	3.4 million hours	0.05 GB/hour	171 TB	171 TB
Total	13,261	19.8 million hours	.05 GB/hour	987 TB	987 TB
Television (2002, FCC)					
Broadcast Stations	1,686	3.1 million hours	1.3 GB - 2.25 GB hour	4,000 TB	6,923 TB
Cable Stations	308	.6 million hours	1.3 GB - 2.25 GB hour	731 TB	1,265 TB
Total	1,994	3.6 million hours	1.3 GB - 2.25 GB hour	4,731 TB	8,188 TB
Total (radio + TV):				5,718 TB	9,175 TB

Source: Raw data from U.S. Federal Communications Commission. Table, How much information 2003.

Table 6.4: United States – Comparison of 2000 sources vs. 2003 sources

Media Type	% change	Year	Number of Stations	Unique Items per Year
Radio stations	5%	1999	12,600	15.8 million hours
		2002	13,261	19.7 million hours
Television (broadcast and cable networks)	6%	1999	1,884	3.4 million hours
		2002	1,994	3.6 million hours

Source: Raw data from U.S. Federal Communications Commission. Table, How much information 2003.

I. Radio

A. 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.)

In the United States we estimate that FM stations broadcast 20 hours per day (7,300 hours per year), AM stations 16 hours per day (5,840 per year) and shortwave stations 12 hours per day (4,380 hours per year).

For the world, we estimate that stations broadcast 16 hours a day.

To determine the percentage of original programming per broadcast hour, we began by looking at broadcasting patterns in the United States. According to the Radio Marketing Guide and Fact Book, from the Radio Advertising Bureau, about 84% of US radio stations have music as their primary focus, and provide little original content. Perhaps 10% (or 5 minutes an hour) of a music station's total broadcast time consists of commentary, weather reports, news updates, and traffic reports; the remaining time is filled by music and advertisements. Stations with non-music formats, such as news, talk, and religious stations, provide, presumably, mostly "original" information each day, perhaps 75% of their broadcast time; the remaining time is filled by advertising. Regardless of format, radio stations average between 12 and 16 minutes of commercials per hour.

For the world percentages, we do not have precise figures on the number of music and non-music format stations. We believe that other countries have more stations with non-music formats and therefore generate a higher percentage of original information. For our calculations below, we estimate 25% original per broadcast hour.

B. World

There are 47,776 active radio stations in the world, according to the CIA World Factbook (online): about 15,800 AM stations, 30,700 FM stations, and 1,300 shortwave stations. Overall this represents an increase of 9%, with the FM format adding 4,800 stations and AM and shortwave losing 881 stations and 143 stations, respectively. **Note: due to the irregular nature of the CIA Factbook data, it is possible that these shifts are caused by reporting errors, rather than any real change.**

Using the daily hours of broadcasting stated above, we estimate there are approximately **280 million** hours of radio programming per year. Applying the 50 MB/hour rule of thumb, one may estimate an annual storage requirement of about **14,000 TB** if one were to record everything broadcast on the radio.

Of this total, we estimate that **70 million hours** are original programming, with an annual storage requirement of about **3,500 TB**.

C. United States

As of 2002, there are **13,261** radio stations in the United States, according to the Federal Communications Commission: 4,811 AM, 6,147 FM, and 2,303 FM Educational stations. Overall, this is a 5% increase from 1999's figures, with additional stations appearing in all 3 categories.

As noted above, AM and FM stations 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 89 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 **4,500 TB**.

Using the above conversion method, we estimate the total "original" programming appearing on the United States airwaves: **19.7 million** hours. (This estimate excludes advertising and music.) The equivalent in bytes is **987 TB**.

Syndicated programming has become a big business, and it is only getting bigger. Howard Stern earned about \$18 million last year, around the same salary as radio personalities Rush Limbaugh (\$30 million eight-year contract until 2009 despite his hearing loss), Dr. Laura Schlessinger (\$13 million) and Don Imus (\$10 million). Stern has signed an annual \$20-million contract with radio syndication firm Infinity, matching what champion golfer Tiger Woods receives to endorse Nike each year.

Satellite radio providers are expanding their markets with car rental companies and automobile makers. The Washington Post reports that XM Satellite Radio on Tuesday added 209,000 subscribers to its digital radio service in the second quarter of 2003; Washington-based XM said it ended the quarter with 692,253 subscribers. XM hopes to reach 1 million subscribers by the end of 2003.

D. European Union (EU)

Unlike in the United States, according to the ENST report, the overall number of EU radio stations is decreasing: from 7,600 stations in 1994 to about 5,500 stations in 2000, representing a decrease of about 28%. However, in a few EU countries the number of stations has increased dramatically (in Greece by 800%, in Luxembourg by 600%, and in Sweden by 400%) since 1990. The 5,500 stations produce approximately 37 million hours of radio programs every year, with **6.8 million hours** of original programming, equivalent to **344 TB**.

E. Online Radio

As of 2001, there were over 35,000 online radio stations broadcasting at varying hours each day, according to online radio advocate Jim Karpen. CNET reports that due to a court ruling in 2001 requiring payment of royalties for each song streamed over the Internet, the number of web radio stations has fluctuated widely, some stations closing permanently or temporarily in protest. As of 2002, the ENST study reports significantly lower numbers: 585 different digital radio services, 400 of which are in the European Union.

II. Television

A. Conversion Factor

The industry rule of thumb for storing video is 1 GB per hour. TiVo and Replay TV encode MPEG-2 at 1.2 GB per hour for video.

B. World

The best source of information on the number of television stations worldwide is the CIA Factbook. Unfortunately, this data comes from different years and is reported in different formats. For instance, the totals for some countries include very low power stations and repeaters, while other countries report these stations in a separate category. Because reporting varies from year to year, it makes it difficult to evaluate time-series data.

There are 20,991 television stations in the world, according to the CIA Factbook. If these stations broadcast 16 hours per day, this would equal about 123 million hours total programming. We estimate about $\frac{1}{4}$ of the programs are "original," – this is **31 million hours** each year. Estimating that one hour of video requires 1.3 GB of storage, then worldwide, program storage would be about **40,000 TB**.

C. United States

As of 2002, there are 1,686 broadcast television stations in the United States, according to the US Federal Communications Commission. This figure includes the major networks (ABC, CBS, NBC, FOX, PBS and newcomers WB, UPN and PAX) and the networks' affiliates, as well as local and public broadcasting stations. In addition, the National Cable Television Association reports that there are 308 cable networks.

If all 1,994 of these stations broadcast 20 hours per day, that would equal about **14.5 million hours** per year. We estimate that about $\frac{1}{4}$ of the television programs broadcast are "original" – this is **3.6 million hours** each year, equivalent to **between 4,700 TB and 8,200 TB**.

D. European Union

There are 659 channels in the EU, as reported by the ENST study. If these stations broadcast 20 hours per day, that would equal **4.9 million hours**. If we estimate that about $\frac{1}{4}$ of the programs broadcast are original, this is 1.2 million hours each year, equivalent to between **1,560 and 2,700 TB**.

E. Consumption

According to Eurostat [2000] as reported by ENST, the average television viewing time in the United States was 4 hours and 20 minutes per day, as of 2000; in the EU viewing time is about 1 hour less: 3 hours and 24 minutes. 68% of the viewers in the United States receive their television by cable, 9% by satellite, and 23% by terrestrial reception. By contrast, the majority (51%) of EU viewers have terrestrial reception, 31% have cable and 19% have satellite.

F. Stock

In 55 years of programming, the networks have accumulated the following stock of material:

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

Source: *Library of Congress Report, Television/Video Preservation Study: Volume 1: Report, October 1997.*

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

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

Source: *Television and Video Almanac 1998*

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.

III. Interesting Facts about Broadcast Media

- At this time, there is no Federal Communications Commission limit on the number of commercials per hour. For many years, TV stations and the networks subscribed to a voluntary broadcasting code that limited prime-time advertising to 9.5 minutes per hour (but allowed more at other times). In 1992, however, pressure to adhere to those guidelines was ruled a violation of antitrust law. An American Association of Advertising Agencies report showed that in the primetime slot on network television, time used by commercials was 16:43 minutes per hour. The daytime rate was 20:53 minutes per hour, while network news showed 18:53 minutes of commercials per hour, and late night news aired 19:06 minutes of ads per hour. The most interrupted program in all of network TV was ABC's Good Morning America.
- Recent internal research by Procter & Gamble indicates that consumers who fast-forward through ads with digital personal video recorders such as TiVo still recall those ads at roughly the same rates as people who see them at normal speed in real time. The popularity of TiVo and Replay TV does not yet seem to be taking off, since consumers may not see the benefit of a digital personal recorder instead of a video cassette recorder.
- The US market for cable and satellite TV services increased 8.8% from \$40.9 billion in 2000 to \$44.5 billion in 2001. Satellite TV remains the dominant source of television for many countries without the infrastructure or technology to support cable television in their area.
- Talk radio has been experiencing significant growth for a number of years and continues to grow. Radio listeners reported just under one and one-half hours (86 minutes) of radio news/talk listening on the average weekday. Adults 18 to 64 listen to an average of three hours of radio per weekday. Therefore this suggests that close to one-half of all radio listening involves news or talk. (News radio listeners account for 93 percent of all persons 18 to 64.)

Table 6.7: Talk Radio Stations (U.S.)

1980	75	1985	100
1987	125	1990	200
1991	350	1992	500
1993	750	1994	1000
1997	1250	1998	1350

Source: *Talker's Magazine Online, The Talk Radio Research Project; American Radio News Audience Survey*

REFERENCES

- AdAge.com, March 17, 2003, "Tivo Less of a Threat? Surprising P & G Findings," <http://www.adage.com/news.cms?newsId=37390>
- [American Radio News Audience Survey](#)
- David Bounie, "The International Production and Dissemination of Information" (Paris: ENST, 2003).

- Cable and Television Overview 2002
- Charlotte Observer, Jan. 11, 2003, "Mute the Sound: TV Commercials Know No Bounds," http://www.charlotte.com/mld/charlotte/news/columnists/jeff_elder/4922737.htm
- CIA Factbook 2002
- CNET, "Senate puts Webcasting Bill on Hold," October 18, 2002, <http://news.com.com/2100-1023-962556.html>
- EOnline, July 17, 2001, "Rush: Nine-Figure Dittohead," <http://www.eonline.com/News/Items/0,1,8552,00.html>
- Euromonitor International analysis <http://www.euromonitor.com/>
- Federal Communication Commission (FCC) - 6/30/02 Figures
- Forbes Magazine, Dec. 15, 2000, "Forbes Faces: Howard Stern," <http://www.forbes.com/2000/12/15/1215faces.html>
- Jim Karpen's Tips http://site.lisco.com/lisco/karpentips/karpen_radio267.htm
- Ron Kaufman, "Ratings and Advertising: What are Americans Really Watching?" <http://www.turnoffyourtv.com/ratingsAds.html>
- *Library of Congress Report, Television/Video Preservation Study: Volume 1: Report, October 1997*
- National Association of Broadcasters
- National Cable Television Association (NCTA)
- *Radio Marketing Guide and Fact Book*, Radio Advertising Bureau 2000-2001
- RadioTalk, Oct. 9, 2001, "Rush Limbaugh's Hearing Loss," <http://www.radiotalk.org/limbaughdeaf.html>
- Talker's Magazine Online, The Talk Radio Research Project; <http://www.talkers.com/talkaud.html>
- *Television and Video Almanac 1998*
- Wired Magazine, Dec. 20, 2002, "Is that a TiVo Under the Tree?" <http://www.wired.com/news/gizmos/0,1452,56828,00.html>
- *The Washington Post*, July 7, 2003, "XM Satellite Shares Rise on Gain in Subscriptions,"

How Much Information? 2003

Summary

[Exec Summary](#)

Stored Information

[Paper](#) | [Film](#) | [Magnetic](#) | [Optical](#)

Information Flows

[Broadcast](#) | [Telephony](#) | [Internet](#)

Wrap-up

[Thanks](#) | [Printable \(PDF\)](#)

Telephony

[I. Standard Fixed Lines](#)

[A United States](#)

[B. World](#)

[II. International Calling](#)

[III. Mobile](#)

[IV. Broadband](#)

[V. Wireless](#)

[References](#)

[Printable Section \(PDF\)](#)

7. TELEPHONY

Telephony, a century-old technology, is expanding rapidly in new directions. In addition to traditional landline phones, telephony now includes a diverse range of technologies, including mobile (cellular, satellite), and broadband (cable, DSL, satellite) connections. These services are used for multiple, overlapping purposes, for transmitting both voice and, increasingly, data traffic. This section highlights some of the trends in telephony and estimates the volume of this information flow.

I. Standard Fixed Lines

A. United States

According to the International Telecommunication Union (ITU), there are 190 million main telephone lines in the United States as of 2002. The average telephone line is used primarily for local calling and is used **over an hour per day** for all types of calls (faxes, modems, etc.)

The number of U.S. landline phones has dropped by more than 5 million, or nearly 3 percent, since 2000, the Federal Communications Commission (FCC) reported in June, 2003. There is a growing trend to use the cell phone as the primary phone, without any need for landlines.

On the other hand, the number of phone lines per household has risen. J.D. Powers and Associates reports that 26 percent of American households now have two or more telephone lines--a 73 percent increase since 1996.

Table 7.1: United States Fixed Telephone Line Usage (in billions)

Year	Local Minutes	Intrastate Minutes	Interstate Minutes	Total Dial Equipment Minutes
1999	3,378	452	585	4,414 billion (74 billion hours)
2000	3,909	472	616	4,998 billion (83 billion hours)
2001	3,784	420	615	4,819 billion (80 billion hours)

Source: *Federal Communications Commission Trends in Telephone Service, August 2003*

A few disclaimers on the numbers reported in Table 7.1: these figures are dial equipment minutes, "measured as calls enter and leave telephone switches so two dial equipment minutes are counted for every conversation minute." Also, the FCC notes that "most calls are not metered and estimates of total calling are subject to wide margins of error" because "most subscribers purchase service with unlimited local calling." Finally, the 2001

figures are extrapolated from previous years because as of that year, companies are no longer required to report these numbers to the FCC.

Using the statistics in Table 7.1 and our conversion factor for audio recordings (64 KB per second), we can calculate the total annual storage for voice traffic in the United States: **9.25 exabytes**. Compression could reduce storage requirements by a factor of 6 to 8, resulting in a total of **1.2 to 1.5 exabytes**.

B. World

There are 1.1 billion main telephone lines in the world as of 2002, according to the ITU. If we estimate, as in the ENST study, that each line carries 3,441 minutes each year, the worldwide total is 3,785 billion minutes, equivalent to **15 exabytes** of data.

II. International Calling

TeleGeography reports that "Consumers and businesses worldwide spent 144 billion minutes—or a combined total of 274,000 years—on calls abroad in 2001." While this represents an increase of 10 percent over 2000, it is also the slowest growth rate in 20 years.

VOICE OVER IP (VOIP)

The ability to send voice messages over the Internet is a rapidly growing area in communication this year and will continue to grow as more companies adopt the technology into their infrastructure. Experts report rapid growth in all areas of this technology, including adoption, investment, and revenue. Frost & Sullivan predict VoIP will account for around 75 percent of world voice services by 2007.

Source: CommWeb 2002

III. Mobile

In number, cell phones are creeping up on landline phones. They already comprise about 43 percent of all U.S. phones, according to the International Telecommunication Union, up from 37 percent in 2000.

Worldwide there were 1.14 billion mobile cellular subscribers in 2002; this figure is slightly higher than the total number of main telephone lines.

According to the Cellular Telecommunications and Internet Association (CTIA), over **600 billion wireless minutes** were used in 2002. This represents an increase of more than 35 percent over 2001 and is equivalent to **2.3 exabytes**.

Recent trends show how wireless text messages (short message services) are becoming more important commodities for wireless communication companies. Many companies are allowing competitors to send text messages to other companies' phones in an effort to gain market share and revenue from the popularity of text-messaging systems. AT&T's recent affiliation with the American Idol television show has helped them reach customers who wish to use telephones, but could not vote for their favorite 'Idol' because lines were busy. Throughout the season, over 7.5 million text messages were sent to the voting system.

Europe still dominates the SMS and mobile communication arena. IDC research found there were 176 million e-mails sent between cell phones in the United States. Next year, they believe there will be 1.5 billion emails sent via mobile phones. In Europe, there are an estimated 30 billion messages exchanged each month by this method.

IV. Broadband

A. United States

Broadband access to the Internet is available through a range of advanced technologies such as asymmetric digital subscriber line (ADSL), wireline other than ADSL, coaxial cable, fiber and satellite and fixed wireless. The number of such high-speed lines has increased by approximately 3 million lines every 6 months during the last 2 years (see Table 7.2).

Table 7.2: United States High Speed Telephone Lines (over 200 kbps in at least one direction)

Date	Number of Lines
June 2000	4.4 million
December 2000	7.1 million
June 2001	9.6 million
December 2001	12.8 million
June 2002	16.2 million
December 2002	19.9 million

Source: Raw data from FCC Trends in Telephone Service, August 2003. Chart, How much information 2003

Naturally, with more high-speed Internet access available, more people are using this method to connect to the Internet. As of May 2003, 39 million, or 13 percent of all Americans connect to the Internet via broadband, according to Nielsen-NetRatings.

The Pew Internet & American Life Project reports that cable modems are the primary broadband technology currently in use in the United States: "In March 2003, 67 percent of broadband users connect using cable modems — up from 63 percent in March 2002 — while DSL had 28 percent of the broadband market in March 2003, down from 34 percent a year earlier."

Pew also reported on how online activities differed between dial-up and broadband users; they report that "broadband users are extraordinarily active information gatherers, multimedia users, and content creators," as shown in Table 7.3.

Table 7.3: Daily Internet Activities: Broadband Users vs. Dial-Up Users

Activity	Broadband Users	Dial-Up Users
News	41%	23%
Research for work	30%	15%
Participation in group	12%	4%
Content creation	11%	3%
Stream multimedia	21%	7%
Download music	13%	3%

Source: Pew Internet & American Life Project

B. World

There are currently **35.9 million subscribers** to broadband services worldwide, with Western Europe as the fastest growing segment of the world adopting broadband DSL (Digital Subscriber Line) technology in homes and businesses. Asia-Pacific areas lead with **15.4 million subscribers**, while Western Europe has been growing at a rate of 121 percent to **9.37 million subscribers** in 2002.

South Korea and Taiwan have the **highest density** of DSL subscribers (per hundred main lines), while the United States and South Korea have the **highest number** of DSL subscribers in the world.

V. Wireless

Cyberatlas reports "A trend to watch is the number of users who connect wirelessly, as their numbers have doubled from 2002 — 4 percent or roughly 1.4 million users now surf without wires."

Researchers believe that in the next 5 years, over 90 million laptops and PDAs ready to access WLAN will enter the market.

HotSpots are locations that are setup to provide Internet access through a wireless network to nearby computers. Market researches indicate that the U.S. WLAN HotSpots count is expected to increase from 4,000 HotSpots today to over 40,000 in 5 years, catering to over 20 million HotSpot users.

According to various reports, Bluetooth chip shipments that touched 10.4 million last year, is expected to cross 40 million this year, and 690 million in 2006. Researchers believe that more than 50 percent of the Bluetooth chips will be embedded in cellular phones, and by 2005, 75 percent of all cellular phones will have an embedded Bluetooth chip.

The adoption of wi-fi in many Starbucks, MacDonald's, and even a few airports in India are helping the public gradually become familiar with this 'new' means of Internet access. They may also generate more revenue for these companies, since they provide access through a subscription service (some are also offered through pre-paid cards).

REFERENCES

- "AT&T Wireless and American Idol Set Global Records in Text Messaging", <http://www.phoneyworld.com/carriers/story.aspx>
- David Bounie, "The International Production and Dissemination of Information" (Paris: ENST, 2003).
- [Cellular Telecommunications and Internet Association](#)
- CIOL, August 20, 2002, "Embedded wireless touches public, enterprise and home", <http://www.ciol.com/content/search/showarticle.asp?arid=37861>
- CISCO Voice over IP Facts and Stats, <http://www.cisco.com/warp/public/779/govtaffs/factsNStats/voip.html>
- CNet News.com, "VoiceStream heeds text-messaging call", http://news.com.com/2100-1033-843900.html?tag=sas_email
- CNN, August 4, 2003, "Bye, bye landline phones" <http://www.cnn.com/2003/TECH/ptech/08/04/cell.only.ap/index.html>
- CyberAtlas, July 24, 2003, "Worldwide Cable Modem Subs Expected to Double By 2007" http://cyberatlas.internet.com/markets/broadband/article/0,,10099_2239451,00.html
- DSL Forum, March 12, 2003, "Western Europe Gets Into Broadband DSL" http://www.dslforum.org/PressRoom/news_dsleurope_3.12.03.html
- Federal Communications Commission, Trends in Telephone Service, August 2003, Industry Analysis Division Common Carrier Bureau, http://www.fcc.gov/Bureaus/Common_Carrier/Reports/FCC-State_Link/IAD/trend803.pdf
- [International Telecommunication Union](#)

Telephony

- J.D. Power and Associates, August 8, 2000, <http://www.jdpa.com/news/releases/pressrelease.asp?ID=2044>
- Telegeography, <http://www.telegeography.com/press/coverage/2002/12-12-2002.html>
- VoIP/IP Telephony Statistics, CommWeb, <http://www.commweb.com/article/COM20021015S0002>

Release date: October 27, 2003. © 2003 Regents of the University of California

How Much Information? 2003

Summary

Exec Summary

Stored Information

Paper | Film | Magnetic | Optical

Information Flows

Broadcast | Telephony | Internet

Wrap-up

Thanks | Printable (PDF)

Internet

[I. General Usage Statistics](#)

[II. World Wide Web](#)

[A. Published Estimates](#)

[B. Web Sampling Study](#)

[III. Blogs \(web logs\)](#)

[IV. Email and spam](#)

[V. Instant Messaging](#)

[VI. P2P File-Sharing](#)

[VII. FTP](#)

[References](#)

[Printable Section \(PDF\)](#)

8. INTERNET

Although the Internet is the newest medium for information flows, it is the fastest growing new medium of all time, and becoming the information medium of first resort for its users. Note that the Web consists of the surface web (fixed web pages) and what Bright Planet calls the deep web (the database driven websites that create web pages on demand).

Table 8.1: The size of the Internet in terabytes.

Medium	2002 Terabytes
Surface Web	167
Deep Web	91,850
Email (originals)	440,606
Instant messaging	274
TOTAL	532,897

Source: *How much information 2003*

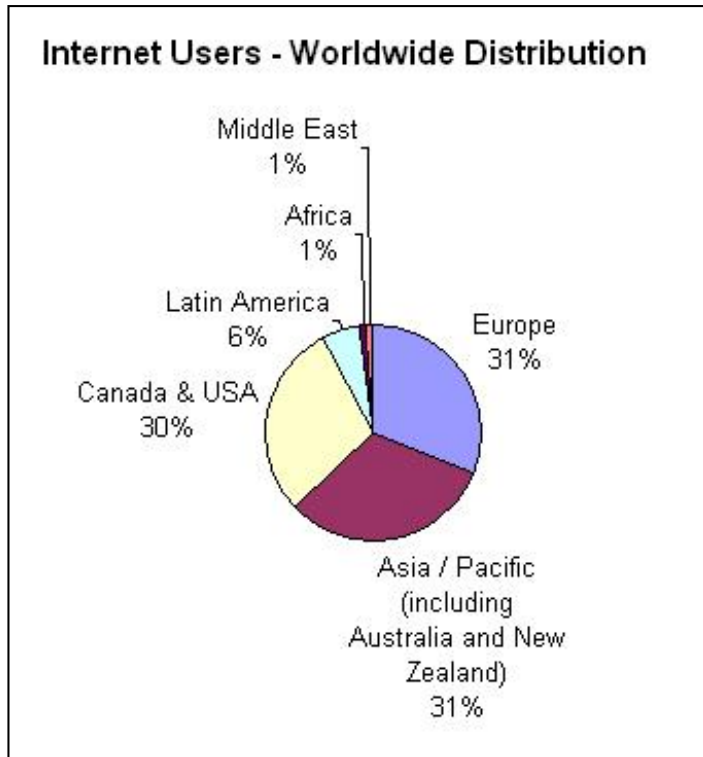
I. General Internet Usage Statistics

The CyberAtlas Stats Toolbox provides pointers to a tremendous assortment of Internet-related usage data. Here are a few topics of general interest.

A. What is the international distribution of Internet users?

According to Nielsen/NetRatings, there is a worldwide Internet population of 580 million users, as of 2002. The International Telecommunications Union provides a 15 percent higher estimate of 665 million users.

Of these, roughly 30 percent are in North America, 31 percent are in Europe, and 38 percent are in other parts of the world, as shown in the following chart.



Source: Nielsen/NetRatings via [CyberAtlas](#)

B. How many hours do individuals spend online?

The average global Internet user spends 11 hours and 24 minutes online per month, according to Nielsen/NetRatings. The average user in the United States spends more than twice that amount of time online: on average, 25 hours and 25 minutes at home and 74 hours and 26 minutes at work.

C. What activities do people do while online?

The Pew Internet and American Life Project reports that on an average day, about 72 million United States users go online. Here are the kinds of activities they do online (top 10 shown here; see the Pew site for the full list).

Table 8.2: Daily Online Activities

Activity	Percent of those with Internet Access	Most recent survey date
Send email	52	March - May 2003
Get news	32	March - May 2003
Use a search engine to find information	29	January 2002
Surf the web for fun	23	March - May 2003

Look for info on a hobby	21	March - May 2003
Do an Internet search to answer a specific question	19	September 2002
Do any type of research for your job	19	November 2002
Research a product or service before buying it	19	December 2002
Check the weather	17	March - May 2002
Send an instant message	14	March - May 2003

Source: [Pew Internet and American Life Project](#)

D. How many web searches are conducted per day?

According to SearchEngineWatch.com, as of January 2003, there were 319 million searches per day at the major search engines. This figure is calculated using the Nielsen//NetRatings "search hours," the total time spent by all visitors searching at each engine.

Table 8.3: Search statistics

Search engine	Search hours per month (in millions)	Search minutes per day (in millions)	Searches per day (in millions)
Google	18.7	37	112
AOL Search	15.5	31	93
Yahoo	7.1	14	42
MSN Search	5.4	11	32
Ask Jeeves	2.3	5	14
InfoSpace	1.1	2	7
AltaVista	0.8	2	5
Overture	0.8	2	5
Netscape	0.7	1	4
Earthlink	0.4	1	3
Looksmart	0.2	0	1

Lycos	0.2	0	1
TOTALS	53.2	106	319

Source: SearchEngineWatch.com, Feb. 25, 2003

II. World Wide Web

A. Published Estimates on Size and Character

Most size studies of the World Wide Web have focused on the number of hosts connected to the network. Preliminary estimates of the amount of data on the web have been made but have not been kept up to date or do not have defined measurement methods. In our last survey we quoted a page size estimate of 18.7 KB from a 1999 article in Nature Magazine that was generated by statistically sampling web servers.

Table 8.4 shows a range of approaches to sizing the Web:

Table 8.4: Methods for Sizing the World Wide Web			
Company	Domains	Method	Frequency
www.whois.net	31,987,198	Domain name registration, this number changes continuously	Continuously
www.netcraft.com	42,800,000	Web servers responding to HTTP request. Each domain name is counted as a server.	Monthly
http://wcp.oclc.org/	9,040,000	IP addresses responding to HTTP request (each IP can maintain many virtual domain names)	Yearly

Source: *How much information 2003*

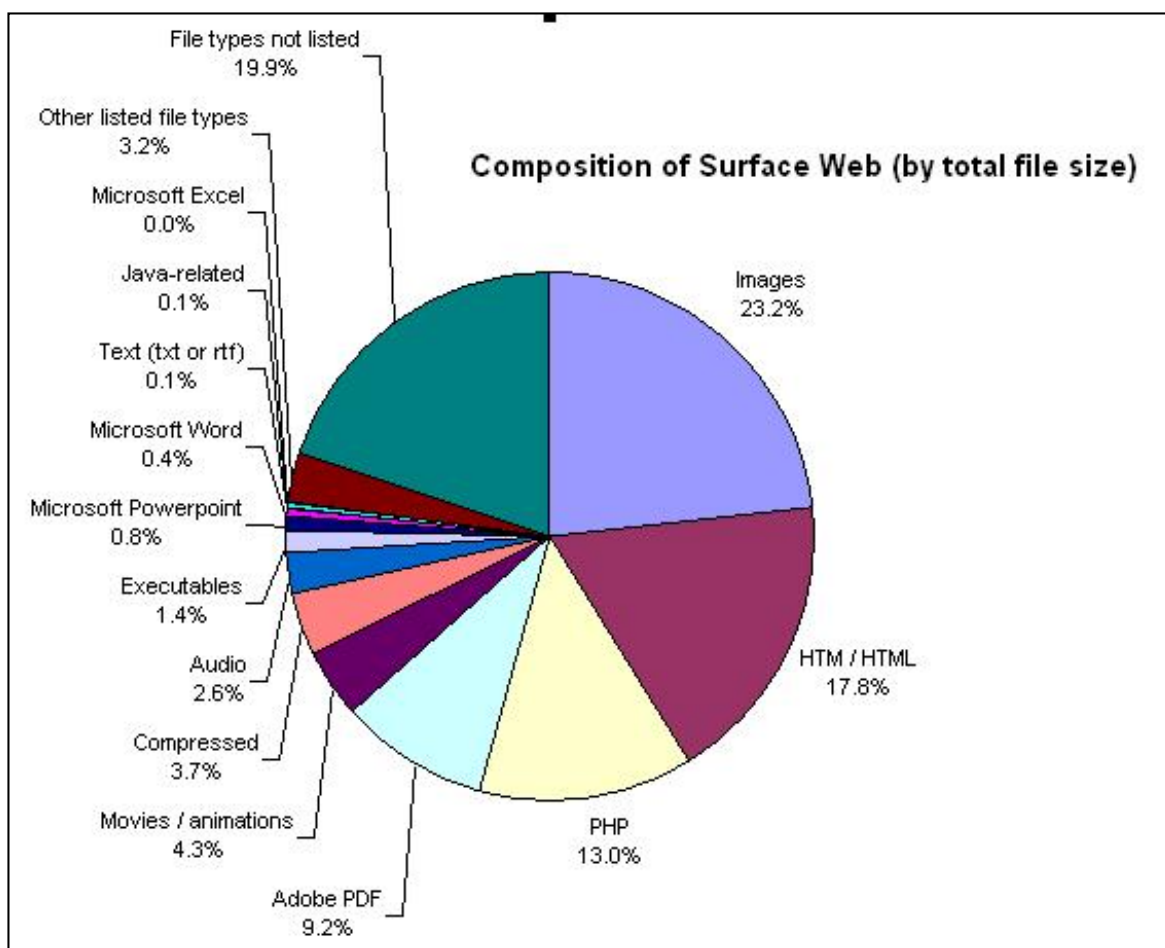
B. Web Sampling Study

We downloaded and analyzed the contents of 9,800 websites in order to estimate the size of an average webpage and the contents of an average website. These sites were chosen randomly from a list of 61 million URLs compiled by the Internet Archive. After confirming that a URL was listed in a DNS registry, we downloaded each registered site in its entirety using "wget," a website mirroring tool. We mirrored each site recursively, following only links that were relative to the original domain name. Once an entire site was completely mirrored we used pattern matching on the file type extensions, and recorded the size of the files. For each site we generated a total size for the site, the total number of files, and file and size totals for a variety of common file types.

Note that we were only able to sample the "surface web"—the static, publicly available web pages which represent a relatively small portion of the entire Web. We were not able to download or measure the dynamic, database-driven websites which comprise the "deep web." As quantified in a landmark study by BrightPlanet in 2000, the "deep Web" is perhaps 400 to 550 times larger than the information on the "surface."

The sum of all the web site file sizes in our sample equals **33.1 GB**. As this sample of 9,806 sites represents 0.02 percent of the 42.8 million web servers (according to the NetCraft Survey, as of August 2003), we may estimate the total size of the surface web as **167 TB** (95% confidence, +/- 1). Therefore the "deep Web" may be between **66,800** and **91,850 TB**.

We counted the number of files of various types (e.g. HTML, images, audio) and found the file types in our sample to be distributed as follows:



Source: *How much information 2003*

We also looked at the functionality and content of the web pages in our sample by matching keyword indicators to text found on the index page of each site. Highlights of our results include:

- **Search.** In 29.8% of the sites sampled, we found the keyword “search” on the index page. Presence of a search function suggests a large or complex site.
- **Form.** The keyword "form" appears in 31.9% of the sampled index pages.
- **Javascript.** The keyword “javascript” appeared in 19.1% of the sites sampled, indicating a degree of website sophistication and interaction.
- **Protection.** In 7.7% of the sites sampled, we found the keywords “password” or “login,” indicating some degree of protection for the content.
- **E-commerce.** On the index page of each site, we checked for words which typically indicate an e-commerce site: cart, shopping, and checkout. In 5.4% of our sampled sites, we found one or more of these words.
- **Porn.** 2,743 sites (or 28%) appeared to contain pornographic content. To generate this statistic, we matched a list of 94 pornographic stopwords to terms in the associated URL and the index page.

For more details on our methodology and findings, see the full writeup (to be submitted for publication

11/2003).

III. Blogs (web logs)

A. What is a blog?

Short for “web log,” a blog is a Web page that serves as a publicly-accessible personal journal for an individual. Typically updated daily, blogs often reflect the personality of the author. A [more complete definition](#) is offered by Jill Walker. In her entry for Routledge’s Encyclopedia of Narrative Theory, Walker writes that a blog is:

“a frequently updated website consisting of dated entries arranged in reverse chronological order so that the reader sees the most recent post first. The style is typically personal and informal. Freely available tools on the World Wide Web make it easy for anybody to publish their own weblog, so there is a lot of variety in the quality, content and ambition of weblogs, and a weblog may have anywhere from a handful to tens of thousands of daily readers. Weblogs first appeared in the mid-nineties and became more widely popular as simple and free publishing tools such as Blogger.com became available towards the turn of the century.

Examples of the genre exist on a continuum from online diaries that relate the writer’s daily activities and experiences to less confessional weblogs that comment and link to other material, discuss a particular theme or function as soapboxes. In addition to the dominant textual form of weblogs there are experiments with adding sound, images and videos to the genre, resulting in photoblogs, videoblogs and audioblogs.

Each entry in a weblog tends to link to further information. Weblog authors also link to other weblogs that have dealt with similar topics, allowing readers to follow conversations between weblogs by following links between entries on related topics. Readers may start at any point of a weblog, seeing the most recent entry first, or arriving at an older post via a search engine or a link from another site. Once reading a weblog, readers can read in several orders: chronologically, thematically or searching by keywords. Weblogs also generally include a blogroll, which is a list of links to other weblogs the author recommend, and many weblogs allow readers to enter their own comments to individual posts.”

Klogs or knowledge-logs are a subset of web logs. According to klogger Spike Hall, a k-log (knowledge log) is “a weblog but also demonstrates or documents a knowledge claim and/or it documents and illustrates the dynamic individual process of a quest for knowledge.”

B. How many blogs are there?

Blogcount.com addresses the “blogosphere,” asking “What is its size, shape, color, true nature?” Blogcount collects and organizes the best reports and analyses on this subject. As of June 23, 2003, Phil Wolff estimates that there are 2.4 to 2.9 million active web logs. He bases his estimate upon statistics from centrally hosted weblogs:

Web log host	Registered	Active	As of
LiveJournal	1,121,464	526,535	23 June 2003
Blogger	1,500,000	705,000	9 June 2003

Diaryland	850,000	400,000	March 2003
Total:		1,631,535	

Source: Blogcount.com

He notes that this figure does not include smaller hosts such as Radio and Moveable Type or blogs hosted behind firewalls on private intranets.

If each blog is 50 KB, then the total size of the active blogosphere is **81 GB**.

C. Who is blogging?

According to Jupiter Research, about 2 percent of Internet users have created a blog. The majority of bloggers use dial-up access to get online, and more than half have a household income below \$60,000 per year. Jupiter also found that blogging is split evenly between the genders and that 70 percent of the bloggers have used the Internet for more than 5 years. (Source: Blogging by the Numbers)

More than 50 percent (350,000) of the 655,000 web logs crawled in National Institute for Technology and Liberal Education (NTILE) web log census are written in English. The rest of the top 10 languages for blogs are (in order): Portuguese, Polish, Farsi, French, Spanish, German, Italian, Dutch and Icelandic.

D. Who is reading web logs?

Jupiter Research estimates that only 4 percent of the online community read blogs. "Blogs seem to be read mostly by men (60 percent vs. 40 percent women), in homes where the total income is more than \$60,000 per year (61 percent). Dial-up remains the connection of choice (54 percent compared to 46 percent broadband), and the majority (73 percent) of blog readers have been online for more than 5 years." (Source: Blogging by the Numbers)

E. More interesting blog facts

- 10,000 domains listed in the whois registry have "blog" in their names
- On average, blogs are updated every 3 days, according to a brief informal study reported at BlogCount.
- About four percent of online Americans report that they have gone to blogs for information and opinions related to the war in Iraq.
- Nielsen/NetRatings says that in May 2003, LiveJournal was the 650th most popular site on the Internet by unique audience. The 184,000 people visiting about every ten days (3.13 visits per person during the period) were so active that LiveJournal was number 213 by pages viewed. When people came, they spent 22 minutes at the site, hitting the back button about 26 percent of the time.
- During the summer of 2003, America Online introduced AOL Journals, allowing members to update blogs by instant message. As of December 2002, AOL has 35.2 million paid members, using 8 languages. The journals will be integrated with AOL community forums, instant messaging, photo albums, and home page construction kits.

IV. Email and Spam

Email is currently one of the most widespread methods of communication. Email users comprise:

- 35% of the total U.S. population (eMarketer)

- 50% of U.S. consumers (Forrester Research)
- 94% of U.S. Internet users (eMarketer)
- 98% of employed Americans with Internet access— 57 million adults (Pew Internet)

Forrester also reports that email use accounts for over 35 percent of all time spent on the Internet, while a PricewaterhouseCoopers survey found people spending 84 percent of their time on the Internet for email.

The Pew Internet and American Life Project recently published their research on email at work. They found the following (through participant self-reports on activity):

- 60% of work emailers receive 10 or fewer messages on an average day; 23% receive more than 20 and only 6% more than 50.
- 78% of work emailers send 10 or fewer messages on an average day; 11% send more than 20.
- 73% of work emailers spend an hour or less per day on their email. That includes 23% of all work emailers who spend fewer than 15 minutes per day handling email.
- 46% of work emailers say their work email volume has stayed the same over the past year.
- 48% say their email volume has increased over the past year.

Since 1995, email volume has increased, though sources differ as to the degree of the increase. A study by Rogen International and Goldhaber Research Associates found that in 1995, employees sent an average of 3 emails a day and received five. As of 2002, employees were sending 20 a day and receiving 30. A 1999 study reported in Newsweek estimated that “a white-collar worker receives about 40 email messages in his office every day.”

Daily email traffic is expected to almost double by 2006, from 31 billion today to 60 billion, according to a new study by International Data Corporation (IDC). If each email is 59 KB (Source: Forrester Research), the daily flow of emails worldwide is currently **1,829 TB**. Over the course of a year, the total would be **3.35 petabytes**.

Table 8.6: Worldwide email messaging

Year	Emails per day	Emails per year
1999	5 billion	1.4 trillion
2000	10 billion	
2001		
2002	14.9 billion or 31 billion	4 trillion
2003		

2004		
2005		
2006	60 billion	

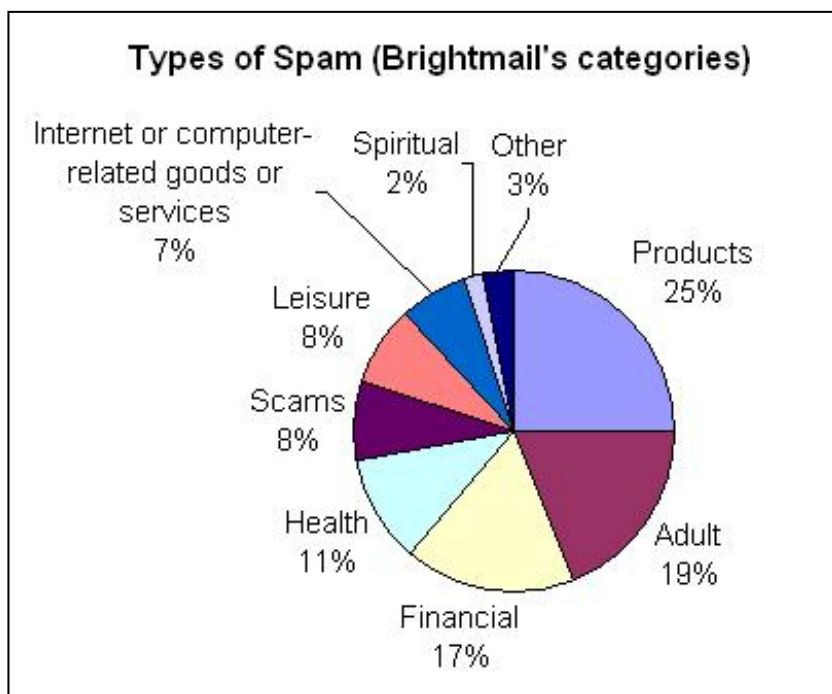
Source: IDC via [Channel One Market Overview](#)>

IDC also estimates that only half of the email traffic will be personal messages. Unsolicited email (also known as spam), commercial notifications and news alerts account for one-third of today's email load and will comprise nearly half of the traffic four years from now, the report said. Therefore we estimate the upper bound of original content in emails as **440,606 terabytes** (uncompressed), lower bound as **333,792 terabytes**.

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 25 million messages per day in approximately 300,000 mailing lists. A sample of mailing lists has shown that 30 percent of them are managed using LISTSERV. Using this information, we would estimate the total number of mailing list messages at 30 billion per year with aggregate volume of 563 terabytes.

A. What is spam?

Spam is unsolicited bulk email. Brightmail, an anti-spam service provider, classifies spam into the following categories (percentages as of May 2003):



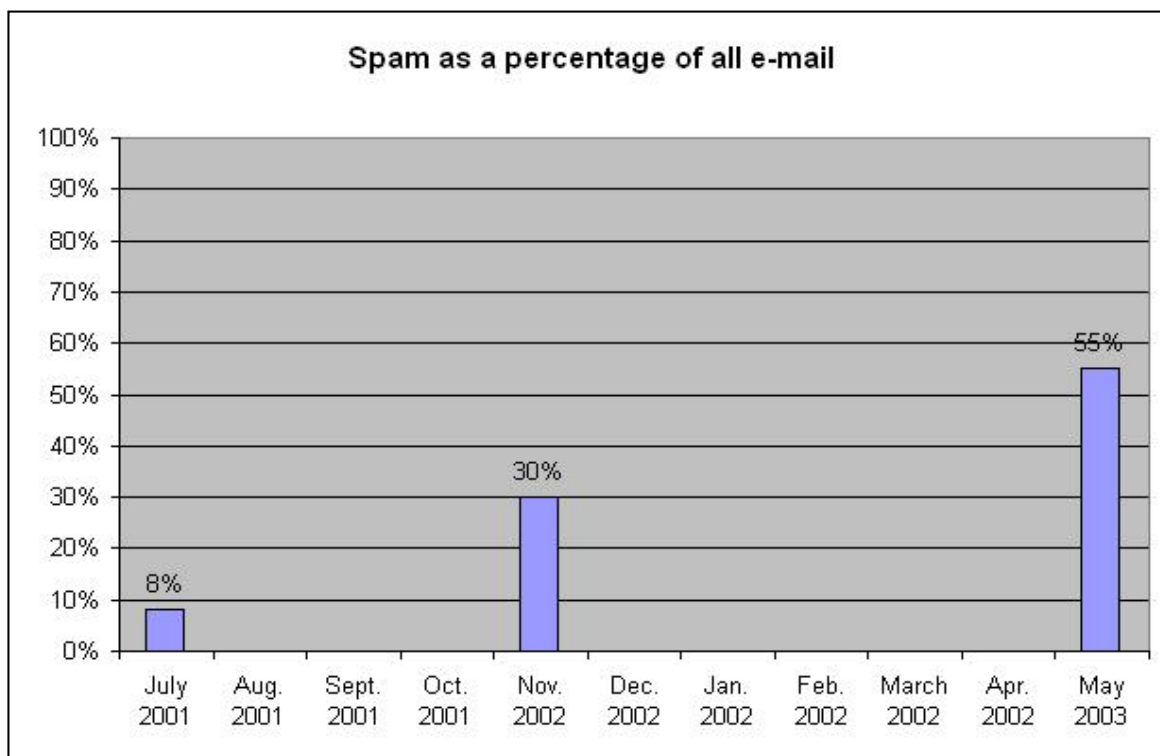
Source: BrightMail <http://www.brightmail.com/spamstats.html>

B. How much spam is sent each year?

Forrester estimates that by 2004, marketers will send more than 200 billion emails in the United States alone.

In December 2002, according to interception figures from Brightmail, a leading spam filtering company, unsolicited bulk email made up 40 percent of all email traveling over the Internet, up from 8 percent in 2001. MessageLabs, a U.K. spam filtering company provides slighter lower figures for 2002 (30 percent spam) but reports an increase to 55 percent spam as of May 2003. As a point of comparison, only 40 percent of United

States Postal Service mail is business marketing.



Source: MessageLabs

According to a recent CNN Techweb article, "Gartner analyst Joyce Graff says spam is a serious problem, but she disputes MessageLabs' estimate that spam will account for half of all email traffic by midyear. She says users often erroneously lump together all their annoying email into the category of spam, including both real spam, such as con games and phony businesses, and business mail that they're copied on by overzealous colleagues."

C. How much spam is blocked or filtered by the ISPs?

The three major email service providers AOL, Microsoft and Yahoo! have more than 200 million email account holders, making them an attractive target for spammers.

The spam problem has shown a tremendous spiral, according to figures provided by AOL (see chart below). "In a single 24-hour period in March 2003, America Online says it trashed a billion spam emails using its software filters. AOL said its members used "report spam" buttons on their email software 5.5 million times during the same period. AOL said it blocks an average of 28 junk emails per account, per day. Graham said "an extremely small fraction" of the messages snagged in AOL's spam filters were legitimate ones. He declined to reveal any figures for that mail."

Yahoo! offers comparable statistics, noting that it intercepts 1 billion spam messages a day.

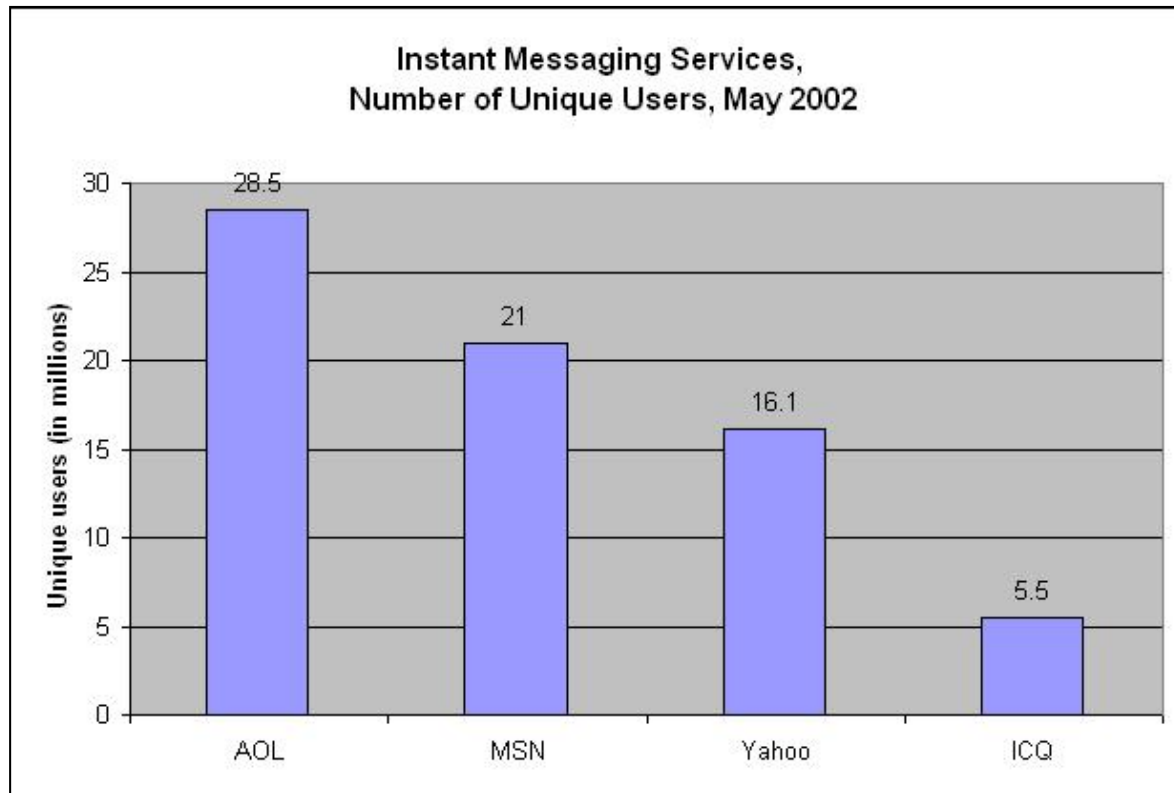
V. Instant Messaging

Instant Messaging Planet provides a concise definition of instant messaging:

"IM is a type of communications service that enables one to create a private chat room with another individual. Typically, the instant-messaging system alerts the user to whenever somebody on their private list is online -- a capability known as "presence." They can then initiate a chat session with that particular individual. People can communicate with each other by typing with a PC, wireless device (cell phone, PDA, etc.) or other Internet appliance/device."

Nearly 40 percent of the active U.S.-based Internet-using population at home logged onto one of the public instant-messaging (IM) networks at least once in May 2002, while 31 percent of U.S. business Internet users used IM in that same time frame, according to a Nielsen//NetRatings study. Gartner Group forecasts 70 percent of all enterprises will use IM in 2003, and that by 2005 IM will represent 50 percent of all business-to-client communication.

The most widely used IM services are AOL, MSN, Yahoo and ICQ, as shown in the figure below.



Source: *InternetNews.com*, June 17, 2002

AOL's messaging statistics illustrate the rapid growth of this communication medium. As of 1999, users were exchanging 400 million messages per day. By the end of 2000, the rate had increased to 660 million messages--double the daily traffic of the U.S. Postal Service. As of June 2003, IEEE reports, AOL messaging has exceeded 2 billion per day.

If we assume similar rates of messaging at the other IM services, we can estimate the total messaging volume: more than 5 billion messages per day. If each text-only message is 0.15 KB (average English-language message length, according to *Expresso Instant Messaging WhitePaper [???)*), then the total daily volume is **750 GB** and the annual volume is **274 terabytes**. All of this can be considered to be unique content.

VI. Peer to Peer (P2P) File-Sharing

A significant new source of storing, creating and exchanging media and data on the Internet is through P2P file sharing networks. P2P file sharing has exploded in popularity since the creation of programs such as Napster in the late 1990's. The first file sharing application, Napster, allowed users to share only music files in the popular MP3 format. Today, there are numerous applications that allow users to share any type of file on the network. In a short period of time, programs such as KaZaA have become the most popular applications on the Internet besides email and Web browsing. KaZaA is the most popular application ever downloaded on the Internet, having recently reached over 230 million downloads worldwide, with an average of 2 million more per week (Source: *Download.com*). In addition, users on KaZaA share almost **5,000 terabytes** of information, including over 600 million files shared by an average 3 million users active at any given time (Source: *KaZaA.com*).

P2P applications tell us about how information is consumed on the Internet, but unlike other parts of this study, relatively little of this information is unique. Looking at a sample of users on the P2P network (approximately 40,000 nodes, 2 million files, 14.4 Terabytes), we were able to determine what types of files were being shared and what kind of files people were sharing on their hard drives. We collected our sample over a period of 24 hours, traversing 400,000 nodes to find approximately 40,000 users were sharing files with others. Using this, we were able to describe how P2P users consume and share information. We looked specifically at the number and size of filetypes being shared per user and on the aggregate, as well as what percentage of these files were unique or similar. We did not look at the content of user files.

A. How P2P users use their hard disks

The dataset was 14.4 terabytes, averaging 7.6 megabytes, with a total file count of almost two million.

Table 8.7: KaZaA total size and count: aggregate size for the dataset

Size	Avg of Size	Min of Size	Max of Size	Std Deviation	Total Files
14.4 TB	7.6 MB	1 BYTE	1.97 GB	44168598.19	1,980,426

Source: *How much information 2003*

The three file types that took up the most storage space were .avi video files, mp3 audio files and .mpg video files; combined they were more than 80 percent of the sample.

Table 8.8: Summary of size by file extension for the dataset: top 25 file types sorted by aggregate file size

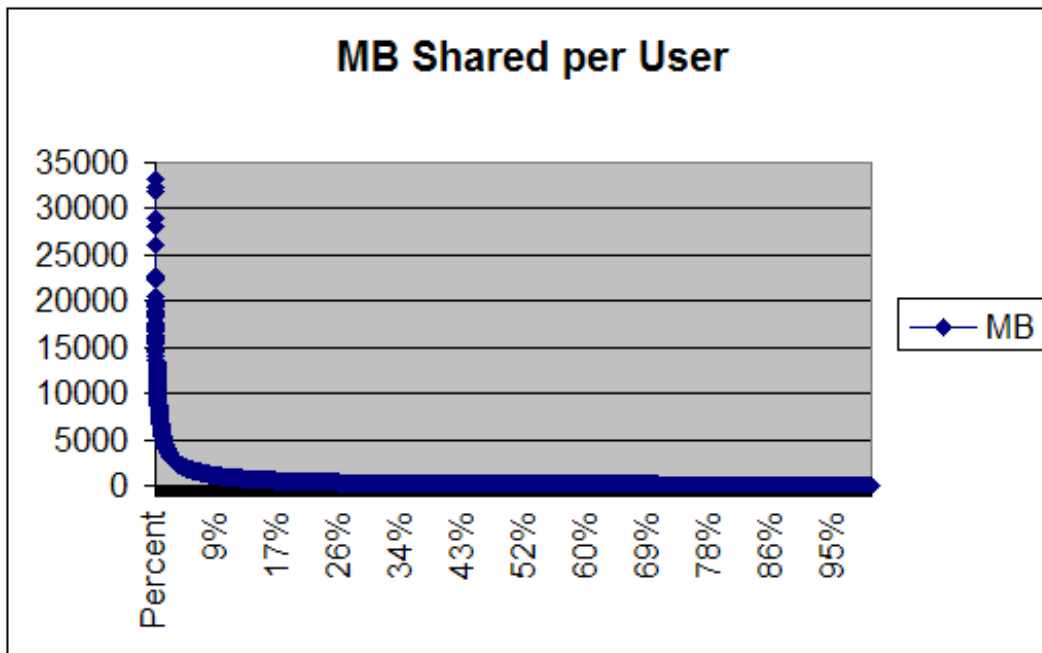
File Extension	GB	Avg of Size (bytes)	Min of Size (bytes)	Max of Size (bytes)	Number of files	Type	% of Total Size
avi	4,851.12	1.73E+08	82	2.12E+09	30,026	video	32.88%
mp3	4,564.79	4,120,597	1	5.58E+08	1,189,488	audio	30.94%
mpg	2,855.88	59,855,823	6	1.6E+09	51,231	video	19.36%
exe	550.76	7,293,783	1	8.16E+08	81,079	software	3.73%
mpeg	311.73	20,693,836	162	9.51E+08	16,175	video	2.11%
wmv	297.00	23,311,552	1316	8.75E+08	13,680	video	2.01%
asf	297.00	34,909,551	13	5.83E+08	9135	video	2.01%
none	248.37	19,767,776	1	1.67E+09	13491	unknown	1.68%
wav	243.44	15,177,476	44	1.24E+09	17,222	audio	1.65%
wma	113.79	2,732,813	111	7.64E+08	44,707	audio	0.77%
zip	63.23	21,470,521	1	1.91E+09	3,162	software	0.43%
mov	48.88	8,790,495	162	3.82E+08	5,970	video	0.33%
bin	34.29	44,311,001	10	8.45E+08	831	software	0.23%
vob	32.71	5.02E+08	8,192	1.07E+09	70	archive	0.22%

iso	27.68	3.81E+08	206,260	7.92E+08	78	archive	0.19%
cab	21.09	2,414,563	19	6.97E+08	9379	system	0.14%
rar	17.54	53,215,891	69	8.53E+08	354	archive	0.12%
pdf	13.22	5,411,649	162	1.34E+08	2,623	document	0.09%
cif	11.18	4.8E+08	38,062	8.36E+08	25	archive	0.08%
dat	8.16	3,,711,050	1	8.14E+08	2,361	system	0.06%
rmj	7.42	4,213,919	101,267	42,755,921	1,891	audio	0.05%
nrg	6.79	3.47E+08	342,172	8.16E+08	21	archive	0.05%
part	5.90	2.18E+08	21,759	6.15E+08	29	e-donkey temp files	0.04%
jpg	5.79	44,829.72	75	29,016,816	138,751	image	0.04%

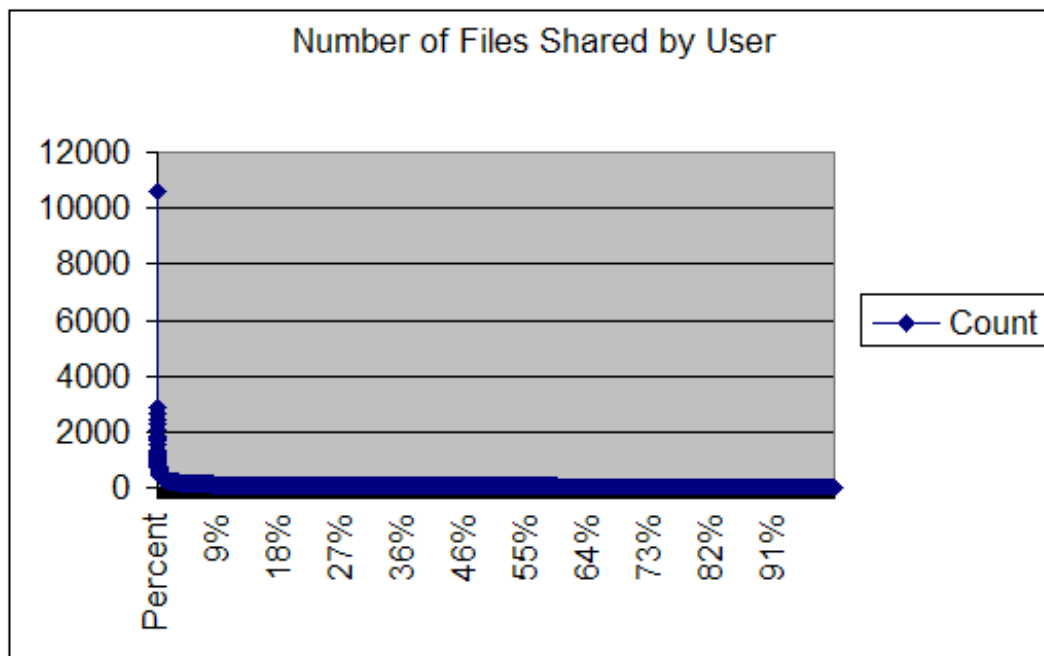
Source: *How much information 2003*

1. How users shared files

In addition to aggregate information, we looked at how many files each individual user had shared and how large their shares were. We discovered that the largest shared files on an individual user's hard disk was over 30 GB, while the smallest was just a couple of bytes. We discovered that 10 percent of the users accounted for about 60 percent of the total size of files shared, and 32 percent of the number of all files shared.



Source: *How much information 2003*



Source: *How much information 2003*

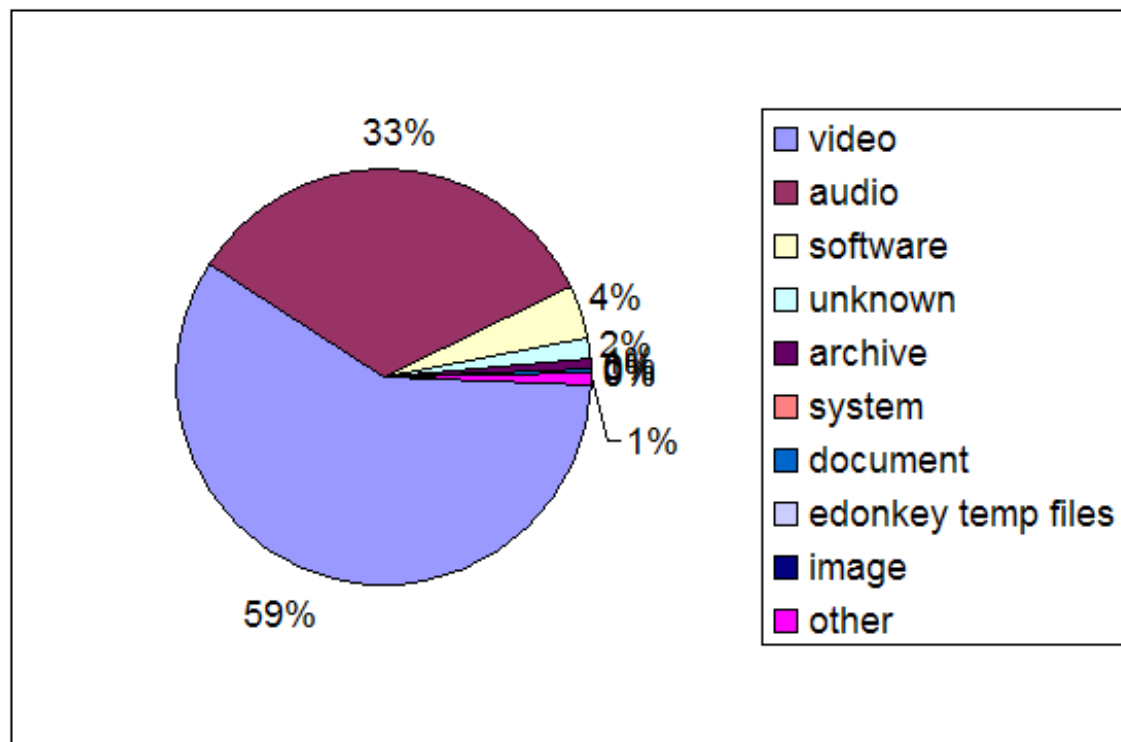
We found that only one user was sharing more than 10,000 files. The mean share size was between 300 and 400 files.

2. Summary of file size by type

We categorized the top 25 files into types and categories to get an idea of the distribution of file categories over our dataset. We found that video and audio data together accounted for about 90 percent of the total size of all files being shared.

Type	Sum of Size (GB)	Number of Files	% of total size
video	8,661.60	126,217	58.70%
audio	4,929.43	1,253,308	33.41%
software	648.28	85,072	4.39%
unknown	248.37	13,491	1.68%
archive	95.91	548	0.65%
system	21.09	9,379	0.14%
document	13.22	2,623	0.09%
'.part' temp files	5.90	29	0.04%
image	5.79	138,751	0.04%
other	124.94	351,008	0.85%

Source: *How much information 2003*



Source: *How much information 2003*

3. What file types are largest?

The largest file types are .AVI video files, followed by archival .ZIP files. AVI files are video files playable on a computer. The range of these in our sample is 82 bytes to 2GB, with most being in the 100-200 MB range. Pornography seems to be a major contributor to this traffic, according to user identification of genre types.

Table 8.10: Largest file types shared on P2P

File Extension	Type of File	Largest File Size (GB)	Total Size (in GB)
avi	Video	1.97	4,851
zip	Archive	1.77	63
mp2	Video/Audio	1.69	3
mpg	Video	1.48	2,855
wav	Audio	1.15	243
vob	Video	1.00	32
mpeg	Video	0.88	311
wmv	Video	0.81	297
ncd	Document	0.80	4

Source: *How much information 2003*

Note on Large Files (over 100 MB): We found 24,947 files of over 80 different file types that were larger than 100 MB in the sample. Although they accounted for only 1 percent of the total number of files in the dataset, they accounted for almost 50% of the size of the collection studied.

4. What file types are most common?

The most common files shared by P2P users are MP3 files, music files encoded using MP3 technology. Images (jpg, bmp) are also popular but take up much less space. The kpl files are KaZaA playlist files. Sixty percent of the files on users' hard disks were MP3 files, taking up about 30 percent of the space.

Table 8.11: Most common file types						
Ext	GB	Avg of Size	Min of Size	Max of Size	Count of KaZaASharesExt	% of Total
mp3	4,564.79	4,120,597	1	5.58E+08	1,189,488	60.06%
kpl	0.36	1,942.186	20	44,789,209	196,797	9.94%
jpg	5.79	44,829.72	75	29,016,816	138,751	7.01%
exe	550.76	7,293,783	1	8.16E+08	81,079	4.09%
mpg	2,855.88	59,855,823	6	1.6E+09	51,231	2.59%
wma	113.79	2,732,813	111	7.64E+08	44,707	2.26%
avi	4,851.12	1.73E+08	82	2.12E+09	30,026	1.52%
wav	243.44	15,177,476	44	1.24E+09	17,222	0.87%
bmp	2.52	160,220.4	54	1.34E+08	16,859	0.85%
mpeg	311.73	20,693,836	162	9.51E+08	16,175	0.82%
wmv	297.00	23,311,552	1,316	8.75E+08	13,680	0.69%
none	248.37	19,767,776	1	1.67E+09	13,491	0.68%
gif	0.13	15,275.78	34	1,598,776	9,394	0.47%
cab	21.09	2,414,563	19	6.97E+08	9,379	0.47%
ini	0.02	1,797.959	1	1,015,477	9196	0.46%
asf	296.10	34,909,551	13	5.83E+08	9135	0.46%
db	3.11	398,867.7	178	85,132,446	8,369	0.42%
dll	1.78	261,220.8	18	40,111,104	7318	0.37%
mov	48.88	8,790,495	162	3.82E+08	5,970	0.30%
txt	0.27	51,415.17	1	1.61E+08	5,553	0.28%
doc	0.49	125,125.3	28	11,816,199	4,189	0.21%
lnk	0.01	3,564.337	104	3,468,472	3661	0.18%
htm	0.06	20,782.51	4	18,030,044	3,226	0.16%
zip	63.23	21,470,521	1	1.91E+09	3,162	0.16%
pdf	13.22	5,411,649	162	1.34E+08	2,623	0.13%

Source: *How much information 2003*

Table 8.12: Top 10 most common file types shared on P2P

File Extension	Type of File	Total # of files
mp3	Audio	1,189,488
kpl	Kazaa playlists	196,797
jpg	Image	138,751
exe	---	81,079
mpg	Video	51,231
wma	Executable	44,707
avi	MP3	30,026
wav	Video	17,222
bmp	Image	16,859
mpeg	Video	16,175
	TOTAL	1,782,335

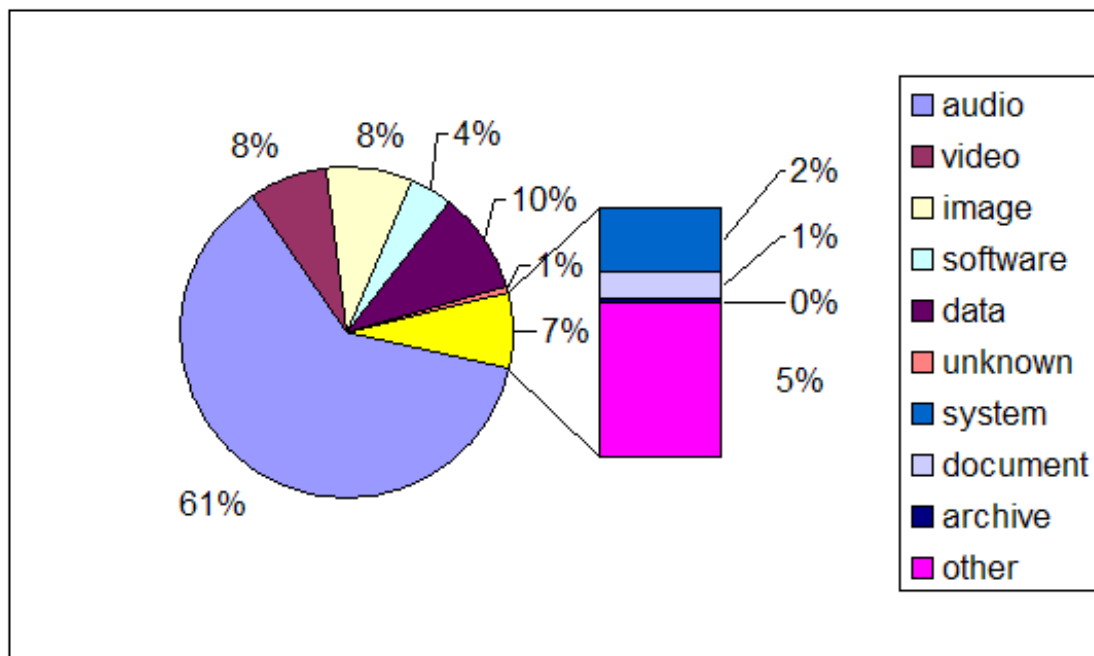
Source: *How much information 2003*

Audio files had the largest number of file (~62 percent) by far, followed by data files (~10 percent), images and video (~8 percent).

Table 8.13: Number of files by type

Type	Number	Percent
audio	1,220,390	61.6%
data	196,797	9.9%
image	165,004	8.3%
video	157,244	7.9%
other	89,745	4.5%
software	81,079	4.1%
system	37,923	1.9%
document	15,591	0.8%
unknown	13,491	0.7%
archive	3,162	0.2%

Source: *How much information 2003*



Source: *How much information 2003*

B. P2P consumption patterns

To look at the consumption of files we analyzed how many 'distinct' files are on the P2P network. By 'distinct' we mean the number of files with the same content, although the same file content often is given many different names. Each file has a hash code associated with it, that associates it with other identical files. To find the top 25 most common files, as defined by their content not their names, we divided the number of unique or distinct file hash codes by the total number of files found, to come up with a number of unique or distinct files being shared by file type.

Distinct Files/ Total Files = Total Distinct Files

<

Table 8.14: Total distinct files, by extension and type				
Extension	Total Files	Distinct Files	Total Distinct	Type
mp3	1,139,302	752,150	66%	audio
kpl	185,614	37,722	20%	audio
jpg	132,398	65,582	50%	image
exe	77,890	18,423	24%	application
mpg	49,147	21,807	44%	video
wma	42,649	26,251	62%	audio
avi	28,785	14,514	50%	video
wav	16,662	12,643	76%	audio
bmp	16,429	7,177	44%	image
mpeg	15,365	7,183	47%	video

none	12,925	6,461	50%	unknown
wmv	12,920	6,156	48%	audio
gif	9,146	5,735	63%	image
ini	8,840	5,516	62%	system
cab	8,814	1,210	14%	system
asf	8,608	3,209	37%	video
db	7,881	7,782	99%	system
dll	7,073	4,114	58%	system
mov	5,845	2,920	50%	video
txt	5,351	3,877	72%	document
doc	3,939	3,146	80%	document
lnk	3,516	3,181	90%	system
htm	3,168	2,280	72%	document
zip	3,071	2,647	86%	document
pdf	2,490	1,917	77%	document

Source: *How much information 2003*

Table 8.15 shows the top twenty music and movie genres, as identified by the users, rank ordered by size of files; percent measures proportion of genre files in the dataset.

Table 8.15: Top 20 genres, by file size			
Genre	Size (GB)	Count	% of Total Count
Erotica	1,516.89	42,273	2.1%
Comedy	1,002.20	22,345	1.1%
Action and Adventure	814.25	3,137	0.2%
Other	798.20	214,391	10.8%
Rock	535.52	133,749	6.8%
Rap	462.09	109,719	5.5%
Music and Musicals	450.25	13,498	0.7%
Science Fiction and Fantasy	332.16	1,427	0.1%
Series	303.61	2874	0.1%
Pop	294.00	89,399	4.5%
Drama	285.12	951	0.0%

Kids and Family	227.59	1,013	0.1%
R&B	200.75	46,714	2.4%
Horror and Suspense	192.77	932	0.0%
Hip-Hop	156.34	51,429	2.6%
Country	153.29	43,266	2.2%
Games	147.29	2,703	0.1%
Anime	143.94	1,409	0.1%
Blues	129.99	34,473	1.7%

Source: *How much information 2003*

VII. FTP Sites

FTP is the abbreviation for File Transfer Protocol. This protocol is used on the Internet for downloading and uploading files to a server.

To get an idea of the scale of the FTP universe we consider anonymous FTP sites. Filewatcher.org reports that as of August 2003, there are 5,700 anonymous FTP sites containing 269,177,202 files, for a total size of **97.04 TB**.

VIII. Usenet

According to Dictionary.com, Usenet (from User's Network) is a distributed bulletin board system that serves millions of readers worldwide. "Originally implemented in 1979 - 1980 by Steve Bellovin, Jim Ellis, Tom Truscott, and Steve Daniel at Duke University, and supported mainly by Unix machines, [Usenet] swiftly grew to become international in scope and, before the advent of the World-Wide Web, probably the largest decentralised information utility in existence." Many newsgroups on Usenet are devoted to sharing files, including software, music, and images.

"As of early 1993, Usenet hosted over 1,200 newsgroups and an average of 40 megabytes (the equivalent of several thousand paper pages) of new technical articles, news, discussion, chatter, and flamage every day. By November 1999, the number of groups had grown to over 37,000."

As of August 2003, approximately **280 GB** of data are posted to newsgroups on Usenet every day, resulting in a total annual flow of **102 terabytes**. According to James Robertson and Emil Sit of MIT, the size of this flow has been doubling every 10 months.

REFERENCES

World Wide Web

- Bergman, Michael K. "The Deep Web: Surfacing Hidden Value" (BrightPlanet White Paper)
<http://www.brightplanet.com/technology/deepweb.asp>
- CyberAtlas, citing the CIA World Factbook
http://cyberatlas.internet.com/big_picture/geographics/article/0,1323,5911_151151,00.html

Blogs (web logs)

- Blogcount <http://dijest.com/bc/>
- Blogging by the Numbers
http://cyberatlas.internet.com/big_picture/applications/article/0,,1301_2238831,00.html
- Day Pop, Top Web Logs <http://www.daypop.com/blogrank/>
- Spike Hall, <http://radio.weblogs.com/0106698/2002/10/26.html>
- LISTSERV Statistics <http://www.lsoft.com/news/statistics.asp>
- NITLÉ Blog Census [http://blogcensus.net./](http://blogcensus.net/)
- Pointblog <http://www.pointblog.com/past/000148.htm>
- USA Today article, July 8, 2003 "Welcome to the Blogosphere"
http://www.usatoday.com/tech/webguide/internetlife/2003-07-08-blogs_x.htm
- Jill Walker http://huminf.uib.no/~jill/archives/blog_theorising/final_version_of_weblog_definition.html

Email and spam

- Brightmail spam statistics <http://www.brightmail.com/spamstats.html>
- Brightmail statistics <http://news.com.com/2100-1001-955842.html>
- Anti-spam technologies <http://research.microsoft.com/displayArticle.aspx?id=414>
- IDC Report on email <http://emailuniverse.com/list-news/2002/10/01.html>
- NFO WorldGroup Survey, <http://one.ie/report/email/marketoverview.asp>
- Pew Internet and American Life Project, "Email at work,"
<http://www.pewinternet.org/reports/reports.asp?Report=79&Section=ReportLevel1&Field=Level1ID&ID=346>
- Forrester Research cited in <http://one.ie/report/email/marketoverview.asp>
- TechWeb, Dec. 12, 2002, "Spam may overtake email in 2003."
<http://www.cnn.com/2002/TECH/biztech/12/12/techweb.email.swamp/index.html>
- Newsweek, Aug 9, 1999, p.51, "Sending AOL a Message."
- Office of Senator Conrad Burns, Monthly Report, April 2003,
http://www.asme.org/gric/Fellows/Reports/PrabhatH_04-03.html
- CNET, News.com, July 1, 2003, "A call for worldwide action on spam," <http://news.com.com/2100-1028-1022552.html>

Instant Messaging and Chat

- Espresso Instant Messaging White Paper
<http://www.virtualthere.com/expresso/ExpressoRevenueModels.pdf>
- IEEE Spectrum Online, June 12, 2003, "AOL Programmer Lays Waste to Employer,"
<http://www.spectrum.ieee.org/WEBONLY/wonews/jun03/waste.html>
- Instant Messaging, Part I: Corporate Productivity Tool or Cool Toy?
http://www.intranetjournal.com/articles/200305/ij_05_01_03a.html
- Instant Messaging Planet <http://www.instantmessagingplanet.com/>
- InternetNews.com, Jun 17, 2002, <http://www.internetnews.com/stats/article.php/1366931>
- Wired News, Jan. 9, 2001, "No Whiners at AOL,"
<http://www.wired.com/news/technology/0,1282,40973,00.html>

FTP

- Filewatcher.org http://filewatcher.org/ftp-list/anonymous_ftp_sites_list.html

Usenet

- ComputerWeekly.com, June 26, 2001, "Usenet - a breeding ground for viruses,"
<http://www.computerweekly.com/Article103439.htm>
- Dictionary.com, Definition of Usenet, <http://dictionary.reference.com/search?q=usenet>
- James Robertson and Emil Sit, "UsenetDHT: Using DHTs for storage in Usenet," MIT Laboratory for Computer Science, 1 August 2003 project-iris.net/isw-2003/papers/robertson.ps
- Statistics for Article Size of Incoming [Usenet Messages] <http://newsfeed.mesh.ad.jp/flow/size.html>

How Much Information? 2003

Summary

Exec Summary

Stored Information

Paper | Film | Magnetic | Optical

Information Flows

Broadcast | Telephony | Internet

Wrap-up

Thanks | Printable (PDF)

Acknowledgements

We wish to thank the following individuals for their assistance in this research project:

Overall

- Jim Gray, Microsoft
- Leslie Shattuck, Intel
- Brewster Kahle, Internet Archive
- Tonya Comer, Hewlett-Packard
- Jeff Ubois
- David Bounie, author of "The International Production and Dissemination of Information," ENST, Paris.

Paper

- Gregg Fales, The Pulp & Paper Report
- Shawn Chen, Bowker Inc.
- Barrie Rappaport, IPSOS-NPD
- Tatiana Repkova, World Association of Newspapers
- Melanie Smuga, UNESCO
- Véronique Rat-Morris ISSN

Film

- Oksana Dykyj, Concordia University (and the Association of Moving Image Archivists)
- Dimitrios Delis, PMAI,
- Michelle Slaughter, InfoTrends Research Group

Magnetic

- Fred Moore, Horizon Information Strategies

Optical

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

Telephony

- Peter Davidson, Davidson Consulting

Internet

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

Release date: October 27, 2003. © 2003 Regents of the University of California

