

# 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?**

<b>Kilobyte (KB)</b>	<i>1,000 bytes OR <math>10^3</math> bytes</i> 2 Kilobytes: A Typewritten page. 100 Kilobytes: A low-resolution photograph.
<b>Megabyte (MB)</b>	<i>1,000,000 bytes OR <math>10^6</math> 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.
<b>Gigabyte (GB)</b>	<i>1,000,000,000 bytes OR <math>10^9</math> 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.
<b>Terabyte (TB)</b>	<i>1,000,000,000,000 bytes OR <math>10^{12}</math> 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.

<b>Petabyte (PB)</b>	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.
<b>Exabyte (EB)</b>	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%
<b>TOTAL:</b>	<b>5,421,221</b>	<b>3,416,281</b>	<b>3,212,731</b>	<b>2,132,238</b>	<b>69%</b>

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
	<b>Subtotal</b>	<b>1,633.8</b>	<b>326.7</b>	<b>1,199.8</b>	<b>240.2</b>	<b>36%</b>

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
<b>Paper</b>	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%
	<b>Subtotal</b>	<b>6,813</b>	<b>6,361</b>	<b>7.1%</b>

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%
	<b>Subtotal</b>		<b>420,254</b>	<b>74,202</b>	<b>431,690</b>	<b>58,209</b>

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%
<b>TOTAL</b>	<b>4,999,230</b>	<b>3,416,230</b>	<b>2,779,760</b>	<b>2,073,760</b>	<b>80%</b>

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%
	<b>Subtotal</b>	<b>102.9</b>	<b>50.9</b>	<b>80.7</b>	<b>28.7</b>	<b>28%</b>

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%
	<b>Subtotal</b>	<b>40.55</b>	<b>20.06</b>	<b>36</b>	<b>15.3</b>	<b>13%</b>

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
<b>TOTAL</b>	<b>17,905,340</b>

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
<b>Total:</b>				<b>43,329</b>	<b>72,443</b>

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
<b>Radio stations (2002, FCC)</b>					
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
<b>Total</b>	<b>13,261</b>	<b>19.8 million hours</b>	<b>0.05 GB/hour</b>	<b>987 TB</b>	<b>987 TB</b>
<b>Television (2002, FCC)</b>					
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
<b>Total</b>	<b>1,994</b>	<b>3.6 million hours</b>	<b>1.3 GB - 2.25 GB hour</b>	<b>4,731 TB</b>	<b>8,188 TB</b>

<b>Total (radio + TV):</b>	<b>5,718 TB</b>	<b>9,175 TB</b>
----------------------------	-----------------	-----------------

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
<b>TOTAL</b>	<b>17,300,000</b>

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
<b>TOTAL</b>	<b>532,897</b>

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](mailto: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.