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

Executive Summary

Abstract

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

Introduction

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

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

Information produced by medium

Most information is stored in four physical media: paper, film, optical (CDs and DVDs), and

magnetic. There are very good data for the worldwide production of each storage medium, and there are reasonably good estimates of how much original content is produced in each of these different formats.

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

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

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

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

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

The second striking fact is the ``democratization of data.'' A vast amount of unique information is created and stored by individuals. Original documents created by office workers are more than 80% of all original paper documents, while photographs and X-rays together are 99% of all

original film documents. Camcorder tapes are also a significant fraction of total magnetic tape storage of unique content, with digital tapes being used primarily for backup copies of material on magnetic drives.

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

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

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

Qualifications

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

Duplication.

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

Compression.

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

Archival Media.

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

World and US production.

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

Growth rates.

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

TV and Radio.

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

Digital Communication

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

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

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

Content	Terabytes
Email	11,285
Usenet	73

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

Although the social impact of the Web has been phenomenal, about 500 times as much email is being produced per year as the stock of Web pages. It appears that about 610 billion emails are

sent per year, compared to 2.1 billion static Web pages. Even the yearly flow of Usenet news is more than 3 times the stock of Web pages. As Odlyzko (2000) puts it, "communication, not content, is the killer app."

Non-digital Communication

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

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

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

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

Consumption of Information

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

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

Item	1992 Hours	2000 Hours	2000 MBytes	% Change
TV	1510	1571	3,142,000	4
Radio	1150	1056	57,800	-8
Recorded Music	233	269	13,450	15
Newspaper	172	154	11	-10
Books	100	96	7	-4
Magazines	85	80	6	-6
Home video	42	55	110,000	30
Video games	19	43	21,500	126
Internet	2	43	9	2,050
Total:	3,324	3,380	3,344,783	1.7

The notable features of this table are 1) the hours spent on TV and radio consumption and their

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

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

Individual and Published Information

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

Table 5: Yearly production of individual information		
Item	Amount	Terabytes
Photographs	80 billion images	410,000
Home videos	1.4 billion tapes	300,000
X-rays	2 billion images	17,200
Hard disks	200 million installed drives	13,760
Total:		740,960

The production of individual information can be compared to the amount of "published" information in Table 6. Note that the amount of "individual" information is over 2,600 times larger thanthe amount of published information.

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

Table 6: Yearly production of published information		
Item	Titles	Terabytes
Books	968,735	8
Newspapers	22,643	25
Journals	40,000	2
Magazines	80,000	10
Newsletters	40,000	.2
Office Documents	7,500,000,000	195
Cinema	4,000	16
Music CDs	90,000	6
Data CDs	1,000	3
DVD-video	5,000	22
Total:		285

Conclusion

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

About this Report

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

About the School of Information Management and Systems

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

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

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

Appendices

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

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

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

  - Film.
    If we used JPEG compression, rather than PhotoCD, we get a much smaller number for the storage requirements for images.

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

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

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

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

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

Bibliography

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

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

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

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

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

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

---