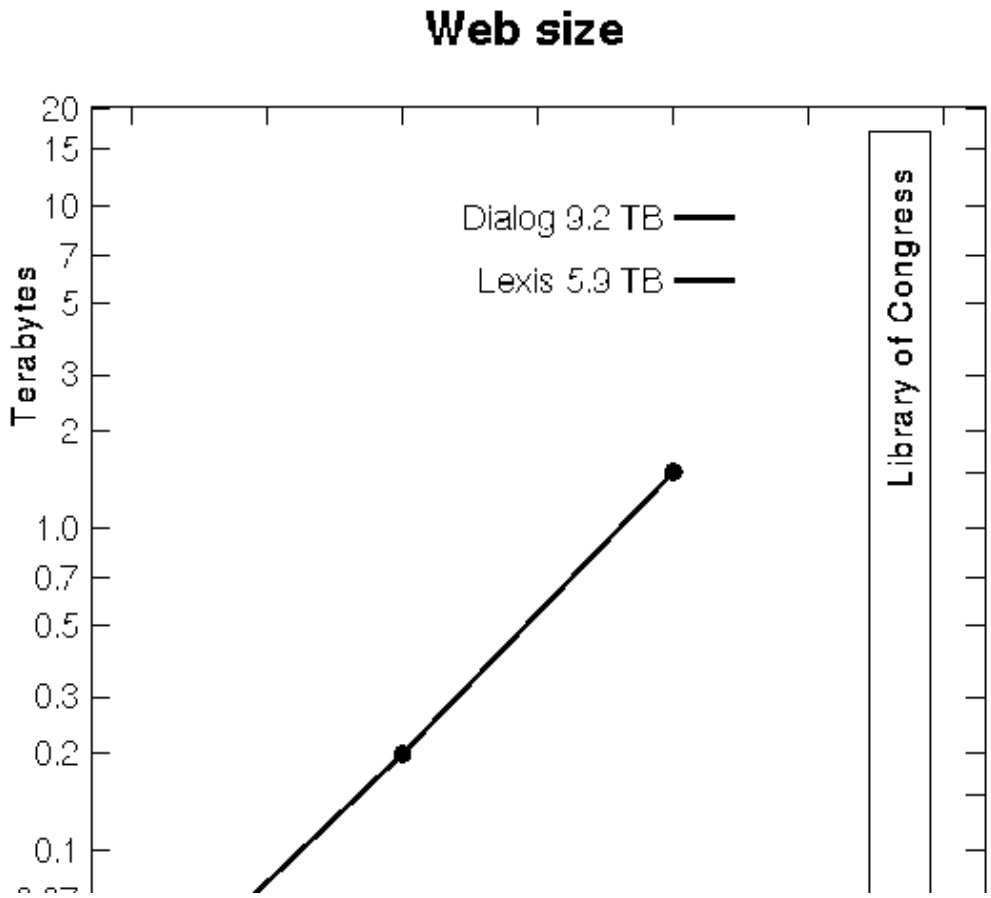
How Much Information Is There In the World?

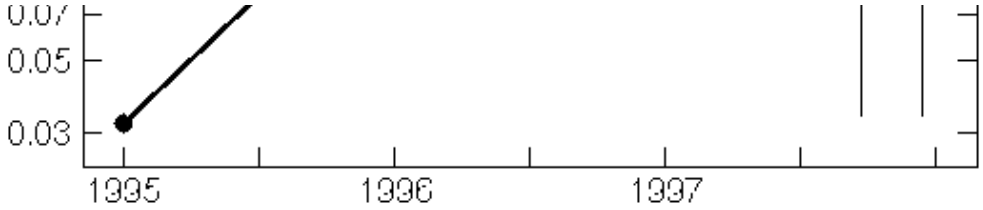
Michael Lesk

Abstract

How much information is there in the world? This paper makes various estimates and compares the answers with the estimates of disk and tape sales, and size of all human memory. There may be a few thousand petabytes [*] of information all told; and the production of tape and disk will reach that level by the year 2000. So in only a few years, (a) we will be able save *everything* \- no information will have to be thrown out, and (b) the typical piece of information will *never* be looked at by a human being.

Here is a chart of the current amount of online storage, comparing both commercial servers [Tenopir 1997]. and the Web [Markoff 1997]. [Mauldin 1995]. with the Library of Congress. These numbers involve Ascii text files only. This chart suggests that next year the Web will be as large as LC.





The Web has been growing 10-fold each year. Can it continue to do so and for how long? Current estimates of the number of Internet users run in the tens of millions, perhaps 50 M, and this might grow to one billion; thus a factor of twenty is available by increasing the number of people on the Web, but not more. Can people put more and more of their life online? Perhaps, but I suspect not more than another factor of 20. This suggests that the amount of Ascii on the Web might increase to 800 terabytes. Is there that much text around? What about images, movies, and sounds?

How much traditional information is there?

The 20-terabyte size of the Library of Congress is widely quoted and as far as I know is derived by assuming that LC has 20 million books and each requires 1 MB. Of course, LC has much other stuff besides printed text, and this other stuff would take much more space.

1. Thirteen million photographs, even if compressed to a 1 MB JPG each, would be 13 terabytes.
2. The 4 million maps in the Geography Division might scan to 200 TB.
3. LC has over five hundred thousand movies; at 1 GB each they would be 500 terabytes (most are not full-length color features).
4. Bulkiest might be the 3.5 million sound recordings, which at one audio CD each, would be almost 2,000 TB.

This makes the total size of the Library perhaps about 3 petabytes (3,000 terabytes).

Of course the most important discrepancy in comparing the Web and the Library of Congress is that the Library of Congress predominantly contains published materials. The Web has more text than LC already, if you only ask for English-language material written in the last 18 months. I tried to guess what fraction of Web material represents something that has been published, however, by sampling fifty random English-language URLS. I found fourteen which looked to me as if they were probably in a large conventional library, or 28%. By contrast most of the contents of Lexis-Nexis and Dialog are versions of published material, albeit much more easily searched.

What other kinds of traditional information might be around? The United States manufactures 38 million tons a year of the kind of paper used for writing and printing. If a typical pound of paper is 220 A4 pages and each sheet held 5000 bytes, that

would be about 8,000 terabytes of text each year. Of course many of the sheets are copies of other sheets, and many of them do not contain words. How much could reasonably be written fresh? Suppose that half the pages have text and that we assume 100 copies of the average sheet; that would be 40 terabytes of fresh information. If 40 million U. S. `knowledge workers' each wrote 1 megabyte a year, that would also be 40 terabytes a year. Since the US gross domestic product of $7T is about one-quarter of the world GDP ($30.8B) I will in general multiply the US by 4 to extrapolate to the earth, and suggest that the entire world's writing amounts to 160 terabytes each year. Of this the published books are about 863,000 (in 1991), plus 9,315 newspapers, [UNESCO 1995]. making perhaps a terabyte of professionally written or refereed material, not even 1% of the total.

Other kinds of information, compared with Ascii text, are bulkier.

1. *Cinema*. There were 4,615 films made world-wide in 1989; at 5MB/sec and 7200 seconds average, that would be 166 terabytes.
2. *Images*. There are about 52 billion (thousand million) photographs taken each year in the world. [Mills 1996]. If each of those is a 10 KB JPG, that is 520,000 terabytes, or 520 petabytes, and these are actually all different. Again, less than 1% represent professionally taken or reviewed pictures, probably less than 0.1%. By comparison even the NASA earth observing project, expected to accumulate 11,000 terabytes, [Fargion 1996]. doesn't affect the numbers.
3. *Broadcasting*. In the US, we have 1593 television stations. If each sends out 5 MB/sec for 30 million seconds per year, that is over 200 petabytes. However, one might expect that only about 1/10 of the programming is actually different for different stations; that is 20 petabytes of distinct programming, and extrapolated to the world would be 80 petabytes. Radio, by contrast, is insignificant; the US has 6,956 radio stations and if each sends out 30 million seconds per year at 8 KB/sec we would have only 1.7 TB in the United States.
4. *Sound*. Sales of recorded music in the US in 1992 were 407 million CDs and 336 million cassettes (and 20 million vinyl disks, still). Assuming 550 MB for each CD and cassette that would be 400 petabytes, much duplicated of course. If the number of different recordings for sale is about 30,000 this would be 15 terabytes in the US and 60 terabytes world-wide.
5. *Telephony* The largest storage requirement would come from converting all telephone conversations to digital form. In the US in 1994 there were 500 billion call-minutes of `interlata toll' and there is about 20 times as much local calling, so at 56 kbits/sec this would be 4,000 petabytes of digitized voice. The only thing I am not considering is consumer videotape, on the grounds

that much of it is used to record off-the-air TV and duplicates the TV stations.

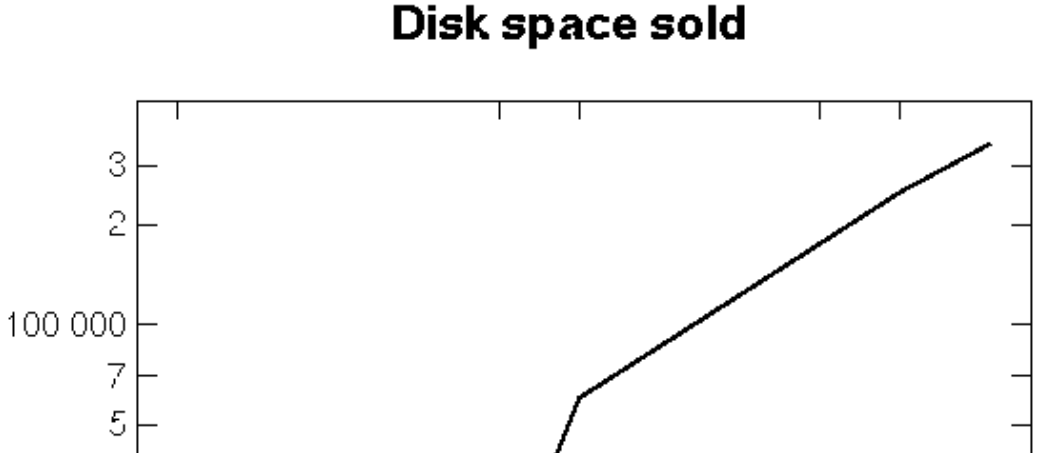
The conclusion is that in terms of text there are terabytes of information and perhaps one terabyte of professional information. Including sounds and images there are thousands of petabytes of information. The letter from Sincerbox which started all of this suggested that there would be 12,000 petabytes of information in the world, perhaps not an unreasonable guess. Only a small part of this, dominated by the TV stations, is commercially produced or validated in some way; perhaps that amounts to 100 petabytes.

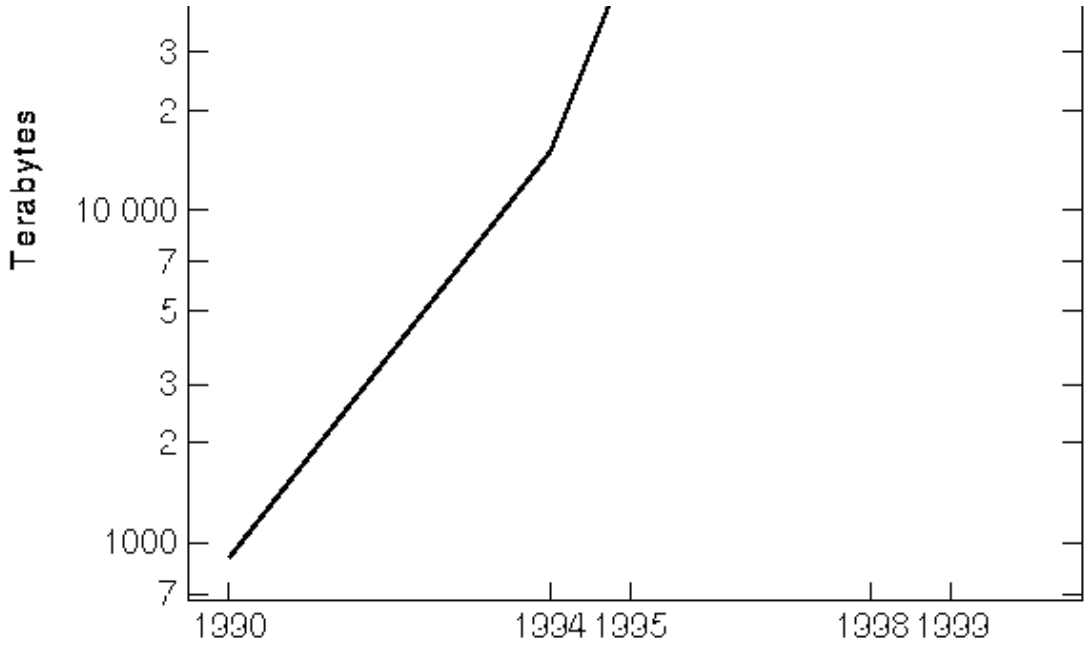
How much computer storage space is there?

The single largest data storage system I have seen described is a year-old description of the Accelerating Strategic Computing Infrastructure project at Livermore, Los Alamos and Sandia Laboratories, which has 75 terabytes of disk, and a plan for hundreds of petabytes of tape archive. [Louis 1996]. The Los Alamos HD-ROM project using scanning electron microscopes to etch bits into stainless steel in a vacuum, which has been transferred to the startup company Norsam Technologies, has achieved 200 GB/square inch. They hope to put 12 terabytes on a single CD-size disk.

One way of guessing the total size of the world's computer storage is simply to view the single largest establishment as one point on a log-normal curve. To oversimplify, the largest city in the world has about 1/300 the population of the world. and the largest company in the world has about 1/300 the world's GDP. So this suggests that if the largest disk farm in the world in 1996 was 75 terabytes, the total disk space in the world was 22,500 terabytes.

Of course, there are statistics on the disk drive industry. The chart below makes a guess at how many terabytes of disk space are sold per year, using data from Computerworld, [Radding 1990]. IBM, [Bell 1994]. and Optitek. [Optitek]. The different uncoordinated sources for this table make it fairly irregular; I've been unable to find good numbers from a single source. But it is clear the answer today is tens of thousands of terabytes of disk sold each year.





Optitek predicts 1998 sales and capacities of different storage media:

Device	Price	Total market	Total size
Magnetic disk	$100/GB	$25B	250 petabytes
RAID disk	$200/GB	$13B	65 petabytes
Optical disk	$20/GB	$0.5B	25 petabytes
Optical jukeboxes	$20/GB	$5B	250 petabytes
Magnetic tape	$1/GB	$10B	10,000 petabytes
Tape stackers	$1/GB	$2B	2,000 petabytes

Both Alan Bell of IBM and Jim Gray of Microsoft estimate that 200 petabytes of tape storage were sold in 1995.

Note that these numbers added up are all comparable to the size of the numbers for the total amount of information in the world. So the implication is that in the year 2000 we will be able to save in digital form everything we want to \- including digitizing all the phone calls in the world, all the sound recordings, and all the movies. We'll probably even be able to do all the home movies in digital form. We can save on disk everything that has any contact with professional production or approval. Soon after the year 2000 the production of disks and tapes will outrun *human* production of information to put on them. Most computer storage units will have to contain information generated by computer; there won't be enough of anything else.

Of course, this has already true despite the lower size of computer memory today. The typical computer disk byte is probably part of some Microsoft object module. After that, it's probably some kind of database. But we still see complaints that relatively little of the data in many large archives (the NASA files or the Palomar sky survey) has ever been looked at by anyone. That will be normal in the future: computer memory will be mostly for other computers. Today this memory is highly duplicative, with tens of millions of copies of popular

programs. Tomorrow, with everyone on-line with high speed connections, and extended use of site license agreements, it may be common for PCs to fetch on demand object modules of software needed once in a while, as we already do at Bellcore. The disks on our machines will then be available for our own personal information. A fast author might write a megabyte a year; not even Trollope wrote 100 MB in his life; but we'll all have at least a gigabyte of personal storage by 2000, when we have about as many petabytes of disk sold as there are millions of computers in the world (300 each, roughly).

How much human memory is there?

And to look at a third measure, how much does human memory hold? Tom Landauer tried to estimate this some years ago and concluded that the brain held about 200 megabytes of information. [Landauer 1986]. He got this number partly by looking at the rate at which people could take in information, both by reading and by looking at pictures. He also studied estimates of the rate at which people forget things, and the amount of information adults need in order to do the tasks they normally do. His numbers (expressed in gigabits, not gigabytes), were 1.8, 3.4, 2.0, 1.4 and .5 gigabits. Averaging these and dividing by 8 yields 227 MB. Since there are between 10e12 and 10e14 neurons, this suggests that the brain contains 1,000 to 100,000 neurons for each bit of memory. Of course, much of the brain is used for perception, motor control, and the like; but even if only 1% of the brain is devoted to memory Landauer pointed out that it looks like your head accepts considerable storage inefficiency in order to be able to make effective use of the information.

With something like 6 billion people on earth, that makes the total memory of all the people now alive about 1,200 petabytes. To the accuracy with which these calculations are being done, the results are comparable. We can store digitally everything that everyone remembers. For any single person, this isn't even hard. Landauer estimated that people only take in and remember about a byte a second; a typical lifetime is 25,000 days or 2 billion seconds (counting time asleep). The result is 2 gigabytes, or something that fits on a laptop drive.

Would it be hard to remember every word you heard in your lifetime, including the ones you forgot? The average American spends 3,304 hours per year with one or another kind of media. [Census 1995]. 1,578 hours are with TV; adding in 12 hours a year of movies, at 120 words per minute that's 11 million words, perhaps 50 megabytes of Ascii. And 354 hours a year of reading newspapers, magazines and books at 300 words per minute reading speed would be another 32 megabytes of text. In seventy years of life you would be exposed to around six gigabytes of Ascii; today you can buy 23 gigabyte disk drives.

Could we simply make a wearable device that would record everything? Yes, if either (a) we had decent speech recognition and OCR, or (b) books move to electronic form and TV sets

provide access to the closed-captioned Ascii form of the scripts. Perhaps both of these choices are likely in the near future. School children no longer need to do arithmetic without calculators; perhaps they will soon no longer need to memorize anything either. If you think this is horrible remember that Plato (in the *Phaedrus*) suggested that writing would `create forgetfulness in the minds of those who learn to use it' and would create `the show of wisdom without the reality.' If writing something down isn't cheating, why is recording it? It is now common for speakers to use transparencies, for a conference to hand out printed proceedings, and for people to sit at talks with cassette recorders. Would it be that terrible if each attendee had a laptop doing speech recognition, and the laptop kept the transcript and provided a small vibration to wake up the attendee when a promising topic was mentioned?

Two years ago I heard Ted Nelson at a conference suggest that we should keep the entire record of everyone's life \- all the home snapshots, videos and the like. Some six-year-old, he said, is going to grow up to be President; and then the historians will wish we knew absolutely everything about his or her life. The only way to do this is to save everything about everyone's life. I laughed, but it's indeed possible. Whether it is worthwhile is another question: are we better off having all possible information and giving it the most sketchy consideration, or having less information but trying to analyze it better? Computers do not use log tables, and chess computers have dictionaries of opening and endgame positions but not whole games. We need to understand our ability to model more complex situations to know how to make best use of stored information.

Conclusion

There will be enough disk space and tape storage in the world to store everything people write, say, perform or photograph. For writing this is true already; for the others it is only a year or two away. Only a tiny fraction of this information has been professionally approved, and only a tiny fraction of it will be remembered by anyone. As noted before the storage media will outrun our ability to create things to put on them; and so after the year 2000 the average disk drive or communications link will contain machine-to-machine communication, not human-to-human. When we reach a world in which the average piece of information is *never* looked at by a human, we will need to know how to evaluate everything automatically to decide what should get the precious resource of human attention.

Today the digital library community spends some effort on scanning, compression, and OCR; tomorrow it will have to focus almost exclusively on selection, searching, and quality assessment. Input will not matter as much as relevant choice. Missing information won't be on the tip of your tongue; it will be somewhere in your files. Or, perhaps, it will be in somebody else's files. With all of everyone's work online, we will have the

opportunity first glimpsed by H. G. Wells (and a bit later and more concretely by Vannevar Bush) to let everyone use everyone else's intellectual effort. We could build a real `World Encyclopedia' with a true `planetary memory for all mankind' as Wells wrote in 1938. [Wells 1938]. He talked of ``knitting all the intellectual workers of the world through a common interest;" we could do it. The challenge for librarians and computer scientists is to let us find the information we want in other people's work; and the challenge for the lawyers and economists is to arrange the payment structures so that we are encouraged to use the work of others rather than re-create it.

Acknowledgment.

This paper was suggested by a query from Glenn Sincerbox of the University of Arizona.

---

* Here are the names of the units of very large storage sizes:

gigabyte	1,000 megabytes
terabyte	1,000 gigabytes
petabyte	1,000 terabytes
exabyte	1,000 petabytes

[Bell 1994]. Alan Bell; *IBM Academy Digital Library Workshop* (Sept 12-13, 1994).

[Census 1995]. United States Census Bureau *Statistical Abstract of the United States* Government Printing Office (1995).

[Fargion 1996]. G. S. Fargion, R. Harberts, and J. G. Masek An Emerging Technology Becomes an Opportunity for EOS From the online file; see the URL: http://ecsinfo.hitc.com/cdwg/datamining/overview.html.

[Landauer 1986]. T. K. Landauer; "How much do people remember? Some estimates of the quantity of learned information in long-term memory," *Cognitive Science,* **10** (4) pp. 477-493 (Oct-Dec 1986).

[Louis 1996]. Steve Louis *Cooperative High-Performance Storage in the Accelerated Strategic Computing Initiative* 5th NASA Goddard Conference on Mass Storage Systems and Technologies (Sept. 17-19, 1996). As reported by Ron Van Meter, http://www.isi.edu/~rdv/conferences/goddard96.html .

[Markoff 1997]. John Markoff; "When Big Brother is a Librarian," *The New York Times* pp. 3, sec. 4 (March 9, 1997).

[Mauldin 1995]. Matt Mauldin, "Measuring the Web with Lycos," *Third International World-Wide Web Conference*, April 1995.

[Mills 1996]. Mike Mills; "Photo Opportunity," *Washington Post* pp. H01 (January 28, 1996).

[Optitek]. The Need for Holographic Storage http://www.optitek.com/hdss_competition.htm.

[Radding 1990]. Alan Radding; "Putting data in its proper place," *Computerworld* pp. 61 (August 13, 1990).

[Tenopir 1997]. Carol Tenopir, and Jeff Barry; "The Data Dealers," *Library Journal* pp. 28-36 (May 15, 1997).

[UNESCO 1995]. *UNESCO Statistical Yearbook* Bernan Press (1995).

[Wells 1938]. H. G. Wells *World Brain* Methuen (1938).