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

Internet - Summary

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

World Wide Web

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

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

estimate of **10 to 20 terabytes** of textual content. At 7.3 million new pages added every day, the rate of growth is [taking an average estimate] 0.1 terabytes of new information [HTML-included] per day.

If we take into account all web-accessible information, such as web-connected databases, dynamic pages, intranet sites, etc., collectively known as "deep" Web, there are **550 billion web-connected documents**, with an average page size of 14 kbytes, and 95% of this information is publicly accessible [2]. If we were to store this information in one place, we would need **7,500 terabytes** of storage, which is 150 times more storage than we would need for the entire "surface" Web, even taking the highest estimate of 50 terabytes. 56% of this information is the actual content [HTML excluded], which gives us an estimate of **4,200 terabytes** of high-quality data. Two of the largest "deep" web sites - National Climatic Data Center and NASA databases - contain **585 terabytes** of information, which is 7.8% of the "deep" web. And 60 of the largest web sites contain **750 terabytes** of information, which is 10% of the "deep" web.

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

More Details

Email & Mailing Lists

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

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

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

More Details

Usenet

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

FTP

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

IRC, Messaging Services, Telnet...

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

References

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