What is the content of the world's technologically mediated information and communication capacity: *how much text, image, audio* and *video*?

Martin Hilbert, Email: martinhilbert@gmail.com
Department of Communication, University of California, Davis, CA, USA and United Nations Economic Commission for Latin America and the Caribbean (UN-ECLAC), Santiago, Chile

This is an author's preprint of an article published as
Hilbert, M. (2014). What Is the Content of the World's Technologically Mediated Information and Communication Capacity: How Much Text, Image, Audio, and Video? The Information Society, 30(2), 127–143. doi:10.1080/01972243.2013.873748

ABSTRACT: This article asks if the global process of digitization has led to noteworthy changes in the shares of the amount of text, images, audio and video in worldwide technologically stored and communicated information content. We empirically quantify the amount of information that is globally broadcasted, telecommunicated and stored (1986 – 2007) and assess the evolution of the respective content shares. Somewhat unexpectedly it turns out that the transfer from analog to digital has not led to toward increasing shares of media-rich audio and video content, despite vastly increased bandwidth. First, there is a certain inertia in the evolution of content, which seems to stick to stable proportions independently from its technological medium (be it analog vinyl and VHS tapes, or digital CDs and hard disks). Second, the relative share of text and still images actually captures a larger portion of the total amount than before the digital age. Text merely represented 0.3 % of the (optimally compressed) bits that flowed through global information channels in 1986 but grew to almost 30 % in 2007. On another level, we are seeing an increasing transition of text and images from one-way information diffusion networks (like newspapers) to digital storage and two-way telecommunications networks, where it is more socially embedded. Both tendencies are good news for big data analysts who extract intelligence from easily analyzable text and image data.

Keywords: content, multimedia, digital, methodology, measurement, Big Data.

Acknowledgements The author is highly indebted with Priscila López, who led the elaboration of the database on which this analysis is based, and thanks the United Nations Economic Commission for Latin America and the Caribbean and USC's Annenberg School of Communication for their support. The views expressed herein are those of the author and do not necessarily reflect the views of the United Nations.

Over recent decades human kind has equipped itself with digital technologies that has led to an unprecedented explosion in available information storage and communication capacities (see Hilbert and López, 2011). The availability of new technological capacity has been expected to change to the kinds of informational content we produce, transmit, and store. For instance, there has been much talk about "multimedia" that combine traditional text and still images, with vibrant audio and animated video content (Ludwig in Vaughan, 2010).[1]

On the one hand, it seems reasonable to expect that the ever increasing amount of bandwidth and storage space has led to an intensified use of bit-heavy audio and video material. In other words, in an age of bandwidth and storage abundance, multimedia content should thrive. Journalists, for example, have started to enrich their static print and image filled articles with additional content (Fortunati, et al., 2009). User-friendly Web 2.0 applications have enabled even non-tech-savvy users to present their personal multimedia content (Jenkins, 2004; O'Reilly, 2005; Burns, 2007; Kalmus, et al., 2009). In an interactive Web space, audio and video content seems to be better received than text-filled Websites: by 2007, between 20 % and 62 % of online teens from the U.S. to Estonia have uploaded videos online (Lenhart, et al., 2007; Kalmus, Pruulmann-Vengerfeldt, Runnel and Siibak, 2009). This contrasts with only 50 % of teens that post written comments on online forums and only 24 % that upload written stories or poems (Kalmus, et al., 2007). This matters for the global communication structure as a whole, as user-created multimedia content platforms have become the centerpiece of the global information and communication infrastructure. In November 2010 the social network Facebook captured 38 % of all unique worldwide web visitors, and the video-sharing platform YouTube another 32 % (Google, 2010), together capturing 7 out of 10 Internet visits. Such anecdotal evidence would suggest that the explosion of bandwidth and storage space, combined with the ease and multimedia user-friendliness of the digital revolution has led to a proliferation of audio and videos, while static text-based content have become increasingly become marginalized.

On the other hand, scholars have also pointed to the outstanding success of text-based SMS (short message service) mobile phone services (Lai, 2004; Brown, Shipman and, Vetter, 2007) and to the importance of digital text on Websites, blogs and eBooks (Schmidt, 2007; Fowler and Baca, 2010). Furthermore, vast and ever increasing databases of financial and alphanumeric scientific data have started to fill ever-expanding server farms in transnational companies and universities (Madnick, Smith and Clopeck, 2009; Manyika, et al., 2011). This would suggest that the share of alphanumeric text has increased.

Here we need to be mindful of the fact that intuition and qualitative assessments on basis of anecdotal examples can be deceptive. The notion of increasing and decreasing shares of content

[1] One of the standard multimedia textbooks suggests: "Use traditional text and graphics where appropriate; add animation when 'still life' won't get your message across; add audio when further explanation is required" (Ludwig in Vaughan, 2010, p. 193).

refers to relative proportions, which are continuously updated by an ever-changing Bayesian conditioning variable – exponentially growing technological capacities to store and communicate information. It is undoubtedly true that there is more audio and video content in absolute terms than before the digital age, simply because there is more content in absolute terms. But has static text and image content increased even more than dynamic audio and video content and led to a shift in relative terms? Our intuition is notoriously bad in grasping both these kinds of Bayesian updating tasks (see Monty Hall problem, 2013) and the exponential rates of change that accompany technological progress (Kurzweil, 2001).

This calls for a more thorough empirical analysis of these tendencies. A deeper understanding of these tendencies not only matters to get the historical record straight, but also at the levels of theory development and practical application. From a theoretical perspective, a rapid change in content with move from analog to digital media infrastructure would suggest a technological driven reality in which "the medium" changes the nature of the message (McLuhan, 1994). If the trends in content production do not change significantly, it would suggest a certain inertia in the development of content, which is rather independent of technological changes.

On a practical level, an understanding of the changes in nature of content carried by world's communication and storage technologies has implications for the so-called "big data" research (Nature Editorial, 2008; Mayer-Schönberger and Cukier, 2013), which promises to enable more intelligent decisions through algorithmic analysis of the available digital data. Until now, the vast majority of the data subject to this kind of analysis has consisted of alphanumerical or static still images, which are much easier to analyze than dynamic videos and audio data (first approximation of any algorithmic analysis of video or audio data converts them into still images and static alphanumeric data, and then uses similar pattern detection algorithms as the ones developed for text and images). In fact the digital age has produced more alphanumerical data than we can currently handle: for now, financial and credit card providers discard around 80-90 % of the mainly alphanumeric data they generate (Zikopoulos, et al., 2012; Manyika, et al., 2011). At the same time, increasingly powerful artificial intelligence applications are being built to extract intelligence from audio and video content (Wang, Liu and Huang, 2000). However, these developments are still in their child shoes and for now less effective.

To inform such analyses, we take inventory of the worldwide evolution of technologically stored and communicated information content. The goal is to provide hard-fact empirical evidence of the historical transition from the analog to the digital age (1986-2007) from a macro perspective.

Taking inventory of the multimedia revolution

Such undertaking poses first and foremost a methodological challenge. We have to gather, combine and aggregate literally thousands of pieces of anecdotal evidence (such as the ones previously mentioned) in a systematic, transparent, and replicable manner. For this we developed

a methodology that allows us to estimate the world's technological capacity to store (in bits) and to communicate information (in bits per second), and gather input from more than 1,100 distinct sources (for details and sources see the online Supporting Appendix that was developed in collaboration with Priscila López: López and Hilbert, 2012; also see Hilbert and López, 2011; 2012a, 2012b). We essentially multiply the number of the diverse technological devices (such as phones, hard-disks and paper books) with their respective informational performance (in optimally compressed bits or bits per second) and then sum up this product:

$$Technological\ information\ processing\ capacity = \\ = \sum_{i\ =\ all\ considered\ technologies} [device_i \times performance\ of\ device_i]$$

The first variable (number of devices) requires statistics on the installed infrastructure. The second variable (performance of device) requires three different kinds of input for each technology: registries on the hardware performance of each device; statistics on the distribution of content handled by the diverse technological devices, and estimations of the most commonly used compression software to compress the different kinds of content (otherwise compression software with different levels of efficiency would confound the inventory). We are thereby able to measure storage and communication (telecom and broadcasting) in optimally compressed bits and bits per second, respectively.[2]

We apply this methodology to the 12 most widely used families of analog storage technologies[3] and the 13 most prominent families of digital memory[4], as well as to 6 analog and 5 digital broadcast technologies[5], and 3 bidirectional analog telecommunication technologies and their 4 most common digital heirs[6]. This allows us to include both, analog and digital information. We focus on the period between 1986 and 2007 and make calculations for 280 subgroups with

---

[2] All of our estimates are yearly averages. We recognize that the installed technological stock of a given year is the result of a process of accumulation over previous years, whereas each year's technologies contribute with different performance rates.

[3] Analog storage technologies consist of: video analog, photo print, audio cassette, photo negative, cine movie film, vinyl LP, TV episodes film, x-rays, TV movie film, newsprint, other paper and print, books.

[4] Digital storage technologies consist of: PC hard-disk, DVD and Blu-Ray, digital tape, server and mainframe hard-disk, CDs and minidisks, other hard-disks (i.e. portable), portable media player, memory cards, mobile phones and PDA, videogames other than hard-disk and optical (mainly ROM and cartages), floppy disks, digital camera and camcorders internal, chip cards.

[5] Broadcasting (unidirectional) analog consist of: TV-Terrestrial, TV-cable, TV-satellite, radio, newspapers, paper advertisement. Broadcasting (unidirectional) digital consist of: TV-terrestrial, TV-cable, TV-satellite, radio, personal navigation GPS.

[6] Telecommunications (bidirectional) analog consist of: fixed (voice) phone, mobile (voice) phone, paper postal letters. Telecommunications (bidirectional) digital consist of: fixed (voice) phone, Internet, mobile (data) phone, mobile (voice) phone.

different performances for a given year (66 for computation, 172 for storage, 23 for telecom, and the rest for broadcasting) (see online Supporting Appendix for details and sources).

The inputs we rely on include databases from international organizations (such as ITU, 20110; UPU, 2007; IFPI, 2007), historical inventories developed by individuals for commercial or academic purposes (such as Coughlin, 2007; Porter, 2005), publicly available statistics from private research firms (such as Morgan Stanley, 2006; IDC, 2008), as well as a myriad of sales and product specifications from equipment producers. Frequently, we compared diverse sources for the same phenomena and strove for reasonable middle grounds in case of contradictions.

Methodological and statistical challenges for content measurement

While other inventories have aimed at the quantification of the amount of technologically-mediated information (Ito, 1981; Pool, 1983; Lyman, et al., 2003; Bohn and Short, 2009; Neuman, Park and Panek, 2012, for an overview see Hilbert, 2012), none of them has focused on the evolution of the distribution of content. In order to create meaningful time series for the world's technological capacity to store and communicate information, we divide the global content into 4 kinds: text (including all alphanumerical information in databases, documents and archives), image (including photos, graphics and drawings), audio (including voice, sounds and music) and video (including animations, streaming and high quality video content).

We exclusively estimate the content of the world's technological capacity, independently from the heterogeneous consumption patterns of this information among users (Hilbert and López, 2012a). For the case of storage we estimate the installed capacity, which means that we account for the entire available hardware capacity, independently of how much of it is actually used (filled up). The simple lack of adequate statistics forces us to this approximation. In reality vinyl records and commercial music CDs are always completely filled up, while VHS video tapes and hard disks are often only partially used, etc. For the case of broadcasting and telecommunication, we estimate what we call the effective usage capacity, which are exclusively those bits that are effectively transmitted (i.e. sent and received by a device, such as a telephone, Internet modem, TV set or radio receiver). This allows us to compare broadcasting and telecom capacities (in the case of telecom, installed and effective capacity are roughly equivalent, given demand and supply dynamics in a shared infrastructure, while broadcasting receivers could receive about eight times more information than they actually do – broadcast receivers do not compete for a shared infrastructure, such as telecom) (for more on these differences see Hilbert and López, 2012a).

The amount of information that is contained in hardware or communicated through bandwidth does not only depend on the hardware, but also on the rate of compression of the information. The same video might be uncompressed and consume 60 megabytes (MB) of hardware storage space (or bandwidth per second), or compressed with MPEG-4 and occupy only 1 MB. Compression rates have changed significantly over recent decades and allow us to store

roughly more than three times as much information in the same hardware in 2007, as in 1986 (Hilbert, 2011). Therefore, in order to be able to create meaningful time series for the quantity of analog and digital content, we need to normalize the content inside the available hardware capacity on a chosen rate of compression (for more see Hilbert and López, 2012a).

Optimally compressed bits as unifying measurement unit

We chose to normalize on the most efficient compression algorithms that were available in 2007, the last year of our estimates. We call this "optimally compressed" information. At its uttermost level of compression, the number of bits in a message approximates the entropy of the source (Shannon, 1948). Entropy is Shannon's famous probabilistic measure of information and one "entropic bit" is defined as the amount of information that probabilistically reduces uncertainty by half (Shannon, 1948).[7] This is useful for our purposes, because it allows us to compare the informational amount of different kinds of content, such as text, images, audio and video: if one pixel can with equal probability be black or white, its revelation returns one bit of (optimally compressed) information; if a sound can with equal likelihood be high or low, its revelation returns one bit of (optimally compressed) information; and if a letter can with equal likelihood be A or B, its revelation returns one bit of (optimally compressed) information. Each of these measures reduces uncertainty exactly by half, and therefore returns the same amount of information in a strictly technical sense.[8] Compression to the entropic level approximates this fundamental measure (for more on information theory see Shannon, 1948; 1951; Pierce, 1980; Massey, 1998; Cover and Thomas, 2006).

Compression reaches the entropic value of information through the elimination of redundancy in the source, which refers to the parts of the message that do not reduce uncertainty (redundantly repeat the uncertainty already reduced by other parts of the message). The amount of redundancy, and therefore the achievable level of compression in a source, depends heavily on the kind of content, and each counts with different compression programs, like RAR, ZIP, GIF, JPEG, GSM, CDMA and MPEG[9] (see the online Supporting Appendix at López and Hilbert, 2012,

---

[7] Not to be confused with hardware binary digits, such as 1 and 0, which are often also referred to as "bits", independently of how much uncertainty they reduce, i.e. independently of the fact whether they are mere redundant data, or real information in Shannon's entropically compressed sense.

[8] This does not mean that each of these bits can have a different value for the receiver and therefore reduce different kinds of uncertainties of the receivers. The question of information value is a different question that is not answered here. Shannon's measure of the bit simply quantifies information, such as the Celsius- and Fahrenheit-scales quantify heat, independent of whether this heat is good or bad for you, or if its valuable or not. This second question depends on its semantic context and any meaningful answer to this second question requires input from the first (e.g. "*how much* heat is there *and* in which context is it *valuable*"?).

[9] Text and sound have one-dimensional redundancy as they are displayed in time: letters, words and sounds depend on the letters, words and sounds that came before (e.g. it is very probable that a "q" is followed by a "u" in English; and that a C-chord is followed by a D-chord in rock music). An image has two-dimensional spatial

section B[10]). Video counts with the highest achievable compression rates (up to a compression factor of 1:60 for the size of the same archive for optimally compressed digital: uncompressed analog content), followed by images (up to 1:16), the audio (up to 1:12) and text (up to 1:7) (see López and Hilbert, 2012, Section B[10]). In reality, those compression factors are only rough approximations and the level of compression differs between different kinds of images and texts (see also Hilbert and López, 2012b).

We normalize on optimally compressed bits (for storage) and bits per second (for communication). We account for uncompressed (analog) archives as if it would be compressed with the most effective compression algorithm available in 2007. When archives are compressed with a less efficient compression algorithm, we recognize this intermediary level of compression and estimate how much more it could be compressed if it were compressed with the most efficient one available (which reduces the file size, but not as much as in the first case) (see Hilbert and López, 2012b). Given that statistics on the most used compression algorithms are scarce, we limit our analysis to the years 1986, 1993, 2000 and 2007, for which we are able to provide justifiable rates of compression (for more see López and Hilbert, 2012).

It is important to point out that this technical definition of information counts every single bit of information in a strictly technical sense, while we do not provide judgment on the value of all the different bits for the user or the amount of uncertainty they reduce for different users. If we compress a movie in a lossy way and loose image quality, or if we even replace the movie by a script in book format, the layperson might say that the movie still contains the "same information", as long as the basic story line is still perceivable and the deleted bits were of little value. This is technically a wrong use of the ambiguous word "information". In the strict technical sense, the movie lost information (e.g. all the little details that the lossy compression algorithm deleted and are non-recoverable, or the director's choices to give life to the script). A historian, for example, would have found the deleted bits in the distant background of the scenery valuable, and they would have reduced uncertainty for historical ends. Value is highly context dependent and therefore subjective to particular situations and circumstances. We are not aware of a possibility to technically account for such subjectivity on a global and historical scale, and therefore do not distinguish between (subjectively) more or less valuable bits.[8] We count (non-redundant = non-compressible) bits independently from having subjective value. We return to this discussion at the final conclusions.

---

redundancy: the receiver can predict a pixel on basis of the pixels to the left and right, and to the pixels above and below it; a video counts with three-dimensional spatial and temporal redundancy: additionally to the redundancy contained in an image, video compressors make use of the fact that the pixel in the following frame depends on the pixels of the previous frame.

[10] Text and sound have one-dimensional redundancy as they are displayed in time: e.g. letters after letter in time. An image has two-dimensional spatial redundancy: e.g. pixel to the left-right, and up-down in space; a video counts with three-dimensional redundancy in space and time.

Statistics on the distribution of content

For most of the analog technologies in this study it is quite straightforward to identify the kind of content and transform it into optimally compressed bits: radio and analog telephony transmit audio; television transmits mainly video, with a small and clearly defined proportion of audio; the same accounts for information storage on magnetic film that is used for TV and cinematic purposes; books contain mainly text; newspapers store a quite stable proportion of text and images; vinyl records and audio cassettes store audio; and printed photos, photo negatives and x-rays contain images, etc. (see López and Hilbert, 2012). Our largest unknown is the content of digital storage devices (e.g. hard disks, floppy disks, CD-ROMs, digital tape, memory cards) and digital networks (e.g. Internet and mobile telephony). We need to obtain an estimate on the distribution of digital content.

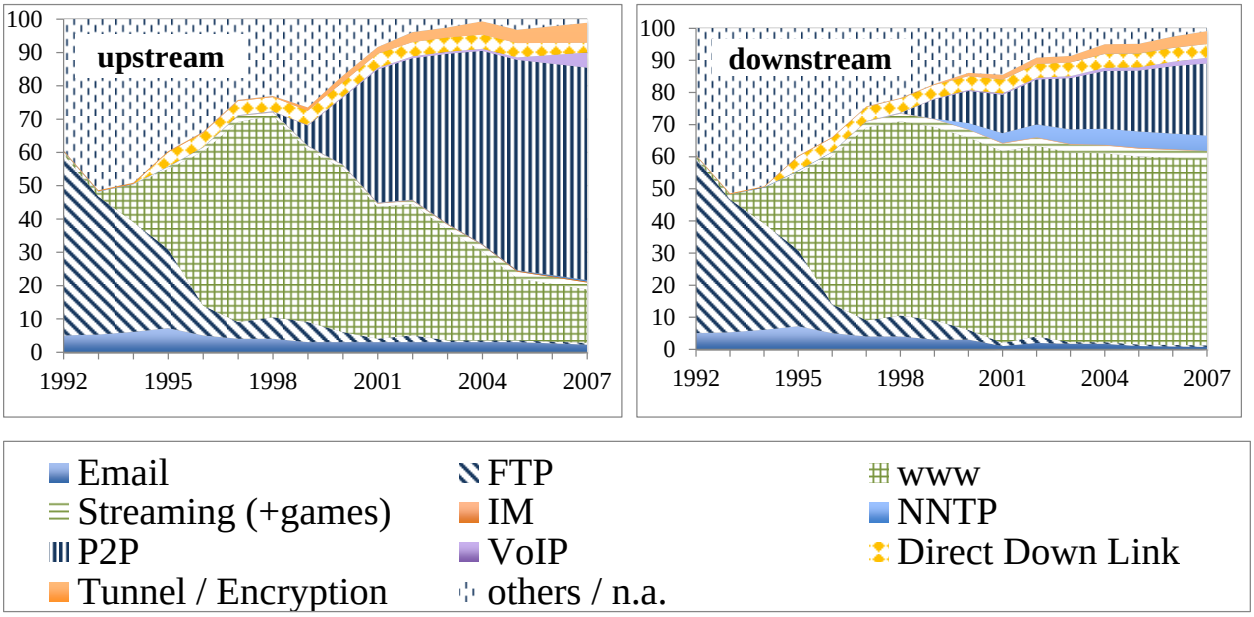
For some digital technologies it was quite straightforward to find an approximation (for more detail on the following explanations see López and Hilbert, 2012, Section C). For example, chip cards that can be found on identification and banking cards only contain alphanumeric text; audio CDs and minidisks only store audio; digital cameras only store images; video DVDs contain stable mix of mainly video and a small part of audio; mobile phone SMS services only transmit alphanumeric text and mobile and fixed-line voice services only audio. We found some useful reports for specific kinds of content, such as the information stored on professional hard-disks of mainframe computers (Madnick, Smith and Clopeck, 2009) or the type of content of the mobile Internet (Kalden, 2004; Ricciato, Hasenleithner and Pilz, 2006; Verkasalo, 2007). Notwithstanding, for others it was necessary to make some assumptions.

For Internet traffic, we basically tracked the distribution of the different Internet protocols and then used statistics that told us the typical distribution of (text, image, audio and video) content in each of those protocols. Statistics on protocols stems from testing equipment that is set up at Internet exchange points and tracks the kind of protocols that traverse the network in real-time (Ipoque, 2006, 2007; Sandvine, 2008). Those protocols include Email, FTP (file transfer protocol), www (worldwide web), media streaming (including gaming), IM (instant messaging), NNTP (network news transfer protocol), P2P (peer-to-peer), VoIP (voice-over-IP), DDL (direct download links), tunnel/encryption, and unknown/others (traffic that cannot be classified in the previous categories or is compressed and unidentifiable).

Based on various sources (see Figure 1), we distinguish between upstream and downstream traffic. This matters as the kind of content differs. In general, content originating from www applications dominate downstream traffic in 2007, while upstream traffic is generated by P2P file sharing (see for example Figure 1). We also create two global profiles, as content depends on the available bandwidth and bandwidth differs among more and less developed countries (see Hilbert, 2013). One profile is based on the statistics from North America and Europe. We apply it to the member countries of the OECD (Organisation for Economic Co-operation and Development),

which is often taken to be representative of the more developed world. Another profile is based on sample statistics from Latin America, Middle East and Africa. We apply it to the rest of the world. We thereby broadly differentiate between countries with different economic profiles. For example, our approximate profiles reveal that www-based traffic is more prominent in the less developed economies from Latin America, Middle East and Africa, while the statistics from North America and Europe show a higher share of P2P content (which often consists of video and music) (see Table D-47 in López and Hilbert, 2012). It would have been desirable to make more fine-grained distinctions (i.e. between countries or regions), but unfortunately the available statistics do not allow for more detail.

Figure 1: Percentage distributions of Internet traffic classified by protocol and application for countries of the OECD (developed regions): upstream and downstream profiles.



Source: authors' own elaboration, based on MacKie-Mason and Varian, 1994; Cisco, 2008; Ipoque, 2006, 2007; Sandvine, 2008; Cano, Malgosa, Cerdan an Garcia, 2001; Leinen, 2001; Leibowitz, Bergman, Ben-Shaul, Shavit, 2002; Fraleigh, et al., 2003; Karagiannis, Broido, Brownlee, Claffy, and Floutsos, 2004; Bartlett, Heidenmann, Papadopoulos and Pepin, 2007. Note: We treat other or unidentified content (n.a.) as we treat text content.

Another set of sources enables us to identify the share of text, image, audio and video content that is typical for each of the previously enlisted protocols (see Ewing, Hall, Schwartz, 1992; Cunha, Bestavros and Crovella, 1995; Arlitt and Williamson, 1996, 1997; Abdulla, Fox, Abrams and Williams, 1997; Arlitt, Friedrich and Jin, 1999; Mahanti, 1999; Pallis, Vakali, Angelis and Said Hacid, 2003; Wang, Makaroff, Edwards and Thompson, 2003; Lacort, Pont, Gil and

Sahuquillo, 2004; Williams, Arlitt, Williamson and Barker, 2005; Guo, et al., 2005; OECD, 2005; Ipoque, 2006, 2007; Nagamalai, Dhinakaran and Lee, 2008).

After reviewing a large number of the available sources, we decided that we would make use of the resulting estimations to also approximate the content of several storage devices, such as hard-disks of personal computers and servers, digital tape, data CD-ROM, flash-drives and the like. This is obviously questionable, since the distribution of Internet traffic content is not necessarily identical with the distribution on digital hard disks. The main reason behind our choice is brutal practicality. There are comparatively more numerous, more credible and more representative sources available to estimate Internet traffic than to estimate any other kind of digital content. On the conceptual level, specific parts of Internet traffic can be expected to be representative of the content of several digital storage devices because this content is fed from and feeds to some kind of Internet content, which leads to a rough proportionality. For example, this is especially true for the relation between P2P traffic and PC hard disks. Therefore, after reviewing various alternatives sources, we assume that the content of PC hard-disks is distributed as per the average of web traffic and P2P traffic, while the content of server hard-disks is proportional to web traffic alone (see López and Hilbert, 2012 for details). In other words, given the lack of better data, we focus on the most representative kind of traffic for specific kinds of storage devices and approximate storage content as a weighted combination of traffic content.

The content of technologically communicated information

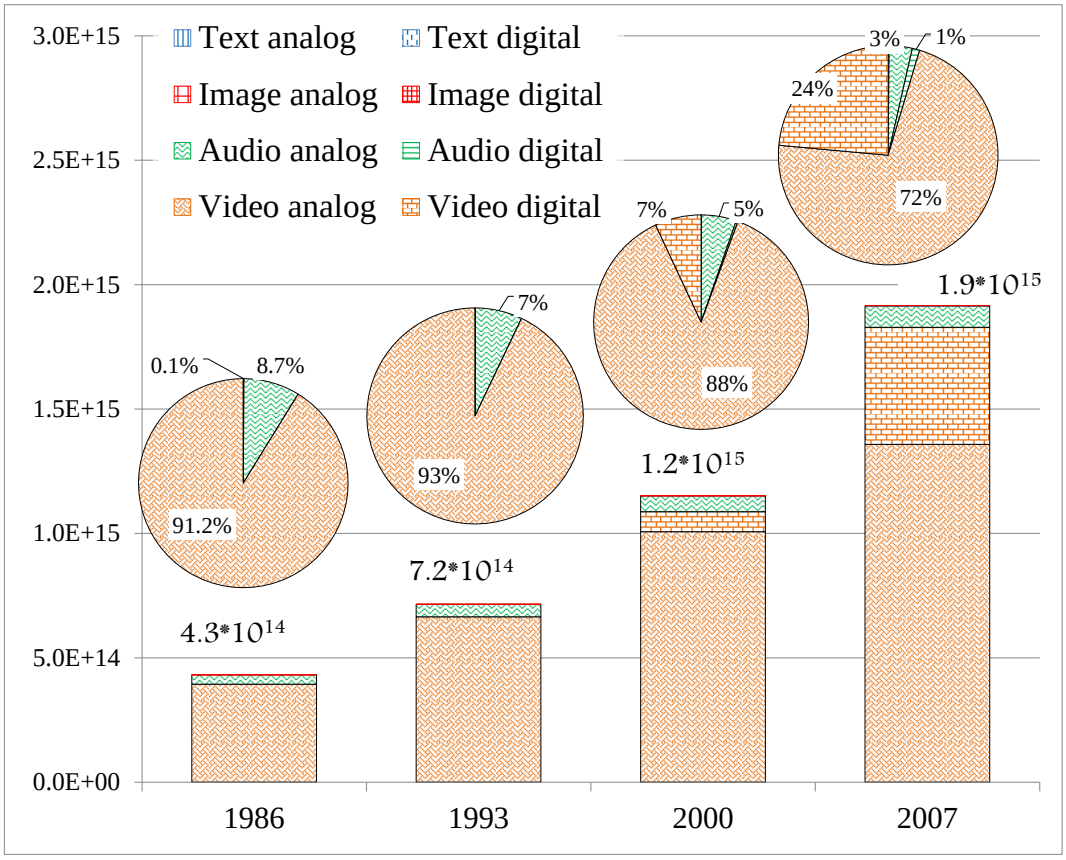
We start with the results of our measurement of the world's communication capacity, for which we have the most solid sources of content distribution. We define the world's technological capacity to communicate as the amount of information that is *effectively received or sent by the user*, *while being transmitted over a considerable distance* (outside the local area). We only measure those bits of information that are effectively communicated (see also Hilbert, 2011). We apply this definition to unidirectional one-way broadcasting, which includes all communication media that uses a mere downstream channel for information dissemination and bidirectional two-way telecommunication networks, which count with both, upstream and downstream channels[11].

[11] From a technological perspective, the distinction is that broadcasting provides a "common-channel" (same content, at the same moment in time), which do not compete for bandwidth (the signal is transmitted independently from the reception of the receiver), while telecom provides "user-defined individual channels", which compete for a shared bandwidth infrastructure (see Hilbert and López, 2012b). The exception in our classification is digital TV, which counts with a small upstream channel, but we nevertheless count it as broadcasting, since this upstream channel was not much in use until 2007 (with the expectation of some sporadic video-on-demand application) (see Hilbert, 2011).

The content of broadcasting

Broadcasting[5] is still the world's most information rich technological information operation, but, as shown in Figure 2, its evolution in the recent past has been relatively uneventful. The world's effective capacity to broadcast only grew at a comparatively low 7 % per year during the last two decades (see Hilbert, forthcoming). Most content has been televised video. This analysis also provides hard-fact evidence confirming the explorative hypothesis formulated in the 1980s that "video killed the radio star" (Buggles, 1979): radio represents a constantly decreasing share of broadcasted audio (83 % of all broadcasted audio in 1986, 71 % in 1993, 60 % in 2000, and 48 % in 2007, the remaining audio is received by TV sets). Images that are diffused through newspapers represent a negligible part of the diffused information (0.1 % in 1986).

Figure 2: Dimension and content distribution of the world's effective technological capacity for broadcast information, in optimally compressed Megabytes (MB) per year, for 1986, 1993, 2000 and 2007 (y-axis refers to bar-graph).



Source: authors' own elaboration, based on various sources, see López and Hilbert, 2012. Note: Shares of text and images are small in unidirectional broadcast networks (less than 0.2%), but do exist (i.e. since we grouped the unidirectional diffusion through analog newspapers into this group; see López and Hilbert, 2012).

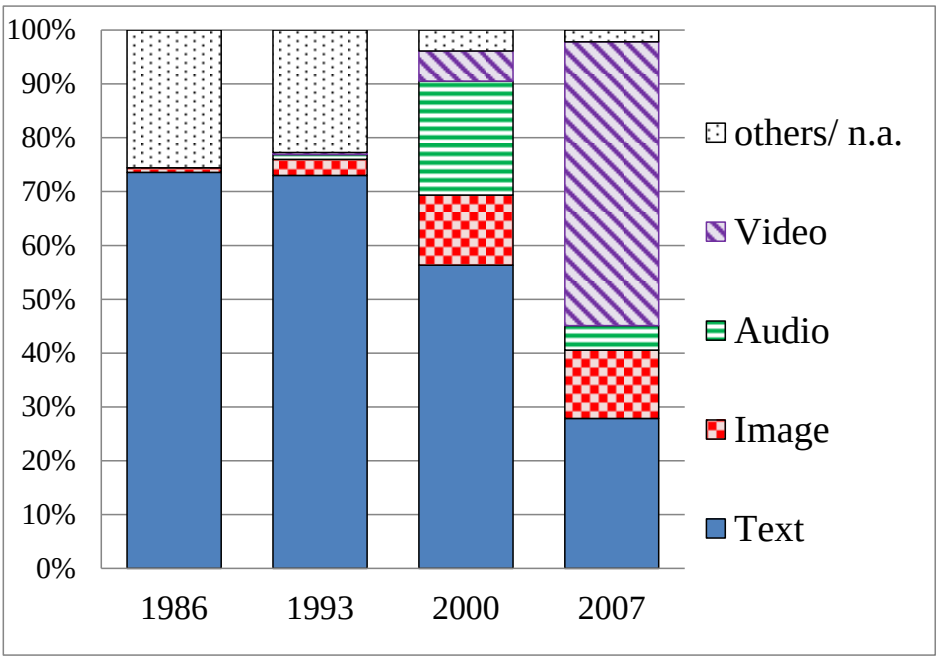
The distinction between analog and digital technologies in Figure 2 also shows the reason for this slow evolution. Most broadcasting was still not digitized by 2007. While digital satellite and cable television lead the way into the digital age (70 % of all digitally broadcasted information was satellite TV in 2000 and 24 % cable; 42 % and 30 % respectively in 2007), only a few countries had started to provide terrestrial over-the-air digital TV by 2007 (merely 40 out of the 217 countries included in our assessment). As a result, only a quarter of the globally broadcasted information is in digital form.

The content of the fixed-line Internet

For two-way telecommunication content, we first focus on the content of the fixed-line Internet only (which is by now the overwhelming part of telecommunication content) and then embed this Internet content into the large historical context of telecom evolution including fixed- and mobile telephony and two-way paper postal letters. Figure 3 presents the result for the case of fixed-line Internet traffic alone. It shows that during the late 1980s and the early 1990s, Internet traffic was dominated by text content (73-74 %), including emails, alphanumerical databases through FTP, and the incipient worldwide web that introduced hyperlinked text-based websites (Berners-Lee, 1998). This changed quickly as bandwidth grew. In 2000 Internet's traffic carried images (13 %) and audio (21 %). Then "MP3 revolution" (McCandless, 1999) started to set in. By 2007, text only made up 27 % of the global flow of information on the Internet. Video content grew from basically 0 % in 1986 and 1993, to represent more than half of the total (54 %). This can be explained by the introduction of P2P applications (like the early Napster, or Kazaa, Emule, Gnutella, Freenet, eDonkey, BitTorrent, etc.) and by the emergence of movie streaming services.[12] This result fits the media richness theory (Daft and Legel, 1984; 1986) and the usual narrative of the evolution of the Internet (Newhagen and Rafaeli, 1996; Morris and Ogan, 1996): while bandwidth grew, the traditional dominance of text was continuously being replaced by images and audio content, and is currently being marginalized by content rich video. We are said to be in midst of an "online video revolution" (Gannes, 2009).

[12] By 2007 the video-sharing platform YouTube was already providing a significant amount of videos (Reuters, 2006), but the provision of real-time video-streaming for movies (like the ones offered by Netflix and Amazon in 2010) was still in an incipient stage then (Falcone, 2008).

Figure 3: Distribution of content for global fixed-line Internet (wireless and wireline) for 1986, 1993, 2000 and 2007 (upstream and downstream).

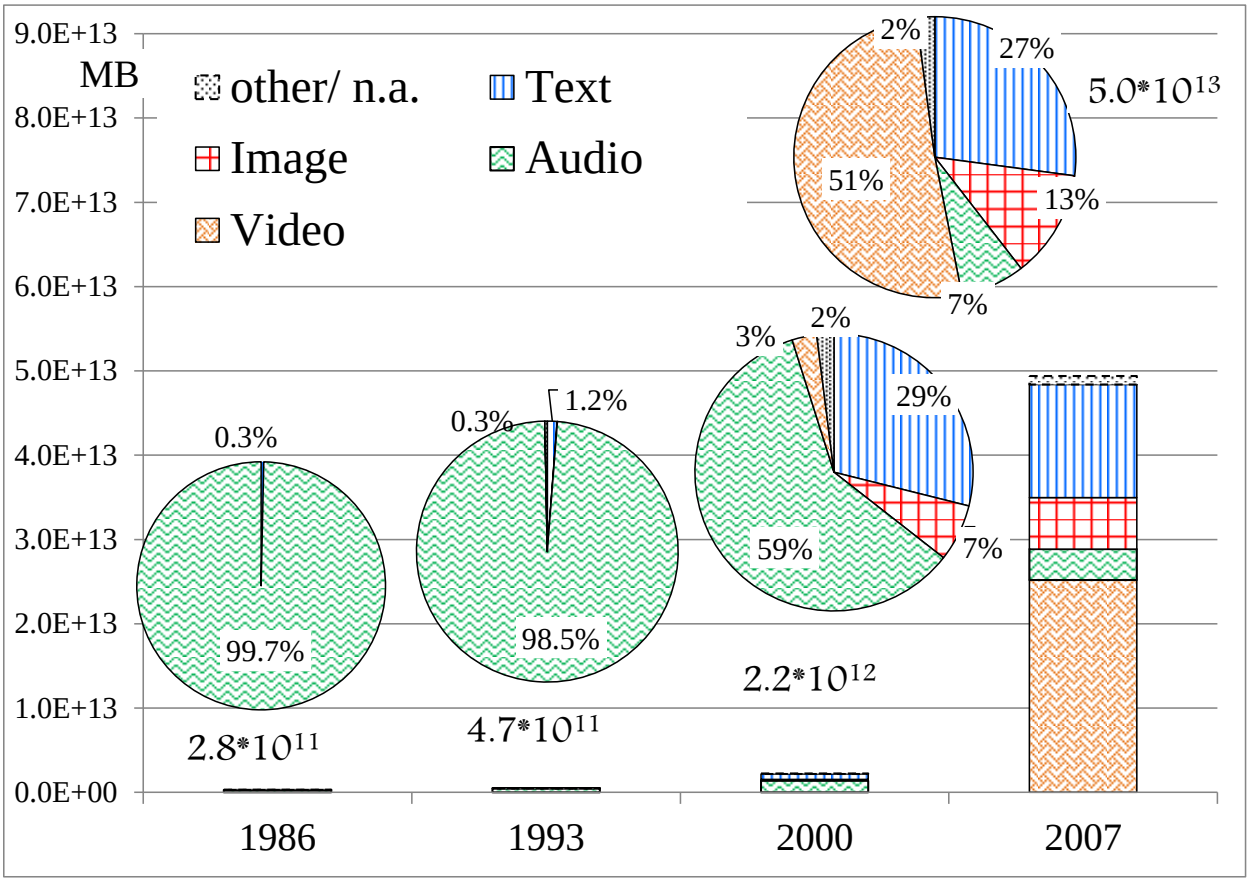


Source: Authors' own elaboration, based on various sources, see above and López and Hilbert, 2012, Section D. Note: We normalize unknown or unidentified content (others/ n.a.) with the same compression factors as text.

The content of telecommunications

It is interesting to see what happens when we situate the evolution of Internet content within the larger context of the most common analog and digital technologies, which include pre-Internet telecom. Figure 4 includes fixed-line Internet, but also fixed and mobile voice telephony (analog and digital), as well as fixed and mobile data services and even analog bi-directional paper postal letters (the pioneer of mediated two-way distance-communication).[6] The first thing to notice is that in comparison with broadcasting (Figure 2), the sum of all telecommunications networks effectively communicate relatively little information, as less than 4 % of the world's communicated bits are effectively passed through telecommunications networks in 2007 (see Hilbert, 2011). Notwithstanding, the technological capacity of telecommunications has grown significantly over the past two decades: 30 % per year (Hilbert, forthcoming). While most telecommunication was implemented through telephone voice traffic during the 1980s, the Internet quickly took over around the year 2000 (see Hilbert and López, 2011).

Figure 4: Telecom: Dimension and content distribution of the world's effective technological capacity telecommunicate, in optimally compressed Megabytes (MB) per year, for 1986, 1993, 2000 and 2007 (y-axis refers to bar-graph).



Source: authors' own elaboration, based on various sources, see López and Hilbert, 2012, Section D

The second thing to notice in Figure 4 is the major changes in content. Almost all content comprised of telephone voice telecommunication in 1986, with a marginal part contributed by text (in 1986 postal letters, and by 1993 the emerging Internet). The large share of text-based content from the Internet (see Figure 3) did not yet make a difference, since the Internet itself presented such a small share of telecommunication back in the late 1980s and early 1990s. Most of this audio content was still passed through analog telephone networks (80%) in 1986.[13] By the year 2000, fixed-line Internet already explained 51 % of all telecommunicated information (see also Hilbert and López, 2011). The increasing share of Internet content also implied an increasing rise of the share of text in the distribution of total content. Since more than half of the Internet content was

[13] The digitization of the fixed-line telephone network took place between the subsequent years, and by 1993, two-thirds of voice telephone traffic was already digitized.

made out of online text in 2000 (i.e. webpages on the worldwide web, see Figures 1 and 2), it results that 29 % of the total amount of telecommunicated information consists of digital text (Figure 4). In other words, while the share of text decreased on the Internet itself (see Figure 3), the Internet still contains more text than previous two-way telecom networks. The increasing share of the Internet in the more general telecommunication landscape led to an increase in the share of text in global telecom networks (Figure 4). These are the previously mentioned kind of conditional shifts in relative shares that are difficult to grasp intuitively without the accompanying data.

The share of images has also grown, representing 13 % of the Internet content in 2000 (Figure 3) and 7 % of the total content of telecommunication networks (Figure 4). By 2007, fixed-line Internet clearly dominated the telecommunication landscape, capturing 97 % of all bits sent (the rest is explained by fixed-line voice telephony, mobile voice and mobile data traffic, each with 1 %) (see Hilbert and López, 2011). By the late 2000s, the world's effective telecommunications capacity nicely reflects the content of the Internet (compare Figure 3, with Figures 7 and 8). Video represented about half of the total, text a quarter, and images one-eighth. Audible information content turns out to be the clear loser in the telecommunication revolution (from almost 100 % in 1986, to only 7 % in 2007).

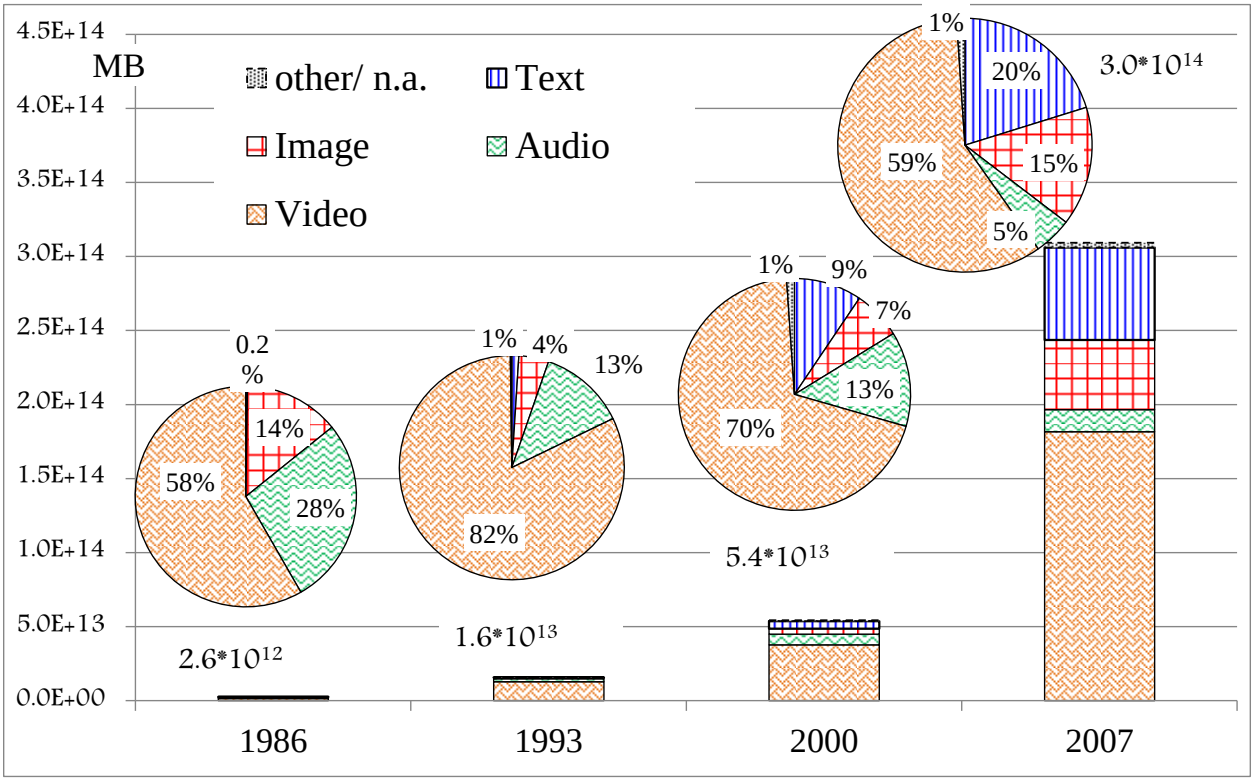
The content of technologically stored information

Let us now turn to the world's storage capacity. We define storage as the *maintenance of information over a considerable amount of time* explicitly for later retrieval and estimate the world's installed (available) storage capacity (not the effectively used capacity). We do not consider volatile storage in this inventory (such as RAM), since the ultimate end of volatile memory is computation, not storage per se. Considering the 25 most prominent storage technologies[3,4], we see that the world's storage capacity grew from 2.6 optimally compressed exabytes in 1986 to some 300 optimally compressed exabytes in 2007 (Figure 5, compare with Hilbert and López, 2011). This implies a compound annual growth rate of 25 % over two decades, which has mainly been driven by the rapid process of digitization (analog storage has only grown at 10 % per year over this period, while digital storage has grown 57 %; see Hilbert, forthcoming).

Analog video tape (such as VHS cassettes) was the predominant storage technology of human kind until the year 2000, storing 58 % of all technologically stored bits in 1986, 86 % in 1993, and 72 % in 2000 (to see how the total is distributed among different technologies, see Hilbert and López, 2011). Many households, especially in the developing world, hosted entire libraries of VHS cassettes. As shown in Figure 5, this translates to a clear dominance of video content, with a complement of audio. Figure 5 provides empirical evidence for the "video-revolution" of the early 1990s (i.e. 1986-1993), with videocassette recorders (VCRs) at its forefront (Wood, 1986; Wood and O'Hare, 1991; Dimmick, 1997). Audio content contributed 28

% in 1986, which mainly consists of Vinyl records (14 %) and audio cassettes (12 %)[14]. Still images represented 14 % of the global storage stock in 1986, which foremost consisted of photographs (8 % in form of printed images, and 5 % in form of negatives).

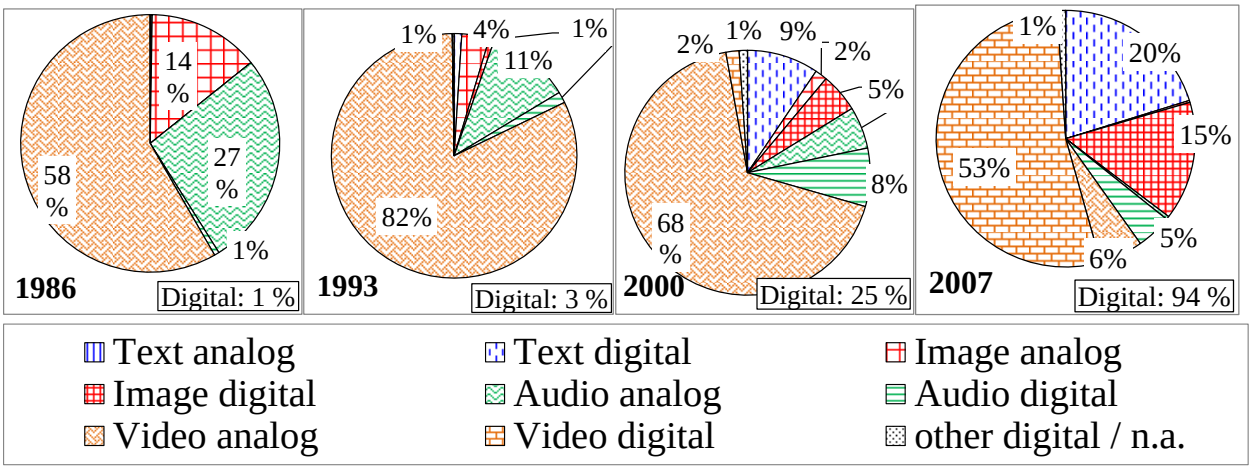
Figure 5: Storage: Dimension and content distribution of the world's installed technological capacity to store information, in optimally compressed Megabytes (MB), for 1986, 1993, 2000 and 2007 (y-axis refers to bar-graph).



Source: authors' own elaboration, based on various sources, see López and Hilbert, 2012, Section C

[14] Vinyl records and audio cassettes quickly lost importance: 1993: 6.2 %; 2000: 1.6 %; 2007: 0.1 %.

Figure 6: Storage: Analog and digital content distribution of the world's installed technological capacity to store information, in optimally compressed Megabytes (MB), for 1986, 1993, 2000 and 2007.



Source: Authors' own elaboration based on various sources, see López and Hilbert, 2012, Sect.C

The transition from 2000 to 2007 is very interesting, mainly because the medium of information storage changed decisively in these years. In Figure 5 it appears that the nature of content only changes slightly, while Figure 6 reveals that the global technological memory has undergone a major revolution during these years: from analog to digital. Figure 6 shows the same distribution, but now separating between digital and analog content. In 2000, still 75 % of all technologically stored information was in analog devices, while only 6 % of analog remained by 2007 (mainly VHS video cassettes, down from 72 % only seven year earlier). Notwithstanding, video continues to represent the vast majority of optimally compressed content in 2007 (59 %). While most video was stored in analog format in the year 2000 (98 % of video), the vast majority of video was digitized by 2007 (90 % of video). Two decades and a technological revolution later, the share of video occupied the same level in 2007 (59 %), as it had twenty years earlier in 1986 (58 %). In other words: the medium changed, the kind of content did not.

Until the year 2000, analog tape (mainly VHS video tapes) represented the lion's share of the world's information storage capacity (71 % in 2000), while PC hard-disks started to dominate by 2007 (42 %, up from 5 % in 2000), followed by optical DVDs (21 %, up from less than half percent in 2000), digital tape (11 %, up from 8 % in 2000) and server hard-disks (8 %, up from half percent in 2000) (Hilbert and López, 2011). Notwithstanding these profound changes in the carrying media, interestingly, the type of content seems relatively unaffected by this revolution. The revolution has taken place in the kind of carrying media, not in the kind of content. The share of video that was previously stored in VHS cassettes was maintained by DVDs and hard disks.

Something similar happened to still images. After its share had been squeezed by the dominance of moving images during the 1990s, it came back to its pre-digital levels by 2007 (14

% in 1986 and 15 % in 2007) (Figure 5). However, almost all images in 1986 were in analog format (mainly printed photographs and negatives, x-rays and newspaper images), while 98 % of all images were in digital format by 2007 (mainly on PC and server hard disks) (Figure 6).

The most stable tendency can be identified for the evolution of the share of audio: it dropped constantly during the past two decades from 28 % in 1986, to 13 % in 1993 and 2000, to only 5 % of the total by 2007 (Figure 5)[15]. Digitization of audio content took place around the year 2000 (58 % of all audio in digital), shortly after the "MP3 revolution" started (McCandless, 1999) (Figure 6), and while increasing in absolute value, it never again reached the share that was help in 1986 by vinyl records (14 % of the total capacity) and audio cassettes (12 %), since other kinds of content grew decisively faster.

The most surprising result is the indisputable victory march of alphanumeric text. Our estimates indicate that text became very important in the digital age. It captures every fifth bit in 2007 (20 % of the total amount of bits stored). This is an unprecedented share, much higher than during the pre-digital age in the 1980s (text represented a mere 0.25 % of the globally stored information in 1986). This is despite the fact that paper-based text (books, newsprint and paper-based advertising) has clearly lost much of its (already very small) share: 0.330 % of the total of stored information in 1986, 0.083 % in 1993, 0.034 % in 2000, and 0.007 % in 2007 (paper-based text represented 28 % of all text content in 1986, 2 % in 1993, 0.1 % in 2000 and 0.01 % in 2007)[16]. In other words, despite the much debated "death of print" (e.g. Orkent, 2000; Kurtz, 2009), there is more information in alphanumerical format nowadays than there was in the age of analog paper-based print. At least in terms of informational content, written text represents more than it ever has during the past two decades. 99.99% of all stored text in the world is in digital format (most of it text archives in PCs, and websites or databases on server hard-disks) (Figure 6).

What do we do with which kind of content?

We can now also compare the magnitudes of the type of content that is stored and communicated. Since storage is measured in bits, and communication in bits per second, this is not straightforward and requires an additional assumption to align both units of measurement. We could compare the amount of communicated information per year (sum of the number of bits communicated during each second of the 365 days of year), with the average storage space available "per year" (number of bits that can be stored on average on each of the 365 days of the year). If both would be equal, we could store all the information that is communicated during one year in the available storage space of a given year. This is not the case, however. In reality, we

[15] It is interesting to note that the share of audio has not decreased significantly when measured in terms of the amount of hardware it occupies.

[16] These relative proportions should not obscure the fact that paper-based books, newsprint and advertising have increased in absolute terms, from some 9 Petabytes in 1986 to 19 PB in 2007.

communicate much more information than we can possibly store. In 1986 we would have filled our available storage capacity every 2 days. Over the last two decades this relation has changed in favor of storage (see marginal shares in Figure 7a), and in 2007 our technological memory could be filled about every 8 weeks with our communication flows.

Figure 7: Cross-tabulations in optimally compressed Megabytes (MB) per year, for 1986, 1993, 2000 and 2007: (a) cross-tabulation with marginal likelihoods; (b) Comparison of content per information sector (amount of information per sector per year, conditioned on kind of content).

	Stored	Broad-casted	Telecom-municated	**Marginal total**
		1986		
Text	0.002%	0.005%	0.00%	**0.01%**
Images	0.08%	0.10%	0.00%	**0.18%**
Audio	0.17%	8.62%	0.06%	**8.85%**
Video	0.35%	90.61%	0.00%	**90.96%**
Others/n.a.	0.00%	0.00%	0.00%	**0.00%**
Marginal total	**0.61%**	**99.33%**	**0.06%**	**100.00%**
		1993		
Text	0.02%	0.00%	0.00%	**0.02%**
Images	0.09%	0.05%	0.00%	**0.14%**
Audio	0.27%	6.86%	0.06%	**7.20%**
Video	1.77%	90.87%	0.00%	**92.63%**
Others/n.a.	0.00%	0.00%	0.00%	**0.00%**
Marginal total	**2.16%**	**97.78%**	**0.06%**	**100.00%**
		2000		
Text	0.43%	0.03%	0.05%	**0.51%**
Images	0.31%	0.03%	0.01%	**0.34%**
Audio	0.60%	5.26%	0.11%	**5.97%**
Video	3.13%	89.99%	0.01%	**93.12%**
Others/n.a.	0.05%	0.00%	0.00%	**0.05%**
Marginal total	**4.51%**	**95.31%**	**0.18%**	**100.00%**
		2007		
Text	2.78%	0.10%	0.60%	**3.48%**
Images	2.09%	0.01%	0.27%	**2.38%**
Audio	0.67%	3.77%	0.16%	**4.60%**
Video	8.08%	80.16%	1.12%	**89.36%**
Others/n.a.	0.14%	0.00%	0.05%	**0.19%**
Marginal total	**13.76%**	**84.05%**	**2.20%**	**100.00%**

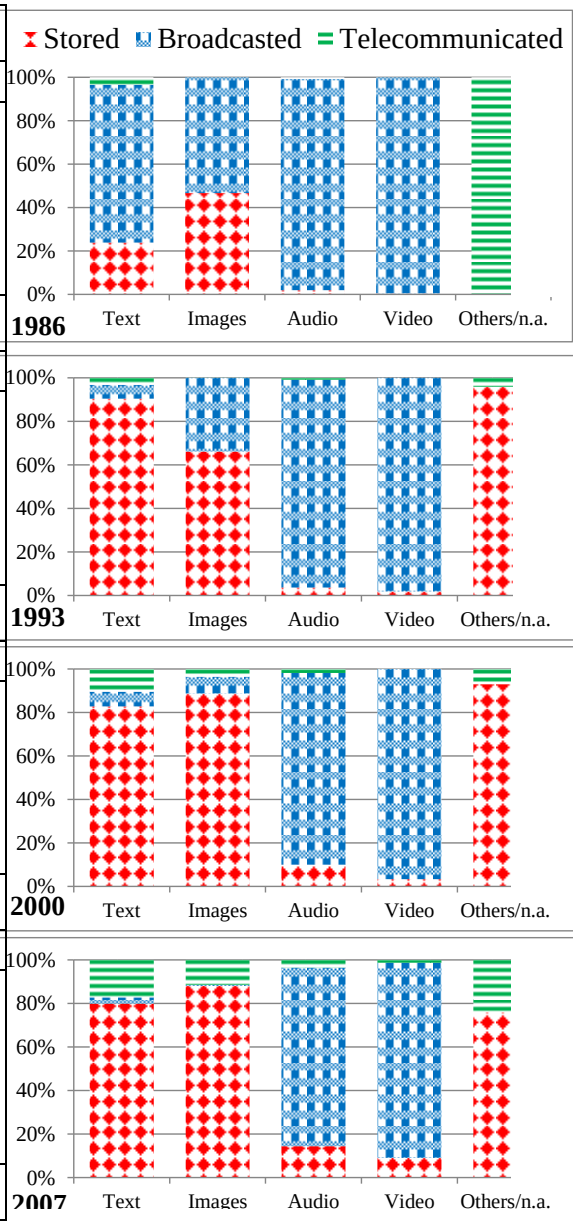


Figure 7a uses this assumption and shows the arising joint shares of content distribution between the two dimensions of our inventory: the basic kind of information operation (storage, one-way broadcasting, and two-way telecommunication), and the kind of content (text, images, audio, video). The left-hand marginal likelihoods on Figure 7a reveal that video content dominates worldwide information processes, representing around 90 % of all bits stored and communicated. This is mainly due to the vast amount of information that is constantly being spread through the world's broadcasting channels and the fact that the underlying assumptions of Figure 7 do not account for information deletion and replacement of stored information during one year (but simply consider the installed capacity, independently of how often the content is replaced). As shown by the lower marginal likelihoods, in 2007 the world's technological capacity to broadcast vs. store vs. telecommunicate information was in a ratio of 38 : 6 : 1. Despite the clear dominance of broadcasting as the most massive information operation, the relative importance of broadcasting has been decreasing over the past two decades (99.3 % in 1986 to 84 % in 2007). Much of the resulting decrease in broadcasted video has been filled with stored alphanumeric text (growing from 0.5 % in 2000 to 3.5 % 2007). Besides, driven by the vast amount of photographs, graphics and drawings in the world's digital storage, the relative weight of images is constantly growing (from 0.2 % in 1986 to 2.4 % in 2007).

While we looked at the joint shares of Figure 7a in a vertical fashion during Figures 2-6 (summing each column of storage, broadcasting and telecom up to 100%), we can now also look at it horizontally through the different rows (summing up the total amount of text, images, audio and video; see Figure 7b). This shows that in 1986 it was most likely to find text content in paper-based newspapers (mainly diffused through unidirectional channels, i.e. broadcasting). This changed quickly, and with the introduction of digital storage devices, more and more text started to be archived over prolonged periods of time. At the same time, an increasing amount of text is being telecommunicated through the Internet (17 % of all text in 2007), even though this part is still small in comparison to the amount of text stored (80 % of all text in 2007). This later number could of course be inflated due to our methodological assumptions. A similar evolution can be observed for the information operation of choice for still images. Following our assumptions, during the late 1980s, images were either stored (i.e. on printed photographs or negatives) or diffused (i.e. through newspapers and paper-based advertising). By 2007, images are either stored (i.e. on PC and server hard-disks) or telecommunicated (i.e. through the Internet). Audio and video content has and still is mainly diffused through unidirectional broadcasting channels, while a rapidly increasing part of it is being stored (from audio: 4 %, and video: 2 % in 2000; to audio: 14 %, and video: 9 % in 2007). Notwithstanding, in comparison with broadcasting, the telecommunication of audio and video content is still comparatively small and incipient in 2007, however, fast growing.

Conclusions, limitations, and outlook

We made the amount of analog and digital content comparable (normalizing on compression rates) and asked how the constellation of worldwide technologically stored and communicated information content changed between 1986 and 2007. In order to do this, it was necessary to make several assumptions. We put much emphasis in explaining them in abundant detail in some 300 pages of the methodological appendix (see López and Hilbert, 2012). We hope that this level of transparency will also serve future generations as input to build on our efforts and improve our analysis. In the future, one could – for example – carry out much more detailed inventories of the effective usage of storage space (not merely the installed capacity as done here) and of the kind of content that is stored on the hard disks of personal computers. This will surely refine the estimates presented here. In this sense, one should not take our results with a grain a salt and allow for a certain margin of error. However, even allowing for such a margin of error, we would be very surprised if the fundamental orders of magnitudes and identified trends would change decisively with further refinements of the underlying empirical data.

Several unexpected findings with theoretical and practical value

Setting digital content into the evolutionary context of its analog predecessor gave us some interesting and unexpected insights. Of theoretical interest is our finding that our macro-perspective on the high-level evolution of informational content revealed a surprising inertia that is independent of the carrying medium. For example, the comparisons between Figures 5 and 6 revealed that the transition from analog VHS and audio tapes and vinyl records to digital CDs and hard disks affected the distribution of content surprisingly little. Analog audio represented 14 % of the world's storage capacity in 1986 and video 58 %, while these shares merely moved to 15 % and 53 % respectively two decades later.

The most surprising finding of this empirical inventory might be that the proportional share of alphanumeric text is larger in the digital "multimedia age" than it has been at the end of the analog age. When focusing exclusively on digital technologies (see Figures 3 and 6), the intuitive notion that medium-richness increases with increasing bandwidth and storage capacity is confirmed. However, when placed into the larger context of the transition for analog to digital media, the proportional share of alphanumeric text and still images turns out to be more prominent than before the digital age. This somewhat surprising result is based on an increasing share of digital content and the fact that digital content contains more text and still images than previous analog content. Even so the share of static text and image content was diminished within digital media (see Figures 1 and 3, also the more detailed Tables D-47 and D-48 in López and Hilbert, 2012; pp. 202-205), at the end of the 2000s their shares were still larger than they were at the end of the 1980s within analog content. The data shows that the 1980s saw a crushing dominance of audio and video content, pushed by handheld video cameras, VHS cassettes, vinyl records and

audio tapes. Alphanumeric text and still images were mainly presented by books, newspapers, photographs (negatives and prints) and some incipient digital tape. They represented a marginal share of the analog universe of the late 1980s. As a result of the combination of increased digital content and the fact that the share of static text and images is (and has been) larger for digital content, the process of digitization did not lead to a proportional increase in the share of media-rich content but to an unprecedented increases of alphanumeric text and still images. Of course, all kinds of content increased vastly in absolute terms, but in relative terms, text and still images gained and video and audio lost share. These results show that the presumed multi-media revolution between 1986 and 2007 actually turns out to be a text and still image revolution. Only the future will tell if this trend is a fundamental characteristic of the digital age, or merely a temporal phenomenon of the current transition toward a full-fledged digital age.

For now, the proportional prominence of alphanumeric text and still images is good news for big data analysts, who aim at creating value from the vast amounts of available information (Hilbert, 2013). Text and still images are much more accessible to currently available algorithms than the dynamic data formats of videos and audio. Figure 7 also revealed that alphanumeric text and still images are nowadays increasingly exchanged through two-way telecommunications networks, and not simply diffused through one-way broadcasting channels as before. The multidirectional two-way nature of telecom networks makes the content more socially embedded than the one-way process of information diffusion from a central node to passive receivers. This is also good news for big data analysts, since it means that the content is increasingly infiltrated by real-time social interactions, which provide hints about behavioral patterns.

Despite the relative increase of alphanumeric text and still images, we have also seen that (mainly broadcasted) video has been, and in 2007 still is the most information intensive kind of content on the planet in absolute terms (9 out of 10 bits), independently of its analog or digital nature. One-way information diffusion through a common channel at the same moment in time (*broadcasting*) is still the most important information operation in the world, and mainly carries video. The much praised role of interactive and bidirectional telecommunications networks over individualized user-defined channels was restricted to a mere 1 out of 34 bits in 2007,[11] even so it is rapidly growing.

Limitations and future research

We focused our inventory on the installed and/or effective capacity of our technologies. This does not automatically lead to insights about media consumption. Media consumption is a subsequent step and different kinds of content can be consumed in different intensities. For the case of the U.S., Bohn and Short (2009) estimate media consumption of bits of hardware capacity, and found that video intensive technologies like TV, computer games, and movies represent 99.2 % of the total number of bits "consumed". Time-budget studies also reconfirm that people read

less: the percentage of young American adults (18-24 years) that regularly read literature has declined from 60 % in 1982, over 53 % in 1992, down to 43 % in 2002 (NEA, 2007). This implies that in terms of bits consumed, it might well be that the share of video increased and the share of text decreased, while the amount of video and text that is available evolved in the opposite direction. Additionally to measuring the amount of bits consumed, Bohn and Short (2009) also estimate consumption in minutes, and find that the same three video intensive categories (TV, computer games, and movies) only occupy 50 % of the media consumption time, while radio listening and computer reading made for most of the rest (unfortunately we do not have time series results of these findings). This shows that capacity and consumption are different (see also Neuman, et. al, 2012) and that different metrics of consumptions (e.g. bits vs. minutes consumed) lead to different results. Future exercises will have to shed more light on these differences.

In the meantime, we have shown that the evaluation of the technological capacity can meaningfully be measured in optimally compressed bits. This allows us to also test for the social impact of different kinds of information content and of different information operations. For example, as suggested by media-richness theory (Daft and Legel, 1984; 1986), in some settings text might be sufficient, while in others, there might be decreasing returns to additional text, and video content might become more relevant (for example, in online business negotiations and online health consultations versus bank transactions and scientific collaborations). In which social settings is text more efficient than video, and what part of the social impact can be explained by simply providing more quantity of a certain kind of content? Media-richness theory has led to inconclusive micro-studies on the organizational level (Dennis and Kinney, 1998). The kind of data presented here can be used for macro-level studies involving statistical regressions, econometric- and structural equation models, and time series impact analysis to answer similar questions: does text or video have differential effects on e-banking and telemedicine? One can also test if one or the other content is more effective when broadcasted or interactively exchanged through telecommunications networks. Such exercises can test for the differential importance of socially embedded bidirectional networks and traditional unidirectional channels of mere information dissemination. The presented variables allow for the testing of a series of related hypothesis. In this sense, the methodology and logic of the presented exercise can be seen as a mere beginning of a much larger research agenda.

This being said, it is obvious that our data end in 2007 and that several interesting technological changes have happened since then, including the proliferation of mobile content and the increasing installation of fiber-to-the-home/business (FTTH/B) for individual households and businesses. It can be expected that some changes happened since then, and surely will continue to happen during the decades to come. The digital revolution, especially in terms of bandwidth, does not show any signs of slowing down (Hilbert, 2011; Hilbert, 2013; forthcoming). The analysis of this article focused on the main transition from the analog to the digital age (less than 1 % of the world's technological memory was digitized in 1986, and more than 94 % in 2007; Hilbert and

López, 2011). For this we elaborated and presented a methodology that allows capturing the main tendencies. It is straightforward to apply this method to more recent and also to future years and to create (maybe continuously updated) future generations of similar information and communication capacity inventories. Hopefully such inventories can also improve and fine-tune the assumptions that had to be made in this exercise (see López and Hilbert, 2012).

References

Abdulla, G., Fox, E. A., Abrams, M., & Williams, S. (1998). WWW Proxy Traffic Characterization with Application to Caching. Computer Science Department at Virginia Tech. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.54.6195

Arlitt, M, & Williamson, C. (1996). Web Server Workload Characterization: The Search for Invariants. In *ACM sigmetics Conference on the Measurement and Modeling of Computer Systems, Philadelphia, PA,*.

Arlitt, M. F., & Williamson, C. L. (1997). InternetWeb servers: workload characterization and performance implications. *Networking, IEEE/ACM Transactions on*, *5*(5), 631–645. doi:10.1109/90.649565

Arlitt, Martin, Friedrich, R., & Jin, T. (1999). Workload Characterization of a Web Proxy in a Cable Modem Environment. *ACM performance evaluation review*, *27*, 25–36. doi:doi:10.1145/332944.332951

Barlett, G., Heideman, J., Papadopoulos, C., & Pepin, J. (2007). *Estimating P2P Traffic Volume at USC* (No. ISI-TR-645). Los Angeles: USC/Information Sciences Institute, 2007.

Berners-Lee, T. (1998, May 7). The World Wide Web: A very short personal history. World Wide Web Consortium (W3C). Retrieved from http://www.w3.org/People/Berners-Lee/ShortHistory.html

Bohn, R., & Short, J. (2009). *How Much Information? 2009 Report on American Consumers*. San Diego: Global Information Industry Center of University of California, San Diego. Retrieved from http://hmi.ucsd.edu/howmuchinfo.php

Brown, J., Shipman, B., & Vetter, R. (2007). SMS: The Short Message Service. *Computer*, *40*(12), 106–110. doi:10.1109/MC.2007.440

Bruns, A. (2007). Produsage, Generation C, and Their Effects on the Democratic Process. Presented at the Media in Transition 5, MIT Boston. Retrieved from http://eprints.qut.edu.au/7521/

Buggles. (1979). *Video Killed the Radio Star*.

Cano, M. D., Malgosa, J. maria, Cedan, J. F., & Garcia, J. M. (2001). Internetmeasurements and data study over the regional network Ciez@net (Vol. Grupo Ingeniería Telemática (GIT)). Presented at the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, Institute of Electrical and Electronics Engineers (IEEE). Retrieved from http://repositorio.bib.upct.es:8080/dspace/handle/10317/916

Cisco Systems. (2008). *Global IP Traffic Forecast and Methodology, 2006–2011* (White Paper). Retrieved from http://www.hbtf.org/files/cisco_IPforecast.pdf

Coughlin, T. (2007). Digital storage technology newsletters. Coughlin Associates. Retrieved from http://www.tomcoughlin.com/

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd Edition.). Hoboken, NJ: Wiley-Interscience.

Cunha, C. R., Bestavros, A., & Crovella, M. E. (1995). Characteristics of WWW Client-based Traces. Computer Science Department Boston University. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.2.1361

Daft, R. L., & Lengel, R. H. (1986). Organizational Information Requirements, Media Richness and Structural Design. *Mangement Science*, *32*(5), 554–571. doi:10.1287/mnsc.32.5.554

Daft, R., & Lengel, R. (1984). Information Richness: A New Approach to Managerial Behaviour and Organizational Design. *Research in Organizational Behaviour*, *6*, 191–233.

Dennis, A. R., & Kinney, S. T. (1998). Testing Media Richness Theory in the New Media: The Effects of Cues, Feedback, and Task Equivocality. *Information Systems Research*, 9(3), 256–274. doi:10.1287/isre.9.3.256

Dimmick, J. (1997). The Theory of the Niche and Spending on Mass media: The Case of the "Video Revolution". *Journal of Media Economics*, *10*(3), 33. doi:10.1207/s15327736me1003_3

Ewing, D. J., Hall, R. S., & Schwartz, M. F. (1992). A Measurement Study of InternetFile Transfer Traffic. University of Colorado at Boulder. Retrieved from http://www.cs.colorado.edu/department/publications/reports/docs/CU-CS-571-92.pdf

Falcone, J. P. (2008, May 9). Netflix Watch Now: Missing too much popular content. *cnet news, News Crave*. Retrieved from http://news.cnet.com/8301-17938_105-9940529-1.html

Fortunati, L., Sarrica, M., O'Sullivan, J., Balcytiene, A., Harro-Loit, H., Macgregor, P., … De Luca, F. (2009). The Influence of the Interneton European Journalism. *Journal of Computer-Mediated Communication*, *14*(4), 928–963. doi:10.1111/j.1083-6101.2009.01476.x

Fowler, G. A., & Baca, M. C. (2010, August 25). The ABCs of E-Reading. *Wall Street Journal online*. Retrieved from http://online.wsj.com/article/SB10001424052748703846604575448093175758872.html

Fraleigh, C., Moon, S., Lyles, B., Cotton, C., Khan, M., Moll, D., … Diot, S. C. (2003). Packet-level traffic measurements from the Sprint IP backbone. *Network, IEEE*, *17*(6), 6–16. doi:10.1109/MNET.2003.1248656

Gannes, L. (2009). YouTube Changes Everything: The Online Video Revolution. In D. Gerbarg (Ed.), *Television Goes Digital* (Vol. 01, pp. 147–155). Springer.

Google, I. (2010, January). The 1000 most-visited sites on the web. Retrieved December 27, 2010, from http://www.google.com/adplanner/static/top1000/

Guo, L., Tan, E., Chen, S., Xiao, Z., Spatscheck, O., & Zhang, X. (2006). Delving into Internetstreaming media delivery: A quality and resource utilization perspective. *InternetMeasurement Conference Proceedings of the 6th ACM SIGCOMM on Internetmeasurement*, 217–230. doi:10.1145/1177080.1177108

Hilbert, M. (2011). Mapping the dimensions and characteristics of the world's technological communication capacity during the period of digitization. In *Working Paper*. Presented at the 9th World Telecommunication/ICT Indicators Meeting, Mauritius: International Telecommunication Union (ITU). Retrieved from http://www.itu.int/ITU-D/ict/wtim11/documents/inf/015INF-E.pdf

Hilbert, M. (2012). How to Measure "How Much Information"? Theoretical, methodological, and statistical challenges for the social sciences. *International Journal of Communication, 6*(Introduction to Special Section on "How to measure 'How-Much-Information'?"), 1042–1055.

Hilbert, M. (2013). *Big Data for Development: From Information - to Knowledge Societies* (SSRN Scholarly Paper No. ID 2205145). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2205145

Hilbert, M. (2013). Technological Information Inequality as an Incessantly Moving Target: The Redistribution of Information and Communication Capacities Between 1986 and 2010. *Journal of the American Society for Information Science and Technology*; online first; 19 Nov, 2013; http://onlinelibrary.wiley.com/doi/10.1002/asi.23020/abstract

Hilbert, M. (forthcoming). How Much of the Global Information and Communication Explosion Is Driven by More, and How Much by Better Technology?. *Journal of the American Society for Information Science and Technology*.

Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science, 332*(6025), 60 –65. doi:10.1126/science.1200970

Hilbert, M., & López, P. (2012a). How to Measure the World's Technological Capacity to Communicate, Store and Compute Information? Part I: results and scope. *International Journal of Communication, 6*, 956–979.

Hilbert, M., & López, P. (2012b). How to Measure the World's Technological Capacity to Communicate, Store and Compute Information? Part II: measurement unit and conclusions. *International Journal of Communication, 6*, 936–955.

International Data Corporation (IDC). (2008). IDC Media Center. Retrieved January 29, 2011, from http://www.idc.com/about/press.jsp

IFPI. (2007). The Recording Industry World Sales 1995-2004. International Federation of the Phonographic Industry. Retrieved from http://www.ifpi.org/content/section_statistics/index.html

Ipoque. (2006). *P2P Survey 2006*. Leibzig: ipoque Internetstudies.

Ipoque. (2007). *Internetstudy 2007* (Hendrik Schulze, Klaus Mochalski). Leibzig: ipoque Internetstudies.

Ito, Y. (1981). The Johoka Shakai approach to the study of communication in Japan. In C. Wilhoit & H. de Bock (Eds.), *Mass Communication Review Yearbook* (Vol. 2, pp. 671–698). Beverly Hills, CA: Sage.

ITU. (2010). Measuring the Information Society 2010. Geneva: International Telecommunication Union, ITU-D. Retrieved from http://www.itu.int/publ/D-IND-ICTOI-2010/en

Jenkins, H. (2004). The Cultural Logic of Media Convergence. *International Journal of Cultural Studies, 7*(1), 33–43. doi:10.1177/1367877904040603

Kalden, R. A. (2004). Mobile Internettraffic measurement and Modeling based on data from commercial GPRS networks. University of Twente. Retrieved from http://doc.utwente.nl/48238/1/thesis_kalden.pdf

Kalmus, V., Pruulmann-Vengerfeldt, P., Runnel, P., & Siibak, A. (2009). Mapping the Terrain of "Generation C": Places and Practices of Online Content Creation Among Estonian Teenagers. *Journal of Computer-Mediated Communication*, *14*(4), 1257–1282. doi:10.1111/j.1083-6101.2009.01489.x

Karagiannis, T., Broido, A., Brownlee, N., Claffy, K., & Faloutsos, M. (2004). Is P2P dying or just hiding? In *IEEE Global Telecommunications Conference 3, 1532-1538*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.60.6280

Kurtz, H. (2009, May 11). The Death of Print? *Washington Post*. Retrieved from http://www.washingtonpost.com/wp-dyn/content/article/2009/05/11/AR2009051100782.html

Kurzweil, R. (2001). The Law of Accelerating Returns. Kurzweil Accelerating Intelligence. Retrieved from http://www.kurzweilai.net/the-law-of-accelerating-returns

Lacort, J., Pont, A., Gil, J. A., & Sahuquillo, J. (2004). A Comprehensive Web Workload Characterization. Presented at the Second International Working Conference on Peformance Modelling and Evaluation of Heterogeneous Networks (HETNETs-04) 2004.

Lai, T. L. (2004). Service Quality and Perceived Value's Impact on Satisfaction, Intention and Usage of Short Message Service (SMS). *Information Systems Frontiers*, *6*(4), 353–368. doi:10.1023/B:ISFI.0000046377.32617.3d

Leibowitz, N., Bergman, A., Ben-shaul, R., & Shavit, A. (2002). Are file swapping networks cacheable? characterizing p2p traffic. In *In Proceedings of the 7th International www Caching Workshop*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.80.602

Leinen, S. (2001). Flow-Based Traffic Analysis at SWITCH. Presented at the Poster at Passive and Active Measurements (PAM) Workshop. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.7068

Lenhart, A., Madden, M., Smith, A., & Macgill, A. (2007). *Teens and Social Media*. Pew Research Center.

López, P., & Hilbert, M. (2012). *Methodological and Statistical Background on The World's Technological Capacity to Store, Communicate, and Compute Information* (online document). Retrieved from http://www.martinhilbert.net/WorldInfoCapacity.html

Lyman, P., Varian, H., Swearingen, K., Charles, P., Good, N., Jordan, L., & Pal, J. (2003). *How much information? 2003*. University of California, at Berkeley. Retrieved from http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/

MacKie-Mason, J. K., & Varian, H. R. (1994). Economic FAQs About the Internet. *Journal of Economic Perspectives*, *8*(3), 75–96. doi:10.3998/3336451.0001.110

Madnick, S., Smith, M., & Clopeck, K. (2009). *How Much Information? Case Studies on Scientific Research at MIT*. Global Information Industry Center at the University of California, San Diego. Retrieved from http://hmi.ucsd.edu/howmuchinfo_research.php

Mahanti, A. (1999). Web Proxy Workload Characterization and Modelling. Master Thesis, University of Saskatchewan, Canada. Retrieved from www.cse.iitd.ernet.in/~mahanti/papers/localitypaper.ps

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey & Company. Retrieved from http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation

Massey, J. L. (1998). *Applied Digital Information Theory: Lecture Notes by Prof. em. J. L. Massey*. Swiss Federal Institute of Technology. Retrieved from http://www.isiweb.ee.ethz.ch/archive/massey_scr/

Mayer-Schönberger, V., and Cukier, K. (2013). Big data: a revolution that will transform how we live, work and think (London: John Murray).

McCandless, M. (1999). The MP3 revolution. *Intelligent Systems and their Applications, IEEE*, *14*(3), 8–9. doi:10.1109/5254.769875

McLuhan, M. (1994). *Understanding Media: The Extensions of Man*. The MIT Press.

Morgan Stanley. (2006). Technology Q1 2006 Global Technology Data Book. Global Technology Team. Retrieved from http://www.morganstanley.com/institutional/techresearch/pdfs/global_techdatabook0306.pdf

Morris, M., & Ogan, C. (1996). The Internetas Mass Medium. *Journal of Communication*, *46*(1), 39–50. doi:10.1111/j.1460-2466.1996.tb01460.x

Monty Hall problem. (2013). Monty Hall problem at Wikipedia. In *Wikipedia, the free encyclopedia*. https://en.wikipedia.org/wiki/Monty_Hall_problem

Nagamalai, D., Dhinakaran, B. C., & Lee, J. K. (2008). An In-depth Analysis of Spam and Spammers. *International Journal of Security and its Applications*, *2*(2). Retrieved from http://arxiv.org/abs/1012.1665

Nature Editorial. (2008). Community cleverness required. *Nature*, *455*(7209), 1–1. doi:10.1038/455001a

NEA (National Endowment for the Arts). (2007). *To Read or Not To Read: a Question of National Consequence* (No. 47).

Neuman, R., Park, Y., & Panek, E. (2012). Tracking the Flow of Information into the Home: An Empirical Assessment of the Digital Revolution in the U.S. from 1960 - 2005. *International Journal of Communication, Special Section on "How to measure 'How-Much-Information'?"*(6), 1022–1041.

Newhagen, J., & Rafaeli, S. (1996). Why Communication Researchers Should Study the Internet: A Dialogue. *Journal of Communication*, *46*(1), 4–13. doi:10.1111/j.1460-2466.1996.tb01458.x

O'Reilly, T. (2005, September 20). What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. O'Reilly. Retrieved from http://oreilly.com/web2/archive/what-is-web-20.html

OECD, (Organization for Economic Cooperation and Development). (2005). *OECD Communications Outlook 2005*. Paris: Directorate for Science, Technology and Industry. Retrieved from www.oecd.org/sti/telecom/outlook

Okrent, D. (n.d.). The Death of Print? *Digital journalist*. Retrieved from http://www.digitaljournalist.org/issue0002/okrent.htm

Pallis, G., Vakali, A., Angelis, L., & Hacid, S. (2003). A study on workload characterization for a Web proxy server. *Proceedings of the 21st IASTED International MultiConference on Applied Informatics, 2003*, 779–784.

Pierce, J. R. (1980). *An Introduction to Information Theory* (2nd Revised ed.). New York, NY: Dover Publications.

Pool, I. de S. (1983). Tracking the Flow of Information. *Science, 221*(4611), 609–613. doi:10.1126/science.221.4611.609

Porter, J. (2005). Disk/Trend Reports 1977-1999. California. Retrieved from http://www.disktrend.com/

Reuters. (2006, July 16). YouTube serves up 100 million videos a day online. *USA Today*.

Ricciato, F., Hasenleithner, E., & Pilz, R. (2006). Composition of GPRS/UMTS traffic : snapshots from a live network. In *4th international workshop on Internetperformance, Simulation, Monitoring and Measurement (IPS-MOME'06)*. Salzburg. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.73.1933

Sandvine. (2008). *2008 Global broadband phenomena* (Research Report). Sandvine Incorporated. Retrieved from http://www.sandvine.com/news/global_broadband_trends.asp

Schmidt, J. (2007). Blogging Practices: An Analytical Framework. *Journal of Computer-Mediated Communication, 12*(4), article 13. doi:10.1111/j.1083-6101.2007.00379.x

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27*, 379–423, 623–656. doi:10.1145/584091.584093

Shannon, C. (1951). Prediction and Entropy of Printed English. *Bell System Technical Journal, 30*, 50–64.

UPU. (2007). Postal statistics. Universal Postal Union. Retrieved from http://www.upu.int/en/resources/postal-statistics/

Vaughan, T. (2010). *Multimedia: Making It Work* (Eighth Edition.). McGraw-Hill Osborne Media.

Verkasalo, H. (2007). Handset-based measurement of smartphone service evolution in Finland. *Journal of Targeting, Measurement and Analysis for Marketing, 16*, 7–25. doi:10.1057/palgrave.jt.5750060

Wang, Q., Edwards, H. K., Makaroff, D., & Thompson, R. (2002). Workload Characterization for an E-commerce Web Site. *In Proceedings of the 2003 conference of the Centre for Advanced Studies conference on collaborative research*, 313–327.

Wang, Y., Liu, Z., & Huang, J.-C. (2000). Multimedia content analysis-using both audio and visual clues. IEEE Signal Processing Magazine, 17(6), 12–36. doi:10.1109/79.888862

Williams, A., Arlitt, M., Williamson, C., & Barker, K. (2005). Web workload characterization: ten years later. In X. Tang, J. Xu, & S. Chanson (Eds.), *Web Content Delivery* (pp. 3–21). Springer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.120.9058

Wood, W. C. (1986). Consumer Spending on the Mass Media: The Principle of Relative Constancy Reconsidered. *Journal of Communication*, *36*(2), 39–51. doi:10.1111/j.1460-2466.1986.tb01422.x

Wood, W. C., & O'Hare, S. L. (1991). Paying for the Video Revolution: Consumer Spending on the Mass Media. *Journal of Communication*, *41*(1), 24–30. doi:10.1111/j.1460-2466.1991.tb02290.x

Zikopoulos, P., Eaton, C., deRoos, D., Detusch, T., & Lapis, G. (2012). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data* (IBM.). New York: McGraw. Retrieved from https://www14.software.ibm.com/webapp/iwm/web/signup.do?source=sw-infomgt&S_PKG=500016891&S_CPM=is_bdebook1&cmp=109HF&S_TACT=109HF38W&s_cmp=Google-Search-SWG-IMGeneral-EB-0508