A review of large-scale '*how much information*' inventories: variations, achievements and challenges

Martin Hilbert
University of California, Davis
www.martinhilbert.net

submitted to *Information Research*: an international electronic Journal (est. 1995)
http://www.informationr.net/ir/

ABSTRACT

Introduction. Pressed by the exploding number and increasing social importance of digital technologies during recent decades, combined with the attention given to the 'big data paradigm', several research projects have taken up the challenge to quantify the amount of information supplied and created, and/or consumed.

Method. This meta-study reviews existing inventories in a descriptive and comparative manner, focusing on methodological differences and challenges.

Analysis. The eight most important information inventory projects are reviewed. It shows that the quantification of information and communication presents a theoretical, methodological, as well as statistical challenge.

Results. Variations in the approaches include differences in how the information realm is conceptualized (e.g. in terms of stocks or flows, or in terms of creation or consumption, etc.); differences in the unit of measurement (words, bits, minutes, etc.); varying geographic and temporal scopes; and diverse additional attributes that highlight complementary aspects of the amount of information (e.g. the kind of technology, the sort of content, the type of user sector, etc.).

Conclusion. Depending on the particular question on the researchers mind and on the subsequent methodological choices, different conclusions and insights are obtained. The review ends with a discussion of the remaining theoretical and practical challenges.

Ever since Aristotle's student Demetrius (367 BC–ca. 283 BC) was asked to organize the Library of Alexandria in order to quantify "how many thousand books are there" (Aristeas, ca. 200 BC; quoted in Charles, 1913, Section 9), people have been curious about how much information there actually is. The increasing role of information and communication technologies (ICTs) as the dominating techno-social paradigm during the second half of the last century and the eventual transition to the digital age over the turn of the century motivated several research projects that have revisited this question more systematically. The quantification of information stocks and flows is driven by the desire to gain a deeper understanding of the social, economic, cultural and psychological role of information in society. To quote Lord Kelvin: "when you can measure what you are speaking about, and express it in numbers, you know something about it" (quoted from Bartlett, 1968, p: 723a). Recent advancements in large scale analysis of the digital wealth of "Big Data" (e.g. Manyika et al., 2011; Mayer-Schönberger and Cukier, 2013; Hilbert, 2015) have provided a more tangible interpretation of data quantifies as a form of socio-economic input, and therefore fueled the interest in the quantification of the availability and consumption of this ever more abundant kind of capital.

Introduction

The first scholars to take up the question in modern times were economist. In 1962, Machlup presented an estimation of "The Production and Distribution of Knowledge in the United States" (Machlup, 1962). From today's perspective he used an output oriented measurement unit, since he did not directly quantify the amount of information or knowledge, but rather the size of the information-intensive industries (in US dollars) and the respective occupational workforce. He followed the logic of national accounting from economics and identified some sectors that he considered information-intensive. Porat (1977) advanced this approach and reached the much-cited conclusion that the value of the composed labor and capital resources of these "information" sectors made up 25% of U.S. gross domestic product in 1967. He measures the economic value of the related "information activity [which] includes all the resources consumed in producing, processing, and distributing information goods and services" (Porat, 1977; p. 2). As information capital he loosely identified a "wide variety of information capital resources [which] are used to deliver the informational requirements of one firm: typewriters, calculators, copiers, terminals, computers, telephones and switchboards . . . microwave antennae, satellite dishes and facsimile machines" (Porat, 1977; pp. 2–3).

Over the decades, the basic notion of the approach evolved and led to the creation of international instruments that institutionalized the definition, harmonization, collection and interpretation of ICT indicators. The most influential heir is the "Working Party on Indicators for the Information Society (WPIIS)" of the OECD (Organisation for Economic Co-operation and Development; an international economic cooperation among 34 industrialized countries) (OECD, 2011). The

Working Party has set a number of global standards for measuring key components of the information society, such as the definition of industries producing ICT goods and services (OECD, 2007), a classification of ICT-, content- and media products, and a definition of electronic commerce and internet commerce transactions (OECD, 2009). Several international organizations from the United Nations have worked on taking such indicators global by fine-tuning them to meet the needs of developing countries (Partnership, 2005; 2008). This statistical groundwork is nowadays feeding an impressive mechanism of institutionalized research production on the advancement of the so-called information society (e.g. ITU, 2007; 2009; 2010; 2011; 2012; UNCTAD, 2005; 2006; 2009; 2010; 2011; 2012; 2013; World Bank, 2006; 2009; 2012), which is accompanied by at least a dozen of international ICT-indexes that rank societies according to their informational readiness (e.g. Minges, 2005). In other words, the worldwide measurement of ICT indicators counts with the commitment of sizable public funds and has already reached a considerable level of institutionalization. This is good news.

Despite the widespread usage and the undoubted usefulness of these kinds of continuous inventories of the digital age, it is important to underline that all of these efforts employ mere proxies of the amount of information and communication. They track indicators like the number of mobile phones and Internet subscriptions, or the amount of money spent or invested into ICT infrastructure, but not the amount of information or communication involved. It can be expected that more ICT infrastructure or more ICT expenditure lead to more information and communication, but this relation is not necessary, nor automatic, and can be deceptive (see Hilbert, 2014c).

Conscious of this limitation, a series of studies have aimed at the direct quantifying the amount of information and communication. These exercises led to new insights and conclusion of considerable public interest, producing attention-grabbing newspaper headlines like “Worldwide Data More Than Doubling Every Two Years” (Storage Newsletter, 2011); “Business Information Consumption: 9,570,000,000,000,000,000,000 Bytes per Year” (HMI News, 2011a); “World's shift from analog to digital is nearly complete” (NBC News, 2011); “All human information, stored on CD, would reach beyond the moon” (Lebwohl 2011); “Data Shows a Digital Divide in Global Bandwidth: Access to the Internet may be going global, but a ‘bandwidth divide’ persists” (MIT Technology Review, 2012); “World’s Total CPU Power: One Human Brain” (Wired, 2011); “New Digital Universe Study Reveals Big Data Gap: Less Than 1% Of World's Data Is Analyzed; Less Than 20% Is Protected” (PR Newswire, 2012); “Disconnect Between U.S. Wireless Demand and Infrastructure Capacity” (HMI News, 2011b), among others.

This article presents a comparative methodological review of the most important of these inventories, discussing their approaches and achieved insights. While collections of articles published elsewhere have provided detailed discussions of the challenges faced by one or the other study (see Special Section on "How to measure 'How-Much-Information'?" published in Volume 6 of the International Journal of Communication; i.e. Hilbert, 2012; Bohn and Short, 2012; Bounie and Gille, 2012; Dienes, 2012; ; Hilbert and López, 2012a; 2012b; Lesk, 2012; Neuman, Park, Panek, 2012; Odlyzko, 2012), this integrative review provides one single comparative overview of the most influential of these inventories in a comparative manner. The idea behind the article is not to normatively argue in favor of one approach or the other (as done elsewhere; see Dienes, 2012). Neither is it the idea to provide the history of the various intents in chronological context (as done elsewhere; see Hilbert, 2012). The main idea is rather to present and stress the complementary nature of the existing approaches in a descriptive manner, providing the reader with a one-stop introduction to the different existent methodological choices.

A comparative overview

Instead of presenting an exhaustive list of the more than two dozen individuals papers and studies undertaken so far, this review presents the distinct flavors of methodological choices by focusing on the most influential studies and grouping them into families (for a more detailed discussion of 26 different studies, see Dienes, 2012). This results in 8 broad families. This aggregation surely compromises historical and conceptual accuracy, but allows for a more clear-cut communication of the main distinctions between approaches as they exist up to date. The single one-stop-shop of this article is presented in Table 1. Throughout the article we will review different aspects of what Table 1 contains, and what is still missing.

Table 1: Comparative overview of eight different methodological families of information quantification inventories

	MPT / Pool	Dienes	Lyman & Varian / Bounie	Neuman, Park & Panek	Short & Bohn	Odlyzko / CISCO	IDC & EMC	Hilbert & López
Main conceptual groups	Supply vs. consumption.	Goods and services; Output, export, import, Consumption.	Stocks and Flows. Unique vs. Duplicate.	Supply vs. consumption.	Consumption; Enterprise production.	Internet traffic.	Created, captured, replicated.	Storage; Communication; Computation.
Unit(s) of measurement	Words, words per minute, words per US$	Bits	Bits; Euros	Minutes	Bits, words, hours; US$.	Bits	Bits	Optimally compressed bits, MIPS
Geographical Scope	Japan; U.S.	Hungary; U.S.; Rest of world.	U.S. with extrapolation to rest of world; Europe.	U.S.	U.S.	World; Five world regions.	World; U.S., Western Europe, China, India, rest of world.	World. 208 different countries for telecom.
Temporal Scope	1960-1977	1980; 1990; 2002; 1945-2010.	2000; 2003	1960-2005	2009; 2010; 2013	1990-2003; 2006; 2012	2007, 2008, 2011, 2012.	1986; 1993; 2000; 2007. (1986-2010 for telecom)
Fine-grained distinctions	Mass vs. point-to-point media. Print vs. electronic media	Corporations, Government, Household, Non-profit. Human vs. machine consumable	Users; Enterprises.	Household.	Consumers, Enterprises.	For 2006 & 2012: Fixed vs. Mobile; Consumer vs; Business.	Consumers vs. Enterprises. Protected vs. non-protected; Cloud vs. decentralized.	Text, images, audio, video. Approximate (hypothetical) users for telecom.

Stylized exemplary findings	* Faster growth of information provision than consumption * End of the hegemony of text * Electronic and point-to-point media became much more price-effective, while analogue mass media stagnated in cost effectiveness	* Flow corporations to households dominates, but households to household is rapidly growing * Decreasing share of Hungary's domestic productions in domestic output	* Electronic channels contain 3.5 times more unique information than storage media * U.S. produces 40% of world's newly created information content in bits, and 60% in Euros. * Paper printing is still increasing	* Ratio between information supply and demand grew from [82 : 1] in 1960, to [884 : 1] in 2005 * Machines will have to help to sort out this information overload	* Consumption grew from 11 to over 14 hours/day from 2008-2013 * Over half of all media bytes are received by computers * Two-thirds of bits are processed by low-end, entry-level servers costing less than US$25,000	* Global mobile data traffic grows 3 times faster than Global fixed IP traffic * Internet video traffic is 64% of all consumer Internet traffic * The average broadband speed grew 30% from 2011 to 2012	* Growth information creation outpaces storage capacity * 70% of information is created and consumed by consumers * Less than a third of info has minimal security or protection	* Share of global digital storage grew from 1 % in 1986, over 25 % in 2000, to 97 % in 2007. * Better compression algorithms contribute as much as more and better hardware. * Digital divides among and within countries continuously evolve
Carrying media included	Mail, direct mail, newspapers, books, magazines, advertising literature, phonograph records, music tapes, outdoor advertising (billboards etc), telephone directories, mailgrams; fixed public and private phone, mobile phone, public and private telegraph, radio, television, wire broadcast,	Education, personal communication, TV and radio, writing reading, phone, cultural services, entertainment, Theatres, museums, concerts; cable TV, TV and radio programming (originals), education; paper-based, videocassettes, records and audiocassettes,	Newspapers, magazines, books (incl. telephone directories), paper-based office and home documents, mails, records, photographic, industrial and cinematographic roll and sheet films, positive and negative, photos, records, magnetic cassettes, hard disk drives, floppy disks, optical disks, Internet, phone,	Newspapers, magazines, books (incl. Telephone directories), mails, records, records, magnetic cassettes, CD, VCR, DVD, DVR, portable audio, videogame; Terrestrial and satellite broadcasting, cable TV, terrestrial and satellite radio broadcasting, theatrical motion picture, wireline,	Newspapers, magazines, books, recorded music; Cable TV SD (Standard Definition), Over-air TV SD, Cable-TV HD (High Definition), Over air TV HD, Satellite SD, Satellite HD, Mobile TV, Other TV (Delayed view), Internet video, satellite radio, AM radio, FM radio, fixed-line voice, cellular voice, high-end computer	Broadband Internet and IP traffic; mobile, cable and wired telecomm services: Internet video to PC, toTV, VoIP, video communications, gaming, P2P, Web/Data	Hard disk drives, optical, tape, flash memory, digital cameras, fixed and mobile phones, PCs, servers; ATMs; RFIC; sensors; MP3 players; GPS; audio players, mobile subscribers, LCD/Plasma TVs, games, security systems, datacenter applications;	Video analog, photo print, audio cassette, photo negative, cine movie film, vinyl LP, TV episodes film, x-rays, TV movie film, newsprint, other paper print, books; PC hard-disk, DVD and Blu-Ray, digital tape, server and mainframe hard-disk, CDs and minidisks, portable hard-disks, portable media player, memory cards, PDA, floppy disks, digital camera camcorders internal, chip cards; TV-

	cable television services, lectures, education, entertainment, outdoor advertising, face-to-face conversations outside the home; movies, data communication. Imagery and music excluded!	magnetic tapes and reels, diskettes, hard disks, fixed and mobile data services, films, manual creation of digital data (keyboarding, mouse).	Radio and TV broadcasting, PC, market software, games software, piracy software.	cellular, IM phone services, dial-up, broadband, WiFi Internet services.	gaming, computer gaming, console gaming, handheld gaming, internet including email, offline programs, movies in theaters, LAN, WiFi; PC, computers, enterprise servers, Processing services of servers.		camcorders; webcams; surveillance; scanners; barcode readers; medical imaging; digitized video.	Terrestrial, TV-cable, TV-satellite, radio, newspapers, paper advertisement, personal navigation GPS; fixed phone, Internet, mobile phone, paper postal; PCs, Videogame consoles, Servers & Mainframe, Supercomputers, Pocket calculators; Microcontrollers; Graphic Processing, Digital Signal Processors.
References	Ito, 1981; Pool, 1983; Pool et al., 1984; Neuman and Pool, 1986; MPT review: Duff, 2000.	Dienes, 1986; 1992; 1994; 2002; 2010.	Lyman, et al., 2000; 2003; Bounie, 2003; Bounie and Gille, 2012.	Neuman, Park, Panek, 2012.	Bohn and Short, 2009; Short, Bohn and Baru, 2012; Short, 2013.	Odlyzko, 2003; 2008; 2010; Cisco Systems, 2008; 2011; 2012; 2013.	Gantz, et al., 2007; 2008; Gantz and Reinsel, 2011; 2012.	Hilbert and López, 2011; 2012a; 2012b; Hilbert, 2011; 2012; 2013; 2014a; 2014b.

Main conceptual groups

Different studies come up with largely varying numbers, which is often confusing to readers of this literature. For example, Lyman et al. (2000; 2003) report that the world produced some 4 exabytes of unique information in the year 2000, while Hilbert and Lopez (2011) estimate that the world's installed capacity of storing and of communicating optimally compressed information in 2000 reached some 1,200 exabytes. The latter number is roughly 300 times larger than the former. Hilbert and Lopez (2011) also report that the amount of globally communicated amount of information sums up to 1.15 zettabytes in 2007, while Bohn and Short (2009) report that only one year later, in 2008, American's alone consume more than three times as much, 3.6 zettabytes. The reason for these differences is of methodological nature. The devil is in the detail. Unique information is not equal to installed technological capacity, and communication is not equal to consumption.

This can also be seen in the resulting growth rates. Focusing on consumption, the growth rates of consumed bytes estimated by Bohn and Short (2009) are in the same order of magnitude than the growth rates of hours of media consumption. They estimate that in the United States, hours of information consumption grew at 2.6 % per year from 1980 to 2008, while bytes consumed increased at 5.4 % per year. Focusing on the installed capacity, Hilbert and López (2011) and Hilbert (2014a) detect annual growth rates of 20-30%.

The other way around, at times similar numbers refer to different things, increasing the existing confusion, especially when secondary literature cites the reported numbers. For example, Gantz, et al. (2008) report that the "digital universe" in 2007 inhabits 281 exabytes, while Hilbert and López (2011) report that the worldwide installed capacity to store information consists of 295 exabytes. While both numbers are similar, the second number refers exclusively to data storage capacity, the first number also includes data creation and communication flows, such as sent SMS and Emails. Besides, the second number refers to optimally compressed bits, while the first number reports uncompressed binary digits. Finally, the second study covers some 60 types of technologies, while the first number tracks some 30 comparable types.

What to measure is a question of research interest, not one of validity. It is true that some studies report their sources and assumption in a more transparent manner than others. For example, López and Hilbert (2011) provide more than 300 pages of Supporting Appendix outlining methodological assumptions and providing the details of their more than 1,100 distinct sources, while Gantz, et al. (2008) present one page of notes on methodology and key assumptions, list 52 sources and declare that additionally internal IDC databases were used. Such differences in style is mainly due to the academic or commercial nature of the study, and does not change the fundamental fact that different researchers are simply interested in different things, which leads to different conclusions.

The first distinction to make is if the amount of information is accounted for in form of a stock (e.g. information storage) or in terms of a flow (e.g. broadcasting or communication). Besides this basic distinction, there are several other aspects, mainly concerning the distinction between information supply (or creation) versus demand (or consumption). Figure 1 differentiates among some broad conceptual groups (see Figure 1). The presentation should only be understood schematically. Particular studies use specific definitions that often crosscut these schematic categories. Some studies compare information supply and demand and have long found an increasing divergence between information provision and consumption, resulting in an increasingly intensified information density per user (Pool, 1983; Neuman, Park, Panek, 2012).

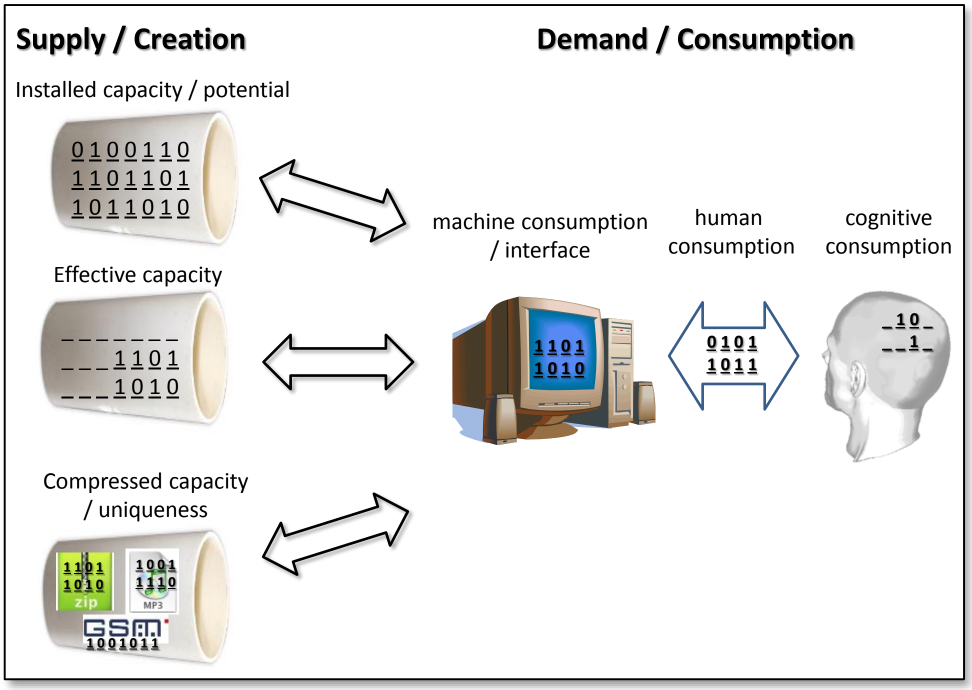
On the supply side, researchers sometimes report the installed capacity, which, in its purest form, simply accounts for the existing technological infrastructure (e.g. Lesk, 1997). The equivalent would be to assume that all hard disks would be filled, all fiber-optic cables run at full capacity, and all PCs and servers would be computing for 24 hours a day. Another alternative is to focus only on the information that is effectively present, which is a subgroup of the former. For example, according to Hilbert and López (2012a), if all broadcast receivers would receive information for 24 hour per day, 15.9 zettabytes could have been transmitted in 2007. Effectively, however, the average broadcasting receiver only runs for some three hours per day, resulting in an effective capacity of 1.9 zettabytes. Besides, and this aspect already jumps ahead to the subsequent section on the measurement unit, when measuring bits, one can measure the brute force number of binary bins existent in a storage device or in a communication channel (often referred to as "binary digits"), or to a more or less sensible compression of the information contains in these space holders (compressed bits) (for a more detailed discussion see Hilbert and López, 2012b). Since compression can largely reduce the numbers of bits of the same content, Lyman, et al. (2000; 2003) present a range of high and low estimates, which responds to different levels of compression available at a certain point in time. Hilbert and López (2011) assume that all content, independent from which year, would be compressed with the optimal compression algorithms available in the year of measurement, which has the benefit of making the amount of content comparable over time.

The general logic of compression leads to the distinction between unique and duplicate information. What compression essentially does is to take redundancy out of the source. This means that five equal pieces of content are not recorded or transmitted five times (i.e. [content], [content], [content], [content], [content]), but rather one time, while adding the marginally negligible informatics remark of duplication (i.e. [content]*5). Optimal compression eliminates duplication asymptotically. Therefore, if it would be possible to make the world's global information capacity subject to one a single compression mechanism, only purely unique information would be identified (see Hilbert and López, 2012b; Box 2). Following the compression logic one could run the compression algorithm not on the global amount of information, but on the amount of information pertaining to an individual. Lyman et al. (2000; 2003) aimed for an approximation of unique content per individual following a less technical methodology. On the contrary, the estimations of Hilbert and López (2011, 2012a) apply compression standards as reported by the industry per file type, such as a song, an average text file, or a movie. This takes out internal redundancy from these standard information entities (predictability and uniqueness within a song, text of movie), measuring only unique information within such entities, while not eliminating redundancy among entire duplicates of the same song, text file or movie.

This created supply of information can then be consumed by a machine and/or human. There are different ways to measure consumption. Bohn and Short (2009) and Neuman, Park and Panek (2012) assess the amount of time an individual interacts with the media and multiply this time with a certain information flow rate. This essentially assumes that every second of interaction has the same average information flow intensity. Something additional that can be done is to apply some kind of "fudge factor" to media interaction time periods, which accounts for a certain "percentage of inattention" (Pool, 1983, p. 610). This suggests distinguishing between a "gross" rate of human consumption and a "net" rate of effective cognitive consumption (see Figure 1). In reality, data sources on the question of attention are scare and ambiguous, which makes this distinction dubious in practice.

Depending on the focus of specific definitions, resulting numbers vary. For example, according to the numbers of Hilbert and López (2012a), if all Internet subscriptions would run at the potential bandwidth promised by Internet network providers for 24 hours per day, the world would need a network infrastructure that could carry 13.6 zettabytes in 2007. At the same time, people report to "use" or "consume" the Internet for 1.6 hours per day on average, which reduces this potential to 907 exabytes of gross media consumption (13,600*1.6/24). Comparing the numbers reported by Odlyzko (2010) and Cisco Systems (2008) about the existing Internet backbone infrastructure that effectively carries information, which is some 68 exabytes, it results that the average user only uses its promised full bandwidth for effectively nine minutes per day. During the remaining 87 minutes of the session, the screen is open, but no telecommunication takes place through the modem.

Figure 1: Broad distinction among conceptual groups



Unit of measurement

Besides the question of conceptualization, there is also the question of the scale of measurement. Usually researchers estimate the number of technological devices, classify these devices into different kinds of device families, and then multiply each kind of device with a respective average performance indicator in a chosen unit that represents information. An alternative approach tracks the amount of US$ spent into the technological infrastructure (instead of tracking the number of devices), and then multiplies the respective spending category with an information performance indicator of a certain unit (Short, Bohn and Baru; 2011).

The first variable that defines the measurement unit is the focus on stocks (information in space), on flows (information per unit), or on some kind of information process (which can refer to some metric that measures information processes in space and time, such as instructions (MIPS) or operation (FLOPS)) (see Table 2).

The pioneering Information Flow Census of Japan's Ministry of Posts and Telecommunications (MPT) from the 1970s and early 1980s (Ito, 1981) initially chose uncompressed binary digits as the unit of measurement. However, they felt that the results did not sufficiently recognize the contribution of text, in relation to data-intensive images and voice, so they decided to introduce the measure of "amounts of words" as the unifying unit. This was effectively implemented by the use of conversion rates that assumed that a minute of speech over radio or a telephone line was equal to 120 words, a picture on a fax machine was equal to 80 words per page, and TV provided 1,320 words per minute (see also Duff, 2000).

Bohn and Short (2009) and Short (2013) have undertaken the effort to present information consumption and in different informational units, namely bits, words and amounts of time. In their (2009) find that in 2008 Americans consumed about 1.3 trillion hours of information outside of work, which totaled 3.6 zettabytes, corresponding to the informational equivalent of 1,080 trillion words. The comparison of the resulting numbers led to interesting insights. Video sources (moving pictures) dominate bytes of information (i.e. from television and computer games). If hours or words are used as the measurement, information sources are more widely distributed, with substantial amounts from radio, internet browsing, and others. This high number of bytes contained in video begs the question of the value of information, i.e. in comparison to the information stemming from radio and internet (more weight in terms of words).

Table 2: Measurement units

		Supply / Creation		
		Storage / Stocks	Communication/ Flows	Computation/ Processes
Demand / Consumption	Technical info metric	bits (compressed)	bits (compressed) / time	instructions/sec (MIPS; FLOPS); bits of output
	Concepts	words	words / time	*“tasks” (?)*
	Time	-	time	*O-notation (?)*

Geographical scope

As shown in Table 1, these kinds of inventories either focus on a global aggregate level, or on a specific country or region (such as the U.S., Japan, Hungary, or Europe). The reason in more practical than methodological and stems from the availability of reliable statistics. For some technologies, such as for the estimation of internet traffic, it is much easier to estimate the global aggregate capacity, while other statistics are only available for specific countries. In order to be able to cover more countries, studies often make inferences on basis of statistics from another country. As subsequent studies have shown, such extrapolations have to be taken with a large grain of salt, since regional and national differences can be surprisingly large. For example, for the estimation of global telephone traffic, Lyman, Varian ad collaborators (2003) follow the lead of

Bounie (2003) in taking the number of minutes per line of France as a representation for the entire world (resulting in some 9.5 minutes per line per day). More detailed data became available later (ITU, 2010) and showed a global weighted average of some 18 minutes per installed line in the world, almost twice as much as the average of France (with more industrialized member countries of the OECD reaching a weighted average of some 21 minutes per line per day and less industrialized non-OECD countries reach some 7 minutes per line per day).

The increasing direct registration of digital information flows through the sampling of IP traffic (e.g. Cisco Systems, 2012) or the testing of broadband bandwidth (e.g. Ookla, 2014) can potentially provide more sustainable and more cost-effective solutions to capture aspects of this worldwide diversity. It is important to notice that related tracking of online usage around the world can go as far as employing illegal practices. One example is the study of Botnet (2012), an anonymous hacker who took over some 420,000 devices to conduct a swift Internet census as the captured routers pinged IP addresses and waited for answers. Another example is the polemic online tracking of the U.S.'s National Security Agency, which publicly states to "touch" about 1.6 % of global Internet traffic, while selecting 0.025 % for more detailed review (including content) (NSA, 2013; p.6). While such extensive 'sampling' provides a wealth of up-to-date information about magnitudes, usage patterns and specific content, it currently takes place in a legal and ethical grey zone with no clear definition about the proportionality and adequateness of means and ends.

Temporal scope

Time series are the key for understanding dynamics and therefore impact. Statistical scarcity is once again the main limitation here. Most studies with extensive reach and long time series (such as Dienes, 2010; or Hilbert and López, 2011) often take more detailed inventories in specific years and then extrapolate between them.

As always when working with time series, methodological consistency is of outmost importance. Even if the very unit of measurement is questioned, methodological consistency can still lead to important insights, since growth rates can reveal relative tendencies independently of the chosen unit of measurement. For example, while the early studies of Japan's Ministry of Posts and Telecommunications (MPT) (Ito, 1981) and of Ithiel de Sola Pool (1983) were criticized for its choice of indicator (focusing exclusively on text, translated in words, while excluding imagery and audio), these pioneering studies were able to show ground-breaking results with regard to the transitions from analogue mass media to electronic point-to-point media during the 1960 and 1970, as well as the diverging trajectories of information provision and consumption.

Fine-grained distinctions

Besides presenting aggregate numbers as results of their inventory, all studies also include some kind of differentiation among different kinds of technologies or users. The nature of this distinction brings us back to the particularity of the question on the researcher's mind.

For example, in the early 1980, Pool (1983) was interested in the transition from mass communication (basically one-way information diffusion technologies), toward (what he called) point-to-point media (basically two-way communication technologies). He achieved to quantify the superiority of point-to-point in terms of cost-effectiveness and therefore the evolution from a broadcasting- to a communication paradigm. Cisco Sytems (2012) distinguishes between wired and wireless traffic and reports that in 2012 wired devices accounted for 59 % of global IP traffic. Other studies distinguish among the kind of content. Cisco Systems (2013) reports that in 2012 IP video traffic accounts for 60 % of all IP traffic. While increasing shares of video content is often seen as one of the characteristics of the digital multimedia age, Hilbert (2014a) reports that the relative share of text actually captures a larger proportion of the two-way communication exchanges than before the digital age. In the late 1980s, most technologically mediated communication exchanges took place in form of voice exchanges (through the telephone) and text represented less than half percent of (optimally compressed) bits that flowed through global information channels in 1986 (in the form of postal letters, etc.). The share of alphanumeric data grew to almost 30 % in 2007, a time when the internet communicates vast amount of written information on the web and people exchange large text files and databases.

Dienes (1994; 2002) distinguishes between the kind of societal sector. He reports that 72 % of the information goods and services output in the U.S. in 1990 is provided by corporations, 16 % by households and 12 % by governments. He also distinguishes between import and export of information goods and service and reports that the United States in 1990 imported 1.7 times more information than it exported.

In principle, there is no limitation to the kind of attribute that can be assigned to the information unit under analysis. In the fourth generation of their digital universe reports, Gantz and Reinsel (2012) became interested in the Big Data paradigm and asked about the share of the total amount that would be useful for informatics analysis. They report that in 2012 some 23% of the information in the digital universe would be useful for Big Data if it were tagged and analyzed, while in practice only 3% of the potentially useful data was tagged at that moment, and even less was analyzed.

Discussion and limitation

One frequent critique of the kind of information quantification studies reviewed here is that they only address the question of 'how much', which foregoes the question of 'meaning' or 'value'. It is important to emphasize that the main goal of the presented studies is the quantification of information, not a value judgment of the quality, impact or value of information. Many of the authors of those exercises even stress that the quantification of information does not necessarily say anything about the quality or value of this information. The assessment of quality or value of information requires the addition of supplementary variables. To quote Shannon's seminal 1948 paper: "Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities" (Shannon, 1948; p.379). This supplementary system allows defining a possible impact, the quality of the information or its value. Per definition, concepts like "impact of information", "value of information" or "quality of information" first of all require a metric for information (in the denominator) and then an additional metric for impact, value, or quality (in the numerator): {[unit of impact] / [unit of information]}; {[unit of value] / [unit of information]}, or {[quality / unit of information]}. In order to create indicators such as [US$ / bit], [growth / bit]; [attention / bit], or [pleasure / bit], one first of all needs to measure the denominator of the ratio: the amount of information. In order to test hypotheses about the value of information, we have to answer the "how much information" question first. Without normalization on the quantity of information, we would helplessly confuse the effects of "more information" with those of "better information." Only if the denominator is fixed, one can start to analyze which kind of the same amount is "better", "more impactful", or "more valuable". In short, information quantity is not equal to information quality or information value, but the second requires the first. Future studies will be required to obtain insights into these additional aspects.

References

Aristeas. (200AD, ca). The Letter of Aristeas to Philocrates. (R. Charles, Trans.)*Translation 1913*. Retrieved from http://www.attalus.org/translate/aristeas1.html

Bartlett, J. (1968). William Thompson, Lord Kelvin, Popular Lectures and Addresses [1891-1894]. In *Bartletts Familiar Quotations* (14th Edition.). Little Brown & Co.

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

Bohn, R., & Short, J. (2012). Measuring Consumer Information. *International Journal of Communication*, *Special Section on "How to measure 'How-Much-Information'?"*(6), 980–1000.

Botnet, C. (2012). *Internet Census 2012: Port scanning /0 using insecure embedded devices*. Retrieved from http://internetcensus2012.bitbucket.org/paper.html

Bounie, D. (2003). *The International Production and Dissemination of Information*. Paris: École Nationale Supérieure des Télécommunications (ENST). Retrieved from http://ses.telecom-paristech.fr/bounie/documents/Recherche/Annex.pdf

Bounie, D., & Gille, L. (2012). International Production and Dissemination of Information: Results, Methodological Issues and Statistical Perspectives. *International Journal of Communication*, *Special Section on "How to measure 'How-Much-Information'?"*(6), 1001–1021.

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

Cisco Systems. (2011). *Cisco Visual Networking Index: Forecast and Methodology, 2010-2015* (White Paper). Retrieved from http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360_ns827_Networking_Solutions_White_Paper.html

Cisco Systems. (2012). *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012–2017* (White Paper). Retrieved from http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html

Cisco Systems. (2013). *The Zettabyte Era—Trends and Analysis* (White Paper). Retrieved from www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/VNI_Hyperconnectivity_WP.html

Dienes, I. (1986). Magnitudes of the Knowledge Stocks and Information Flows in the Hungarian Economy. In *Tanulmányok az információgazdaságról* (in Hungarian., pp. 89–101). Budapest. Retrieved from http://infostat.hu/publikaciok/86-nagysr.pdf

Dienes, I. (1992). A Comparative Analysis of Communications Flow in Hungary, the United States and Japan. In *Proceedings of the III. Conference "Information Economy and Policy."* Budapest. Retrieved from http://infostat.hu/publikaciok/91-desola.pdf

Dienes, I. (1994). *Accounting the Information Flows and Knowledge Stocks in the U.S.: Preliminary Results* (Presentation in the University of California in Berkeley). Berkeley. Retrieved from http://infostat.hu/publikaciok/94-berkeleyreport.pdf

Dienes, I. (2002). *How much information? Chapters from the history of the accounting of information flows and knowledge stocks* (Cég-INFO). Retrieved from http://free.x3.hu/infostat

Dienes, I. (2010). *Twenty Figures Illustrating the Information Household of Hungary Between 1945 and 2008* (in Hungarian). Retrieved from http://infostat.hu/publikaciok/10_infhazt.pdf

Dienes, I. (2012). A Meta Study of 26 "How Much Information" Studies: Sine Qua Nons and Solutions. *International Journal of Communication*, *Special Section on "How to measure 'How-Much-Information'?"*(6), 874–906.

Duff, A. S. (2000). *Information society studies*. New York, NY: Psychology Press.

Gantz, J., Chute, C., Manfrediz, A., Minton, S., Reinsel, D., Schlichting, W., … McArthur, J. (2007). *The expanding digital universe: A Forecast of Worldwide Information Growth Through 2010*. Framingham: IDC (International Data Corporation) sponsored by EMC. Retrieved from http://www.emc.com/leadership/digital-universe/expanding-digital-universe.htm

Gantz, J., Chute, C., Manfrediz, A., Minton, S., Reinsel, D., Schlichting, W., & Toncheva, A. (2008). *The Diverse and Exploding Digital Universe: An Updated Forecast of Worldwide Information Growth Through 2011*. Framingham:

IDC (International Data Corporation) sponsored by EMC. Retrieved from http://www.emc.com/leadership/digital-universe/expanding-digital-universe.htm

Gantz, J., & Reinsel, D. (2011). *Extracting value from chaos*. IDC (International Data Corporation) sponsored by EMC. Retrieved from www.emc.com/collateral/analyst.../idc-extracting-value-from-chaos-ar.pdf

Gantz, J., & Reinsel, D. (2012). *The digital universe in 2020: Big Data, Bigger Digi tal Shadows, and Biggest Growth in the Far East*. IDC (International Data Corporation) sponsored by EMC. Retrieved from http://www.emc.com/leadership/digital-universe/iview/index.htm

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

Hilbert, M. (2012a). How much information is there in the "information society"? *Significance*, *9*(4), 8–12. doi:10.1111/j.1740-9713.2012.00584.x

Hilbert, M. (2012b). 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. (2015). Big Data for Development: A Review of Promises and Challenges. *Development Policy Review*.

Hilbert, M. (2014a). 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. (2014b). What is the content of the world's technologically mediated information and communication capacity: how much text, image, audio and video? *The Information Society: An International Journal*.

Hilbert, M. (2014c). Technological information inequality as an incessantly moving target: The redistribution of information and communication capacities between 1986 and 2010. *Journal of the Association for Information Science and Technology*, 65(4), 821–835. doi:10.1002/asi.23020

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.

HMI News. (2011a). Business Information Consumption: 9,570,000,000,000,000,000,000 Bytes per Year. *UCSD*. Retrieved from http://hmi.ucsd.edu/howmuchinfo_news_10_26_11.php

HMI News. (2011b). Disconnect Between U.S. Wireless Demand and Infrastructure Capacity. *UCSD*. Retrieved from http://hmi.ucsd.edu/howmuchinfo_news_10_26_11.php

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 (International Telecommunication Union). (2007). *Measuring the Information Society 2007*. Geneva: International Telecommunication Union, ITU-D. Retrieved from http://www.itu.int/ITU-D/ict/publications/idi/

ITU (International Telecommunication Union). (2009). *Measuring the Information Society 2009*. Geneva: International Telecommunication Union, ITU-D. Retrieved from http://www.itu.int/ITU-D/ict/publications/idi/

ITU (International Telecommunication Union). (2010). *Measuring the Information Society 2010*. Geneva: International Telecommunication Union, ITU-D. Retrieved from http://www.itu.int/ITU-D/ict/publications/idi/

ITU (International Telecommunication Union). (2011). *Measuring the Information Society 2011*. Geneva: International Telecommunication Union, ITU-D. Retrieved from http://www.itu.int/ITU-D/ict/publications/idi/

ITU (International Telecommunication Union). (2012). *Measuring the Information Society 2012*. Geneva: International Telecommunication Union, ITU-D. Retrieved from http://www.itu.int/ITU-D/ict/publications/idi/

Lebwohl, B. (2011, February 10). Martin Hilbert: All human information, stored on CD, would reach beyond the moon. *Earthsky: A Clear Voice for Science*, *Human world*. Retrieved from http://earthsky.org/human-world/martin-hilbert-all-human-information-stored-on-cd-would-reach-beyond-the-moon

Lesk, M. (1997). How Much Information Is There In the World? lesk.com. Retrieved from http://www.lesk.com/mlesk/ksg97/ksg.html

Lesk, M. (2012). One in a Million: Information vs. Attention. *International Journal of Communication, Special Section on "How to measure 'How-Much-Information'?"*(6), 907–919.

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. R., Dunn, J., Strygin, A., & Swearingen, K. (2000). *How much information 2000*. University of California, at Berkeley. Retrieved from http://www2.sims.berkeley.edu/research/projects/how-much-info/

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/

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. http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation

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

Minges, M. (2005). *Evaluation of e-Readiness Indices in Latin America and the Caribbean*. Santiago: United Nations ECLAC. Retrieved from http://www.cepal.org/SocInfo/publicaciones

MIT Technology Review. (2012). Data Shows a Digital Divide in Global Bandwidth: Access to the Internet may be going global, but a "bandwidth divide" persists. *Mike Orcutt*. Retrieved from http://www.technologyreview.com/view/427142/data-shows-a-digital-divide-in-global-bandwidth/

NBC News. (2011). World's shift from analog to digital is nearly complete. *Andrea Leontiou*. Retrieved from http://www.nbcnews.com/id/41516959/ns/technology_and_science-innovation/#.UtP767R0nKc

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.

Neuman, R., & Pool, I. de S. (1986). The flow of communications into the home. In S. Ball-Rokeach & M. Cantor (Eds.), *Media, audience and social structure* (pp. 71–86).

NSA (National Security Agency). (2013). *The National Security Agency: Missions, Authorities, Oversight and Partnerships*. Retrieved from www.nsa.gov/public_info/_files/speeches_testimonies/2013_08_09_the_nsa_story.pdf

Odlyzko, A. (2003). Internet traffic growth: Sources and implications. In B. B. Dingel, W. Weiershausen, A. K. Dutta, & K.-I. Sato (Eds.), *Optical Transmission Systems and Equipment for WDM Networking II* (pp. 1–15). Retrieved from http://adsabs.harvard.edu/abs/2003SPIE.5247.....D

Odlyzko, A. (2008, February 25). Threats to the Internet: Too Much or Too Little Growth? *Internet Evolution*. Retrieved from http://www.internetevolution.com/author.asp?section_id=592&doc_id=146747&

Odlyzko, A. (2010). *Minnesota Internet Traffic Studies (MINTS)*. University of Minnesota. Retrieved from http://www.dtc.umn.edu/mints/

Odlyzko, A. (2012). The Volume and Value of Information. *International Journal of Communication, Special Section on "How to measure 'How-Much-Information'?"*(6), 920–935.

OECD (Organisation for Economic Co-operation and Development). (2007). *Information Economy – Sector definitions based on the International Standard Industry Classification (ISIC 4)*. Paris: Working Party on Indicators for the Information Society. Retrieved from www.oecd.org/sti/sci-tech/38217340.pdf

OECD (Organisation for Economic Co-operation and Development). (2009). *Guide to Measuring the Information Society 2009*. Paris: Working Party on Indicators for the Information Society. Retrieved from http://www.oecd.org/sti/sci-tech/43281062.pdf

OECD (Organisation for Economic Co-operation and Development). (2011). *Guide to Measuring the Information Society 2011*. Paris: Working Party on Indicators for the Information Society. Retrieved from http://www.oecd.org/bookshop?9789264095984

Ookla. (2014). *NetIndex source data*. Retrieved from http://www.netindex.com/source-data/

Partnership, U. N. (2005). *Measuring ICT: the Global Status of ICT Indicators*. Geneva: ITU, OECD, UNCTAD, UNESCO Institute for Statistics, UN Regional Commissions (UNECLAC, UNESCWA, UNESCAP, UNECA), World Bank, EUROSTAT.

Partnership, U. N. (2008). *The Global Information Society: a Statistical View*. Geneva: ITU, OECD, UNCTAD, UNESCO Institute for Statistics, UN Regional Commissions (UNECLAC, UNESCWA, UNESCAP, UNECA), World Bank, EUROSTAT. Retrieved from http://www.unctad.org/en/docs/LCW190_en.pdf

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

Pool, I. de S., Inose,, H., Takasaki, N., & Hurwitz, R. (1984). *Communication Flows: A Census in the United States and Japan*. Amsterdam: North-Holland and University of Tokyo Press.

PR Newswire. (2012, December 11). New Digital Universe Study Reveals Big Data Gap: Less Than 1% Of World's Data Is Analyzed; Less Than 20% Is Protected. Hopkinton. Retrieved from http://www.prnewswire.com/news-releases/new-digital-universe-study-reveals-big-data-gap-less-than-1-of-worlds-data-is-analyzed-less-than-20-is-protected-183025311.html

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

Short, J. (2013). *How Much Media? 2013 Report on American Consumers*. San Diego: Global Information Industry Center of University of California, San Diego. Retrieved from http://hmi.ucsd.edu/howmuchinfo.php

Short, J., Bohn, R., & Baru, C. (2011). *How Much Information? 2010 Report on Enterprise Server Information*. San Diego: Global Information Industry Center at the School of International Relations and Pacific Studies, UC San Diego. Retrieved from http://hmi.ucsd.edu/howmuchinfo_research_report_consum_2010.php

Storage Newsletter. (2011, June 29). WW Data More Than Doubling Every Two Years. Retrieved from http://www.storagenewsletter.com/rubriques/market-reportsresearch/idc-digital-universe-emc/

UNCTAD (United Nations Conference on Trade and Development). (2005). *Information Economy Report 2005*. Geneva: United Nations Conference on Trade and Development. Retrieved from http://unctad.org/en/pages/PublicationArchive.aspx?publicationid=1540

UNCTAD (United Nations Conference on Trade and Development). (2006). *Information Economy Report 2006: The development perspective*. Geneva: United Nations Conference on Trade and Development. Retrieved from http://unctad.org/en/pages/PublicationArchive.aspx?publicationid=1549

UNCTAD (United Nations Conference on Trade and Development). (2009). *Information Economy Report 2009, Trends and Outlook in Turbulent Times* (No. UNCTAD/IER/2009) (p. 153). Geneva: United Nations Conference on Trade and Development. Retrieved from http://www.unctad.org/Templates/webflyer.asp?docid=11915&intItemID=2068&lang=1

UNCTAD (United Nations Conference on Trade and Development). (2010). *Information Economy Report 2010: ICTs, Enterprises and Poverty Alleviation*. Geneva: United Nations Conference on Trade and Development. Retrieved from http://unctad.org/en/pages/PublicationArchive.aspx?publicationid=1575

UNCTAD (United Nations Conference on Trade and Development). (2011). *Information Economy Report 2011: ICTs as an Enabler for Private Sector Development*. Geneva: United Nations. Retrieved from http://unctad.org/en/PublicationsLibrary/ier2011_en.pdf

UNCTAD (United Nations Conference on Trade and Development). (2012). *Information Economy Report 2012: The Software Industry and Developing Countries*. Geneva: UNCTAD. Retrieved from http://unctad.org/en/pages/PublicationWebflyer.aspx?publicationid=271

UNCTAD (United Nations Conference on Trade and Development). (2013). *Information Economy Report 2013: The Cloud Economy and Developing Countries*. Geneva: UNCTAD. Retrieved from http://unctad.org/en/pages/PublicationWebflyer.aspx?publicationid=710

Wired. (2011, February 11). World's Total CPU Power: One Human Brain. *Ars Technica, Wired Science*. Retrieved from http://www.wired.com/wiredscience/2011/02/world-computer-data

World Bank. (2006). *Information and Communications for Development: Global Trends and Policies in 2006*. Washington D.C.: Pierre Guislain, Christine Zhen-Wei Qiang, Bruno Lanvin, Michael Minges, and Eric Swanson. World Bank. (2009). *Information and Communications for Development: IC4D 2009, Extending Reach and Increasing Impact*. Washington D.C.

World Bank. (2012). *Information and Communications for Development: IC4D 2012, Maximizing Mobile*. Washington D.C.