I

Information Quantity

Martin Hilbert
Department of Communication, University of California, Davis, Davis, CA, USA

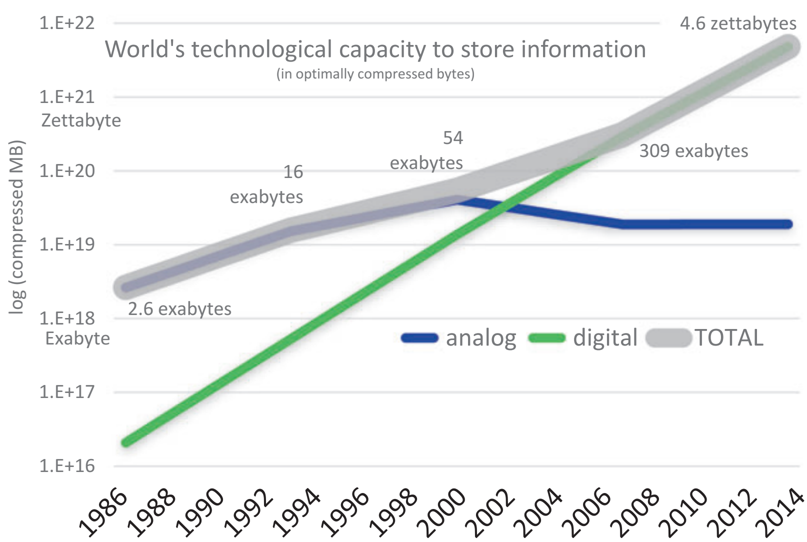
The question of "how much information" there is in the world goes at least back to the times when Aristotle's student Demetrius (367 BC – ca.283 BC) was asked to organize the Library of Alexandria in order to collect and quantify "how many thousand books are there" (Aristeas 200AD, sec. 9). Pressed by the exploding number of information and communication technologies (ICT) during recent decades, several research projects have taken up this question again since the 1960s. They differ considerably in focus, scope, and measurement variable. Some used US$ as a proxy for information (Machlup 1962; Porat 1977), others number of words (Ito 1981; Pool 1983; Pool et al. 1984), some focused on the national level of a single country (Dienes 1986, 2010), others broad estimations for the entire world (Gantz and Reinsel 2012; Lesk 1997; Turner et al. 2014), some focused on unique information (Bounie 2003; Lyman et al. 2000), and others on a specific sector of society (Bohn and Short 2009; Short et al. 2011) (for a methodological comparison and overview see Hilbert 2012, 2015a).

The big data revolution has provided much new interest in the idea of quantifying the amount of information in the world. The idea is that an important early step in understanding a phenomenon consists in quantifying it: "when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind" (Lord Kelvin, quoted from Bartlett 1968, p. 723). Understanding the data revolution implies understanding how it grows and evolves.

In this inventory, we mainly follow the methodology of what has become a standard reference in estimating the world's technological information capacity: Hilbert and López (2011). The total technological storage capacity is calculated as the sum of the product of technological devices and their informational storage performance. Technological performance is measured in the installed binary hardware digits, which is then normalized on compression rates. The hardware performance is estimated as "installed capacity" (not the effectively used capacity), which implies that it is assumed that all technological capacities are used to their maximum. For storage this evaluates the maximum available storage space ("as if all storage were filled"). The normalization on software compression rates is important for the creation of meaningful time series, as compression algorithms have enable to store more information on the same hardware infrastructure over recent decades (Hilbert 2014a; Hilbert and López 2012a). We normalize on "optimally compressed

L.A. Schintler, C.L. McNeely (eds.), *Encyclopedia of Big Data*,
https://doi.org/10.1007/978-3-319-32001-4_511-1

Information Quantity, Fig. 1 World's technological capacity to store information 1986–2014 (log on y-axis) (Source: Based on the methodology of Hilbert and López (2011), with own estimates for 2007–2014)



bits" (as if all content were compressed with the best compression algorithms possible in 2014 (Hilbert and López 2012b). It would also be possible to normalize on a different standard, but the optimal level of compression has a deeper information theoretic conceptualization as it approaches the entropy of the source (Shannon 1948). For the estimation of compression rates of different content, justifiable estimates are elaborated for 7-years intervals (1986, 1993, 2000, 2007, 2014). For more see Hilbert (2015b) and López and Hilbert (2012).

For the following result, the estimations for the period 1986–2007 follow Hilbert and López (2011). The update for 2007–2014 follows a mix of estimates, including comparisons with more current updates (Gantz and Reinsel 2012; Turner et al. 2014).

Figure 1 shows that the world's technological capacity to store information has almost reached 5 zettabytes in 2014 (from 2.6 exabytes in 1986 to 4.6 zettabytes in 2014). This results in a compound annual growth rate of some 30%. This is about five times faster than the world economy grew during the same period. The digitalization of the world's information stockpile happened in what is a historic blink of an eye: in 1986, less than 1% of the world's mediated information was still stored in digital format. By 2014, less than 0.5% is stored in analog media. Some analog storage media are still growing strongly today. For example, it is well known that the long-promised "paperless office" has still not arrived. The usage of paper still grows with some 15% per year (some 2.5 times faster than the economy), but digital storage is growing at twice that speed. The nature of this exponential growth trend leads to the fact that until not too long ago (until the year 2002) the world still stored more information in analog than in digital format. Our estimates determine the year 2002 as the "beginning of the digital age" (over 50% digital).

It is useful to put these mind-boggling numbers into context. If we would store the 4.6 optimally compressed zettabytes of 2014 in 730 MB CD-ROM discs (of 1.2 mm thickness), we could build about 20 stacks of discs from the earth to the moon. If we would store the information equivalent in alphanumeric symbols in double-printed books of 125 pages, all the world's landmasses could have been covered with one layer of double-printed book paper back in 1986. By 1993 it would have grown to 6 pages and to 20 pages in the year 2000. By 2007 it would be one layer of books that covers every square centimeter of the world's land masses, two layers by 2010/2011, and some 14 layers by 2014 (letting us literally stand "knee-deep in information"). If we would make piles of these books, we would have about 4500 piles from the Earth to the sun.

Estimating the amount of the world's technological information capacity is only the first step. It can and has been used as input variable to investigate a wide variety of social science questions of the data revolution, including its international distribution, which has shown that the digital divide carries over to the data age (Hilbert 2014b, 2016); the changing nature of content, which has shown that the big data age counts with a larger ratio of alphanumeric text over videos than the pre-2000s (Hilbert 2014c); the crucial role of compression algorithms in the data explosion (Hilbert 2014a), and the impact of data capacity on issues like international trade (Abeliansky and Hilbert 2017).

Further Readings

Abeliansky, A. L., & Hilbert, M. (2017). Digital technology and international trade: Is it the quantity of subscriptions or the quality of data speed that matters? *Telecommunications Policy, 41*(1), 35–48. https://doi.org/10.1016/j.telpol.2016.11.001.

Aristeas. (200AD, ca). The letter of Aristeas to Philocrates. 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 ed.). Boston: 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.

Bounie, D. (2003). *The international production and dissemination of information* (Special project on the economics of knowledge Autorità per le Garanzie nelle Comunicazioni). Paris: École Nationale Supérieure des Télécommunications (ENST).

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

de S. Pool, I., 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.

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.

Dienes, I. (2010). Twenty figures illustrating the information household of Hungary between 1945 and 2008 (in Hungarian). http://infostat.hu/publikaciok/10_infhazt.pdf

Gantz, J., & Reinsel, D. (2012). *The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the Far East*. IDC (International Data Corporation) sponsored by EMC.

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. (2014a). How much of the global information and communication explosion is driven by more, and how much by better technology? *Journal of the Association for Information Science and Technology, 65*(4), 856–861. https://doi.org/10.1002/asi.23031.

Hilbert, M. (2014b). 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. https://doi.org/10.1002/asi.23020.

Hilbert, M. (2014c). 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. https://doi.org/10.1080/01972243.2013.873748.

Hilbert, M. (2015a). A review of large-scale 'how much information' inventories: Variations, achievements and challenges. *Information Research, 20*(4). http://www.informationr.net/ir/20-4/paper688.html.

Hilbert, M. (2015b). *Quantifying the data deluge and the data drought* (SSRN scholarly paper no. ID 2984851). Rochester: Social Science Research Network. https://papers.ssrn.com/abstract=2984851

Hilbert, M. (2016). The bad news is that the digital access divide is here to stay: Domestically installed bandwidths among 172 countries for 1986–2014. *Telecommunications Policy, 40*(6), 567–581. https://doi.org/10.1016/j.telpol.2016.01.006.

Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science, 332*(6025), 60–65. https://doi.org/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.

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: Sage.

Lesk, M. (1997). How much information is there in the world? lesk.com. http://www.lesk.com/mlesk/ksg97/ksg.html

López, P., & Hilbert, M. (2012). Methodological and statistical background on the world's technological capacity to store, communicate, and compute information (online document). 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.

Machlup, F. (1962). *The production and distribution of knowledge in the United States*. Princeton: Princeton University Press.

Porat, M. U. (1977, May). *The information economy: Definition and measurement*. Superintendent of Documents, U.S. Government Printing Office, Washington, DC. 20402 (Stock No. 003-000-00512-7).

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423, 623–656. https://doi.org/10.1145/584091.584093

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, University of California, San Diego. http://hmi.ucsd.edu/howmuchinfo_research_report_consum_2010.php.

Turner, V., Gantz, J., Reinsel, D., & Minton, S. (2014). *The digital universe of opportunities: Rich data and the increasing value of the internet of things*. IDC (International Data Corporation) sponsored by EMC.