Internet Sectoral Overview

Number 3
September, 2023
Year 15

Data economy: Concepts and challenges for measurement

Governing the data economy

By Marielza Oliveira[1]

Introduction

A Web search for the phrase "data is the new oil" returns 2.12 billion responses in 0.58 seconds,[2] demonstrating how deeply it has embedded itself in our lexicon. Coined in 2006 by British mathematician Clive Humby, it is often repeated because it is evocative, conjuring the image of a mysterious natural substance, hidden somewhere, waiting to be unearthed and refined, so that it can reveal the meaning of everything.

The "data-as-oil" analogy succinctly conveys a series of underlying notions. For starters, the word "data" is the plural form of the Latin word *datum*. The singular form is rarely used, as *datum* is perceived as only being useful in quantity: While one cannot do much with a single unit, having multiples, collected over time, across entities or categories, is what enables the application of analytical methods to extract insights from the relationships between individual *datum*. Thus, embedded in the concept of data is an assumption that "more is better."

As in the Ubuntu philosophy, which posits that the self can only be understood in relation to the community in which it is embedded, *datum* is perceived as without meaning until extracted and placed in contrast and comparison to other *datum*. This point is humorously illustrated by Douglas Adams, in his novel *The hitchhiker's guide to the galaxy*, in which a computer, when asked about the "meaning of life, the universe, and everything," responds with "42," rendering its answer useless due to the lack of context. Forming data by gathering increasing quantities of *datum* enriches the number and variety of connections from which valuable insights can be derived (Morrell, 2021). Data is thus considered relational in nature, by definition, and yields the greatest value when insights can be derived about the entire community that it represents.

Data is understood as owning its very existence to its collector, who decides what questions to investigate, which pieces of information to gather, how to label and classify retained data points, and

[1] Director of the Division for Digital Inclusion, Policies and Transformation in the Communications and Information Sector at the United Nations Educational, Scientific and Cultural Organization (UNESCO). She was the Director of UNESCO Beijing (2015-2020) and previously served as Global Results Manager at the United Nations Development Programme (UNDP), where she also managed a portfolio of Latin American countries (2001-2015). She holds a Ph.D. in Business Administration (1995) and a master's degree in Finance (1990) from the University of Illinois, in the United States.

[2] Search performed using Google tool on September 8, 2023.

Photo: UNESCO archive

Marielza Oliveira
Director of the Division for Digital Inclusion, Policies and Transformation in the Communications and Information Sector at the United Nations Educational, Scientific and Cultural Organization (UNESCO).

how to use resulting datasets. The corollary to this view that value is created by accumulating, organizing, and comparing pieces of information is that the value derived from data should, therefore, rightfully accrue entirely to those performing these tasks. Individual *datum* merits no consideration in this framework, beyond being (as the translation of this Latin noun indicates) "a thing that is given" to be transmuted into data, and from there into knowledge.

Under this perspective, *datum* is assumed to be both freely available and intrinsically worthless, and it is the very act of capturing and imprisoning as many of them as possible into (aptly named) table "cells" that form valuable data. The very word "in-formation" alludes to the importance attributed to this organizing process. For instance, Wikipedia defines data economy as the digital ecosystem in which so-called raw data is "gathered, organized, and exchanged by a network of companies, individuals, and institutions to create economic value" and earn a reward. No reference is made to any tasks required to create *datum*, and no consideration is given to compensating *datum* creators for their labor or for any harm they may suffer in the process of extraction and organization.

From the humble *datum* to Big Data

These views on digital *datum* and data are derived from how early computerized database management systems operated. Starting in the late 1960s, information technologies enabled the quick storage and retrieval of data from databases and the localization and exchange of specific *datum* in large data sets. Such databases were pre-designed – with clear specifications in terms of tables, columns, indexes, and other parameters – to contain and rapidly process well-structured data, generated and collected according to expressed research methods and protocols, and in adherence to ethical standards requiring neutrality, no bias, and no harm to *datum* providers.

The advent of faster, much more powerful, and increasingly interconnected devices – desktop computers, tablets, mobile phones, smart appliances, sensors, smart TVs, fitness trackers, street cameras, smart wearables, cars, and others – catalyzed a massive surge in the amounts and varieties of information available in digital format. According to Statista (2021), the total amount of data created globally reached 64.2 zettabytes by 2020 and is projected to grow to more than 180 zettabytes by 2025.

Storage capabilities are also growing. For instance, the installed base of storage capacity reached 6.7 zettabytes in 2020 and has been growing at a compound annual growth rate (CAGR) of 19.2% since then. Market value grew at a 5.5% CAGR between 2017 and 2021, reaching US$101 billion in 2022, and is projected to grow at a 7.5% CAGR in the 2022-2032 period (Future Marketing Insights, 2022).

The vast majority of such data is unstructured, consisting of the digital traces – text, audio, image, video – left by human interactions in social media, e-commerce, the shared economy, public service sites, and other Internet platforms. People's *datum* – exchanges with friends, searching and browsing patterns, online shopping carts, "likes" and rants in social media, family pictures,

geolocation, faces and palms, heartbeats, gaits and voices, and other intimate and revealing details – are being collected, including through sophisticated surveillance technologies, into Big Data sets for real-time processing. A significant proportion of *datum,* including click rates, Internet Protocol (IP)-specific location data, and search logs, is considered as the mere "exhaust" of online activity (Snyder & Castrounis, 2018), which would be wasted if not collected and processed for economic gain.

The advent of Big Data has appended these views on *datum* and data.

MUCH MORE IS EVEN BETTER

Recent evolutions in predictive analytics and machine learning have rendered very large datasets capable of not only yielding insights for decision-making but also enabling the creation of innovative technologies such as Artificial Intelligence (AI). Users may not even know, at the time of data collection, all the ways they may be able to use the collected data in the future. Businesses and governments are increasingly dependent on Big Data for insights that enable rapid responses to changing socioeconomic, political, cultural, and environmental conditions, turning the companies that collect and process the most Big Data into the highest valued ones. Seven out of the ten largest companies in the world, as measured by market capitalization, are key players in the data economy (Statista, 2023b). Therefore, the multipurpose potential of semi-structured and unstructured Big Data makes it more valuable than structured data, as reflected in the exponential growth of the global Big Data analytics market value: It reached US$272 billion in 2022, is projected to reach US$308 billion by end-2023, and to surpass US$745 billion by 2030 (Fortune Business Insights, 2022).

The same logic underpins decisions made by generative AI developers. The Big Data sets on which generative AI is built originate from Internet scraping, aiming to collect as much of its content as possible (Washington Post, 2023). With exceptions, so far, an obvious trend in this technology niche is the increase in the size of datasets used for training foundation models, in the expectation that adding parameters boosts model versatility (i.e., enables fine-tuning for additional tasks) as well as the "emergence" of new properties and capabilities. Bidirectional Encoder Representations from Transformers (BERT), a foundation model released in 2018, had 340 million parameters trained on a dataset of 3.3 billion tokens within a 16 GB dataset. GPT-4, launched in 2023, is rumored to possess 1.76 trillion parameters trained on trillions of tokens within a 45 GB dataset (Amazon, n.d.).

The technologies and economic incentives are, therefore, in place to accelerate datafication; thus societies must intensify efforts to govern this process for societal gain.

DATUM HAS INTRINSIC MEANING

The simplest definition of Big Data characterizes it as possessing "3 Vs": Volume (size as measured by the number of records in the dataset), velocity (the rate at which new information is added), and variety of sources and types of dataset content (IT Chronicles, n.d.). These huge datasets are filled with

Businesses and governments are increasingly dependent on Big Data for insights that enable rapid responses to changing socioeconomic, political, cultural, and environmental conditions, turning the companies that collect and process the most Big Data into the highest valued ones.

Personal *datum* is so closely tied to the notion of selfhood as to be almost indistinguishable from it.

units of content purposefully created to convey specific ideas. For instance, a recent Washington Post article revealed the contents of Google's *Colossal Clean Crawled Corpus* (C4) dataset (Goodwin, 2023), used to train generative AI foundation models such as Google's T5 and Meta's LLaMA. It contains scrapings from 15.7 million websites, particularly text collected from those dedicated to journalism, sciences, academia, marketing, patents, and others, with material sourced from Wikipedia, Coursera, major newspapers (such as New York Times and Washington Post), personal blogs, posts from social media (such as Reddit and X, formerly known as Twitter), pirated books and unpublished novels, and millions of other pieces of content created with the express intent of delivering signifying information to their users. Differently from how *datum* within a structured dataset is interpreted, each *datum* in such datasets requires no association with other *datum* for intrinsic significance and to convey information. This changes the value that should be attributed to *datum*, including to its creation, and how such value should be apportioned.

PROTECTING *DATUM* IS AS CRITICAL AS PROTECTING DATA

The United Nations Department of Economic and Social Affairs (UN DESA) states that "The value chain in the data economy begins with collecting personal and non-personal data and making them available for storage and eventual analysis" (UN DESA, 2019, p. 2). Personal *datum* such as name, identification number, geolocation, email or IP address, cookie identification, biometrics, and other *datum* that can identify a person, by itself or in combination with other data, is considered the most valuable type, as it enables companies and governments to identify and target specific individuals with customized digital solutions. For the same reason, it is also the most sensitive, being afforded special protections under general and sector-specific (such as in the health sector) data privacy norms. Personal *datum* is so closely tied to the notion of selfhood as to be almost indistinguishable from it.

Article 17 of the International Covenant on Civil and Political Rights (United Nations, 1976) enshrined the right to privacy into international law. United Nations Conference on Trade and Development (UNCTAD) research shows that 137 out of 194 countries have enacted legislation protecting data and privacy, and another 17 countries have draft legislation under consideration (UNCTAD, 2021). In addition, given the pervasiveness of cross-border data flows, international and regional bodies such as the Organisation for Economic Co-operation and Development (OECD) and the European Union have issued supranational guidelines and other normative frameworks to offer the necessary privacy protections for personal data, such as the influential General Data Protection Regulation (GDPR).

The World Economic Forum (WEF) recognizes that information technologies pose a challenge to privacy, even though people generally understand why the tracking and sharing of their information is essential to materialize the benefits of connectivity (Schwab, 2016). This challenge has made the market for data anonymization, masking, and security a fast-growing one, with a predicted CAGR greater than 13% in the 2021-2028 period (Data Bridge Market

Research, 2021). Nevertheless, with so much data collected about Internet users, anxieties keep rising about protection of personal data against exploitation, breach, and exposure. A recent survey by PrivacyHawk showed that 45% of United States (U.S.) Internet users are extremely concerned with their online privacy, and 40% are actively using tools to obscure their digital traces and protect their identity (Business Wire, 2023).

These anxieties are well-founded. The International Standardization Organization (ISO) defines data anonymization as the "process by which personal data is altered in such a way that a data subject can no longer be identified directly or indirectly, either by the data controller alone or in collaboration with any other party" (ISO, 2017, p. 7). Permanent anonymization is, however, a very difficult task. Rocher et al. (2019) have develop a method to calculate the likelihood of re-identification of individuals whose data is contained in datasets deemed as having been anonymized and found that it is generally quite high. Further advancements in digital technologies and data collection mechanisms are likely to exacerbate privacy risks. Their work indicates, therefore, that current anonymization methods are insufficient for the challenge of technically protecting privacy, making it more urgent to strengthen its legal protections.

European norms around data protection and privacy are grounded in Human Rights, a normative framework that emphasizes the protection of individuals' dignity and worth, and thus focus on the protection of personal *datum*. American norms, on the other hand, are grounded in commercial law and focus on regulating how organizations keep and use consumer data. A holistic framework is needed to abridge these perspectives, which considers how to simultaneously value and protect both *datum* and data.

The stakes are higher with Big Data, which, in addition to aiding or even automating decisions, also serves to train AI models. For instance, generative AI – encompassing Large Language Models (LLM) and Denoising Diffusion Probabilistic Models (DDPM) – utilizes deep learning methods to identify underlying patterns in Big Data sets based on a probability distribution and, when prompted by users, generate outputs that recreate the original data distribution. As generative AI models are trained on vast quantities of data, it is difficult to ascertain whether the *datum* included in the dataset, or the generated outputs, are in violation of copyright or privacy laws. UNESCO's Recommendation on the Ethics of Artificial Intelligence (2021), adopted by all its Member States, specifically includes the right to privacy and data protection among the principles for development and use of AI technologies. The Recommendation establishes that data protection frameworks and governance mechanisms should be carried out with a multistakeholder approach and under the protection of judicial systems, enabling data subjects to fully exercise their rights in regard to personal data throughout the AI lifecycle.

Nevertheless, with so much data collected about Internet users, anxieties keep rising about protection of personal data against exploitation, breach, and exposure.

RELATIONSHIPS BETWEEN *DATUM* IN BIG DATASETS CREATE HIGHER VALUE, BUT ALSO HIGHER CONCERNS

Big Data analytics are designed to generate group insights that can be applied to all individuals who share the same characteristics, behavior patterns,

Huge amounts of unstructured data are more useful than structured data for the purpose of finding similarities among the members of a population, since the characteristics to apply for grouping individuals can be unlimitedly and flexibly determined (...).

or preferences. Huge amounts of unstructured data are more useful than structured data for the purpose of finding similarities among the members of a population, since the characteristics to apply for grouping individuals can be unlimitedly and flexibly determined through analytics, rather than being somewhat limited by population sampling.

Group identity and relationships reveal information about individuals that they may not have given directly, which, in turn, facilitates decision-making about them. This capability of Big Data has prompted Kitchin (2014) to include "indexical in identification" and "relational in nature" among its defining characteristics; and Viljoen (2021) to assert that "This relational aspect of data production drives much of the social value and harm of data collection and use in a digital economy" (p. 573). The distillation of insights on in-group individuals has reached such an extent that a recent "leak" revealed 650 thousand ways in which people are segmented and labeled by AdTech companies (Keegan & Eastwood, 2023).

Early references to Big Data described the collection and processing of people's behavioral, transactional, and demographic data in social media and e-commerce platforms, with the intention of deriving insights from which to predict or influence the behavior of that specific user, but, more importantly, of others with similar characteristics and/or preferences. Valuable services could be developed, such as recommender systems that help customers to find interesting products through offerings such as "customers who bought this item also bought...," and fraud detection algorithms that reduce financial services risk by spotting anomalous patterns that deviate from the activities of typical users. However, so could harmful systems, including algorithmic-driven distribution of disinformation that target people looking for innocuous social media content; and predictive policing systems flagging innocent people as potential criminals on the basis of characteristics shared with those convicted of the relevant crimes. Since Big Data enables the distillation of population-level insights from how people relate to one another, individuals have an interest in *datum* about them and how it was obtained. Thus, supra-individual legal and economic interests should also be considered in ethical and legal normative frameworks about data.

Concerns over the datafication of human experiences and the harms caused by data extraction, analysis, and use are rising. Indiscriminate collection of Internet content brings not only information, but also harmful content such as misinformation, disinformation, and hate speech into datasets. Digital divides in access to Internet infrastructure and content generate imbalanced Big Data in terms of representation of various social groups, which in turn lead to biased and discriminatory algorithms, services, and other outputs. To minimize the potential for harms, it is essential that ethical standards be applied throughout the entire data value chain, from its creation to collection, labelling, analysis, use, sharing, and disposal.

ALTHOUGH *DATUM* HAS ECONOMIC VALUE, ITS CREATORS DO NOT EARN A FAIR SHARE

The advent of generative AI has evidenced the distinction between *datum* and data creators, and Big Data set collectors. That this is problematic is becoming increasingly evident as news continue to emerge about the distress felt by artists, scientists, journalists, and other content creators, by having their intellectual efforts harvested into colossal datasets built on any data publicly available, often without their meaningful consent or even awareness, with the justification that this creates socioeconomic value when such datasets are used for innovation. Underpayment for creative work is also rampant; for instance, performers recently reported being asked to scan their bodies and voices, to generate digital replicas owned by others who can then use them "for eternity" for no additional payment beyond the 15 minutes of labor performed while scanning (Allyn, 2023).

Heated exchanges have taken place across Internet platforms, with AI developers claiming that it is fair (use?) that creators to give away their *datum* "for the benefit of humanity." In response, creators have pointed out the hypocrisy of AI developers for asking them to give away their creations for free while expecting to profit from the resulting AI models. A number of class action lawsuits have been filed over the use of copyrighted materials found within generative AI training datasets. The Terms of Use of digital news, scientific journals, and social media platforms have been updated to prohibit scraping, and many such websites, which until recently were open, have instituted paywalls to limit access to their content, reducing access to information particularly for people unable to afford these additional costs.

Datum, including personal *datum*, is increasingly viewed as assets over which creators should be able to exercise privacy, property, and other rights. Arrieta-Ibarra et al. (2018) call attention to the neglected role of users in creating digital data, proposing this data task should be recognized as labor so that people may be adequately compensated for it. However, they warn that this approach "may run against the near-term interests of dominant data monopsonists who have benefited from data being treated as 'free'" (p. 38).

At the heart of conflicts between *datum* creators and users are contrasting conceptualizations of value: Intrinsic versus instrumental, attributed to data without reference to *datum*, value-in-exchange versus value-in-use. These elements must be reconciled.

At the heart of conflicts between *datum* creators and users are contrasting conceptualizations of value: Intrinsic versus instrumental, attributed to data without reference to *datum*, value-in--exchange versus value-in-use. These elements must be reconciled.

From Big Data to economic value

The study of value creation and appropriation is the purview of Economics, which examines how a society uses scarce resources for the production, allocation, and consumption of goods and services. A central concept is the production function, which describes how economic inputs, called factors of production, can be combined to generate a given quantity of specific outputs. The classical production function contemplates assets such as equipment ("capital"), work

done by humans ("labor"), and natural resources (generically represented as "land"), as its factors. Economic agents who provide these factors to the production process are remunerated respectively through profits, wages, and rent, while companies employ the production function to determine how much output they should produce at each given price, and what combination of inputs they should use given their costs and availability.

Production functions depend on the technology used. Industrial revolutions represent major downward shifts in the proportion of labor needed to produce a given quantity of output. Steam engines augmented human muscle with machines, and thus set in motion the first Industrial Revolution; electricity and conveyor belts powered the second one; and automation that removed physical labor altogether brought about the third. Machines capable of learning from vast amounts of data have extended automation to cognitive labor, unleashing the fourth Industrial Revolution.

The data-as-oil analogy enshrines data, particularly Big Data, as the new source of power that moves the global economic engine to create prosperity. It appears in production functions in two ways: As a production factor, when economic agents use knowledge obtained through data processing and analysis to improve decision-making and behavior; or as an output, when economic agents treat data as the main valuable, tradable asset in the data economy.

Thus, data is recognized not only as a new factor of production (including in formal national economic planning, such as done by China since 2020) but actually as the most important one in the Information Age, being as essential to the digital economy as oil is to the industrial economy. According to the World Bank, throughout the last decade the digital economy has annually contributed over 15% of global Gross Domestic Product (GDP), while also growing 2.5 times faster than the physical economy (Hayat, 2022). Physical assets no longer comprise the bulk of business value: Intangibles, including data, already account for 90% of the total market value of the top-500 companies in the world. Yet, discussions of data in Economics are still relatively incipient.

Machines capable of learning from vast amounts of data have extended automation to cognitive labor, unleashing the fourth Industrial Revolution.

One reason is that data, unlike oil and other physical goods, is not a depletable resource, making it difficult to determine its economic value, and even harder to ascertain who should be compensated for any value created from it. Data belongs to a category of goods called "non-rival," i.e., goods that can be simultaneously stored, shared and (re)used by multiple users, without reducing either their quality or the quantity available to others. In economic terms, data has an opportunity cost equal to zero, i.e., economic agents do not have to choose between alternative uses: All the data available can be employed in all productive opportunities at once. In fact, the more uses one finds for the same data, the faster the costs of producing and storing a dataset are defrayed, leading to lower costs of producing each good or service for which that data is a production factor. Additionally, data generates scale effects: The more data creators that are in the economy, the more data is available from which a data user can derive insights, which raises that user's own productivity. But every user benefits similarly, leading to an overall increase in economic efficiency and productivity. Societies thus benefit from the widest possible use of data, limited only by its availability and quality.

Availability of Big Data depends on people being able to safely create and exchange information online, which in turn depends on several elements: People's access to affordable and reliable Internet connections and digital devices; how well Internet platforms are able to accommodate creators with disabilities or those who use different languages; and people's skills in seeking and imparting information over digital ecosystems. The creation of high-quality Big Data, on the other hand, depends on the degree to which Human Rights (including privacy, freedoms of expression and association, access to information, non-discrimination, and others) are protected on the Internet. When people feel safe online, when they are able to identify and disregard harmful content, when there are no barriers (such as walled gardens or fragmentation) to wide information exchanges, when they trust information ecosystems to safeguard their *datum* and rights, the quality of data generated through these exchanges increases.

The elements of information creation and exchange are reflected in UNESCO's Internet Universality framework, launched in 2018, which established five key principles for Internet development summarized in the acronym "ROAM-X": The internet should be human Rights-based, Open to all, Accessible by all, governed through Multistakeholder participation, and uphold cross-cutting elements such as online safety and gender equality (UNESCO, 2018). The ROAM principles were implemented in more than 40 countries, to diagnose potential improvements to digital ecosystems that can lead to inclusive access to information for all, and the building of true knowledge societies (UNESCO, n.d. a).

Availability further depends on the data's degree of openness, i.e., on who controls it, who is legally entitled to use it, and under which conditions. Technical standards for data interoperability are also important, since quite a lot of available data remains unused simply because it is held in formats and structures that digital systems are unable to process, including in outdated legacy formats such as paper, tapes, and CD-ROM. Finally, data must also be easy to find, i.e., it must be "discoverable" by those wishing to collect, analyze and use it. These requirements are neatly encapsulated in the FAIR guiding principles for scientific data (GO FAIR, n.d.) – Findability, Accessibility, Interoperability, and Reuse.

Further unleashing the power of data to maximize socioeconomic well-being requires economic models and data governance regimes that equitably empower and reward all *datum* and data creators, users, and societies, avoiding privatization or hoarding of data in ways that are socioeconomically inefficient.

From economic value to societal value

People, businesses, and governments are using data to reduce search and transaction costs and derive insights on which to make informed decisions. Data is making societies more efficient and economies more productive, while also improving the efficacy of public policy, delivery of public services, transparency, and accountability. Further unleashing the power of data to maximize socioeconomic well-being requires economic models and data governance regimes that equitably empower and reward all *datum* and data creators, users, and societies, avoiding privatization or hoarding of data in ways that are socioeconomically inefficient.

When data is captured and privatized, asymmetries arise in the capacity to derive knowledge and to innovate, increasing inequality. Thus, a key issue for modern economies is how the rights to own, use, or profit from data are defined.

In the digital age, increasing access to opportunities of fair participation in data value chains is key to promoting socioeconomic development.

In his treatise on *The problem of social costs* (1960), economist Ronald Coase proposes that potential users of a non-rival good should pool their resources, thus augmenting the quantity available to all, distributing the costs of nurturing its quality, and facilitating findability. UNESCO promotes open data (UNESCO, n.d. b), particularly of research and government data, so that data can play its role in improving quality of life for all.

FOSTER A DATA CULTURE BY EMPOWERING *DATUM* CREATORS, DATA PRODUCERS AND USERS

Breaking down data silos and making data more widely accessible enables it to permeate the entire economy and yield higher socioeconomic gains. A new and important role for governments is to nurture data pools by fostering national open government/open data platforms that address national and local challenges, as well as through international cooperation for transborder pooling of data resources that can address global challenges. Governments can also incentivize the participation of private sector in sectoral data pools through which proprietary data is shared, increasing sectoral productivity and opportunities for innovation for all market participants, as well as leveling the playing field for startups. Last but not least, governments can empower people with control over their own *datum* to become active agents in the data economy, by fostering licensing mechanisms such as data commons for social causes, and data trusts for data commercialization. Appropriate mechanisms must also be enacted to curb both *datum* and data free riding, which occurs when users of a data pool are able to avoid paying or to under-pay for that resource, leading to lower quantity and/or quality from which all data users suffer.

ELIMINATE DIGITAL DATA INEQUALITIES THAT INHIBIT PARTICIPATION IN THE DATA ECONOMY

In the digital age, increasing access to opportunities of fair participation in data value chains is key to promoting socioeconomic development. At the national level, greater efforts must be made to expand meaningful connectivity; develop digital skills and competencies among people, governmental institutions, and businesses; and to properly regulate both privacy and property rights to *datum* and to data. At the international level, efforts should be intensified to support developing countries in more effective usage of data to fight poverty and improve well-being, including sharing good practices in regulatory mechanisms to recompense data creation and to protect personal data and privacy, enacted in only 48% of least developed countries (Chakraborty, 2022).

INCENTIVIZE UPDATES TO LEGACY DATA SYSTEMS

A significant amount of data remains in legacy formats such as paper, film, tape, and other analog media. Digitizing legacy archives ensures access to information that could otherwise be lost due to deterioration, natural disasters, conflict, and other causes. Improving the technical infrastructure increases the ability of economic agents to create, collect, store, and analyze fast-growing

data sets in a way that is secure and cost effective. Unlocking the value of data also requires updating legal arrangements to protect data rights, facilitate data transactions, and provide mechanisms for dispute resolution.

PROMOTE THE GREENING OF DIGITAL DATA VALUE CHAINS

Similar to oil drilling, data "mining" and processing cause damage to the natural environment and exploit finite natural resources. Economics has so far treated such negative impacts as externalities, but new approaches are emerging on how to account for environmental impacts. As data centers rapidly become some of the most voracious consumers of natural resources, consuming water, energy, and rare minerals required for chip production; and old digital devices clogging landfills and leaking toxic materials, it is imperative that societies account for the environmental impacts of data systems, including through necessary regulatory measures.

EXTEND NATIONAL ACCOUNTING STANDARDS AND APPROACHES

Economic models focus on households and individuals as consumers or providers of labor and fail to acknowledge them as producers who use their own labor and capital. Digital technologies are making it more urgent to examine the ways in which labor usage is shifting due to automation, with many tasks previously carried out by businesses (e.g., self-service supermarket checkouts, automatic bank tellers, and Internet shopping) are being redirected to individuals and households, and from governments to people (via digital public services). Conversely, some tasks are moving from individuals to businesses (for example, via outsourcing of household food production to digital platforms, such as iFood). In this sense, economists need better tools for measuring and modeling the value-creating impact of the data economy, including individual *datum* production factors, intermediate inputs, and final outputs. Excellent household surveys such as the ICT Households, conducted by the Regional Center for Studies of the Development of the Information Society (Cetic.br), department of the Brazilian Network Information Center (NIC.br), which can be a model source about both the equipment (devices and infrastructure) used and the labor employed in *datum* creation (NIC.br, n.d.). Global estimates, such as the one performed by Statista, can also contribute to the analysis: For instance, their most recent research indicates that in the fourth quarter of 2022, the average time spent on the internet per person was 395 minutes (six hours and 35 minutes) per day (Oberlo, 2023), with social media accounting for 151 minutes per day (Statista, 2023a). This could also enhance official labor statistics, which ignore the contributions of the household economy to the market economy and underrepresent gig work and other tasks performed in the data economy.

(...) economists need better tools for measuring and modeling the value-creating impact of the data economy, including individual *datum* production factors, intermediate inputs, and final outputs.

References

Allyn, B. (2023, August). Movie extras worry they'll be replaced by AI. Hollywood is already doing body scans. *NPR*. https://www.npr.org/2023/08/02/1190605685/movie-extras-worry-theyll-be-replaced-by-ai-hollywood-is-already-doing-body-scan

Amazon. (n.d). *What are Foundation Models?* https://aws.amazon.com/pt/what-is/foundation-models/

Arrieta-Ibarra, I., Goff, L., Jiménez-Hernández, D., Lanier, J., & Glen Weyl, E. (2018). Should we treat data as labor? Moving beyond "free". *American Economic Association Papers and Proceedings*, *108*, 38-42.

Brazilian Network Information Center. (n.d.). *Survey on the use of information and communication technologies in Brazilian households*. https://cetic.br/en/pesquisa/domicilios/

Business Wire. (2023). *PrivacyHawk releases 2023 personal data, privacy and AI report highlighting consumer alarm and sentiment about privacy and AI*. https://www.businesswire.com/news/home/20230822244117/en/PrivacyHawk-releases-2023-personal-data-privacy-and-AI-report-highlighting-consumer-alarm-and-sentiment-about-privacy-and-AI

Chakraborty, T. (2022). Is data the new oil? *Yubi*. https://www.go-yubi.com/blog/data-new-oil/

Data Bridge Market Research. (2021). *Global data masking market – Industry trends and forecast to 2028*. https://www.databridgemarketresearch.com/reports/global-data-masking-market

Fortune Business Insights. (2022). *Market research report*. https://www.fortunebusinessinsights.com/big-data-analytics-market-106179

Future Marketing Insights. (2022). *Data center market report*. https://www.futuremarketinsights.com/reports/data-center-market

GO FAIR. (n.d.). *FAIR Principles*. https://www.go-fair.org/fair-principles/

Goodwin, D. (2023). Search the 15.7 million websites in Google's C4 dataset. *Search Engine Land*. https://searchengineland.com/search-websites-google-c4-dataset-395820

Hayat, Z. (2022, August). Digital trust: How to unleash the trillion-dollar opportunity for our global economy. *World Economic Forum*. https://www.weforum.org/agenda/2022/08/digital-trust-how-to-unleash-the-trillion-dollar-opportunity-for-our-global-economy

International Organization for Standardization. (2017). *ISO 25237:2017 Health informatics: Pseudonymization*. https://www.iso.org/standard/63553.html

IT Chronicles. (n.d.). *What is Big Data*. https://itchronicles.com/what-is-big-data/

Keegan, J., & Eastwood, J. (2023, June). From "heavy purchasers" of pregnancy tests to the depression-prone: We found 650,000 ways advertisers label you. *The Markup*. https://themarkup.org/privacy/2023/06/08/from-heavy-purchasers-of-pregnancy-tests-to-the-depression-prone-we-found-650000-ways-advertisers-label-you

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1). https://doi.org/10.1177/2053951714528481

Morrell, J. (2021). Does more data equal better analytics? *Datameer*. https://www.datameer.com/blog/does-more-data-equal-better-analytics/

Oberlo. (2023). *How much time do people spend online?* https://www.oberlo.com/statistics/how-much-time-does-the-average-person-spend-on-the-internet

Rocher, L., Hendrickx, J. M., &. Montjoye, Y. A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications, 10*, 3069. https://doi.org/10.1038/s41467-019-10933-3

Schwab, K. (2016, January). The Fourth Industrial Revolution: What it means, how to respond. *World Economic Forum*. https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/

Snyder, S., & Castrounis, A. (2018). How to turn 'data exhaust' into a competitive edge. *Knowledge at Wharton*. https://knowledge.wharton.upenn.edu/article/turn-iot-data-exhaust-next-competitive-advantage/

Statista. (2021). *Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025*. https://www.statista.com/statistics/871513/worldwide-data-created/

Statista. (2023a). *Daily time spent on social networking by Internet users worldwide from 2012 to 2023*. https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/

Statista. (2023b). *The 100 largest companies in the world by market capitalization in 2023*. https://www.statista.com/statistics/263264/top-companies-in-the-world-by-market-capitalization/

United Nations. (1976). *International Covenant on Civil and Political Rights*.

United Nations Conference on Trade and Development. (2021). *Data Protection and Privacy Legislation Worldwide*. https://unctad.org/page/data-protection-and-privacy-legislation-worldwide

United Nations Department of Economic and Social Affairs. (2019). Data Economy: Radical transformation or dystopia? *Frontier Technology Quarterly*. https://www.un.org/development/desa/dpad/wp-content/uploads/sites/45/publication/FTQ_1_Jan_2019.pdf

United Nations Educational, Scientific and Cultural Organization. (n.d. a). *Internet Universality Indicators*. https://www.unesco.org/en/internet-universality-indicators

United Nations Educational, Scientific and Cultural Organization. (n.d. b). *Open Data*. https://www.unesco.org/en/open-solutions/open-data

United Nations Educational, Scientific and Cultural Organization. (2018). *UNESCO's Internet Universality Indicators: A framework for assessing Internet development – draft for the consideration of the Intergovernmental Council of IPDC*. https://unesdoc.unesco.org/ark:/48223/pf0000265830

Viljoen, S. (2021). A relational theory of data governance. *Yale Law*. https://heinonline.org/HOL/LandingPage?handle=hein.journals/ylr131&div=12&id=&page=

Washington Post. (2023). *Inside the secret list of websites that make AI like ChatGPT*. https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/

Interview I

Photo: Personal archive

Marcos Dantas
Full professor at the Federal University of Rio de Janeiro (UFRJ) and counselor of the Brazilian Internet Steering Committee (CGI.br).

The data economy in Brazil

Marcos Dantas is a full professor at the School of Communication at the Federal University of Rio de Janeiro (UFRJ). He holds a Ph.D. in Production Engineering from COPPE-UFRJ and is a professor in the postgraduate programs in Communication and Culture (PPGCOM-ECO) and Information Science (PPGCI-ECO/IBICT-UFRJ). He is a member of the Board of Directors of the Brazilian Network Information Center (NIC.br) and the Brazilian Internet Steering Committee (CGI.br). He co-authored (with D. Moura, L. Ormay, and G. Raulino) the book *O valor da informação: de como o capital se apropria do trabalho social na era do espetáculo e da internet* (Boitempo, 2022). In this interview, he discusses the importance of the "data economy" in the current context and possible ways of distributing the value obtained from data among the society that produces it. Finally, he addresses the challenges of measuring the data economy in Brazil.

Internet Sectoral Overview (I.S.O.)_ How important is "data economy" in the context of the digital transformation we are currently experiencing?

Marcos Dantas (M.D.)_ A 2017 edition of *The Economist* magazine stated, "data is the oil of the 21st century". We know that the economy of the 20th century was fueled by oil. We cannot do anything without oil: Not just as a source of energy but it is present in almost every type of utensil we use in our daily lives. In addition to being economically decisive – for this very reason – oil has also been and still is the reason for wars, coups d'état, and even assassinations of political leaders, in other words, oil also means power and the struggle for power. In this sense, as a basic economic resource and an instrument of political and geopolitical power, we can admit that the analogy made by the magazine is valid. In the 21st century, the economy and the political struggle will revolve around data. You only have to look at the size of some corporations, such as Alphabet or Amazon, to confirm this economic dimension. One need only consider the revelations made by Edward Snowden and Julian Assange, or the electoral activities carried out by Cambridge Analytica, to get a glimpse – and I stress "a glimpse" – of how access to and manipulation of data can already be replacing the old methods of violent coups and assassinations that were common in the days of oil hegemony. However, despite its decisive importance, we still know little about this new economy. Even the definition of "data" is in dispute. This ignorance contributes greatly to the fact that we still do not have a solid public policy to promote and regulate the data economy, unlike the one we have had for many decades in Brazil and worldwide in relation to the oil economy and energy resources in general.

I.S.O._ Considering that the data economy has digital platforms as an important part of its structure, what are the possible ways to rethink the monetization of user-generated data?

M.D._ Social digital platforms have invented a business model based on offering their users' data to advertisers through auctions. Users naively engage in intense activity, providing this data for free, allowing the platforms to profit from it. This data comprises records, in electro-electronic format, of users' social activities. Similar to how the content of a book represents a writer's activity in ink and paper form (an activity inherently social, as no writer is a Robinson Crusoe), data is produced by billions of "writers," so to speak, whose "readers" are the advertisers. As data is the product of a social activity, since it always originates from interpersonal interactions or the social needs of individuals, it is first and foremost a social resource – like oil or water. If we have reached a point in history where digitized social data has become an economic resource, the first point to address would be how to make this wealth generate benefits for society as a whole. For the time being, it only benefits a small group of major shareholders of those platforms.
Nonetheless, within the production process, or to be straightforward, in individual and collective labor, the true producers work for free. A portion of the revenue pays for the work of scientists, engineers, technicians, and platform employees, who develop the systems and algorithms for data mining. In certain business models, some of the revenue may also reward the most successful data producers, the "influencers." However, they are successful precisely because of their ability to attract the unpaid labor of millions, or even billions, of data providers. These influencers assist (and it is a substantial assistance!) platforms to mine the socially available data concerning the individuals' bodies and actions.
Within a capitalist economy, socio-digital platforms assume an essential role: They eliminate spatial constraints on time, thereby exponentially increasing the turnover of capital. The shorter the period required to make an investment (the more times the same capital can be reinvested), the greater the profit for any given producer. Therefore, once these platforms have emerged, a capitalist economy will no longer be able to live without them. And the rest of us will have to learn to live with it. The question is *how*? How can we ensure that the practically incalculable value contained in the social data they appropriate can also be shared with the society that generates it? This is how I understand the question.
Perhaps the most obvious answer is to compensate data producers. Meta reports an average revenue of US$40 per user. Users use Meta's platforms at no cost but generate US$40 for Meta. In other words, the time this user has spent in Meta's service, typing their mobile phone screen or the computer mouse, is valued at US$40 for Meta. Why not remunerate users for

"If we have reached a point in history where digitized social data has become an economic resource, the first point to address would be how to make this wealth generate benefits for society as a whole."

"In theory, measuring data should not be complex, as we are talking about physical quantities. (...) However, the public authorities don't have the tools to control the data flow, as they can, for example, monitor the flow of oil leaving or entering the country."

their time? Why not share this revenue with them? Why do only Jeff Bezos, Elon Musk, Black Rock Inc., Vanguard Group Inc., State Street, and the like continue to accumulate profits from such a wealth?

I.S.O._ In your opinion, what are the challenges associated with measuring the data economy in Brazil?

M.D._ The challenges are the same in Brazil and worldwide. Perhaps in Brazil, like in other countries within the capitalist periphery, these common challenges are added to those specific to our subordinate condition and our great social inequalities. In theory, measuring data should not be complex, as we are talking about physical quantities. As I write, I am aware that I have 245 gigabytes of data registered on my computer's hard disk. The problem is that society is really unaware of the data amount that circulates and is recorded on the servers of major platforms. We may even have a broad sense of the overall data traffic volume. The Brazilian Network Information Center (NIC.br), for example, provides us with this statistic for Brazil. However, the public authorities don't have the tools to control the data flow, as they can, for example, monitor the flow of oil leaving or entering the country.
The Central Bank and the Internal Revenue Service can identify suspicious financial transactions because banks are obliged to report all their financial transactions to the public authorities. Nevertheless, neither in Brazil nor around the world do we have public authorities with powers and means to control the data flow, a resource which, as we have seen, already deserves to be treated like oil. To the extent that we can better understand this economy, there is hope that Brazilian society will eventually realize that "data is ours," much in the same way that, decades ago, our parents' (or grandparents') generation understood that "the oil is ours!"

Article II

Policy options for the data economy: A literature review[3]

By Ramy El-Dardiry,[4] Milena Dinkova,[5] and Bastiaan Overvest[6]

Introduction

In this article, we discuss the literature on the economics of data. The rising importance of data in our economy and society has prompted more research into these topics in recent years. Data and the resulting digitalization of our society present tremendous opportunities and challenges. On the one hand, digitalization might introduce and sustain a new period of economic growth and help overcome societal challenges, for example, by enabling personalized education or preventive medicine. On the other hand, there are uncertainties about the future of privacy and our democracy and worries about the power of a handful of technology firms. Thereby, new research provides a better understanding of how to think about these opportunities and challenges.

Given the wide and profound impact of digitalization, it is perhaps not surprising that research on data involves many disciplines. Economists have recently begun to better understand how data function as a factor of production. Legal scholars have leveraged concepts from economics to study externalities associated with data and privacy. Philosophers, sociologists, and political scientists have been concerned with the new power balance that emerges from the digital era. Meanwhile, computer scientists are constantly inventing new ways to better safeguard privacy or to enable exchanges of data.

In this article, we define and characterize what is meant with data. We also review the economic properties of data, such as non-rivalry and low replication costs, and the estimates for the value of data in our economy.

On the one hand, digitalization might introduce and sustain a new period of economic growth and help overcome societal challenges (...). On the other hand, there are uncertainties about the future of privacy and our democracy and worries about the power of a handful of technology firms.

[3] The edited version of the homonymous work published by the CPB Netherlands Bureau for Economic Policy Analysis. Available at: https://www.cpb.nl/en/brave-new-data-policy-pathways-for-the-data-economy-in-an-imperfect-world#docid-160570

[4] Netherlands Bureau for Economic Policy Analysis.

[5] Netherlands Bureau for Economic Policy Analysis.

[6] Netherlands Bureau for Economic Policy Analysis.

Data differ from ideas: Both are forms of information, but they serve different purposes. (...) that ideas are pieces of information that provide instructions on how to create output (...). Data on the other hand are used in the production process, either to create products or services or new ideas.

Data characteristics

Data come in many different shapes and are used in a variety of ways. Understanding these differences is important for designing policies that balance opportunities and challenges. For example, using anonymized income statements for an academic paper on the financial performance of small and medium-sized enterprises touches upon different issues than using someone's social media profile to target advertising. At the same time, some economic properties of (digital) data are independent of the data type. In this article, we first aim to get a better understanding of data by defining data, categorizing their differences, and identifying their common denominators. Then, we study the data economy in more detail and discuss the value of data.

DEFINITIONS AND CATEGORIZATION

Carrière-Swallow and Haksar (2019) define "data" as a "factual representation of a characteristic, action, or natural occurrence" (p. 7). Furthermore, they make a distinction between qualitative and quantitative data and the way they are stored (digital versus analog). Hilbert and López (2011) show how data have become increasingly digitized during the last decades. Data are now predominantly stored digitally.[7]

Data differ from ideas: Both are forms of information, but they serve different purposes. According to Jones and Tonetti (2020), "an idea is a production function whereas data is a factor of production" (p. 2821). Concretely this means that ideas are pieces of information that provide instructions on how to create output from a certain set of inputs (Romer, 1990). Data on the other hand are used in the production process, either to create products or services or new ideas.

Several classification schemes for data have emerged in the literature (Wdowin & Diepeveen, 2020). Crémer et al. (2019) make a distinction between personal and non-personal data and classifies them as volunteered, observed, or inferred based on the channel through which the data have been acquired. Furthermore, they propose to distinguish between four categories of use cases: Applications and analyses can use individual-level data, bundled individual--level data, aggregate-level data, or contextual data. Individual-level data refers to data from a specific user or machine. When these data are combined to come up with movie or music recommendations, e.g., they use the term "bundled individual-level data." Without additional information, it is not possible to trace aggregate data back to the individual level. Examples include frequency tables showing the distribution of digital skills levels of a population group or profit and loss statements. Contextual data are not derived from individual-level data. Typical examples are satellite data, mapping data, or earthquake data.

[7] For some fascinating ancient ways to store data see e.g. BBC news article about the world's first accountants (Harford, 2017).

Statistics Canada (2019) suggests organizing data according to what they are about or what they represent – for instance weather data, sports data, or economic data. In a report on international data transfers, the Swedish National Board of Trade (2015) classifies data based on how they are used in the production process of companies. Examples include employment data, quality data, and customer data.

ECONOMIC PROPERTIES

In this section, we study the economic properties of data. First, we focus on nonrivalry and partial excludability. Second, we discuss the impact of data on economic costs and their marginal returns.

NONRIVALRY

One of the most distinctive features of data is nonrivalry. An economic good is nonrival when it can be used by multiple consumers or firms at the same time, without diminishing its quantity or quality. Jones and Tonetti (2020) use an illustrative analogy with rival goods to explain what it means that at "the technological level, data is infinitely usable" (p. 2819). Because of rivalry, workers typically need their own desk and computer, and every warehouse relies on its own and exclusive collection of forklifts. If we would assume this capital to be nonrival, however, then all workers could use all desks and computers at once and all warehouses would be able to use any forklift in the industry. This is the case with data. Due to non-rivalry, all data could theoretically be used by all firms at the same time which implies that economic gains would remain untapped as long as this nonrivalry is not exploited. Carrière-Swallow and Haksar (2019) note that policies and private interests determine whether data will be nonrival in practice.

Goldfarb and Tucker (2019) generalize the nonrivalry of data to products and services by comparing goods made of atoms and goods made of bits. Unlike goods made of atoms, bits are nonrival, because the replication costs of digital information are almost zero – you can copy-paste software code but not a Ferrari.

One of the most distinctive features of data is nonrivalry. An economic good is nonrival when it can be used by multiple consumers or firms at the same time, without diminishing its quantity or quality.

PARTIAL EXCLUDABILITY

Some types of data are excludable, i.e., denying others access is not prohibitively costly. When data collectors exclude others, data takes on the features of a club good (Buchanan, 1965). When others cannot be prevented from accessing data, data is non-excludable and can be regarded as a public good.

Coyle et al. (2020) provide a short overview of the excludability of different data types. For instance, administrative (like tax returns or patient records) or planned data (like work schedules or budgets) are types of data where others can easily be excluded from. In contrast, environmental data, such as rainfall or geospatial data, are accessible to anyone since everyone can collect their own data on publicly observable phenomena – although the private costs of

Digital products and services have made it easier to verify identities and create reputation systems. Digital platforms, such as Uber and Airbnb, leverage the reduction in verification extensively to build trust in their two-sided marketplaces.

measurement may be too high to actually do it. A common way to make data excludable is by putting data behind a paywall (often tied to account registration). Think of newspaper articles or datasets for researchers. Offline storage is probably the easiest way for limiting access – only breaking physically into the device or space where the data is stored can lift the lock.

Data collectors and data processors face different incentives when deciding the level of access to data. They can, for example, restrict access in order to secure their competitive advantage and maintain their current market position (Carrière-Swallow & Haksar, 2019). Privacy legislation could be another reason for an organization to exclude others from access.

IMPACT OF DATA ON ECONOMIC COSTS

Goldfarb and Tucker (2019) describe how digitalization reduces five economic costs (search, replication, transportation, tracking, and verification costs), which are all connected to the properties of digitized data.

First, digitalization decreases search costs. Search engines for example have made it much easier to find relevant information, whether it concerns products, knowledge, or data itself. Second, the replication costs of digital products are close to zero: In other words, marginal costs are negligible. Moreover, as we have seen, the reproduction of data does not impact others due to its nonrival nature. Although marginal costs almost vanish, rolling out successful digital products often requires significant upfront investment – e.g., to establish a large enough network or to build a solid data infrastructure. Third, data are associated with near-zero transportation costs, and can be transferred across the globe without much effort, therefore digital business models have increasingly become global while businesses are able to scale at a more rapid pace. Fourth, tracking costs are lowered: Digital data makes it easier to keep track of transactions, people, and firms. Digitalization has therefore led to increasing levels of personalization. Examples include price discrimination and personalized advertisements, both of which have the potential to facilitate the matching of supply and demand. Fifth, lower tracking costs have enabled a reduction in verification costs. Digital products and services have made it easier to verify identities and create reputation systems. Digital platforms, such as Uber and Airbnb, leverage the reduction in verification extensively to build trust in their two-sided marketplaces.

INCREASING AND DECREASING RETURNS

In their book *Radical markets* (2018) Eric Posner and Glen Weyl discuss the marginal value of data in depth. The marginal value of an extra data point can either be decreasing or increasing with the number of data points collected, depending on the context.

To understand how this works, first consider a standard statistics problem. Let's say for example that you are interested in determining average household savings. Although the uncertainty in mean household savings decreases with the number of data points collected, the marginal decline becomes increasingly smaller as more data points are added. Thus, data lose their value over

volume and variety. Moreover, there is always a level of uncertainty that suffices for the application at hand. Gathering more data once this uncertainty level is reached is inefficient.

Posner and Weyl (2018) explain how in the data economy, where machine learning algorithms play an increasingly important role, the marginal values of data can be increasing. The underlying reason for these increasing returns of data is that different algorithms require different amounts of data. Typically, more data are needed the more complex a problem is. For a single learning problem, data again exhibit diminishing value of return, but collecting more data might now enable *new* problems to be solved causing a jump in the value of data collected. Whether data have an increasing or diminishing value of return is then determined by the value of the different problems. When most value resides with the most complex problem, it is likely that data have increasing value of return. In contrast, when most value resides with the simplest problems, data are likely to have diminishing value of return.

(...) the data economy, where machine learning algorithms play an increasingly important role, the marginal values of data can be increasing. (...) Typically, more data are needed the more complex a problem is.

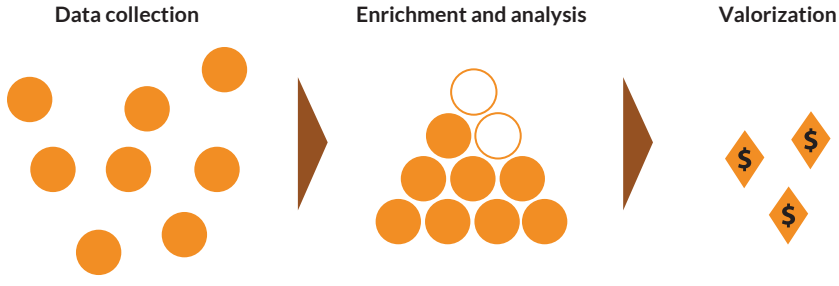
THE VALUE OF DATA IN THE ECONOMY

EXTRACTING VALUE FROM DATA: THE DATA VALUE CHAIN

Data are input to production processes. The data value chain describes how it contributes to production. In our article, we split up the data value chain into three (Figure 1). First, data need to be collected and stored. Second, they are analyzed and combined to create insights. Third, the insights translate into products and services.

Firms and institutes in the data economy either focus on a part of the chain or control the entire value chain for their business. Cloud services and big data consultants are examples of firms that specialize in offering products for a particular part of the value chain. The activities of tech firms, e.g., Alphabet, Amazon, and Apple, span the entire value chain. In the literature, different versions of the value chain appear, which originate from the number of chains or slightly different terminology.

Figure 1 – DATA VALUE CHAIN



Source: Prepared by the authors.

More and more economic activities take place within the data value chain. Those activities and the connected supply chains are thereby becoming more important parts of the overall economy.

Often, economic agents who play a role in the value chain are referred to as data subjects, data collectors, and data processors (Carrière-Swallow & Haksar, 2019). In the case of personal data, the person whose information details have been recorded is referred to as the data subject. A data collector collects and stores data, consequently incurring costs. On the demand side, the data processor uses the data, and aggregates and analyzes them. In practice, the data collector and data processor could be the same organization.

THE DATA ECONOMY

More and more economic activities take place within the data value chain. Those activities and the connected supply chains are thereby becoming more important parts of the overall economy. To monitor the impact of the data economy, the European Commission uses the following definition:

> The data economy measures the overall impacts of the data market on the economy as a whole. It involves the generation, collection, storage, processing, distribution, analysis elaboration, delivery, and exploitation of data enabled by digital technologies. The data economy also includes the direct, indirect, and induced effects of the data market on the economy. (European Commission, 2017, p. 2)

Using this definition, the size of the data economy in 2019 was estimated to be 2.6% of the gross domestic product (GDP) for the European Union (325 billion euro, excluding United Kingdom). Moreover, the data economy is expanding rapidly. In a conservative scenario the data economy is forecasted to grow to 430 billion euro in 2025 (3.3% GDP), while in the most aggressive outlook its size is forecasted to become 827 billion euro by 2025 (5.9% of GDP) (European Commission). In a recent complementary effort to define the size of the digital economy, the Organisation for Economic Co-operation and Development (OECD) stresses that there “remains some subjectivity or ‘fuzziness’” in turning definitions into numbers (OECD, 2020). Thus, the absolute numbers of these estimates depend on how the definition is translated in practice; therefore, they are somewhat arbitrary.

References

Buchanan, J. M. (1965). An economic theory of clubs. *Economica*, *32*(125), 1-14.

Carrière-Swallow, M. Y., & Haksar, M. V. (2019). *The economics and implications of data: An integrated perspective*. International Monetary Fund.

Crémer, J., Montjoye, Y. A., & Schweitzer, H. (2019). Competition policy for the digital era. *Report for the European Commission.*

European Commission. (2017). *Communication from the commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Building a European data economy*. https://digital-strategy.ec.europa.eu/en/library/communication-building-european-data-economy

European Commission. (2020). *The European data market monitoring tool: Key facts & figures, first policy conclusions, data landscape and quantified stories: d2.9 final study report*. Publications Office. https://data.europa.eu/doi/10.2759/72084

Coyle, D., Diepeveen, S., Wdowin, J., Kay, L., & Tennison, J. (2020). *The value of data – policy implications report*. Bennet Institute for Public Policy, Cambridge and Open Data Institute.

Goldfarb, A., & Tucker, C. (2019). Digital Economics. *Journal of Economic Literature*, *57*(1) 3-43.

Harford, T. (2017). *How the world's first accountants counted on cuneiform*. BBC World Service, 50 Things That Made the Modern Economy. https://www.bbc.com/news/business-39870485

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

Jones, C. I., & Tonetti, P. (2020). Nonrivalry and the economics of data. *American Economic Review*, *110*(9), 2819-2858.

Organisation for Economic Co-operation and Development. (2020). *A roadmap toward a common framework for measuring the digital economy*. OECD publishing.

Posner, E. A., & Weyl, E. G. (2018). *Radical markets – uprooting capitalism and democracy for a just society*. Princeton University Press.

Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy*, *98*(5), S71-S102.

Statistics Canada. (2019). Measuring investment in data, databases and data science: Conceptual framework. *Series: Latest Developments in the Canadian Economic Accounts*. https://www150.statcan.gc.ca/n1/en/catalogue/13-605-X201900100008

Swedish National Board of Trade. (2014). *No transfer, no production - a report on cross-border data transfers, global value chains, and the production of goods*. https://ec.europa.eu/futurium/en/system/files/ged/publ-no-transfer-no-production.pdf

Wdowin, J., & Diepeveen, S. (2020). *The value of data - accompanying literature review*. Bennet Institute for Public Policy, Cambridge and Open Data Institute.

Interview II

Photo: Manon Voland

Michael Kende
Internet Policy expert and chairman of Board of Datasphere Initiative.

Data and the free nature of the Internet

In this interview, Michael Kende, an Internet Policy expert and chairman of the Board of the Datasphere Initiative, discusses some of the themes debated in his book *The flip side of free: Understanding the economics of the Internet* (The MIT Press, 2021), including: The reasons for the "free" nature of the Internet; the economic implications arising from this; the risks for Internet users in the context of the use and reuse of personal data; policies to mitigate them; and research agendas for understanding the economics of the Internet.

Internet Sectoral Overview (I.S.O.)_ In your book **The flip side of free: Understanding the economics of the Internet,** ***some central questions in your analysis are why the Internet is free and what are the consequent economic implications of that. Why are these questions so important nowadays?***

Michael Kende (M.K.)_ The book highlights three ways in which the Internet is free. First, many of the standards and software underlying the Internet are open and freely available for developers. Second, increasingly broadband packages are unlimited, meaning that increased usage is free once the monthly charge has been paid, and often, public Wi-Fi can be used completely free of charge. Finally, many Internet services are available for free.
This is unique in many ways, compared with more traditional services, and has driven many of the economic and social benefits of the Internet that users enjoy today. The benefits of this were perhaps most clear during the pandemic when our reliance on the Internet reached a peak.
However, as indicated by the title of the book, there is a flip side that needs to be addressed.
Open standards and open-source software may not get sufficient attention from developers, given the lack of financial return. As a result, for instance, there is a famous case in which the OpenSSL software library for securing transactions was vulnerable to what was known as the Heartbleed bug. It cost US$500 million to fix the bug. While it is true that commercial software is also vulnerable to bugs, it became apparent that far too few resources had been available to develop the OpenSSL software. As our online reliance continues to increase, so should the resources to minimize addressable vulnerabilities.
The use of unlimited data packages has driven an ever-growing variety of uses for work, health, education, and entertainment. However, recently a downside emerged: Large broadband providers, in Europe and elsewhere, are complaining to policymakers about the amount of traffic they are carrying to end users, and in particular video traffic. They are reluctant to raise

rates or impose charges based on data usage and, instead, are asking the largest content providers to pay their "fair share" of the costs they attribute to the traffic. This would fundamentally threaten the voluntary interconnection agreements that are used today to create the "network of networks" that is the Internet.
Finally, free services provide significant benefits to users, for communications, social media, and entertainment. "Free" is a special price, which delivers significant value to users and developers who do not have to factor cost into their decision to adopt services and standards. However, the flip side must be addressed and it should be done in ways that do not fundamentally change the nature of the Internet.

I.S.O._ One of the conclusions of your book is that many Internet services are available to us at no cost, in return for personal data, that can be used and reused. What are the main risks for Internet users in this scenario? What policies could effectively tackle these risks while preserving the advantages of Internet services for users?

M.K._ Free services are paid through advertising, which is targeted with personal data from the users. This personal data, in turn, creates concerns about privacy and is at risk of being exposed through data breaches. The result is a deficit of digital trust, that must be addressed to fully benefit from the impact of the Internet and new services that keep emerging.
There is no good analogy for data because it has unique properties. One common analogy for data is oil. While it is true that data is driving the digital revolution the way that oil drove the industrial revolution, the similarities end there. Oil is non-renewable, its use unavoidable creates negative climate effects, and it is tangible. Data, on the other hand, is generative, and it can be combined and used simultaneously; careful usage does not need to create negative effects. Besides that, data is virtual and not tangible.
These properties have created risks for Internet users. Even if we knew all the data that we were making available to service providers, and even if we closely read the terms of service, we cannot fully know how our data is being used, or by whom. This requires privacy regulations that reasonably limit how data are used and put more control in the hands of users, but also service provider terms of service that are easier to understand and provide more user controls. No single stakeholder can address the issues of privacy alone.
There are also significant challenges with cybersecurity, resulting in unauthorized and illegal use of our personal data. While some data breaches are not avoidable, based on unknown vulnerabilities or insider actions, many are avoidable. Here the question is: Why companies are not better able to protect the data of their own users? In economic terms, there are two reasons for this. First, it is difficult for companies to prove that they have implemented tighter cybersecurity, so there is too little upside to invest in data protection. Second, companies have limited liability if there is a data breach; consequently, there is too little downside to not investing.

"Even if we knew all the data that we were making available to service providers, and even if we closely read the terms of service, we cannot fully know how our data is being used, or by whom."

"In particular, there are the paradoxes: Users who are increasingly aware of privacy issues still provide personal data online, and even after a trust breach becomes known, the use of existing services does not seem to decrease."

The situation is analogous to automobile safety in the 1960s when cars could be sold with minimum protection. Since then, governments have begun to impose some regulations, such as for crash protection, and cars are tested and rated for safety, resulting in demand for safety beyond what is required. The same process must take place for cybersecurity, with new regulations and standards, creating awareness and demand for greater protections and increased liability when companies fall short.

I.S.O._ How has the debate proposed in your book advanced since its publication? In your opinion, what are the main topics that should be investigated to better understand the economics of the Internet?

M.K._ Privacy and personal data protection issues have impacted digital trust levels, which would benefit from increased investigations. In particular, there are the paradoxes: Users who are increasingly aware of privacy issues still provide personal data online, and even after a trust breach becomes known, the use of existing services does not seem to decrease. This can best be understood through the lens of behavioral economics, which seeks to understand actions that do not seem "rational" in an economic sense but are common.
These paradoxes can arise for a number of reasons. First, there may be a lack of awareness about the true privacy risks for personal data, which can result in a seeming overuse of relevant services. Second, even with awareness, users may overestimate their own ability to avoid problems or underestimate the likelihood of them arising. Another issue common to behavioral economics is the preference for current benefits from Internet services over any future costs, even if the risks are well understood.
There is another time-related issue as well. It appears that users are unlikely to give up existing services, even ones that resulted in a significant breach of trust. For instance, the website Ashley Madison – created for the express purpose of enabling adultery – was attacked and the data of all 36 million users was exposed, with significant personal costs to the users. Nonetheless, the website not only survived but now claims twice as many users as before. If anything, the publicity around the breach seems to have created demand for its services.
So, users are reluctant to give up existing services. On the other hand, new services – which are not yet providing benefits – may be easy to forgo in the name of privacy. COVID-19 contact tracing apps, whose widespread use would have limited the spread of the virus, did not reach more than 30% of the population in most countries. One reason was worry over privacy, as the apps kept track of users' location to determine if they were exposed to someone who tested positive. While the apps tended to be privacy-preserving, keeping far less data than many other sites with arguably fewer benefits, they could not overcome the reluctance.
As a result, the adoption and use of online services may not be rational but has significant impacts. Understanding users' awareness of privacy and data protection concerns, and how they drive behavior, will help to address them, and increase the general level of trust.

Domain Report

Domain registration dynamics in Brazil and around the world

The Regional Center for Studies on the Development of the Information Society (Cetic.br), department of the Brazilian Network Information Center (NIC.br), carries out monthly monitoring of the number of country code Top-Level Domains (ccTLD) registered in countries that are part of the Organisation for Economic Co-operation and Development (OECD) and the G20.[8] Considering members from both blocs, the 20 nations with the highest activity sum more than 90.76 million registrations. In June 2023, domains registered under .de (Germany) reached 17.59 million, followed by China (.cn), the United Kingdom (.uk), and Netherlands (.nl), with 9.51 million, 7.37 million and 6.3 million registrations, respectively. Brazil had 5.22 million registrations under .br, occupying 5th place on the list, as shown in Table 1.[9]

[8] Group composed by the 19 largest economies in the world and the European Union. More information available at: https://g20.org/

[9] The table presents the number of ccTLD domains according to the indicated sources. The figures correspond to the record published by each country, considering members from the OECD and G20. For countries that do not provide official statistics supplied by the domain name registration authority, the figures were obtained from: https://research.domaintools.com/statistics/tld-counts. It is important to note that there are variations among the date of reference, although the most up-to-date data for each country is compiled. The comparative analysis for domain name performance should also consider the different management models for ccTLD registration. In addition, when observing rankings, it is important to consider the diversity of existing business models.

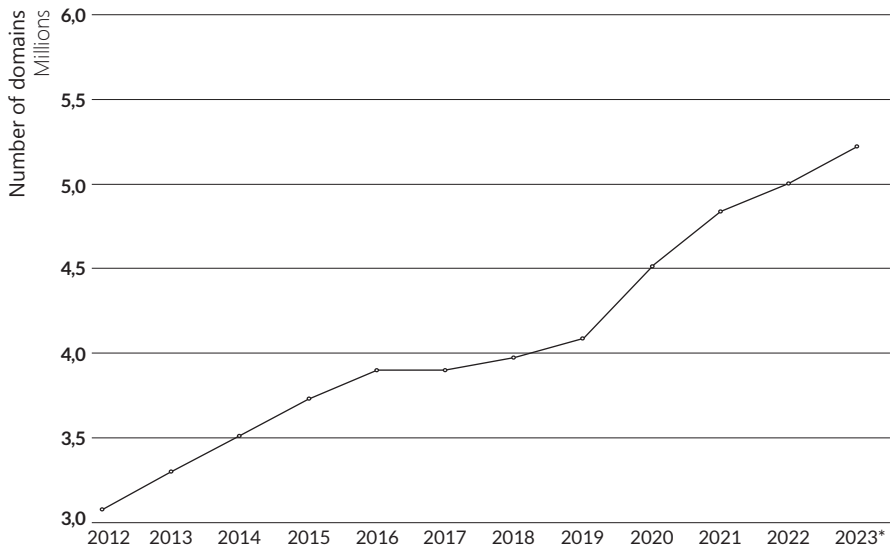
Table 1 – TOTAL REGISTRATION OF DOMAIN NAMES AMONG OECD AND G20 COUNTRIES

Position	Country	Number of domains	Date of reference	Source (website)
1	Germany (.de)	17,598,905	01/09/2023	https://www.denic.de
2	United Kingdom (.uk)	9,511,231	31/07/2023	https://www.nominet.uk/news/reports-statistics/uk-register--statistics-2023
3	China (.cn)	7,370,599	01/09/2023	https://research.domaintools.com/statistics/tld-counts
4	Netherlands (.nl)	6,333,731	01/09/2023	https://stats.sidnlabs.nl/en/registration.html
5	**Brazil (.br)**	**5,223,034**	**31/08/2023**	**https://registro.br/dominio/estatisticas**
6	Russia (.ru)	5,073,407	01/09/2023	https://cctld.ru
7	Australia (.au)	4,263,841	01/09/2023	https://www.auda.org.au
8	France (.fr)	4,081,492	31/08/2023	https://www.afnic.fr/en/observatory-and-resources/statistics
9	European Union (.eu)	3,679,990	01/09/2023	https://research.domaintools.com/statistics/tld-counts
10	Italy (.it)	3,482,097	01/09/2023	http://nic.it
11	Colombia (.co)	3,465,280	01/09/2023	https://research.domaintools.com/statistics/tld-counts
12	Canada (.ca)	3,364,108	01/09/2023	https://www.cira.ca
13	India (.in)	2,954,328	01/09/2023	https://research.domaintools.com/statistics/tld-counts
14	Switzerland (.ch)	2,554,894	15/08/2023	https://www.nic.ch/statistics/domains
15	Poland (.pl)	2,525,714	01/09/2023	https://www.dns.pl/en
16	Spain (.es)	2,069,267	09/08/2023	https://www.dominios.es/dominios/en
17	United States (.us)	1,986,914	01/09/2023	https://research.domaintools.com/statistics/tld-counts
18	Japan (.jp)	1,748,824	01/09/2023	https://jprs.co.jp/en/stat
19	Belgium (.be)	1,739,915	01/09/2023	https://www.dnsbelgium.be/en
20	Portugal (.pt)	1,735,197	01/09/2023	https://www.dns.pt/en/statistics

Collection date: September 1, 2023.

Chart 1 shows the performance of .br since 2012.

Chart 1 – TOTAL NUMBER OF DOMAIN REGISTRATIONS FOR .BR – 2012 to 2023*



* Collection date: August 31, 2023.
Source: Registro.br
Retrieved from: https://registro.br/dominio/estatisticas

In August 2023, the five generic Top-Level Domains (gTLD) totaled more than 189.77 million registrations. With 159.14 million registrations, .com ranked first, as shown in Table 2.

Table 2 – TOTAL NUMBER OF DOMAINS AMONG MAIN gTLD

Position	gTLD	Number of domains
1	.com	159,149,727
2	.net	12,796,114
3	.org	10,780,717
4	.info	3,753,027
5	.xyz	3,293,663

Collection date: September 1, 2023.
Source: DomainTools.com
Retrieved from: research.domaintools.com/statistics/tld-counts

Internet Markers in Brazil

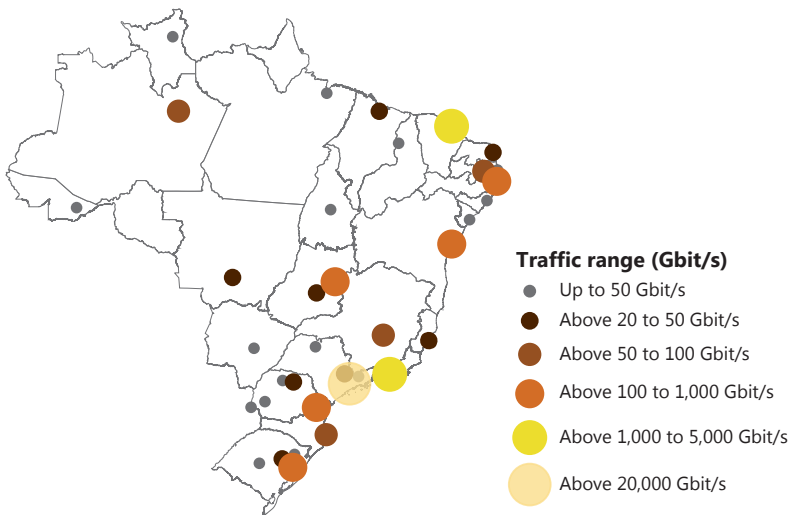
IX.br: Data on Internet Exchange Points

IX.br (Brazil Internet Exchange) is an initiative of the Brazilian Network Information Center (NIC.br), supported by the Brazilian Internet Steering Committee (CGI.br), which promotes and implements Internet Exchange Points (IXP), the necessary infrastructure for direct interconnection between the networks, also known as Autonomous Systems (AS), which make up the Internet in Brazil.

The interconnection of several AS in an IXP simplifies Internet transit, establishing more direct traffic to a given destination. This improves quality, reduces costs and increases network resilience.

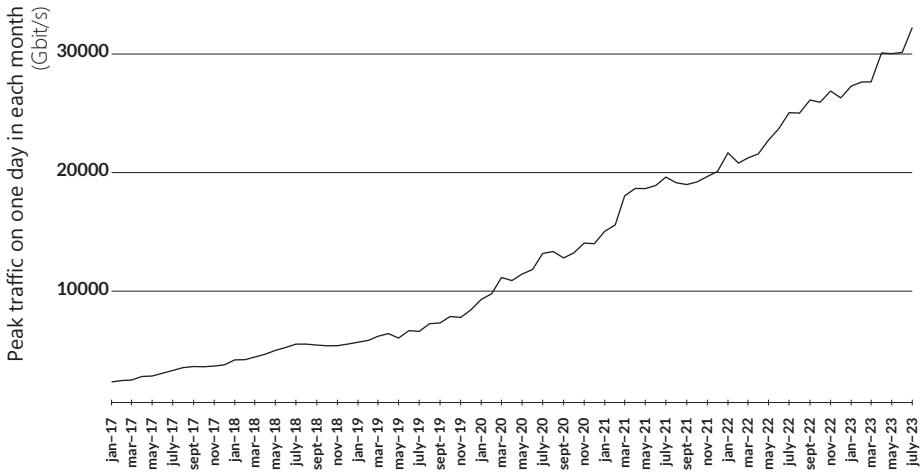
The initiative currently encompasses 36 independent IXP, distributed throughout Brazil (Figure 1), and is one of the most important clusters of IXP worldwide. Chart 1 shows the continuous traffic growth of the IXP cluster that comprises IX.br over the past five years.

Figure 1 – TRAFFIC EXCHANGE POINTS (IXP) IN BRAZIL, BY TRAFFIC RANGE



Reference period: July 2023.
Source: IX.br | NIC.br
Retrieved from: https://ix.br/trafego/agregado/

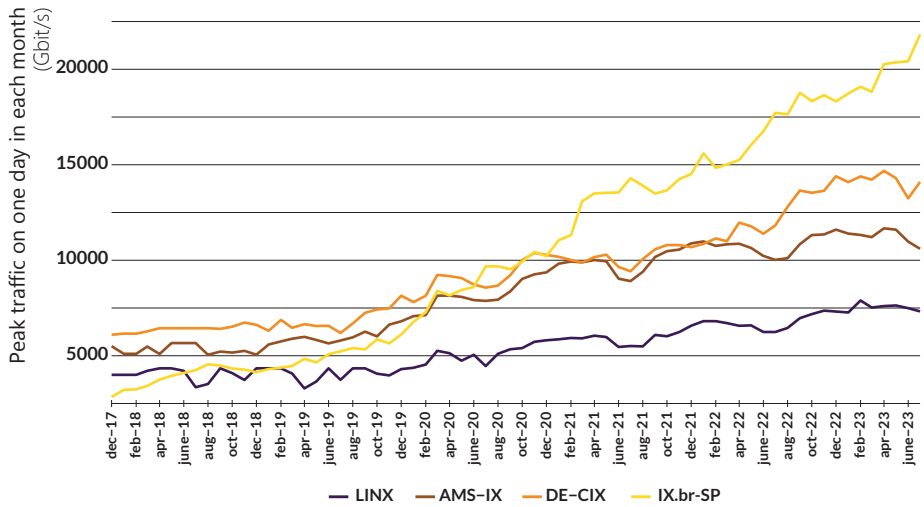
Chart 1 – TRAFFIC PEAK FOR THE IX.br INTERNET EXCHANGE POINT CLUSTER – 2017 to 2023



Source: IX.br|NIC.br
Retrieved from: https://ix.br/agregado/

Chart 2 compares the peak traffic of the São Paulo IXP, the largest in Brazil, with the three largest in Europe: LINX (London, England), AMS-IX (Amsterdam, Netherlands), and DE-CIX (Frankfurt, Germany), between 2017 and 2023.

Chart 2 – LONDON (LINX), AMSTERDAM (AMS-IX), FRANKFURT (DE-CIX) AND SÃO PAULO (IX.br-SP) IXP, BY TRAFFIC PEAK - 2017 to 2023



Source: IX.br|NIC.br
Retrieved from: https://www.de-cix.net/en/locations/frankfurt/statistics;
https://www.ams-ix.net/ams/documentation/total-stats;
https://portal.linx.net/services/lans-snmp; https://ix.br/trafego/agregado/

Here you can find more information about IX.br's activities and statistics.

Online activities and the digital crumbs

Data is a fundamental asset in the context of the data economy, contributing to the decision-making processes of organizations, as well as serving as the basis for new business models.

Online activities and services leave behind a series of digital "crumbs" or data.[10] Thus, the growing presence of digital platforms in the daily lives of millions of individuals results in the production of an enormous amount of data.

In 2022, 149 million (81%)[11] individuals in Brazil aged 10 years and older were Internet users.[12] The following data show some of the activities and services[13] carried out on the Internet by this population:

Activities

- **93%** sent instant messages.
- **80%** used social networks.
- **80%** watched videos, shows, movies, or series online.
- **69%** shared content on the Internet.

Services

- **40%** ordered cab rides or private drivers, such as Uber or 99.
- **38%** paid for series or movies streaming services, such as Netflix or Globoplay.
- **33%** ordered meals on sites or applications, such as iFood or Rappi.

[10] Letouzé, E. (2018). Big Data & development: An overview. https://cetic.br/media/docs/publicacoes/6/20191211163913/internet-sectoral-overview-x-1-big-data.pdf

[11] Data from the ICT Households 2022 survey by Cetic.br|NIC.br. Available at: https://cetic.br/en/pesquisa/domicilios/

[12] A "user" is defined as someone who has been using the Internet for less than three months at the time of the interview, as defined by the International Telecommunications Un ion (ITU).

[13] Other activities and services carried out on the Internet by Internet users, collected by the ICT Households 2022 survey, can be found at: https://cetic.br/en/tics/domicilios/2022/individuos/

/Credits

TEXT

DOMAIN REPORT

Thiago Meireles (Cetic.br | NIC.br)

INTERNET MARKERS IN BRAZIL

Julio Sirota (IX.br | NIC.br) and Milton Kaoru Kashiwakura (NIC.br)

GRAPHIC DESIGN

Giuliano Galves, Larissa Paschoal, and Maricy Rabelo (Comunicação | NIC.br)

PUBLISHING

Grappa Marketing Editorial

ENGLISH REVISION AND TRANSLATION

Ana Zuleika Pinheiro Machado

EDITORIAL COORDINATION

Alexandre F. Barbosa, Graziela Castello, Javiera F. M. Macaya, and Mariana Galhardo Oliveira (Cetic.br | NIC.br)

ACKNOWLEDGMENTS

Marielza Oliveira (UNESCO)

Ramy El-Dardiry, Milena Dinkova, and Bastiaan Overvest (Netherlands Bureau for Economic Policy Analysis)

Marcos Dantas (UFRJ and CGI.br)

Michael Kende (Datasphere Initiative)

ABOUT CETIC.br

The Regional Center for Studies on the Development of the Information Society – Cetic.br (https://www.cetic.br/en/), a department of NIC.br, is responsible for producing studies and statistics on the access and use of the Internet in Brazil, disseminating analyzes and periodic information on the Internet development in the country. Cetic.br acts under the auspices of UNESCO.

ABOUT NIC.br

The Brazilian Network Information Center – NIC.br (http://www.nic.br/about-nic-br/) is a non-profit civil Entity in charge of operating the .br domain, distributing IP numbers, and registering Autonomous Systems in the country. It conducts initiatives and projects that bring benefits to the Internet infrastructure in Brazil.

ABOUT CGI.br

The Brazilian Internet Steering Committee – CGI.br (https://cgi.br/about/), responsible for establishing strategic guidelines related to the use and development of the Internet in Brazil, coordinates and integrates all Internet service initiatives in the country, promoting technical quality, innovation, and dissemination of the services offered.

*The ideas and opinions expressed in the texts of this publication are those of the respective authors and do not necessarily reflect those of NIC.br and CGI.br.

unesco
Centre under the auspices of UNESCO

cetic.br
Regional Center for Studies on the Development of the Information Society

nic.br
Brazilian Network information Center

cgi.br
Brazilian Internet Steering Committee

STRIVING FOR A BETTER INTERNET IN BRAZIL

CGI.BR, MODEL OF MULTISTAKEHOLDER GOVERNANCE

www.cgi.br

nic.br cgi.br

FOTOS: RICARDO MATSUKAWA E SHUTTERSTOCK