29

DATA STORAGE ETHICS: SECURITY AND RESPONSIBILITY AT ALL LEVELS

Gilles Bach, Switzerland[365]

29.1 Data Explosion

Digital Data is today at the center of our lives. Personal lives. Professional lives. Our lives as citizen. Through the internet and its myriad of websites and applications used by about 5 billion users worldwide[366], we all leave daily - whether we are aware of it or not – dozens to thousands of digital footprints in this vast and complex digital mechanism and world called cyberspace.

---

[365]Gilles Bach is a Strategy Consultant in Data and Digital Transformation. Teacher of the Cyber Ethics Online Course of Globethics. | DOI: 10.58863/20.500.12424/4276061 |

[366] Digital 2022 Global Report. 2022. DataReportal, https://datareportal.com/reports/digital-2022-global-overview-report.

Even when we enter a physical venue and we interact and transact with a brick-and-mortar store, business, corporation or public organization for any type of purpose we are creating electronic records in their systems most of the time.

The overall quantity of data being generated is exponential. In 2021[367], 79 zettabytes of data were created globally (79 followed by 21 zeros). This has to be compared to only two zettabytes in 2010, about 16 in 2015, 175 being forecasted in 2025, and 570 potentially estimated in 2030[368]…

These huge amounts of data, and the way they are collected and in particular stored before, during or after their processing and uses, generate in return important ethical responsibilities towards the individuals and the organizations to which they belong as well as their surrounding environments and eco-systems.

As we will see these responsibilities encompass a very large range of dimensions at very different levels that have to be considered and that are affected in many various ways: technology itself, geopolitics, privacy, law, human rights, IT security, governance and organization, sovereignty, economics, sustainability and even sociology and philosophy.

This is what we will scrutinize in the subsequent parts of this article by setting the various contexts, explanations, examples and possible solutions to raise the awareness and self-reflection, improve the understanding, and ignite the implementation of these fundamental ethical responsibilities.

---

[367] See from 2010 to 2020: Petroc, Taylor. 2022. Amount of data created, consumed, and stored 2010-2020, with forecasts to 2025, Statista, https://www.statista.com/statistics/ 871513/worldwide-data-created/

[368] See beyond 2025: Balnojan, Sven. 2020. The Future of Good Data — What You Should Know Now!, Towards Data Science, https://towardsdatascience.com/the-future-of-good-data-what-you-should-know-now-f2a312a0e469

29.2 Data Growth: Four Factors

As we have seen, the curve of the yearly data generation is following an exponential trend. This phenomena has been actually facilitated by four factors that started to reach a sufficient level of sophistication around the end of the first decade of 2000 to enable the accelerated development of increasingly data -heavy, -hungry and -related technologies and applications: more powerful and affordable computing power, higher and significant storage capabilities, sophisticated software and data architecture frameworks, and last but not least the spread of fast and reliable internet services.

In terms of data types being generated, the current and continuous growth of data volume comes from the multiplication of videos and photos, the rise of social media, the ever increasing level of digitization of businesses and organizations in the world across all sectors and geographies, and from the explosion of smartphones, IoT (Internet of Things) devices, sensors, machines and vehicles capturing data.

80% of the data growth is due to unstructured type of data[369] (mostly pictures and videos) as compared to structured data (inputs and information to fill forms and complete most transactions and processes) and takes therefore much more storage space.

The Covid pandemic lockdowns did obviously accelerate this digitalization thanks to a forced transformation of many physical processes and types of transactions that became banned and inapplicable literally overnight across all areas of life and business. Also, the expected development and increasing sophistication of big data, advanced data science, automation and AI algorithms requires more data for both systems training purposes and generating accurate results - and this is only the beginning.

---

[369] Artificial Intelligence. To Unlock the Hidden Value of Unstructured Data. NRoad, accessed from the blog VentureBeat, https://venturebeat.com/data-infrastructure/report-80-of-global-datasphere-will-be-unstructured-by-2025/

Not forgetting about the upcoming Metaverse developments and machine-generated data that will increasingly contribute to this expansion. But let's start first with some history summarizing the evolution of storage mediums.

29.3 Data Storage: History, Types and Evolution

As a reminder – even if we tend to forget about it – data and information can be obviously stored on traditional types of mediums, and in particular on paper. This makes this data much less accessible and much more difficult to retrieve and to leverage. Which is not really compatible with the current era of digital transformation and digitization and limits all the benefits that can be derived from analyzing and using data more systematically and in a faster and automated way. On the other hand, in a time of frequent digital data hacking and breaches, this can make such data potentially more secure as not everybody can access it.

The physical nature of this medium also necessitates a significant physical footprint and is particularly vulnerable to calamities. The fate of the famous Universal Library of Alexandria, one of the largest and most significant library of the ancient world reminds us of the particular fragility of this type of storage in the event of calamities like fire, floodings or similar. Hence this medium only being limited nowadays to very specific niches and usages.

Then, in the late 1950s mainframe computers started to appear and used sets of punched cards, paper tape, or magnetic tape to store and transfer data and programs. This is how IT started in some large corporations with such systems installed in big server rooms that were connected to dozens or more terminal stations used by employees to access or input data into the system. This corresponds to the image that we can see in some movies from the seventies in particular.

At the end of the seventies and in the eighties, micro-computers, later also called personal computers (PC) where on the rise and permitted

individual users, at work or at home, to process information and use software decoupled from any server for accessing and storing data and programs, with no need for any communication network. They were equipped with independent local medium storage systems like tapes, floppy discs or hard drives.

In the 1990s the appearance and rise of internet in a mainstream fashion brought the interconnectedness between local computers and remote web servers operating websites and applications as well as file sharing servers and protocols.

It is only in 2006 that the current notion of Cloud computing started to emerge when Amazon launched[370] a cloud storage service available from anywhere on the web, helping 'free developers from worrying about where they are going to store data, whether it will be safe and secure, if it will be available when they need it, the costs associated with server maintenance, or whether they have enough storage available'.

This was at the very origin of this huge rise that brought the spending on public cloud services globally from scratch to an estimated 495 billion[371] USD in 2022 as compared to 411 in 2021 and a forecast of 924 billion by 2027.

The share of corporate data stored in the cloud increased from 30% in 2015 to 60%[372] in 2022 with 3.6 billion cloud users in the world.

---

[370] Initial Amazon Web Services (AWS) cloud storage service launch press release, on March 14, 2006 - https://press.aboutamazon.com/2006/3/amazon-web-services-launches

[371] Gartner Forecasts Worldwide Public Cloud End-User Spending to Reach Nearly $500 Billion in 2022. Gartner, Stamford, Press Release, 2022, https://www.gartner.com/en/newsroom/press-releases/2022-04-19-gartner-forecasts-worldwide-public-cloud-end-user-spending-to-reach-nearly-500-billion-in-2022

[372] Sujay Vailshery, Lionel. 2022. Share of corporate data stored in the cloud in organizations worldwide from 2015 to 2022, Statista, https://www.statista.com/statistics/1062879/worldwide-cloud-storage-of-corporate-data/

29.4 Data Storage in the Cloud: A New Era of Possibilities

Cloud storage and cloud computing are more precisely about the remote storage of data and the remote access to computing power and applications services over the internet, provided by a third party vendor usually labeled as a CSP – Cloud Services Provider. These services are available both for organizations and individuals. It replaces the model where the data and applications are stored and processed locally, either on a computer, or on on-premise servers.

There are globally 3 types of clouds. Private clouds, that are dedicated and operated solely for a single organization. It can be managed internally or through a third-party vendor. Public clouds use shared infrastructure to deliver services to many organizations or individuals. And Hybrid cloud is a storage infrastructure using a combination of public cloud services and on-premise or private cloud resources.

Public clouds are seen by organizations as 'appropriate only for less critical data and less sensitive workloads'[373], therefore the strategic choice to keep those particularly critical data sets within a private cloud when appropriate.

There are many benefits though for organizations to moving their data initially stored on their own managed, on-premise servers to the public cloud: the advantages relating to outsourcing, like reducing the directly managed and held resources in terms of IT hardware and infrastructures, implying Capex costs (capital expenditures) savings, as well as reducing IT staff costs. The CSPs do therefore offer in comparison much better prices thanks to the economies of scale of shared infrastructures. On top of that they are able to offer state of the art cybersecurity protection to the stored data, workloads and services, by applying the very best and

[373] Brinda, Mark and Kate Woolley. 2019. Public vs. Private Cloud? The Market Says Hybrid, Bain & Company, https://www.bain.com/insights/public-vs-private-cloud-the-market-says-hybrid/

elaborated, up-to-date practices and technologies with highly skilled specialists.

Cloud services are also enabling many new types of services supporting enhanced collaboration between workers and organizations all over the world, improving the speed and flow of information being exchanged, of transactions, the delivery of services from one end of the globe to another, including real-time operations, advanced data analyses, an increased level of big data and AI sophistication, and much more. All this in a very scalable, flexible, and cost-effective way, which confirms definitely Data as being the '4th production factor in this era of the fourth industrial revolution'[374] on top of the traditional factors that are natural resources, labour and capital.

29.5 Cloud Storage: Macro Risks

As we have seen the cloud creates tremendous benefits and opportunities for the development of businesses and organizations in the world. It can even be considered as the 'powerhouse that drives today's digital organizations'[375]

Electronic data storage requires however by definition electrical power to store and retrieve data. And the cloud requires access to a working, reliable and fast telecommunication network linked to the internet. This is of no surprise and seems very basic and obvious in our days. But at a time in which conflicts with worldwide repercussions are surging or threatening to surge again, raising concerns and tensions in terms of energy cost, and even about energy supply at all, this is not benign.

---

[374] Stückelberger, Christoph / Duggal, Pavan (Eds.) 2018. Cyber Ethics 4.0: Serving Humanity with Values, Global Series, Globethics Publications, 45, https://repository.globethics.net/handle/20.500.12424/169317

[375] Gartner Forecasts Worldwide Public Cloud End-User Spending to Reach Nearly $500 Billion in 2022, loc. cit.

And if we look at real or potential threats - whoever perpetrating them – against central and strategic telecommunication infrastructures absolutely indispensable to most of our digitized ways of communicating, informing, transacting and delivering products and services, this adds even more to the concerns.

We must also be aware that 99% of the internet runs through a network of 420 submarine cables connecting all continents to each other. 'It is estimated that over USD 10,000 billion in financial transactions run today through these "seabed highways". This is especially the case of the main global financial exchange system, SWIFT (Society for Worldwide Interbank Financial Telecommunications). The security of these transactions is a political, economic, and social problem.'[376]

In the aftermath of the September 2022 Nordstream natural gas pipelines spectacular sabotage, and knowing that actual incidents and attempts on internet cables already happened in the past, a massive disruption caused by coordinated attacks could become a nightmare scenario for the overall world economy. The satellites that come to mind as a possible alternative for sustaining the internet connectivity in such an occurrence account actually only for 1% of the data exchanges. They are too costly and the connection speed is much lower. Not mentioning the fact that an old, defunct satellite has been destroyed in November 2021[377], just 3 months before the invasion of Ukraine, by an ASAT, meaning an anti-satellite missile test.

These two enabling conditions for the cloud, electricity and internet networks may also be put at risk by natural disasters, other human factors actions and even – though with a low but not impossible level of

[376] Rona Rita David, Submarine Cables: Risks and Security Threats, Energy Industry Review, March 25, 2022, https://energyindustryreview.com/analysis/submarine-cables-risks-and-security-threats/

[377] Foust, Jeff. 2021. Russia destroys satellite in ASAT test, SpaceNews, https://spacenews.com/russia-destroys-satellite-in-asat-test/

probability – by an event as distant as an extremely powerful solar flare which could potentially render all electronic devices inoperative.

These factors represent clearly a big weakness and risk in relation to the correct execution of data storage access and activities depending on the cloud. With 60% of the global GDP estimated to rely on digital communications in 2022[378], such a scenario would be absolutely disastrous. This has to do with a black swan type of event, with initially a very low probability of occurrence but a potentially enormous, disastrous impact. In the very current context however, we are just realizing that these options are not anymore merely a theoretical figment of imagination.

29.6 Cloud Storage, Local Storage, Metadata, Background Analysis and Processes

If you take many pictures with your smartphone you might have noticed that with models from the last 3-4 years, these photos are now very often labeled and tagged automatically thanks to machine learning image recognition algorithms. For example, category tags like food, landscape, sport, document, etc… or even people names can be generated and added automatically, as well as up to the geographic position and coordinates if your GPS is activated and if you enabled and set pro-actively the last two options in particular.

---

[378] Digital Development, World Bank. https://www.worldbank.org/en/topic/digitaldevelopment/overview. Amid this war context, the European Parliament and Commission have respectively released an in-depth analysis on the security threats to undersea communication cables and infrastructure in June 2022 and a five-point plan in October 2022: European Parliament, Directorate-General for External Policies of the Union, Bueger, C., Liebetrau, T., Franken, J., Security threats to undersea communications cables and infrastructure : consequences for the EU : in-depth analysis, Publications Office of the European Union, 2022, https://data.europa.eu/doi/10.2861/35332

Some of this automatic tagging is done by default, like the general categories, some others need to be enabled. Such data is called 'metadata', meaning 'data that describes data', in this case data describing pictures.

This functionality can be very helpful to categorize pictures in your gallery, in particular if you make a lot pof them, and makes it easier later on to search pictures in a faster and relevant manner. At the same time this initiative to deploy automation of such functionalities could create a clear feeling of intrusion into the private sphere of people. This raises then questions about how this type of automated analysis of stored content might be accessible or leveraged by various layers of stakeholders.

Does this data remain locally stored on the smartphone? Is it somehow accessible and leveraged – for commercial purposes or other – by one or several suppliers of the digital chain: the application creator, the operating system of the device, the phone maker? Is it even shared further with third parties? What happens also to this information when photos are synchronized and backed-up in the cloud?

The question expands actually to the real-time or later analysis of any stored content, may it be stored on the cloud or stored locally on PCs or any device connected to telecommunication and internet networks. The messages in our e-mail boxes, social media profiles and accounts information even when set to private, and of course documents and productivity suites software solutions and applications either based on cloud solutions or even classically stored locally on a computer, etc... are concerned.

Very often such information and authorizations are part of the Terms & Conditions of the respective listed types of suppliers. But who reads and fully understands those very long types of texts? According to a study the 'combined terms and conditions of 13 top apps including TikTok, WhatsApp and Zoom would take 17 hours and five minutes to read'. On top of that a privacy researcher mentioned that 'both the length and language used in such statements made them difficult for adults and

children to be able to make informed consent about what they were agreeing to'[379].

There are exactly the same kind of lack of clarities with specified authorized actions granted to installed mobile applications that often do not make any sense given the initial purpose and functionalities of the program, with no possibility of deselecting one single permission and providing no information about what exactly is collected or not, and to which purpose and type of processing.

With such information being too often burried in the middle of complex and boring statements, or not available at all in a detailed manner, it just reflects, irrespective of pure legal aspects, a kind of lack of transparency that the industry in general should address to re-establish and increase trust and respect between users, citizens, businesses and suppliers along the data storage and processing chain. Upgrading these standards, showing genuinely more transparency while enabling a true and better understanding of what is done or not with our data will be ultimately beneficial to all.

29.7 Information Security Principles and International IT Standards

Ensuring cybersecurity, and in particular the protection of digital data is obviously one of the major ethical responsibility for any organization managing and storing sensitive or voluminous amounts of data and systems and their leaders.

When it comes to computer data security, there are three fundamental principles and dimensions of data - adopted by all major IT Security standards and certifications - that need to be safeguarded at anytime: their confidentiality, their integrity and their accessibility. ISO 27000

[379] Kleinman, Zoe, Popular app T&Cs 'longer than Harry Potter', November 2020, https://www.bbc.com/news/technology-54838978

standards on IT Systems Security Management do cover the necessary application of those principles.

Ensuring Confidentiality in this context is about controlling and preventing the unauthorized use, disclosure of information or its unauthorized access.

Integrity is about ensuring that the data has not been modified, altered or deleted in any way, without permission, justification and in a monitored way. Availability is about ensuring a reliable and timely access to data whenever needed.

These principles do cover both the prevention and fight against external cyberattacks as well as against insider threats and non-authorized internal accesses.

The full and strict implementation and respect of those principles are absolutely vital from an ethical standpoint. Steve Wozniak, Apple's co-founder, raised warnings on those risks and on the fact that "the more we transfer everything onto the web, onto the cloud, the less we're going to have control over it"[380] as soon as 2012, a time when the cloud just started to spread more into mainstream users.

29.8 Data, Privacy, International Law and Human Rights

In terms of international laws and regulations on cybercrime and data protection, the 2001 Budapest Convention paved the way by defining and clarifying the criminalization of conducts relating to crimes committed in or through the cyberspace[381]. This international treaty served as a

[380] Meyer, David. 2012. Wozniak: 'I really worry about everything going to the cloud', Zdnet, https://www.zdnet.com/article/wozniak-i-really-worry-about-everything-going-to-the-cloud/

[381] Stückelberger, C., Duggal, P. 2018. Cyber Ethics 4.0: Serving Humanity with Values, op. cit., 368.

base to develop domestic laws around the world, as well as building up international cooperation on the matter.

It resulted in the production of many cyber criminal laws, and subsequently also in the development of more specific national data protection and privacy laws on the various continents.

As per the UN Conference for Trade and Development (UNCTAD), 137 countries out of 194 had put in place as of the end of 2021[382] legislation to secure the protection of data and privacy. Privacy is here a crucial notion that is close to the principle of confidentiality already mentioned but that relates more specifically to the confidentiality of personal data.

It refers fundamentally to the individual's right to control and maintain their own data, underpinned directly by the article 12 of the 1948 Universal Declaration of Human Rights on privacy, that has been detailed even further in the International Covenant on Civil and Political Rights (ICCPR) multilateral treaty from April 1988[383] in the following way: 'Article 17 provides for the right of every person to be protected against arbitrary or unlawful interference with his privacy, family, home or correspondence as well as against unlawful attacks on his honour and reputation. In the view of the Committee this right is required to be guaranteed against all such interferences and attacks whether they emanate from State authorities or from natural or legal persons. The obligations imposed by this article require the State to adopt legislative and other measures to give effect to the prohibition against such interferences and attacks as well as to the protection of this right.'

---

[382] Data Protection and Privacy Legislation Worldwide, United Nations Conference on Trade and Development, https://unctad.org/page/data-protection-and-privacy-legislation-worldwide

[383] UN ICCPR Source: UN Human Rights Committee (HRC). 1988. CCPR General Comment No. 16: Article 17 (Right to Privacy), The Right to Respect of Privacy, Family, Home and Correspondence, and Protection of Honour and Reputation, https://www.refworld.org/docid/453883f922.html

29.9 The EU GDPR Regulation: Ambitious Data Privacy Standard and Model

In the context of the rising digitalization of the world, generation of data and related services, the European Union decided to upgrade in 2016, its previously existing Data Protection Directive from 1995 to adapt it to the new, evolving technological landscape, usages and challenges since the emergence of the internet and the cloud.

The General Data Protection Regulation became enforceable in May 2018 and has a primary focus on protecting personal data and privacy. It is considered as the toughest, most comprehensive and advanced data privacy and security law in the world.

Historically, the right to privacy is part of the 1950 *European Convention on Human Rights*[384], and the development of such legislation represents the natural manifestation of fundamental values that are common across European societies.

This regulation rules the individuals' controls and rights over their personal data, as well as data obligations for organizations if the data controller (the organization collecting the data), or the data processor or the data subject is based in the EU. As processors, the Cloud Service Providers operating in the EU and/or managing personal data of EU citizens leaving in the EU are fully subject to the law.

In a nutshell, the GDPR clarifies the following rights for the data subjects over their data: right to consent, right to be informed, right to access, rectify or erase their information, restriction of processing, data portability to list the main ones.

It also sets specific, new principles and requirements for the storage and processing of personal data. On top of implementing the three fundamental IT Security Management principles - Confidentiality, Integrity

---

[384] European Convention on Human Rights, Official texts, https://www.echr.coe.int/Pages/home.aspx?p=basictexts&c=

& Availability, two new principles with an impact on data storage are introduced: data minimisation and data storage limitation.

Data minimisation imposes the collection and processing of data only strictly necessary to fulfil the announced purpose. Data storage limits means that organizations should not keep data for longer than needed. Although there are no specific time references indicated within the regulation organizations have to create and implement a data retention policy and to perform periodic reviews to identify, and address, data stored beyond intended use.

Another particular key requirement is about pseudonymization of personal data, meaning de-identifying personal data, e.g. by replacing it by artificial identifiers. Encryption, anonymization and data obfuscation can be part of that same process.

Strong requirements in terms of global Cybersecurity and informing quickly and systematically data subjects and supervisory authorities in the event of data breaches are as well - among many more requirements - part of this very exhaustive framework.

Interestingly, GDPR enforcement is based on the principle of Accountability. Organizations have to take responsibility for how they process and store personal data. They have the obligation to demonstrate compliance, meaning keeping all the necessary records and document evidences to show that they have created, implemented and applied all the related policies, processes and routines to comply with the law.

For that matter, in the case of 'regular and systematic monitoring of data subjects on a large scale' it is even mandatory to designate a Data Privacy Officer (DPO) that will be in charge of overseeing the internal compliance with the law.

The last layer of the law is logically about fines and penalties, with non-compliant organizations or at the origin of data breaches, facing sanctions of up to EUR 20 million or 4% of their global sales, whichever is greater.

Overall, the very comprehensive scope of the GDPR, its forward-thinking principles and requirements, its very demanding enforcement rules and the underlying values that led to its writing actually represent an outstanding model to be replicated in other countries and parts of the world to care and protect individuals and their privacy.

Many of these principles and requirements can also be leveraged for inspiring general approaches to protect critical data, even non-personal data, and to build more resilient and better protected data systems and data storage systems. Following such a model even as an organization not subjected at all to the EU context would be the manifestation of a great ethical responsibility and respect towards users and stakeholders and becomes then truly part of the scope of Corporate Social Responsibility policies and strategies.

The fully detailed GDPR regulation and its related explanations are available on a specifically, dedicated EU portal[385]: including numerous relevant checklists[386].

29.10 Data Sovereignty

In a sensitive geopolitical context, competition between nations - in economics, politics and influence - is getting even fiercer in the era of digitalization and upcoming developments in data science and AI technologies, with sophisticated use cases expected to gradually and exponentially appear in the next 5 to 15 years.

Data is the fuel for all these technologies with ever-increasing impacts and becomes therefore a strategic asset for any organization and for any country. This confirms even more the status of data as the new, fourth production factor, as we have already seen.

With the possible reconstitution of geopolitical blocs with very opposed views of the world and even the come-back of the war and materi-

[385] Complete guide to GDPR compliance, https://gdpr.eu/

[386] The GDPR full text, https://gdpr.eu/compliance/

alisation of conflicts in various parts of the world, it becomes then a strategic imperative to protect the data of a state, of it's individuals, and of it's economic agents and institutions.

Data sovereignty was already and becomes then an even more critical question.

One of its translations is the legislative concept that data an organization collects, stores, and processes are subject to the laws, regulations and governance structures of the country where the data is collected, where the data is stored and/or relating to the nationals or organizations of that state. The notion is complex with various definitions, various scopes and various obligations linked to this type of laws according to the country.

The European GDPR regulation – already mentioned – is a major example of data sovereignty law as it rules the individuals' controls and rights over their personal data and data obligations for organizations if the data collector, or the data processor or the data subject is based in the EU. This also applies then to organisations based outside the EU if they collect or process personal data of individuals located inside the EU.

Another example, even anterior and with even more implications at global levels, is the US Patriot Act introduced in October 2001. Under this act, officials were granted possibilities of access to any information physically within the United States, regardless of the information's origin. The 2018 CLOUD Act, clarified and extended the possibilities of access by compelling 'U.S.-based technology companies via warrant or subpoena to provide requested data stored on servers regardless of whether the data are stored in the U.S. or on foreign soil.'[387]

This means that any information collected by an American server could also be potentially accessed, even, as specified, if the server was not located in the United States but owned, managed or handled by any

[387] Swire, Peter. 2018. The CLOUD Act and its Impact on Cross-Border Access to the Contents of Communications, https://www.alstonprivacy.com/cloud-act-impact-cross-border-access-contents-communications/

American organization somewhere along any part of the data collection, storing or processing chain.

This raises obviously questions and concerns, beyond the initial, appropriate purpose of this law, regarding potential direct or indirect abuses or misuses of such possibilities of accessing data for various reasons, and in particular economic competition. In that case how about the cloud storage offerings and the productivity tools in the cloud, proposed and hosted by providers originating from the United States?

These services offer tremendous possibilities in terms of storage, productivity and efficiency both to companies, organizations and individuals all over the world.

The EU GDPR actually regulates most of these questions, at least regarding data relating to individuals. This regulation also became a model for about a dozen other countries from various continents. In other countries there might be - or not – existing laws and regulations covering even partially those questions. If not, the respective Terms of Service defined by each provider along the data flow and storage chain do apply. While most of the time the data is encrypted from end-to-end with strong encryption technologies, the more secured practice for critical data and corporate clients is to manage encryption keys independently[388] from the cloud service provider where the data is located. The data owner owns in that case the encryption keys.

The localization of data becomes then an important factor, knowing in addition that not every country might have currently existing or sufficient local data centers capabilities installed and local cloud services providers.

On top of that the global concentration of cloud providers is quite impressive, with AWS (Amazon Web Services), Microsoft and Google

[388] Musthaler, Linda. 2013. Cloud encryption: control your own keys in a separate storage vault, Network World, https://www.networkworld.com/article/2170564/cloud-encryption-control-your-own-keys-in-a-separate-storage-vault.html

representing 66% of the global cloud spending[389] in the third quarter of 2022 coming from 61% one year earlier.

There are also so-called data localization or data residency laws, which are intended to keep in particular citizens' personal data in-country and subject to local regulation. This type of laws are often initially thought and enacted to protect citizens data privacy from other laws and regulations abroad that might not be as strict as the ones in the initial country or area. Data dissemination, in its broad meaning, and the fact that data is often replicated partially or totally across various places or data centers does indeed not help to ensure full transparency on what happens to this data and what are the related risks and consequences.

Some countries do however implement such laws for motivations relating more to economic protectionism, political reasons, or even directly – openly or not – surveilling their own citizens and exerting information and data control, including censorship. Anupam Chander even speaks of 'Data Nationalism'[390] (2015).

These laws may also create difficulties from various perspectives: conflict between political motivations and technical and operational efficiency (access, reliability, security, energy), conflicting or overlapping data sovereignity laws, complexifying the data flows internationally and thus complexifying the delivery of international services, up to feeding the 'ongoing struggle between democracy and totalitarianism'. In the latter case the 'business versus ethical' dilemma takes on its full meaning for global technology players in the field of data cloud storage and services when they operate in countries with lower standards in terms of democracy or human rights. What is acceptable, what is not? 'Should I

[389] Haranas, Mark. 2022. Top 5 Cloud Market-Share Leaders: AWS, Microsoft, Google In Q3 2022, https://www.crn.com/news/cloud/top-5-cloud-market-share-leaders-aws-microsoft-google-in-q3-2022

[390] Anupam Chander & Uyên P. Lê, Data Nationalism, 64 Emory L. J. 677, 2015, https://scholarlycommons.law.emory.edu/elj/vol64/iss3/2

stay or should I go?' Should we accept abusive requirements putting, directly or indirectly, fundamental human rights and freedom at risk?

This is why on top of national data privacy and protection regulations that need to become even more widespread and exhaustive in their content, and in order to avoid and manage conflicting and overlapping laws, working towards an international agreement on the matter is a vital task. Both for protecting citizens, human rights, organizations, and for providing a clear path to a safe technological and economical development based on data as a resource for progress.

29.11 Health Data Ethics: History of a Precursor Domain

Health Data is one of the most sensitive type of data to be collected, handled and stored. It touches directly the most intimate part of our physical incarnation as well as direct or indirect psychological and behavioural facets of who we are and what we are facing in this life.

Like all the other types of data, health related digitalized data, meaning patients and healthcare data up to clinical trials, genetic information and now also the new sphere of data generated by personal or medical connected devices and related applications, are exploding in terms of volume being generated.

This type of data is obviously highly sensitive and as such needs to remain fully confidential like the Hippocratic Oath already stated it very clearly 2.400 years ago: 'And whatsoever I shall see or hear in the course of my profession, as well as outside my profession in my intercourse with men, if it be what should not be published abroad, I will never divulge, holding such things to be holy secrets.'

The health sector has therefore been a precursor in many aspects of data protection and privacy.

From the Oath that many students are swearing to actual national and local regulations and policies relating to medical privacy, the health

sector is probably one that has among the highest sense of awareness and established processes in terms of data protection and careful use.

A list of some notable national health and medical privacy frameworks and initiatives of interest:

- The US HIPAA[391], Health Insurance Portability and Accountability Act of 1996, sets standards for uses and disclosures of *protected health information (PHI)*[392], and provides civil and criminal penalties for violations.

- The EU GDPR regulation classifies health data not only as a personal data, but as a sensitive personal data[393], among six other categories like e.g. ethnicity, religion, political opinions, genetic information, and biometric information about an individual. These sensitive personal data categories can initially not be recorded, treated and stored outside of a list of very specific exemptions, or the direct, explicit consent of the person. The list of exemptions relating to health data is obviously directly related to the operational needs for delivering medical and health services.

- NHS Digital, the IT & Data Department of the NHS (National Health Service) in UK applies also the GDPR and created purposely a widespread and systematic training program[394] requiring the NHS staff to complete appropriate annual data protection and security training and pass a mandatory test

- On top of applying the GDPR regulation as a member of the EU, France has developped an official mandatory governmental label,

---

[391] HIPAA Portal: https://www.hhs.gov/hipaa/index.html

[392] Protected health information, Wikipedia.org, https://en.wikipedia.org/wiki/Protected_health _information

[393] Art. 9 GDPR Processing of special categories of personal data, https://gdpr-info.eu/art-9-gdpr/

[394] Data Security Standard 3 - Staff training, https://digital.nhs.uk/cyber-and-data-security/guidance-and-assurance/data-security-and-protection-toolkit-assessment-guides/guide-3---staff-training

'Health Data Hosting'[395], to certify specifically each single health data hosting provider. It is based on the ISO 27001 standard requirements for Information Security Management Systems, with a 3-years duration and a yearly intermediate audit.

- France has also developed in 2019 the concept of a centralised, public 'Health Data Hub' (HDH)[396], storing or more precisely aggregating copies of many public health data sources for research purposes and the aim of improving the quality of care and treatment. The objective of the HDH is to 'enable project coordinators to easily access non-nominative, unified data on a secure platform, in compliance with regulations and citizens' right. They will be able to cross-reference and analyse the data in order to improve the quality of care and patient support'.

On this HDH platform, data sources are as diverse as hospital data, social security data, clinical data, death registers, public health data, national blood bank database, specific diseases monitoring, etc.

The idea is to leverage and cross-pollinate this precious data through exploratory analysis, data science statistical techniques and AI algorithms to foster medical innovation and discovery of hidden correlations or patterns, e.g. improving symptoms and diseases detection, providing answers to rare pathologies, predicting individual patient trajectories, improve pharmacovigilance and drug safety, etc...

The projects accessing and leveraging the data are listed publicly online in a fully transparent manner. 'The data, within a well-defined scope, are accessible to project coordinators contributing to the public interest, following an approval process involving in particular *a scientific committee, an ethical committee,* and the National Commission for Data Protection and Liberties (CNIL).'

---

[395] Health Data Hosting (HDS), Ministère de la Santé. Agence du numérique en santé, https://ue.esante.gouv.fr/information-systems-security-pre-condition-trust/health-data-hosting-hds

[396] Health Data Hub, https://www.health-data-hub.fr/page/faq-english

This really shows that it is actually possible with very strict protocols, processes and controls at the organizational, administrative and technical levels, from design to implementation, to both combine the realization of tremendous value, derived from massive, diverse sources of initially very sensitive data, and all this in the full and plain respect of regulation, numerous ethical considerations and individuals rights.

The very demanding level of requirements and specific characteristics in terms of data ethics and storage in the health sector could represent a model to inspire many other sectors, starting from raising the awareness very early in the education cursus to diligent rules and processes at all levels of the data collection, storage and processing chain, and all this under close monitoring and scrutiny by specific ethical committees. The idea is not necessarily to replicate 100% of all these practices but to have a high standard model that can be a starting point to set best practices within an organization, an industry, or a certain geography.

29.12 Data Governance: Improved Data Practices and New Ethical Developments?

In a context characterized by this exponential growth of data volumes, rapidly evolving technologies and use cases, the complexification of demanding compliance requirements, increasing pressure due to cyber-risks, and the multiplication of data storage types and locations, having a clear view and deep understanding of the overall data owned by an organization is not a luxury.

This is why Data Governance is of structuring importance.

Data Governance is commonly referred as a 'data management function to ensure the quality, integrity, security and usability of the data collected by an organization'[397] through its complete lifecycle. This

[397] O'Reilly. 201. Data Governance, the Definitive Guide.

involves also ensuring the compliance and conformity to regulations. This definition, its scope and reach is however still somehow loose and not yet completely mature and definitely shaped within the industry.

Being in complete awareness, knowledge, understanding and control of all the data owned, stored and managed by the organization should be a primary objective of any organization through it's data governance structures and policies.

This might be implied but is currently very often not sufficiently formalized and applied. And even less an actual reality within organizations. Establishing a plain, comprehensive, structured Data Governance (DG) function and program is still too often overlooked in many corporations.

According to a 2022 study[398], only '62% of organizations have had a data governance program in place for at least five years.'

Many organizations do not have complete, formal DG structures, roles, teams, policies. This creates all the conditions for possible lacks or gaps in terms first of accountability[399], and then of complete data understanding and mapping, leading in particular to loopholes in terms of correct and proper data categorization and protection.

Very often simple things like just having a complete, up-to-date data catalog mapping all the data of the organization, and data dictionaries specific to all databases and systems allowing the understanding and correct categorization of data are very rarely existing, up-to-date or complete. It is true that there are so many systems and data everywhere, in particular in big corporations that it becomes rapidly complex to implement and maintain such inventories systematically.

---

[398] The Strategic and Evolving Role of Data Governance in 2022 and Beyond, an Enterprise Strategy Group (ESG) Institute Study: https://www.mega.com/press-releases-strategic-and-evolving-role-data-governance-2022-and-beyond

[399] A data governance program must address accountability, in 'Designing Data Governance Structure', Journal on Computing (JoC) Vol.2 No.4, January 2013, http://dl6.globalstf.org/index.php/joc/article/download/576/2106

Very fortunately specific software solutions can help, as well as the application of the GDPR that forces many organizations that are concerned to rethink their data management policies and organization.

On top of creating and formalizing DG structures, roles and policies, like a Data Governance Council or Steering Committee, and formalizing further the usual data management roles from an IT perspective, it can make sense to create across the non-IT functions, departments and teams specific data stewards roles in charge of the very specific data linked to their daily operations, as they are in the best position to both understand this data and to control it. This is a 'simple' practice but still not sufficiently implemented. It must be a shared responsibility across departments.

Adding when possible an independent ethical committee in charge of advising the organization on the ethical dimensions of their ongoing operations or upcoming projects in terms of data handling and storage, beyond the pure compliance with existing regulations, is clearly another sound and recommended practice.

Considering Data Governance through an holistic and forward-thinking lens is also a great opportunity to add new responsible criteria and principles like Data Sustainability on the agenda in order to create and fully implement specifically dedicated policies (see subsequent section).

29.13 Data Long-term Sustainability

According to a report from the International Energy Agency (IEA), energy 'demand from data centers and data transmission networks accounted each in 2021 for 1 to 1.5% of global electricity use'[400]. These two elements are the main parts of the cloud services infrastructure

[400] Data Centres and Data Transmission Networks, EIA: https://www.iea.org/reports/data-centres-and-data-transmission-networks

chain. This equates to about 1% of global energy-related greenhouse gas emissions (GHG).

Despite the huge growth of data between 2015 and 2021 (data generation x5[401] and data storage x2.5[402]), data centers energy use is only estimated to have grown in the meantime by 10 to 60% (IEA). This is due to strong efficiency improvements relating to hardware, infrastructure and cooling optimizations as well as technical developments.

That being said, digitalization in general is both a source of CO2 emissions and at the very same time a part of the solution, as it also creates substantial savings through the avoidance of CO2 emissions.

Employees working from home (WFH) thanks to digitalization, with no commute necessary to work, can prevent the emission of about 270kg of CO2[403] on average a year per person for one WFH day a week up to 900 kg[404] a year for a full-time remote position according to various estimates.

Similarly '90% of business travel emissions (outside daily commutes) are from air'[405], meaning that using video-conferencing to replace short, medium and long-haul distance journeys whenever possible

---

401 Petroc Taylor, Volume of Data Created 2015-2025, Statista, https://www.statista.com/statistics/871513/worldwide-data-created/

402 Global Data Stored (IDC): https://www.businessinsider.com/sc/this-data-storage-solution-could-reduce-co2-emissions-2021-11?r=US&IR=T

403 French Environment Agency – 09-2020: https://presse.ademe.fr/wp-content/uploads/2020/09/ADEME_InfographieTT.pdf

404 See #7 in: Burgess, Andrea. 2023. Work from Home Models Predicted to Cause 34.3 Million Tons of Greenhouse Gas Emissionshttps://www.alliancevirtualoffices.com/virtual-office-blog/work-from-home-co2-emissions-statistics/

405 Carbon budgets for corporate business travel: https://www.reuters.com/business/aerospace-defense/corporate-business-travel-carbon-budgets-loom-airlines-2021-10-10/

can help to dramatically reduce GHG emissions by a factor of up to 66 in the case of large conferences[406].

However, at the same time data, its processing and its storage are going to continue their exponential growth. And as every single industry, data management and storage activities have to contribute to the overall 2050 net-zero objective set in the Paris COP 21 Agreement.

As part of their Corporate Social Responsibility policies (CSR), Corporations are increasingly defining carbon budgets and planning carbon reductions attached to their activities. This is where these efforts of reduction meet.

If we consider the top three Cloud Service Providers (CSP) – Amazon, Microsoft and Google – they have all pledged[407] to become plainly carbon neutral for their cloud activities between now and 2040 at the latest. This is happening in particular through powering their data centers with renewable or decarbonized energy, through continued optimisation of their efficiency and thanks to the purchase of RECs (Renewable Energy Credits).

Organizations in general, storing their data on their own on premise servers and data centers might find it easier to rely on CSP providers through their pooling of resources – when possible - to reduce the carbon footprint of their data and to mitigate related environmental impacts.

Another strategy to reduce drastically data center's carbon emissions is to move 'cold data' (data rarely accessed and not needing instant, fast

---

[406] Two studies: Lynch, Jim. Video conference CO2 emissions quantified in new study, University of Michigan Ann Arbor, https://news.engin.umich.edu/2021/02/video-conference-co2-emissions-quantified-in-new-study/; Elizabeth Claire Alberts-Mongabay, Making conferences virtual or hybrid could significantly mitigate climate change, says new study. 2022. Green News, Euro News, https://www.euronews.com/green/2022/01/12/making-conferences-virtual-or-hybrid-could-significantly-mitigate-climate-change-says-new-

[407] Amazon, Google, Microsoft: Here's Who Has the Greenest Cloud, https://www.wired.com/story/amazon-google-microsoft-green-clouds-and-hyperscale-data-centers/

access) to less energy-consuming tape drive storage technologies, that can cut emissions by up to 95%[408] as compared to HDD technologies. Too much data is stored indefinitely and is kept "Just In Case" (JIC) or turns to "Write Once Read Never" (WORN) data status.

Cold data could represent about 60 to 75%[409] of stored data according to various industry experts. This means that there is an opportunity to save much GHG emissions by implementing ambitious data retention and storage tiering policies with a focus on archiving cold data on tapes, and also simply deleting them definitely if there is no related compliance requirements, and if there are insufficient to no reasons to keep it.

Driving digital data decarbonization implies to go to the very roots of the question, meaning reducing the volume of data to be stored in the very first place. Upfront, during, and after processing.

Data collection minimization, data purposes limitation, data retention policies in particular are principles – already mentioned – that have their part in a pro-active manner of managing this question, beyound any compliance and regulatory aspect.

Regular data cleanings and reflections on which data to delete would become best practices. Adding more systematic data deduplication (eliminating duplicate copies of repeating data) and data compression techniques to this list and to the focus on decarbonating data centers infrastructures and there is an holistic Data Sustainability framework and vision, with both technological and organizational levers that is emerging.

Last but not least, we must all be aware, as individuals, that we are also fully responsible for the generation of data and this exponential

---

[408] Here's how data centers can become truly green. 2021. Fujifilm with Insider Studios, https://www.businessinsider.com/sc/this-data-storage-solution-could-reduce-co2-emissions-2021-11?op=1&r=US&IR=T

[409] Active Archives and the State of the Industry 2020. 2020. https://activearchive.com/wp-content/uploads/2020/06/AAA-Annual-Report-2020.pdf

growth that is happening. Each average user out of the 5 billion people worldwide using internet creates about 1.7 MB of data every second[410]. We shall therefore think, self-reflect and act about our own daily digital habits that are contributing to this increased data volumes. Keeping numbers of pictures and videos taken reasonable, using acceptable, eco-friendly resolutions when filming or watching videos online, cleaning our e-mail boxes, etc. will also concretely help mitigate the data storage related CO2 footprint.

29.14 The Future of Data Storage, the Permanence of Data and Related Questions

Current storage technologies can store data for up to 50 years at a maximum (and actually probably much less on average) and with energy consumptions that despite their optimisation and efficiency are still challenging given the continuous data explosion and while the planet is leaning towards 2050 zero-net GHG emissions targets.

In this context, one new – surprising - storage medium could represent a relevant long-term alternative, addressing both of these limitations: synthetic DNA. Encoding synthetic DNA with data and encapsulating it in silica could actually store up to 700 terabytes of information per gram of medium[411] and last for thousands to millions of years with little to no power consumption.

While the first experiments started in the 80s, based on concepts imagined in the 60s and are still being developed in laboratories necessitating expensive and advanced equipment, a DNA Data Storage Alliance has been formed in 2020 with the objective to 'jumpstart the standards

---

[410] Petrov, Christo. 25+ Impressive Big Data Statistics for 2023, https://techjury.net/blog/big-data-statistics/

[411] Isaacson, Betsy, Storing Digital Data for Eternity, Newsweek, 06/22/15, https://www.newsweek.com/2015/07/03/storing-digital-data-eternity-345557.html

development for DNA data storage' and the 'mission to create an interoperable ecosystem for DNA-based data storage solutions.'[412]

Low speeds of writing and reading, high cost levels and managing error-free data replication are still the main challenges to overcome though.

While getting such a technology mature and operational for everyday use would be an incredible progress in terms of keeping archived data safe from many risks of losses and dramatically lowering environmental impacts, the question of storing data permanently or for the eternity also raises very specific questions.

Protecting humanity's knowledge and memory in general, making contact with extraterrestrial life through time capsules, or preserving 'the most essential information to sustain or rebuild civilization in the event of an apocalypse (digital or other)'[413], these are some of very high-level projects and purposes for which such an almost endless retention of data can make sense without a doubt.

However, beyond those specific purposes, is it acceptable and does it make sense to retain data indefinitely or for too long? Even without going into the centuries duration that we have just mentioned, just considering the general life of an individual.

At a time when/where our digital footprint across all our devices makes it possible in a way or another to establish very detailed individual profiles of each of us, literally called digital twins, sometimes even updated in real-time, there is the question of not only collecting all this data in the first place, but also to keeping it and aggregating it indefinitely.

---

[412] SNIA Announces DNA Data Storage Technology Affiliate, SNIA, https://www.snia.org/news_events/newsroom/snia-announces-dna-data-storage-technology-affiliate

[413] Campbell-Dollaghan, K. 2016. We Have The Technology To Store Data For Eternity. Now What? https://www.fastcompany.com/3056762/we-have-the-technology-to-store-data-for-eternity-now-what

We all evolve through life and its cycles: childhood, adolescence, adulthood, senior stage. For digital natives in particular, born with digital devices all around and having been using them and exposed to them since sometimes very early ages, is the perspective acceptable to have their data, being aware of it or not, being aggregated, retained, creating a full historical archive and profile of their whole life? A 2021 French study[414] figured out that even 39% of babies had a digital footprint (although usually without a name and an account) *before even being born*, through their parents posting news and ultrasound scanning on social media during pregnancy...

Should we be defined, profiled, assessed, recommended, guided, advised, based on aggregated, persistent data, dating back sometimes to many years? Is there a risk in terms of data-based recommendations centered around either general demographics, and/or on past activities and centers of interests to potentially keep us circumscribed, trapped indefinitely in a given state or stage, self-reinforcing artificially some of our already existing patterns or biases? Is this generalized approach not narrowing down who we are, who we could become, limiting the perspective of getting exposed to ideas, concepts, and anything existing that is not yet in our direct environment of thought and life? Do we ultimately shape data? Or is data shaping us? That's the question.

And when you consider the question not only at the level of single individuals but at a generational level up to the overall society, is there not a risk to increase issues and problems like social reproduction, determinism, reinforced self-repeating patterns, communitarianism and related?

This brings the question down to Dataism, this mindset or philosophy, maybe even becoming kind of a religion for some, where data and

[414] See study: Sondage exclusif: la digitalisation de la vie familiale. Faireparterie, https://www.faireparterie.fr/etude-enfants-rapport-digital/

information flows become the 'supreme value'[415] by trying/attempting to characterize, understand and model absolutely anything in any field and dimension of life.

While aggregating and leveraging data to feed algorithms will undoubtedly allow major positive innovation and progress, providing more clarity, better understanding thus enabling more objectivity in many decision-making processes, we need at the same time to remain human-centered in our purposes.

As the quote says: "Not everything that can be counted counts, and not everything that counts can be counted." And what we must absolutely not forget to value in this increasingly technological and data-driven world are our qualities of humanness and humaneness, that make humanity and all of us intrinsically unique and our quality of being human unquestionably precious.

29.15 Conclusion

As we have seen data storage is a multi-faceted subject involving many various dimensions intertwined with ethical considerations: technology and technicalities, corporate organization, regulation and law, privacy, human rights, geopolitics, economics, health & data, IT governance, corporate social responsibility, sustainability, sociology…

The current possibilities created by the flexibility, simplicity, speed and sophistication of cloud storage and processing are tremendous. It is up to all of us to make the best usage of these technologies and of the precious fuel that powers them, data, for the highest purposes as well as the good of humanity while carefully protecting data itself and values like privacy, fairness, democracy, openness, transparency, equality of rights, and respect of all among others.

---

[415] Harari, Yuval Noah (2017). Homo Deus: A Brief History of Tomorrow. UK: Vintage Penguin Random House. p. 428. See also article in Christoph Stückelberger/ Pavan Duggal, Eds., Cyber Ethics 4.0, op. cit.

This responsibility has to be shared across all levels of stakeholders along the data storage and management chain.

It starts with the:

- CSP providers that are at the core of this industry to offer the highest standards for protecting data, privacy, equity of access, democratic behaviours and to minimize impacts on the environment, all in a full transparent manner
- Technologists, to design evolving technologies and systems that take into consideration the needs for privacy, for managing and storing data within ethical limits and in a sustainable way
- All organizations that generate, store and process data to do so in a responsible and focused way, minimizing their data needs and with a constant control of their operations to pro-actively prevent risks, and showing positive corporate moral responsibility in the field of data
- Government, lawmakers and authorities to regulate in order to avoid data privacy and protection loopholes on their territory, while not themselves legislating on data and data control for their own political interests
- International organizations and bodies to establish international regulation and principles to fight overall data and data storage inappropriate uses and practices and to articulate and resolve conflicting national laws in the most ethical manner
- IT and Data Professionals becoming fully aware of the complete spectrum of their responsibilities in terms of ethical data storage and data management practices, raising their level of professional excellence and professional ethics on this matter
- Universities and higher education institutions to equip systematically upfront their students in IT and Data specializations with dedicated ethical trainings
- Citizens to be aware and in control of their own data, to be vigilant about corporation practices and to act in a sustainable way

- All of the above to educate and educate plus update themselves on the matter of ethics and data

The overall realisation by everyone of the extent of their own responsibility, crossed with the pro-active consideration of the very diverse aspects and dimensions of ethical data storage and practices is the way leading to a safer, more human and more trusted technological development. Let's take this path all together!