Exploring Digital Sovereignty Through Data Flows : Empirical Evidence from the Backbone of the Internet

Enxhi LEKA*

March 19, 2026

Abstract

This study investigates the achievability of digital sovereignty within the European Union by examining website data flows, mainly focusing on non-personal data. Amid growing concerns over data governance and the resurgence of digital sovereignty as a central theme in EU policies, this research uniquely addresses firms' data storage decisions and their implications for EU users. Utilizing an original dataset of the most visited websites in France, this paper analyzes data location preferences across various sectors, revealing a complex interplay between privacy regulations, firm size, sector-specific tendencies, and the underlying internet infrastructure. The findings suggest that firms prioritize data storage in countries with strong privacy regulations and tend to locate data closer to consumers to minimize latency. However, variations are observed based on sector-specific needs and firm size, with larger and tech-oriented firms showing less sensitivity to distance. The study also highlights the significant role of the Internet's backbone infrastructure in shaping data storage strategies, pointing to potential challenges in aligning with digital sovereignty goals.

Keywords: Digital sovereignty, Cloud adoption, Firm location, Data flows

JEL Codes: L52, L86, O25

*enxhi.leka@imt-bs.eu
Institut Mines-Télécom Business School, 9, rue Charles Fourier - 91000 Évry, France.

1 Introduction

This paper examines the concept of digital sovereignty within the European Union, detailing the intricate landscape of website data flows and their implications for autonomy over digital information. By investigating how these data flows operate, the study aims to identify the challenges and opportunities associated with establishing greater control over digital assets in the EU. In the context of EU policies and data regulations, digital sovereignty has resurfaced as a central theme, drawing attention to data governance dynamics. Though it might appear novel, its roots trace back to historical EU data protection laws crafted to address concerns about information control. Digital sovereignty denotes Europe's capacity for autonomous action within the digital sphere. It encompasses defensive measures and proactive strategies aimed at promoting digital innovation, including collaboration with companies outside the EU (Madiega, 2020). Digital sovereignty is resurging with a focus on securing data governance, boosting EU tech sectors, and enforcing stringent regulations (Madiega, 2020). The crux lies in the EU's ability to regulate according to its values and rules, fostering a digital sovereignty space that stretches beyond data privacy to encompass various technological domains.

Although the market for personal data and its regulation has attracted considerable academic attention, there is a lack of focus on non-personal data flows and their significance in the European Union's quest for digital sovereignty. In this article, we aim to address this gap by posing the following questions. How do firms decide where to store their data for EU users? Do data storage strategies differ according to the type of data? How do different origin firms vary in terms of storage decisions? Does the current state of the Internet's backbone allow for a successful EU digital sovereignty? Our findings show that all firms tend to get as close to the consumer as possible in order to reduce latency. However, different sectors and firms of different sizes rely on the Internet's backbone on different levels. In addition, heavier and high-priority data tend to be stored closer to the consumer. Hence, certain data flows could be contained within Europe in this context. Nonetheless, Intenet's backbone infrastructure could be more difficult to fit within the sovereignty requirements.

Digital sovereignty in the European Union encapsulates a strategic response to the challenges of data governance, cybersecurity, and the preservation of autonomy in the face of technological

interdependence.[1] Initiatives to assert this sovereignty began with the Data Protection Directive[2] (Directive 95/46/EC adopted in 1995), which set the early framework for data privacy and protection across the continent. This directive laid the groundwork for the General Data Protection Regulation[3] (GDPR), which further empowers individuals with greater control over their personal data and serves as a foundational element shaping the region's digital governance framework. Furthermore, the Free Flow of Non-Personal Data Regulation[4] complements GDPR by ensuring the seamless movement of non-personal data across the EU, thus reinforcing digital sovereignty by eliminating data localization restrictions. This regulation is vital for a crucial approach to data mobility, enabling businesses and services to operate more efficiently across national borders. Moreover, the Cybersecurity Act[5] is crucial in fortifying Europe's digital sovereignty by establishing a cybersecurity certification framework for digital products, services, and processes. By enhancing digital infrastructure security and increasing trust in digital solutions, this act supports Europe's strategic autonomy in the digital domain. At its core, digital sovereignty in Europe involves the quest for greater control over data storage, processing, and flow within national boundaries. After the Schrems II[6] ruling in July 2020 declaring the Privacy Shield invalid, a second version[7] was adopted on July 10, 2023, aiming to accentuate further the importance of scrutinizing cross-border data transfers, compelling a reevaluation of data protection mechanisms and adequacy standards. In addition, the Data Act[8] and the Data Gover-

[1]Digital sovereignty for Europe. Available at: https://www.europarl.europa.eu/RegData/etudes/BRIE/2020/651992/EPRS_BRI(2020)651992_EN.pdf, last accessed December 8, 2023.

[2]Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31995L0046, last accessed December 8, 2023.

[3]General Data Protection Regulation GDPR. Available at: https://gdpr-info.eu/, last accessed December 8, 2023.

[4]Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union (Text with EEA relevance.). Available at: https://eur-lex.europa.eu/eli/reg/2018/1807/oj, last accessed December 8, 2023.

[5]Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2013 (Cybersecurity Act) (Text with EEA relevance). Available at: https://eur-lex.europa.eu/eli/reg/2019/881/oj, last accessed December 8, 2023.

[6]The CJEU judgment in the Schrems II case. Available at: https://www.europarl.europa.eu/RegData/etudes/ATAG/2020/652073/EPRS_ATA(2020)652073_EN.pdf, last accessed December 8, 2023.

[7]EU-US data transfers. Available at: https://commission.europa.eu/law/law-topic/data-protection/international-dimension-data-protection/eu-us-data-transfers_en, last accessed December 8, 2023.

[8]Data Act. Available at: https://digital-strategy.ec.europa.eu/en/policies/data-act, last accessed December 8, 2023.

nance Act[9] focus on enabling data to move freely within the internal market and maximize the value of data in the economy. As the European Union navigates the intricacies of interconnected digital services and platforms, legislative initiatives such as the Digital Markets Act (DMA)[10] and the Digital Services Act (DSA)[11] are poised to redefine the regulatory landscape, addressing challenges associated with data flows and setting the stage for a more secure and competitive digital environment.

In order to obtain information on the location of the server where the data is stored, we collect monthly data flows for the most visited websites from French users for 13 months, from November 2022 to November 2023. This list includes websites headquartered in France and abroad. We simulate a French user by visiting the first page of each website from Paris, France, and intercepting the data flow for each piece of data on the webpage.[12] Our dataset includes 341 websites headquartered in 32 countries, leading to 265,031 transactions. We define the origin of the website from the location of their headquarters. Websites are spread across 23 categories, such as e-commerce and shopping, and science and education. The data collection shows that the data is stored in 22 countries worldwide.

We are able to identify the location where each piece of data included in a given webpage is stored. Each website contains different types of data, such as text, media, or applications, that can be their own property or that of an external provider (such as web analytics or image banks). We consider the geographic distance from the consumer's location (Paris, France) to the data storage location as the primary explicatory variable to the choice of data centers. Our results show that data owned by an external provider is less distance-sensitive. This could be due to firm size; external providers are mainly large firms with several data center collaborations, so it is less costly to replicate the data and locate it as close to the consumer as possible. Due to the heterogeneity in firm size, the results remain valid when considering website size.[13]

As we are interested in studying the location choice for data storage focusing on data

[9]European Data Governance Act. Available at: `https://digital-strategy.ec.europa.eu/en/policies/data-governance-act`, last accessed December 8, 2023.

[10]The Digital Markets Act. Available at: `https://digital-markets-act.ec.europa.eu/index_en`, last accessed December 8, 2023.

[11]The Digital Services Act package. Available at: `https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package`, last accessed December 8, 2023.

[12]The flow of the request and response for each piece of data is called a transaction.

[13]We can approximate the website's size thanks to the Alexa ranking.

flows between the customer of the website and the data center, we apply a Poisson pseudo-likelihood regression model with multiple levels of fixed effects to evaluate the choice of data storage location made by Internet firms. Initially, data centers built takes into account many conditions, including data regulation (**?**). Likewise, data location implies security and privacy issues (Goldfarb and Trefler, 2018) that we attempt to control. Even though we are able to consider several controls and fixed effects, we might be unable to avoid identification issues. We collect the data monthly for the same websites, which should lead to the same storage location each month as we do not change the visiting location. However, we observe a variation in storage location for the same piece of data. This can be due to several reasons. First, high traffic toward the same server (DNS) requires the DNS load balancing procedure to distribute incoming network traffic or computing workload across multiple servers or resources to ensure optimal utilization, minimize response times, and prevent overload on any single server (Aversa and Bestavros, 2000; Hong et al., 2006). Second, different types of incidents, such as submarine cables or data center issues, can lead to the unavailability of the usual route (Hawkins et al., 2000; Javaraiah, 2011). For instance, a fire in one of the OVH data centers in 2021,[14] broken undersea cables in Africa,[15] and damaged ones in the Baltic Sea[16] in 2023. Under such conditions, a secondary backup data center or cables are used mainly in locations different from the initial ones.

This work builds upon prior literature and contributes to three streams of literature. Firstly, we build upon the scarce literature on digital sovereignty, which has recently highlighted the increasing relevance of this concept in economic and managerial contexts, particularly with the rise of decentralized technologies and data governance frameworks. Heess et al. (2025) propose a multi-agent approach for secure information flows in decentralized energy systems, emphasizing self-sovereign identities and zero-knowledge proofs as enablers of digital sovereignty in energy markets. Similarly, Hulkó et al. (2025) analyze how the European Union's legislation, including

[14]Millions of websites offline after fire at French cloud services firm. Available at: `https://www.reuters.com/article/idUSKBN2B20NT/`, last accessed December 8, 2023.

[15]Major African Internet connections disrupted as two vital undersea cables fail. Available at: `https://dig.watch/updates/major-african-internet-connections-disrupted-as-two-vital-undersea-cables-fail`, last accessed December 8, 2023.

[16]Sweden investigating damage to Baltic undersea cable. Available at: `https://www.bbc.com/news/world-europe-67138269`, last accessed December 8, 2023.

the Digital Markets Act and the Digital Service Act, shapes digital sovereignty policies and geopolitical and economic crises. Meanwhile, Kaur (2025) discusses India's shift towards digital sovereignty through data localization laws and the implications of the Digital Personal Data Protection Bill on economic competitiveness. Jamshed et al. (2025) propose a blockchain-based identity management framework to reinforce data sovereignty and security within digital asset management and IoT applications. Additionally, Babkin and Shkarupeta (2024) explore the role of technological sovereignty in the evolving data economy, advocating for multi-agent systems to drive economic and industrial transformation. Further, Fang et al. (2024) examine privacy-enhanced distributed identity management systems, demonstrating how self-sovereign identity frameworks can enhance user control over digital identities in cloud computing environments. Digital goods, such as data, are bound to be traded and subject to international exchange. Even though digital trade has transportation and search costs as well as almost null traditional barriers, it is impacted by distance, similarly to traditional goods trade (Blum and Goldfarb, 2006). In addition, Goldfarb and Trefler (2018) and Sun and Trefler (2023) look at the transformative impact of AI on international trade in digital goods and services. This impact is highly relevant to economies of scale, knowledge externalities, and the need for nuanced regulations to navigate this evolving landscape. Data flows raise questions on technology sovereignty. Edler et al. (2020) define technology sovereignty as the ability of a country to provide the technologies it deems critical for its welfare, competitiveness, and ability to act and to develop these or source them from other economic areas without one-sided structural dependency. While EU digital sovereignty efforts emphasize stricter regulations to encourage local data storage, economic incentives play a key role in shaping firm behavior. Studies on access regulation in network industries suggest that setting lower access barriers can delay infrastructure investment, as firms find it more cost-effective to rely on existing systems rather than build independent networks (Bourreau et al., 2014). Data centers can be located worldwide, resulting in different location options at different price points for international clients. This scenario results in country interdependence frameworks presenting the close relationship between geopolitical risks and technology (Khan et al., 2022; da Ponte et al., 2023; Edler et al., 2020). We contribute to this work by highlighting the role of website data flows specifically for EU users in digital

sovereignty.

The second stream of literature is related to cloud industries. A study by Brynjolfsson and McElheran (2016) finds that cloud computing adoption positively impacts firm productivity, with companies that adopt cloud technology experiencing an increase in output and revenue growth. Another study by Melville et al. (2004) finds that cloud computing adoption leads to higher IT flexibility, which results in higher business agility and productivity. Nevertheless, the impact of cloud usage on firms is unequal in size. In addition, studies do not necessarily agree on their findings. On the one hand, Jin and McElheran (2017) argue that older firms show a deficient level of benefit from IT services, whereas young firms benefit more in survival, growth, and performance. Similarly, Khayer et al. (2020) explore the factors that influence the adoption of cloud computing in small and medium enterprises (SMEs) and investigate how cloud adoption affects SMEs' performance. The study uses a dual-stage analytical approach combining structural equation modeling and artificial neural networks. It identifies several critical predictors of cloud adoption, including relative advantage, service quality, and top management support. The study also confirms the positive impact of cloud adoption on firm performance and suggests that managerial actions should focus on improving perceived risk, relative advantage, and top management support. Similarly, while focusing on SMEs, Khayer et al. (2021) find that cloud adoption positively influences firm performance directly and through organizational agility. On the other hand, a study using the 2018 Annual Business Survey (ABS) conducted by the US Census Bureau and the National Center for Science and Engineering Statistics (NCSES) shows that the adoption of advanced technologies is rare and skewed towards larger and older firms, while digitization and some use of cloud computing are widespread (Zolas et al., 2021). Furthermore, the DeStefano et al. (2020) investigates the effect of cloud computing on firms through the lens of geographic organization. Their results show an impact on incumbent firms by decentralizing their organizations and increasing their geographic spread. However, they do not find similar results in young firms. Nevertheless, Haug et al. (2016) highlight the importance of analyzing the effects of cloud computing at the industry level, the type of output firms produce, and the market in which they operate. For instance, by using a value relevance model, a study finds that an unanticipated increase in the share of a firm's revenues from cloud computing

has a positive effect on excess stock returns and a negative effect on idiosyncratic risk (Nezami et al., 2022). Yet, the effects vary depending on market structures and firms, with increases in market maturity intensifying the positive effect of moving to the cloud on excess stock returns and increases in advertising intensity strengthening the negative effect of shifting to the cloud on idiosyncratic risk. Cloud computing is transforming the digital economy, creating new challenges related to market concentration, interoperability, and data governance. While firms and policymakers increasingly focus on digital sovereignty, the economic literature on cloud computing remains relatively sparse, as highlighted by Crémer et al. (2024), who emphasize the need for further research into its competitive dynamics and policy implications. Our paper extends this analysis by studying cloud usage in the website market, specifically which categories are more prone to use cloud storage rather than on-premise technologies.

The third stream of literature focuses on firm location decisions and, specifically, data storage location. On the one hand, data center providers base the decision of location choice on multiple criteria such as the population, qualified labor, home values, land prices, climate conditions, and electricity prices (**?**). This choice can also be based on other factors such as historical factors (proximity of the headquarters with the first data center providers), power and fiber network (types of energy, fiber grid, proximity to undersea cables and Internet Exchange Points), and governmental policies (incentives, tax exemption). Similar to telecommunications network investments, data storage decisions depend on both regulatory constraints and economic incentives. As Bourreau et al. (2014) illustrate, regulatory interventions must carefully balance access conditions with long-term investment incentives to avoid excessive reliance on existing infrastructures. On the other hand, firms build strategies for choosing the best location for their needs, whether dealing with personal or website data. Contrary to common belief, Blum and Goldfarb (2006) find that digital goods such as data are not free of trade costs. Hence, data localization greatly impacts cost constraints, privacy, and security issues (Goldfarb and Trefler, 2018). In addition, it implies direct and significant effects on cross-border data flows (Svantesson, 2020). n this context, Wang et al. (2025) employ a Negative Binomial model to examine the spatial distribution of digital enterprises, showing that firms cluster in regions with strong digital infrastructure, market accessibility, and agglomeration economies. Their findings suggest that

digital firms' location choices are shaped by a mix of economic and technological factors, reinforcing the importance of digital infrastructure and connectivity in shaping firm behavior. The literature has documented this question from a personal data perspective. Rochelandet and Tai (2016) study Internet firms and show that the stricter the regulation of a country regarding data collection, the less attractive it is for firms to locate their businesses. However, to our knowledge, there is limited academic inquiry on website data flows. Our paper builds on this literature by trying to uncover the storage location decisions for Internet firms.

Our results show that the distance of the website customer to the data storage location is negatively associated with the choice of location. Furthermore, large and tech firms are less sensitive about distance. In terms of type of data, large and high-priority data are stored closer to the consumer to offer the shortest latency. Even though 68% of the websites are from the EU, the infrastructure of the Internet's backbone is dominated by US firms. These findings suggest that digital sovereignty has favorable ground regarding website data storage but less regarding infrastructure elements.

The rest of this chapter is structured as follows. Section 2 describes the Internet's backbone infrastructure. Section 3 presents the data sources. Section 4 presents the dataset and variables of interest. Section 5 depicts the empirical specification used throughout the study and the main results. Section 6 supports the results through robustness checks, and Section 7 concludes.

2 Internet's Backbone Infrastructure

The Internet backbone is a critical component of global communication infrastructure, comprising a network of high-capacity data routes and core routers that interconnect various large-scale networks. These backbones are typically fiber optic lines that span significant distances and are capable of carrying large amounts of data. This "network of networks" has seen considerable evolution, especially in its relationship with urban hierarchies and the expansion of fiber-optic technologies. Malecki (2002) highlights the emergence of the Internet as a defining technology of the 21st century, emphasizing its rapid development and its impact on urban areas. The Internet backbone has shown a clear bias towards world cities, serving as critical hubs in the global network (Malecki, 2002). O'Kelly and Grubesic (2002) provide insight into the spatial

organization of commercial internet backbones, which reflects a competitive, privatized market for service provision. Their work underscores the importance of the backbone's infrastructure in facilitating digital information transport across locations. Grubesic and O'Kelly (2002) further explore the accessibility of commercial Internet services, focusing on the role of fiber-optic backbone points of presence (POPs) established by internet service providers. They highlight how larger metropolitan areas maintain dominant shares of telecom infrastructure, which influences the accessibility of various regions. Knieps (2003) discusses the periphery of the Internet and the provision of Internet services, particularly focusing on Internet access and the backbone. This paper sheds light on the evolving dynamics of internet access technologies and the internet backbone's role in this context. Riezenman (2001) addresses the increasing traffic on the internet backbone and the adoption of new technologies to enhance the capacity of fiber-optic systems. This work is crucial in understanding how the backbone infrastructure has adapted to the growing demands of Internet usage. Lastly, Labovitz et al. (1999) provide an experimental study on the stability of major internet paths and backbone failures, offering valuable insights into the reliability and performance of the Internet backbone. The Internet backbone is an intricate and dynamic system, crucial for global connectivity and digital information flow. Its development and management have been influenced by various factors, including technological advances, market competition, and urban development.

The seamless flow of data across the Internet's backbone is critical. This infrastructure, comprising interconnected components such as fiber-optic cables, hosts, routing equipment, network nodes, Internet Service Providers (ISP) organizations, data centers, and end-users, forms the bedrock of the Internet backbone. The backbone's foundation is its extensive network of fiber-optic cables, vital for high-speed data transfer. These cables are intricately connected through a sophisticated array of routers and switches. O'Kelly and Grubesic (2002) provide a detailed examination of the spatial organization and accessibility of commercial fiber-optic backbones, a crucial element in understanding the backbone's capacity for digital information transport. Additionally, Sengupta et al. (2003) discuss the architecture of optical backbones, emphasizing the importance of scalable and flexible interconnections in core IP networks. Data centers are integral to the backbone, hosting server farms for services like cloud computing and

web hosting. Their strategic placement is vital for minimizing latency. Network nodes, particularly internet exchange points (IXPs), manage the routing of data across networks. Grubesic and O'Kelly (2002) delve into the role of fiber-optic backbone points of presence (POPs) in enhancing the accessibility of cities to the Internet, highlighting how these nodes facilitate traffic management and data accessibility. ISPs are crucial in connecting users to the backbone, managing the delivery of data from end to end. Host organizations, such as corporations and academic institutions, often maintain their dedicated network infrastructures. Castillo-Velázquez and Delgado-Villegas (2020) highlight the critical role of ISPs in backbone network services, pointing out the complexities involved in managing such networks. The backbone's resilience and security are paramount. Guven et al. (2019) examine the vulnerabilities in backbone infrastructure, emphasizing the importance of robust security measures to safeguard against threats. Furthermore, the work of Markopoulou et al. (2008) offers an analysis of failures in an operational IP backbone network, underlining the necessity for reliable infrastructures to maintain internet dependability.

3 Data Collection

We rely on website and Internet traffic data to study the factors impacting firms' decisions on data storage. This paper collects data on websites in France using a combination of data collection and open-source methods. The dataset comprises three sources: website-related data, transaction-related data, and country-related data.

The first data source is website-related data collected in November 2019 through data collection or open-source methods. The dataset includes the website category from SimilarWeb. The Alexa website provides audience measurement information such as the ranking for French users, daily page views per visitor, daily time spent on the website, and the traffic rate from search and ranking.[17] Moreover, to determine the website's origin, we collect the city and country location of the headquarters for each website.

The second data source comes from a data collection process based on the Alexa ranking

[17]The Alexa ranking has no longer been available since May 2022.

of the most visited websites from France.[18] Based on these websites, we run a monthly data collection process from November 2022 to November 2023 consisting of interception data flows. For every website, we visit its first page from Paris, France, to simulate an EU user and stay for 30 seconds to avoid uploading issues. Each transaction consists of a request and a response stage. The data collection program specifically intercepts each transaction's request and response stages. These two stages bear all the information and code needed to display a website. Through intercepting data flows, we can distinguish different types of data (objects) relative to every transaction, such as the code of the website itself, image, text, advertising plug-in, fonts, etc. Appendix A shows an example of a website and the data it contains. Moreover, the data collection provides information such as the owner of the transaction, which can be the website itself, an advertiser, a music bank, or an image bank. In addition, we collect all the information relative to the request and response content, providing our dataset with the specificities of each object, such as the domain name, the uploading (the time between the request and response stage) time, internet service provider (ISP), the organization as well as all data on the storage location as the city, country, zip code, latitude, longitude.

The third source of data is country-related data. We collect open-access data relative to the country, such as the number and location of Internet exchange points (IXP) from the World Bank Group,[19] privacy regulation from the Commission Nationale de l'Informatique et des Libertés (CNIL),[20] and profit taxes from Trading Economics.[21]

To ensure the validity of the dataset, we dropped all websites with under 13 transactions (corresponding to the lower percentile) and those that gave inconclusive responses as we considered that the data collection had not worked or the website had not uploaded correctly. In addition, the websites for which the geographic location could not be determined or for which other information from the data collection was missing are dropped. This gives a total of 341 websites and 265,031 observations.

[18]The list of the websites is presented in Appendix B

[19]Internet Exchange Points. Available at: https://datacatalog.worldbank.org/search/dataset/0037932, last accessed January 15, 2023.

[20]Data protection around the world. Available at: https://www.cnil.fr/en/data-protection-around-the-world, last accessed January 15, 2023.

[21]Taxes On Income, Profits And Capital Gains (% Of Revenue) By Country. Available at: https://tradingeconomics.com/country-list/taxes-on-income-profits-and-capital-gains-percent-of-revenue-wb-data.html, last accessed January 15, 2023.

4 Description of the Dataset

The dataset includes data on the countries where data comes from, such as privacy regulation, political stability, sales tax, and the number of Internet Exchange Points (IXPs). In addition, we have website-related data such as the category, the website's headquarters location, and ranking. Last, we cover information on the data type (media content, application, or text), the uploading time, and the data's owner for each transaction.

4.1 Website Data

The 341 websites of the dataset are listed under 23 different categories[22], such as Art and Entertainment, News and Media, Finance, etc. Table 1 shows the distribution of websites and transactions across the categories, and Table 15 in Appendix B lists the sample of websites used in the study for each category. Furthermore, by identifying the location of the headquarters for each website, we define their origin. Headquarters are located in 32 countries with 196 websites from France, 63 from the USA, 16 from Russia, and 66 from other countries. Table 2 shows the distribution of websites and transactions according to the headquarters locations and Table 3 shows the distribution of the transactions between EU and US websites.

4.2 Transaction's Location

The dependent variable, *Transaction's location*, measures the number of transactions of a given category coming from the same country, meaning the number of pieces of data stored in the same country. By intercepting data flows, we can pinpoint the exact location through transaction information such as the city, country, longitude, and latitude. Table 4 presents the descriptive statistics for this variable, showing that on average 475.72 pieces of data are stored in the same country. However, the data for transaction locations are widely dispersed, with considerable variation among individual observations.

We find thus data coming from 22 countries and 233 cities. Figure 1 depicts the distribution of the transactions across countries and Figure 2 the statistical distribution of the dependent

[22]Website category data comes from SimilarWeb.

	Websites	Transactions
Adult	10	11,689
Arts and Entertainment	38	25,103
Business and Consumer Services	9	7,201
Community and Society	1	444
Computers Electronics and Technology	68	43,839
E-commerce and Shopping	16	26,734
Finance	22	17,006
Food and Drink	11	8,171
Gambling	2	2,058
Games	15	11,968
Health	6	4,516
Heavy Industry and Engineering	3	2,178
Home and Garden	7	5,770
Jobs and Career	4	2,932
Law and Government	10	8,346
Lifestyle	9	6,920
News and Media	27	18,687
Pets and Animals	2	2,490
Reference Materials	13	9,376
Science and Education	49	35,297
Sports	6	6,113
Travel and Tourism	6	4,346
Vehicles	4	3,847
Total	341	265,031

Table 1: Distribution of websites and transactions over the categories

variable.

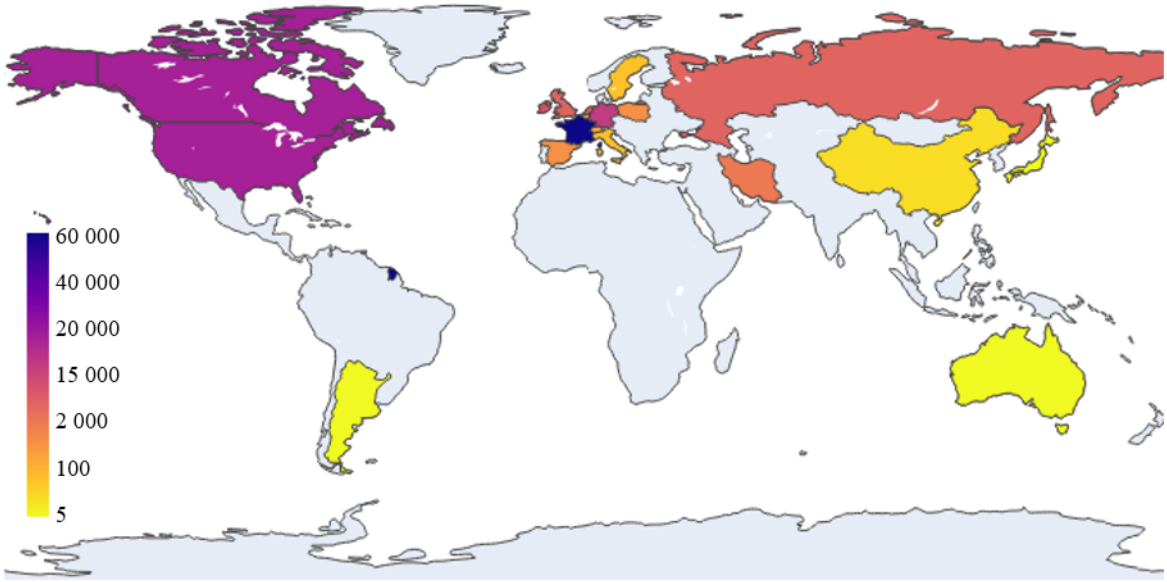

Figure 1: Storage location of the transactions (the legend depicts the number of transactions).

	Websites	Transactions
Armenia	1	2,635
Belgium	1	1,178
Canada	1	1,299
China	6	6,438
Cyprus	4	2,985
Czech Republic	1	1,214
Côte d'Ivoire	1	759
Denmark	1	983
France	196	146,249
Georgia	1	391
Germany	5	2,767
Iran	7	4,579
Israel	1	370
Italy	1	880
Lithuania	1	1,035
Luxembourg	2	544
Malta	1	1,748
Mexico	1	747
Morocco	1	928
Netherlands	5	3,721
Norway	1	1,160
Poland	1	373
Russia	16	15,719
Saudi Arabia	1	1,231
Singapore	1	971
South Korea	1	1,178
Spain	4	3,150
Sweden	2	1,374
Switzerland	5	2,013
USA	63	45,441
Ukraine	1	1,206
United Kingdom	7	9,765
Total	341	265,031

Table 2: Distribution of websites and transactions over the headquarters country

	Mean	Std.Dev.	Min	Max	Count
Headquarters					
French website	0.55	-	0	1	265,031
EU website	0.68	-	0	1	265,031
US website	0.17	-	0	1	265,031

Note: The variable EU website includes French websites.

Table 3: Descriptive statistics of transactions according to the headquarters

	Mean	Std.Dev.	Min	Max	Count
Dependent variable					
Transaction's location	475.72	455.09	1	1,820	265,031

Table 4: Descriptive statistics of the dependent variable

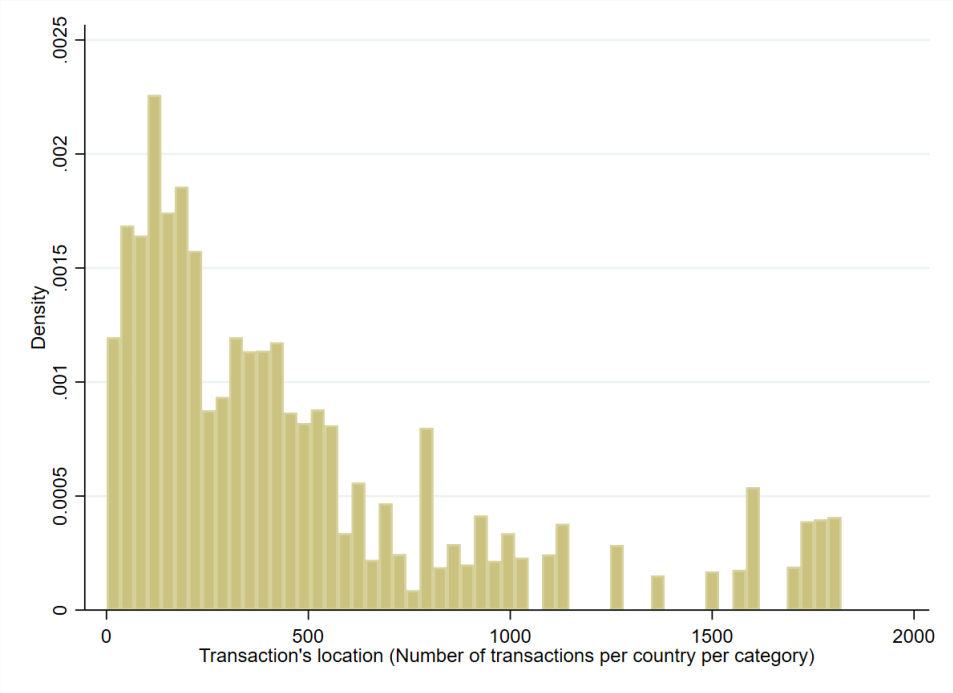

Figure 2: Statistical distribution of the dependent variable.

4.3 Other Variables of Interest

Each transaction, from the request to the response, represents the route of each piece of data on the website. We will categorize the data according to their into three groups: applications, media content (images, audio and videos) and texts.[23] We create a dummy variable to identify each type of piece of data.

Built upon prior academic works, the weight of the data is a key factor in the uploading time. Thus, we categorize the data into heavy or not through the dummy variable *Heavy data*. Besides their content, websites can use third-party providers for services such as picture banks, advertising, traffic, etc. In addition, identifying the data provider makes it possible to identify

[23]The *Application* type means that data being transferred is in a structured format used for application-specific purposes such as javascript, json, woff, woff2. *Images, audios and videos* include various types of multimedia, either visual or auditive data that can be either visible or invisible on the website, such as png, jpg, webp, gif, mp4, mpeg. *Text* data encompasses various types of plain text data that can be transferred over HTTP, such as css, javascript, html.

the decision maker regarding storage. The variable *External Provider* is a dummy equal to one when indicating that the data owner is an external provider and not the website itself. Table 5 shows the statistics for these data.

	Mean	Std.Dev.	Min	Max	Count
Transaction related					
Image/Video/Audio	0.44	-	0	1	265,031
Application	0.38	-	0	1	265,031
Text	0.19	-	0	1	265,031
Uploading time	0.82	0.72	0.004	17.87	265,031
Heavy data	0.44	-	0	1	265,031
External provider	0.34	-	0	1	265,031

Table 5: Descriptive statistics at the transaction level

4.4 Explanatory Variables

As mentioned in the previous section, we visit the websites using a virtual machine located in Paris, France, for which we have the longitude and latitude. Using the same information on the server location for each piece of data, we are able to calculate the *Exact distance* between the storage location and the simulated visitor. Aside from distance, privacy regulations can also significantly influence firms' data storage decisions, particularly concerning personal data encompassing customer and employee information. Adhering to regulations like GDPR and HIPAA prevents penalties, fosters trust, and upholds reputation. This adherence ensures continuity in global operations and maintains positive public perception. Ethical data use, prioritized for data-driven growth, enhances customer loyalty. Moreover, economies of scale can be achieved by efficiently storing personal and non-personal data in the exact location and streamlining operations while maintaining rigorous privacy measures. The variable *Privacy regulation* is a dummy equal to one if, according to the CEPII database, there is regulation dedicated to privacy and zero otherwise. In our data, 60.74% of transactions come from EU of EEA member countries; 15.18% from authority and law(s); 14.02% come from partially adequate countries; 4.94% come from adequate countries; 3.79% from countries having data protection law(s); 1.33% from countries having no specific law.

The existing literature underscores the substantial influence of political stability on firms'

location decisions (Alesina and Perotti, 1996; Henisz, 2000; Besley and Burgess, 2002; Rodrik et al., 2004; Javorcik, 2004; Melitz and Ottaviano, 2008; Berman and Couttenier, 2015). Numerous studies have examined the link between political stability and localization choices, including government stability, regulations, and political risks. Research reveals that political stability fosters an environment conducive to business growth and investment, reducing uncertainties and risks associated with storage decisions. Firms favor politically stable environments with clear legal frameworks, predictable regulations, and low corruption, which enhance operations and profitability. Political stability encourages firms to invest in storage facilities such as data centers, driving economic development, job creation, and productivity. It attracts foreign investment, improves logistics efficiency, and enhances overall competitiveness, benefiting both firms and society. Additionally, secure storage locations bolster supply chain resilience, ensuring vital goods and services during crises. The variable *Political stability* is an ordinal variable from 1 to 4, representing weak to strong political stability, respectively. This information is provided by CEPII.

Even though sales taxes might not directly impact storage location choices it is taken into account by the data center providers, which, among other factors (**?**), can impact the price of data storage. Many studies have demonstrated that higher sales tax rates tend to discourage firms from choosing certain storage locations, particularly in regions with higher tax rates. This suggests that sales tax serves as an essential determinant in the location choice of firms, with lower sales tax rates being more favorable (Kolko, 2000). Additionally, the research shows that the impact of sales tax on the probability of a firm choosing a specific location is not uniform across industries (Devereux and Griffith, 1998; Newman and Sullivan, 1988). Studies have found that industries with higher profit margins are more sensitive to changes in sales tax rates regarding location decisions (Bartik, 1989). This indicates that the effect of sales tax is contingent upon the specific characteristics of the industry in question. Furthermore, regional disparities in sales tax rates can also influence firms' location preferences, with lower-tax regions being more attractive for businesses (Kemsley, 1998). From Trading Economics, we retrieve the continuous variable *Sales tax*, which gives the sales tax applied to firms in different countries.

The literature on internet exchange points' impact on firms' storage location choices of-

fers insightful perspectives on digital infrastructure's role in shaping business decisions. This research holds societal significance by unveiling how technology infrastructure and economic activities intertwine, moulding the digital landscape's effects on businesses and individuals. Firstly, studies emphasize the positive link between the number of Internet exchange points (IXPs) and firms' choice of storage location. IXPs serve as vital internet hubs, enhancing data exchange between networks. More IXPs in a region create a robust digital ecosystem, offering improved accessibility, lower latency, and higher bandwidth. This attracts businesses to these locations, fostering local economies and spurring innovation. Moreover, the literature suggests that clustering multiple IXPs nearby generates significant agglomeration effects, amplifying storage location appeal. Network effects, economies of scale, knowledge sharing, and collaboration opportunities contribute to this phenomenon. Concentrated IXPs stimulate competition, technological progress, and synergistic firm relationships, advancing economic development. The variable IXPs expresses the number of Internet Exchange Points in a country. These facilities tend not to be equally spread across a country, which might misrepresent the results. Hence, we include the *weighted IXP* by the country's surface in the model.[24] Table 6 presents the descriptive statics for these variables and Table 7 presents the correlation matrix between them. The sources of all the variables used are presented in Table 16 in Appendix C.

	Mean	Std. Dev.	Min	Max	Count
Explanatory variables					
Exact distance	2,353.86	2,959.97	1	16,959	265,031
Privacy regulation	0.99	-	0	1	265,031
Political stability	3.93	0.37	2	4	265,031
Sales tax	14.53	8.40	0	25	265,031
IXP weighted	144,424.04	245,259.35	104	705,696	265,031
Robustness checks variables					
Distance between capitals	2,268.32	2,526.63	262	16,938	265,031
Distance between most populated cities	2,218.21	2,550.52	262	16,975	265,031

Table 6: Descriptive statistics of the explanatory variables

[24] IXP weighted = $\frac{\text{Surface of country i}}{\text{Number of IXPs in country i}}$

	Transaction's location	Exact distance	Privacy regulation	Political stability	Sales tax	Ixp weighted
Transaction's location	1					
Exact distance	-0.266	1				
Privacy regulation	0.104	-0.0747	1			
Political stability	0.0976	-0.00678	-0.0228	1		
Sales tax	0.204	-0.960	0.0920	-0.125	1	
IXP weighted	-0.139	0.559	-0.141	-0.172	-0.493	1
N	265,031					

Table 7: Correlation of the explanatory variables

5 Empirical Analysis

Based on the existing literature on the optimization of network latency and server location, which highlights the crucial role of physical distance between the user and the server (Lakhina et al., 2002; Draves et al., 2004; Dhakal et al., 2007; Akhoondi et al., 2012; Guo et al., 2015); the closer the server is to the user, the lower the latency; we will compute this distance for our data. In this paper, we employ a Poisson pseudo-likelihood regression model with multiple levels of fixed effects to study the factors influencing the number of transactions of a given category in a given country. The choice of this model is driven by its suitability for count data, where the dependent variable inherently follows a discrete and non-negative distribution (Figure 2). The Poisson distribution assumption, where the mean and variance of count data are equal, aligns well with our research context of studying transaction counts. This model allows us to investigate the impact of various explanatory variables on transaction frequency while accounting for potential overdispersion. Further, we employ a Poisson Pseudo-Maximum Likelihood (PPML) estimator instead of Ordinary Least Squares (OLS) due to its superior handling of count data and heteroskedasticity. OLS, when applied to log-linear models, can produce biased estimates in the presence of heteroskedasticity (Silva and Tenreyro, 2006). Moreover, PPML can handle zero observations without requiring ad hoc transformations (Correia et al., 2020). The dependent variable is the number of transactions q of a given category c in a given country i at period t. The Poisson pseudo-likelihood estimation assumes that $q_{i,c,t}$ follows a Poisson distribution with a mean $\lambda_{i,c,t}$, which is determined by explanatory variables $X_{i,c,t}$ and model parameters β:

$$q_{i,c,t} \sim \text{Poisson}(\lambda_{i,c,t}) \tag{1}$$

The mean $\lambda_{i,c,t}$ is modeled as:

$$\lambda_{i,c,t} = exp(\beta X_{i,c,t} + \gamma_c + \mu_s + \delta_t) \tag{2}$$

Where $X_{i,c,t}$ is a vector of explanatory variables capturing characteristics such as physical distance between users and servers, regulatory indicators, tax rates, and digital infrastructure; γ_c, μ_s, δ_t are fixed effects for category, website, and time period, respectively, allowing us to control for unobserved heterogeneity across these dimensions. Country-level variables are included directly in the model and are identified through variation across countries and over time. Through this methodology, our study aims to gain insights into the determinants of transaction frequency in different countries and categories, providing an understanding of transaction behaviors within an economic context. The Poisson pseudo-likelihood regression with multiple levels of fixed effects aligns with the specific characteristics of our data and research objectives, enabling us to make robust inferences about the factors influencing transaction patterns in different countries for the chosen category.

5.1 Main Results

Table 8 depicts the results for the determinants of the location choice for storage. The dependent variable is the number of transactions of a given category in a given country. Column (1) shows the results for the exact *Distance* between the user's and the server's location. In line with the expectations, the coefficient is negative and statistically significant at the 1% level. The longer the distance between the user and the server in a specific country, the less data will be stored in this country. Column (1) includes website and period fixed effects.

Progressively, we include other control variables to reduce the biases on the magnitude of the distance. Column (2) adds the dummy variable *Privacy regulation* with a positive and statistically significant coefficient at the 1% level. As expected, data are likely to be stored in countries with privacy regulations. Even though the dataset does not include personal data, firms

might tend to store all of their data in the exact location and not differentiate by the sensitivity of the data but take advantage of the economies of scale. By adding other control variables through Columns (3) to (5), the coefficient for privacy regulation will increase but remain consistently positive and statistically significant.

Column (3) adds the variable *Political stability*, which is an index going from 1 to 4, with 4 being a strong level of political stability. Unlike previous literature (Koop and Tole, 2008; Rochelandet and Tai, 2016) looking at firms' location decisions, the positive and statistically significant coefficient does not show a trade-off between privacy regulation and political stability. A higher level of political stability ensures firms the safety of their servers and data, leading to a more considerable amount of data stored in the country. In Columns (4) to (5), this coefficient gets stronger from 0.203 to 0.266 and stays statistically significant.

Column (4) adds the *Sales tax* measures the local taxes. Data centers are subject to frequent turnovers in servers, which can impact the price of the services procured by the websites. The coefficient is negative and statistically significant; the higher the sales taxes, the fewer transactions will come from the country. The coefficient gets weaker even though the sign and statistical significance are consistent through regression (5).

The presence of IXPs is a sign of a highly developed technological environment. However, IXPs are servers that tend to be located mainly along the coastal areas rather than inland. The correlation between IXPs and the dependent variable is unexpectedly negative. However, considering the number of IXPs and the surface of a country (Number of IXPs/Surface in sq. km), *IXP weighted*, the coefficient found in Column (5) is positive and statistically significant.

5.2 High-ranked Websites are more Sensitive to Distance

Different-sized firms can have different approaches to data storage strategies. Table 9 shows that geographic distance plays a consistently significant and negative role in the choice of data center location across both subsamples, the websites that are on the *Top 50* ranking and those *Under 50*. The coefficient on *Ln(Exact distance)* is -0.121 for the *Top 50* websites and 0.094 for the *Under 50* group, significant at the 1% level. This result suggests that both large and small firms tend to prefer locating their data infrastructure closer to their users, likely to reduce latency and enhance

	All				
	(1)	(2)	(3)	(4)	(5)
Ln(Exact distance)	-0.112***	-0.111***	-0.110***	-0.151***	-0.143***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)
Privacy regulation=1		1.087***	1.089***	1.060***	1.292***
		(0.030)	(0.030)	(0.028)	(0.029)
Political stability			0.203***	0.088***	0.266***
			(0.010)	(0.009)	(0.010)
Sales tax				-0.028***	-0.017***
				(0.000)	(0.000)
Ln(IXP weighted by surface)					0.100***
					(0.001)
Category FE	Yes	Yes	Yes	Yes	Yes
Website FE	Yes	Yes	Yes	Yes	Yes
Period FE	Yes	Yes	Yes	Yes	Yes
Pseudo R2	0.796	0.797	0.798	0.812	0.818
Observations	265,031	265,031	265,031	265,031	265,031

Notes: Poisson pseudo-likelihood regression model with category, website and period fixed effects. Transaction's location is the dependent variable. Columns (1)-(5) include the entire sample. Significance at 1%; 5% and 10% indicated respectively by ***, ** and *.

Table 8: Poisson pseudo-likelihood regression with multiple fixed effects (main results)

performance. The stronger effect observed for highly ranked websites suggests that large firms are more capable of optimizing infrastructure placement to minimize latency, potentially due to more abundant resources, greater technical capabilities, and a broader presence in global cloud networks.

Beyond this consistent distance effect, the regression reveals two institutional variables, *Political stability* and *Sales tax*, for which the sign of the coefficient changes between subsamples, while remaining statistically significant in both. This divergence provides insight into how firms of different sizes respond differently to the institutional environment of potential hosting locations.

First, the coefficient on *Political stability* is negative for the *Top 50* websites (0.026) but positive and substantially larger for *Under 50* websites (0.515), significant at the 1% level. This suggests that smaller firms are more likely to choose politically stable countries when locating data infrastructure. The rationale is intuitive: smaller firms may lack the internal capacity, legal, operational, or financial, to manage risks associated with unstable environments. In contrast,

larger firms may tolerate or even exploit politically less stable jurisdictions, potentially because these offer cost advantages, regulatory leniency, or strategic geographic positioning. Their broader risk management capabilities and operational diversification allow them to absorb or mitigate instability-related risks more effectively.

Second, the variable *Sales tax* shows a negative and significant effect for *Top 50* websites (-0.016), but a positive and significant coefficient for the *Under 50* group (0.008). This implies that larger firms actively avoid high-tax jurisdictions, aligning with expectations about global tax optimization strategies. These firms often have the legal and organizational means to incorporate tax considerations into location decisions and to benefit from arbitrage opportunities across jurisdictions. On the other hand, smaller firms may exhibit less sensitivity to local tax regimes, either because they are less exposed to tax burdens at scale or because other factors, such as regulatory clarity, data protection, or political stability, are more salient in their decision-making. It is also possible that smaller firms face a narrower choice set in terms of available infrastructure providers, limiting their ability to respond to tax incentives.

5.3 Tech Firms are more Global

The owner of the piece of data, either the website itself or an external provider, decides on the storage location. However, different pieces of data (such as the code of the page, images, text, audience tracking, etc.) that constitute a website are not necessarily owned by it. Hence, we can compare the approach toward data storage location between the website and an external provider by identifying the data owner. When the data owner is an external provider, this company is frequently a tech firm such as Google, Amazon, Cookielaw, Doubleclick, Facebook, Strpst, etc. Table 10 depicts the model's results applied to the sub-sample consisting of when the owner of a given object is not the website but an external provider. In our dataset, 34% of the data is owned by an external provider. The results show a weaker coefficient for the distance variable in Column (1). Column (2) depicts the results of the model applied to the sub-sample of data owned by the website. We interpret this lower coefficient as a result of multinational tech firms with a wide geographic spread, allowing them to get closer to consumers in different countries easily.

	Top 50	**Under 50**
	(1)	(2)
Ln(Exact distance)	-0.121***	-0.094***
	(0.002)	(0.001)
Privacy regulation=1	-0.007	1.708***
	(0.099)	(0.010)
Political stability	-0.026	0.515***
	(0.022)	(0.004)
Sales tax	-0.016***	0.008***
	(0.001)	(0.000)
Ln(IXP weighted)	0.143***	0.139***
	(0.006)	(0.001)
Category FE	Yes	Yes
Website FE	Yes	Yes
Period FE	Yes	Yes
Pseudo R2	0.758	0.742
Observations	33,749	231,282

Notes: Poisson pseudo-likelihood regression model with category, website, and period fixed effects. Transaction's location is the dependent variable. Column (1) applies the model to the *Top 50 websites* and Column (2) to the *Under 50 websites*. Significance at 1%; 5% and 10% indicated respectively by ***, ** and *.

Table 9: Poisson pseudo-likelihood regression with multiple fixed effects (according to ranking)

In addition to the overall lower distance sensitivity observed for data owned by external providers, the results in Table 10 reveal a striking contrast in how political stability influences storage location decisions, depending on who owns the data. When the data owner is an external provider, the coefficient on *Political stability* is positive and significant (0.115), indicating a preference for politically stable countries. In contrast, when the data is owned by the website itself, the coefficient becomes negative and significant (-0.202), suggesting a preference for less politically stable jurisdictions.

This reversal in sign, while both effects remain statistically robust, points to distinct strategic logics between these two types of actors. External providers, such as multinational tech firms (e.g., Google, Amazon, Facebook), are likely to place greater emphasis on political stability due to their large-scale operations, legal exposure, and need for long-term predictability in hosting environments. They may prioritize institutional reliability, compliance frameworks, and infrastructure protections that are more readily available in stable jurisdictions. This behavior

aligns with risk management strategies typical of globally regulated firms operating across diverse legal environments.

By contrast, when the website itself owns the data, the negative coefficient suggests a greater willingness to place data in less politically stable countries. This could reflect several non-mutually exclusive mechanisms. First, websites may be pursuing cost efficiencies by exploiting regulatory arbitrage in jurisdictions with weaker oversight or lower operational costs, which may correlate with political instability. Second, firms may be responding to latency or market access priorities, situating infrastructure close to specific user bases regardless of the political context. Finally, the website-controlled data may be more transactional or ephemeral in nature, lowering the perceived risks associated with politically volatile locations.

	External provider	**Website**
	(1)	(2)
Ln(Exact distance)	-0.186***	-0.145***
	(0.001)	(0.001)
Privacy regulation=1	1.041***	1.099***
	(0.061)	(0.117)
Political stability	0.115***	-0.202***
	(0.012)	(0.020)
Sales tax	-0.042***	-0.020***
	(0.001)	(0.000)
Ln(IXP weighted)	0.082***	0.059***
	(0.002)	(0.002)
Category FE	Yes	Yes
Website FE	Yes	Yes
Period FE	Yes	Yes
Pseudo R2	0.727	0.901
Observations	91,227	173,801

Notes: Poisson pseudo-likelihood regression model with category, website, and period fixed effects. Transaction's location is the dependent variable. Column (1) applies the model to the sub-sample when the data owner is an external provider. Column (2) applies the model to the sub-sample of data owned by the website. Significance at 1%; 5% and 10% indicated respectively by ***, ** and *.

Table 10: Poisson pseudo-likelihood regression with multiple fixed effects (external provider)

5.4 Large Data is Located Closer to the User

Data present on a website differ in terms of size, which has a direct impact on the uploading time. First, data size imposes bandwidth constraints. Thus, situating sizable data closer to users is essential to minimize network strain and ensure timely access. Additionally, cost-efficiency is a compelling factor. Research highlights that large datasets' storage and maintenance expenses can be exorbitant. By strategically locating these datasets near users, organizations can curtail data transfer costs and optimize resource utilization, fostering economic sustainability. Furthermore, big data analytics and machine learning have accentuated the importance of close data proximity. Studies like Chiang and Zhang (2016) illustrate how the real-time processing of substantial datasets necessitates their localization near end-users to enable rapid insights and decision-making. In a nutshell, large pieces of data should be stored closer to the user. For instance, applications and text are smaller and weigh less than media content (images, audio, and videos); the latter should be stored closer to the user. To verify it we split the sample into three sub-samples according to the type of data, either multimedia (image / audio / video), application or text. Table 11 shows the regression results for the different data types. Column (1) shows the results for the sub-sample of the type of data multimedia content (image, audio or video), and Column (2) shows the results for the sub-sample of the data type application and Column (3) depicts the results for the data type text. Results show that, indeed, websites are more sensitive to distance when dealing with larger pieces of data. However, the opposite effect is found for the variable privacy regulation, which shows that, as expected, data such as images, audio and video are less privacy sensitive.

5.5 Finance and Science and Education Firms are more Reliant on On-premise Technologies

Different sectors have different needs for data storage. Sectors dealing with sensitive personal data, such as banks or insurance firms, must comply with strict regulations requiring them to store data on their own on-premise data centers. On the other hand, media firms tend to focus on reducing latency; thus, distance plays an important role. We focus our analysis on the

	Image/Audio/Video	Application	Text
	(1)	(2)	(3)
Ln(Exact distance)	-0.151***	-0.145***	-0.143***
	(0.001)	(0.001)	(0.002)
Privacy regulation=1	0.969***	1.444***	1.251***
	(0.106)	(0.037)	(0.041)
Political stability	0.378***	0.226***	0.155***
	(0.019)	(0.016)	(0.021)
Sales tax	-0.023***	-0.021***	-0.017***
	(0.001)	(0.001)	(0.001)
Ln(IXP weighted)	0.105***	0.086***	0.088***
	(0.002)	(0.002)	(0.004)
Category FE	Yes	Yes	Yes
Website FE	Yes	Yes	Yes
Period FE	Yes	Yes	Yes
Pseudo R2	0.853	0.805	0.815
Observations	116,032	99,762	49,213

Notes: Poisson pseudo-likelihood regression model with category, website, and period fixed effects. Transaction's location is the dependent variable. Column (1) applies the model to the sub-sample when the data is an *Image/Audio/Video*. Column (2) to the sub-sample when the data is an *Application*. Column (3) to the sub-sample when the data is *Text*. Significance at 1%; 5% and 10% indicated respectively by ***, ** and *.

Table 11: Poisson pseudo-likelihood regression with multiple fixed effects (large data)

categories in which each website is a part of. Table 1 in Section 4.1 presents the distribution of the websites and transaction according to each category. Table 12 shows the results of a Poisson pseudo-likelihood regression model with period-fixed effects on the entire sample. In Column (1), we include the six categories with the most websites; the other categories are the omitted reference group. In Column (2), we include an interaction effect between the categories and the distance. Results show that categories that rely more on public cloud providers are more sensitive to distance. For instance, for News and Media, latency is crucial; thus, distance plays a more important role. On the other hand, categories such as Finance and Science, and Education tend to rely less on public data centers. These results could be due to two reasons. First, regulated sectors must often use on-premise storage solutions for security issues. Second, due to historical reasons such as firms having had their own storage facilities for a long time and restricted needs for private cloud providers.

	All	
	(1)	(2)
Ln(Exact distance)	-0.094***	-0.018***
	(0.000)	(0.001)
Category (ref. Other)		
Arts and Entertainment	0.572***	0.913***
	(0.003)	(0.006)
Computers Electronics and Technology	1.191***	1.442***
	(0.003)	(0.005)
E-commerce and Shopping	0.910***	1.462***
	(0.003)	(0.005)
Finance	0.932***	1.394***
	(0.004)	(0.005)
News and Media	0.605***	0.795***
	(0.004)	(0.006)
Science and Education	1.723***	2.237***
	(0.003)	(0.005)
Interaction effect		
Arts and Entertainment × Ln(Exact distance)		-0.065***
		(0.001)
Computers Electronics and Technology × Ln(Exact distance)		-0.047***
		(0.001)
E-commerce and Shopping × Ln(Exact distance)		-0.131***
		(0.001)
Finance × Ln(Exact distance)		-0.103***
		(0.001)
News and Media × Ln(Exact distance)		-0.033***
		(0.001)
Science and Education × Ln(Exact distance)		-0.105***
		(0.001)
Category FE	No	No
Period FE	Yes	Yes
Pseudo R2	0.659	0.686
Observations	228,121	228,121

Notes: Poisson pseudo-likelihood regression model with period fixed effects. Transaction's location is the dependent variable. Columns (1)-(2) apply the model to the entire sample. The omitted reference is *Other* for the website category. Significance at 1%; 5% and 10% indicated respectively by ***, ** and *.

Table 12: Poisson pseudo-likelihood regression with multiple fixed effects (categories)

6 Robustness Checks

6.1 Estimates with Alternative Measures of Distance

Considering the statistical significance levels found, we control whether the results hold for different measures of distance. Table 13 addresses these estimations. Column (1) represents the initial estimates, including the variable *Ln(Exact distance)*, which measures the exact distance between the consumer and the storage location. Column (2) estimates the same model, but considers the variable *Ln(distance capitals)*, the distance between the capital of the country where the consumer is located (Paris, France) and the capital of the country where the data center is located. Descriptive statistics for these variables are presented in Table 6. The coefficient of the distance is negative and significant. All other variables are consistent except political stability, which is negative. Column (3) estimates the same model by considering the distance between the most populated cities in each country (the country where the consumer is located and the country where the data is stored), through the variable *Ln(distance most populated cities)*, which is negatively associated with the dependent variable. Like Column (2), political stability is the only variable whose coefficient changes from positive to negative. As mentioned previously, distance is crucial for latency, hence uploading time. Overall, the main variable holds a negative association with the dependent variable.

6.2 Estimates with Alternative Functional Forms

This study's primary analytical framework is the Poisson regression model. However, including another regression model in the analysis serves as a rigorous methodological measure to validate the results' reliability and robustness. The selection of the Poisson regression model as the principal analytical tool derives from the characteristic nature of the dependent variable under investigation, typically encompassing count data, specifically the quantification of events or incidents within a designated time interval. The Poisson model aligns effectively with the discrete and non-negative attributes inherent to count variables, offering an appropriate method to assess independent variables' influence on the event occurrence rate.To assess whether this assumption holds, we conduct robustness checks using both Ordinary Least Squares (OLS) regression and

	All		
	(1)	(2)	(3)
Ln(Exact distance)	-0.143***		
	(0.001)		
Ln(distance capitals)		-0.843***	
		(0.003)	
Ln(distance most populated cities)			-1.010***
			(0.004)
Privacy regulation=1	1.292***	1.236***	1.245***
	(0.029)	(0.029)	(0.028)
Political stability	0.266***	-0.039***	-0.113***
	(0.010)	(0.010)	(0.010)
Sales tax	-0.017***	-0.089***	-0.108***
	(0.000)	(0.000)	(0.001)
Ln(IXP weighted)	0.100***	0.203***	0.292***
	(0.001)	(0.001)	(0.001)
Category FE	Yes	Yes	Yes
Website FE	Yes	Yes	Yes
Period FE	Yes	Yes	Yes
Pseudo R2	0.818	0.833	0.833
Observations	265,031	265,031	265,031

Notes: Poisson pseudo-likelihood regression model with category, website, and period fixed effects. Transaction's location is the dependent variable. Columns (1)-(4) apply the model to the entire sample. Column (2) includes the distance between the capitals of the two countries instead of the exact distance. Column (3) includes the distance between the most populated cities of the two countries. Significance at 1%; 5% and 10% indicated respectively by ***, ** and *.

Table 13: Robustness checks with alternative measures of distance

a Negative Binomial regression model (NBREG). The Negative Binomial model is particularly relevant in cases where overdispersion is present, when the variance of the dependent variable exceeds the mean, allowing for greater flexibility in the distributional assumptions of the error term. Prior literature has demonstrated the effectiveness of Negative Binomial regression in modeling count data in digital economics (Wang et al., 2025).

Table 14 shows the robustness of the results to different functional forms. Column (1) reports the original estimations through a Poisson pseudo-likelihood regression with multiple fixed effects (category, website and period). Column (2) reports estimating the number of pieces of data stored in a given country through an OLS model with category, website, and period fixed effects. The distance between the consumer and the storage location is negatively associated with

the dependent variable. The results hold for the rest of the variables. Considering the descriptive statistics of the dependent variable, an additional model that could provide a robustness check is negative binomial regression. Column (3) shows these results while accounting for category, website, and period fixed effects. The coefficient of the distance is negative and statistically significant. Overall, the main results hold through these additional alternative estimates.

	Poisson	**OLS**	**Negative binomial**
	(1)	(2)	(3)
Ln(Exact distance)	-0.143***	-63.565***	-0.149***
	(0.001)	(0.291)	(0.001)
Privacy regulation	1.292***	193.023***	1.351***
	(0.029)	(7.367)	(0.028)
Political stability	0.266***	56.228***	0.332***
	(0.010)	(2.938)	(0.011)
Sales tax	-0.017***	-7.157***	-0.026***
	(0.000)	(0.117)	(0.000)
Ln(IXP weighted)	0.100***	36.076***	0.104***
	(0.001)	(0.531)	(0.001)
Category FE	Yes	Yes	Yes
Website FE	Yes	Yes	Yes
Period FE	Yes	Yes	Yes
Pseudo R2	0.818		0.101
Adjusted R2		.8123894	
Log-Likelihood	-9276869	-1776255	-1705150
Observations	265,031	265,031	265,031

Notes: Column (1) applies a Poisson pseudo-likelihood regression model with category, website, and period fixed effects to the entire sample. Column (2) applies an OLS model with category, website, and period fixed effects to the entire sample. Column (3) applies a Negative binomial model with category, website, and period fixed effects to the entire sample. Significance at 1%; 5% and 10% indicated respectively by ***, ** and *.

Table 14: Robustness checks with alternative estimates

7 Conclusions and Policy Implications

This paper provides empirical evidence through website data flows that data storage location decisions are negatively associated with the distance from the data center to the customer. By identifying the data owner and leveraging the website ranking, we can demonstrate that large firms are more sensitive to distance, while tech-specialized ones are less sensitive to it. Distance

plays a different role according to the data type. Our findings show that large data tends to be stored closer to the consumer to optimize latency and performance. The same goes for high-priority data, such as the code of a web page that must be uploaded first. Categories with high-intensity use of data centers, such as news and media, and arts and entertainment, are more sensitive to distance. On the other hand, categories that include websites that are required by regulation to have on-premise storage facilities (e.g., Finance) or have a history of using on-premise storage (e.g., Science and Education) rely less on public data centers, thus are less sensitive to distance.

The policy implications derived from our study offer valuable insights into the intricate relationship between data flows and digital sovereignty, particularly in the context of websites providing services to EU users. Firstly, our empirical findings provide concrete evidence regarding the geographic origins of websites catering to EU users. By examining the patterns of data flows, policymakers gain a nuanced understanding of where these digital services originate, allowing for more informed decision-making in shaping regulations and policies. Secondly, our study emphasizes distance's significant role in data flows. The observed sensitivity to distance underscores the practical implications for digital sovereignty. Policymakers can leverage this insight to formulate regulations and strategies that capitalize on the geography of data storage, promoting a more resilient and sovereign digital landscape. Recognizing the impact of distance on data storage decisions prompts policymakers to consider localized solutions, potentially incentivizing domestic data infrastructure development and enhancing control over critical data assets. Our results suggest that, from a data flow perspective, the current landscape presents opportunities for bolstering digital sovereignty. Policymakers can leverage this understanding to craft targeted regulations and incentives that align with the data flow dynamics and contribute to the overarching goal of strengthening digital sovereignty in Europe.

By embracing policies that strategically address the geographic aspects of data flows, EU countries can assert greater control over their digital infrastructure, enhance data governance, and fortify their position in the global digital ecosystem. However, the design of regulatory interventions is crucial. While efforts to promote digital sovereignty through data localization policies can strengthen control over critical infrastructure, they must also balance short-term

access benefits with long-term investment incentives. Moreover, the impact of EU regulations on data localization mirrors challenges faced in other digital markets, where firms strategically navigate cross-border restrictions impacting market competition, consumer welfare, and firm behavior. As shown in the framework of network industries, high access costs can either accelerate investment in new infrastructure or create a disincentive for firms to transition from legacy systems (Bourreau et al., 2012; Bourreau and Manenti, 2023). In the same vein as Crémer et al. (2024), this highlights the need for well-calibrated policies that encourage investment in European cloud infrastructure while ensuring that data flows remain efficient and competitive.

The results of our study need to be understood in light of certain limitations. First, our dataset does not allow for control of data storage prices. Thus, we can not compare the cost of storing in different locations. Second, we do not have information on whether the websites of our dataset possess on-premise storage locations. Third, as we only collect the first page of each website, our data is limited to the existence of other storage solutions, hence, storage locations. It is worth highlighting that websites can use multi-cloud solutions or multi-location storage according to the data type and local regulations.

We would like to extend this study by replicating the data collection setting by simulating users from different countries in order to better validate our findings. Second, we plan to investigate data replication to better understand the data storage strategies between different data types. Future research will also focus on exploring the dataset better. For instance, studying other components of digital sovereignty with a focus on the component of the Internet's backbone infrastructure. Extending our current study with firms' headquarters (such as ISPs, CDNs, DNSs, or data owners) that allow data flows to EU users. This could offer a more complete picture of digital and data sovereignty through country interdependency. Nonetheless, this study offers a first attempt at understanding data flows and the feasibility of digital sovereignty.

References

Akhoondi, M., Yu, C., and Madhyastha, H. V. (2012). Lastor: A low-latency as-aware tor client. In 2012 IEEE Symposium on Security and Privacy, pages 476–490. IEEE.

Alesina, A. and Perotti, R. (1996). Income distribution, political instability, and investment. European economic review, 40(6):1203–1228.

Aversa, L. and Bestavros, A. (2000). Load balancing a cluster of web servers: using distributed packet rewriting. In Conference Proceedings of the 2000 IEEE International Performance, Computing, and Communications Conference (Cat. No. 00CH37086), pages 24–29. IEEE.

Babkin, A. and Shkarupeta, E. (2024). Industry 6.0: the essence, trends and strategic opportunities for russia. Russian Journal of Industrial Economics, 17(4).

Bartik, T. J. (1989). Small business start-ups in the united states: Estimates of the effects of characteristics of states. Southern economic journal, pages 1004–1018.

Berman, N. and Couttenier, M. (2015). External shocks, internal shots: the geography of civil conflicts. Review of Economics and Statistics, 97(4):758–776.

Besley, T. and Burgess, R. (2002). The political economy of government responsiveness: Theory and evidence from india. The quarterly journal of economics, 117(4):1415–1451.

Blum, B. S. and Goldfarb, A. (2006). Does the internet defy the law of gravity? Journal of international economics, 70(2):384–405.

Bourreau, M., Cambini, C., and Doğan, P. (2012). Access pricing, competition, and incentives to migrate from "old" to "new" technology. International Journal of Industrial Organization, 30(6):713–723.

Bourreau, M., Doğan, P., and Lestage, R. (2014). Level of access and infrastructure investment in network industries. Journal of Regulatory Economics, 46:237–260.

Bourreau, M. and Manenti, F. M. (2023). Selling cross-border in online markets: The impact of the ban on geoblocking strategies. International Journal of Industrial Organization, 86:102892.

Brynjolfsson, E. and McElheran, K. (2016). The rapid adoption of data-driven decision-making. American Economic Review, 106(5):133–139.

Castillo-Velázquez, J.-I. and Delgado-Villegas, A. (2020). Gns3 limitations when emulating connectivity and management for backbone networks: A case study of canarie. In 2020 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), pages 1–4. IEEE.

Chiang, M. and Zhang, T. (2016). Fog and iot: An overview of research opportunities. IEEE Internet of things journal, 3(6):854–864.

Correia, S., Guimarães, P., and Zylkin, T. (2020). Fast poisson estimation with high-dimensional fixed effects. The Stata Journal, 20(1):95–115.

Crémer, J., Biglaiser, G., and Mantovani, A. (2024). The economics of the cloud. Technical report, Toulouse School of Economics (TSE).

da Ponte, A., Leon, G., and Alvarez, I. (2023). Technological sovereignty of the eu in advanced 5g mobile communications: An empirical approach. Telecommunications Policy, 47(1):102459.

DeStefano, T., Kneller, R., and Timmis, J. (2020). Cloud computing and firm growth.

Devereux, M. P. and Griffith, R. (1998). Taxes and the location of production: Evidence from a panel of us multinationals. Journal of public Economics, 68(3):335–367.

Dhakal, S., Hayat, M. M., Pezoa, J. E., Yang, C., and Bader, D. A. (2007). Dynamic load balancing in distributed systems in the presence of delays: A regeneration-theory approach. IEEE transactions on parallel and distributed systems, 18(4):485–497.

Draves, R., Padhye, J., and Zill, B. (2004). Comparison of routing metrics for static multi-hop wireless networks. ACM SIGCOMM Computer Communication Review, 34(4):133–144.

Edler, J., Blind, K., Frietsch, R., Kimpeler, S., Kroll, H., Lerch, C., Reiss, T., Roth, F., Schubert, T., Schuler, J., et al. (2020). Technology sovereignty: From demand to concept [technologiesouveränität: Von der forderung zum konzept]. Technical report, Fraunhofer Institute for Systems and Innovation Research (ISI).

Fang, J., Feng, T., Guo, X., and Wang, X. (2024). Privacy-enhanced distributed revocable identity management scheme based self-sovereign identity. Journal of Cloud Computing, 13(1):154.

Goldfarb, A. and Trefler, D. (2018). Ai and international trade. Technical report, National Bureau of Economic Research.

Grubesic, T. H. and O'Kelly, M. E. (2002). Using points of presence to measure accessibility to the commercial internet. The Professional Geographer, 54(2):259–278.

Guo, C., Yuan, L., Xiang, D., Dang, Y., Huang, R., Maltz, D., Liu, Z., Wang, V., Pang, B., Chen, H., et al. (2015). Pingmesh: A large-scale system for data center network latency measurement and analysis. In Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication, pages 139–152.

Guven, E. Y., Yagci, M. Y., Boyaci, A., Yarkan, S., and Aydin, M. A. (2019). A survey on backbone attack. In 2019 7th International Symposium on Digital Forensics and Security (ISDFS), pages 1–5. IEEE.

Haug, K. C., Kretschmer, T., and Strobel, T. (2016). Cloud adaptiveness within industry sectors–measurement and observations. Telecommunications policy, 40(4):291–306.

Hawkins, S. M., Yen, D. C., and Chou, D. C. (2000). Disaster recovery planning: a strategy for data security. Information management & computer security, 8(5):222–230.

Heess, P., Holly, S., Körner, M.-F., Nieße, A., Radtke, M., Schick, L., Stark, S., Strüker, J., and Zwede, T. (2025). A multi-agent approach with verifiable and data-sovereign information flows for decentralizing redispatch in distributed energy systems. Energy Informatics, 8(1):24.

Henisz, W. J. (2000). The institutional environment for multinational investment. Journal of Law, Economics, and Organization, 16(2):334–364.

Hong, Y. S., No, J., and Kim, S. (2006). Dns-based load balancing in distributed web-server systems. In The Fourth IEEE Workshop on Software Technologies for Future Embedded and

Ubiquitous Systems, and the Second International Workshop on Collaborative Computing, Integration, and Assurance (SEUS-WCCIA'06). IEEE.

Hulkó, G., Kálmán, J., and Lapsánszky, A. (2025). The politics of digital sovereignty and the european union's legislation: navigating crises. Frontiers in Political Science, 7:1548562.

Jamshed, H., Waheed, U., Iqbal, S., Faheem, M., Ashraf, M. W., and Mansoor, Y. (2025). Dynamic smart contracts framework on ethereum private blockchain for real estate management. The Journal of Engineering, 2025(1):e70063.

Javaraiah, V. (2011). Backup for cloud and disaster recovery for consumers and smbs. In 2011 Fifth IEEE International Conference on Advanced Telecommunication Systems and Networks (ANTS), pages 1–3. IEEE.

Javorcik, B. S. (2004). Does foreign direct investment increase the productivity of domestic firms? in search of spillovers through backward linkages. American economic review, 94(3):605–627.

Jin, W. and McElheran, K. (2017). Economies before scale: survival and performance of young plants in the age of cloud computing. Rotman School of Management working paper, (3112901).

Kaur, N. (2025). Decoding the digital personal data protection bill: Strengths, weaknesses, and the road ahead. Weaknesses, and the Road Ahead (January 04, 2025).

Kemsley, D. (1998). The effect of taxes on production location. Journal of Accounting Research, 36(2):321–341.

Khan, K., Su, C.-W., Umar, M., and Zhang, W. (2022). Geopolitics of technology: A new battleground? Technological and Economic Development of Economy, 28(2):442–462.

Khayer, A., Jahan, N., Hossain, M. N., and Hossain, M. Y. (2021). The adoption of cloud computing in small and medium enterprises: a developing country perspective. VINE Journal of Information and Knowledge Management Systems, 51(1):64–91.

Khayer, A., Talukder, M. S., Bao, Y., and Hossain, M. N. (2020). Cloud computing adoption and its impact on smes' performance for cloud supported operations: A dual-stage analytical approach. Technology in Society, 60:101225.

Knieps, G. (2003). Competition in telecommunications and internet services: a dynamic perspective. In Internet, Economic Growth and Globalization: Perspectives on the New Economy in Europe, Japan and the USA, pages 217–227. Springer.

Kolko, J. D. (2000). Essays on information technology, cities, and location choice. Harvard University.

Koop, G. and Tole, L. (2008). What is the environmental performance of firms overseas? an empirical investigation of the global gold mining industry. Journal of Productivity Analysis, 30:129–143.

Labovitz, C., Ahuja, A., and Jahanian, F. (1999). Experimental study of internet stability and backbone failures. In Digest of Papers. Twenty-Ninth Annual International Symposium on Fault-Tolerant Computing (Cat. No. 99CB36352), pages 278–285. IEEE.

Lakhina, A., Byers, J. W., Crovella, M., and Matta, I. (2002). On the geographic location of internet resources. In Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment, pages 249–250.

Madiega, T. (2020). Digital sovereignty for europe. EPRS: European Parliamentary Research Service. Belgium.

Malecki, E. J. (2002). The economic geography of the internet's infrastructure. Economic geography, 78(4):399–424.

Markopoulou, A., Iannaccone, G., Bhattacharyya, S., Chuah, C.-N., Ganjali, Y., and Diot, C. (2008). Characterization of failures in an operational ip backbone network. IEEE/ACM transactions on networking, 16(4):749–762.

Melitz, M. J. and Ottaviano, G. I. (2008). Market size, trade, and productivity. The review of economic studies, 75(1):295–316.

Melville, N., Kraemer, K., and Gurbaxani, V. (2004). Information technology and organizational performance: An integrative model of it business value. MIS quarterly, pages 283–322.

Newman, R. J. and Sullivan, D. H. (1988). Econometric analysis of business tax impacts on industrial location: what do we know, and how do we know it? Journal of Urban Economics, 23(2):215–234.

Nezami, M., Tuli, K. R., and Dutta, S. (2022). Shareholder wealth implications of software firms' transition to cloud computing: a marketing perspective. Journal of the Academy of Marketing Science, 50(3):538–562.

O'Kelly, M. E. and Grubesic, T. H. (2002). Backbone topology, access, and the commercial internet, 1997–2000. Environment and Planning B: Planning and Design, 29(4):533–552.

Riezenman, M. J. (2001). Optical nets brace for even heavier traffic. IEEE Spectrum, 38(1):44–46.

Rochelandet, F. and Tai, S. H. (2016). Do privacy laws affect the location decisions of internet firms? evidence for privacy havens. European Journal of law and Economics, 42:339–368.

Rodrik, D., Subramanian, A., and Trebbi, F. (2004). Institutions rule: the primacy of institutions over geography and integration in economic development. Journal of economic growth, 9:131–165.

Sengupta, S., Kumar, V., and Saha, D. (2003). Switched optical backbone for cost-effective scalable core ip networks. IEEE communications magazine, 41(6):60–70.

Silva, J. S. and Tenreyro, S. (2006). The log of gravity. The Review of Economics and statistics, pages 641–658.

Sun, R. and Trefler, D. (2023). The impact of ai and cross-border data regulation on international trade in digital services: A large language model. Technical report, National Bureau of Economic Research.

Svantesson, D. (2020). Data localisation trends and challenges: Considerations for the review of the privacy guidelines.

Wang, Y., Zhou, J., and Zhang, R. (2025). Market accessibility, agglomeration, and spatial location of digital enterprises. International Review of Economics & Finance, 98:103842.

Zolas, N., Kroff, Z., Brynjolfsson, E., McElheran, K., Beede, D. N., Buffington, C., Goldschlag, N., Foster, L., and Dinlersoz, E. (2021). Advanced technologies adoption and use by us firms: Evidence from the annual business survey. Technical report, National Bureau of Economic Research.

A Appendix - Example of the data on a website

This is a screenshot of the online page of a French journal accessed by a user located in Paris, France, as part of the data collection. The page contains data visible to the user, including text, images, and advertisements. Additionally, the web page contains invisible data, such as GIFs or trackers, and the code of the page. The data collection intercepts all the data that constitutes this web page.

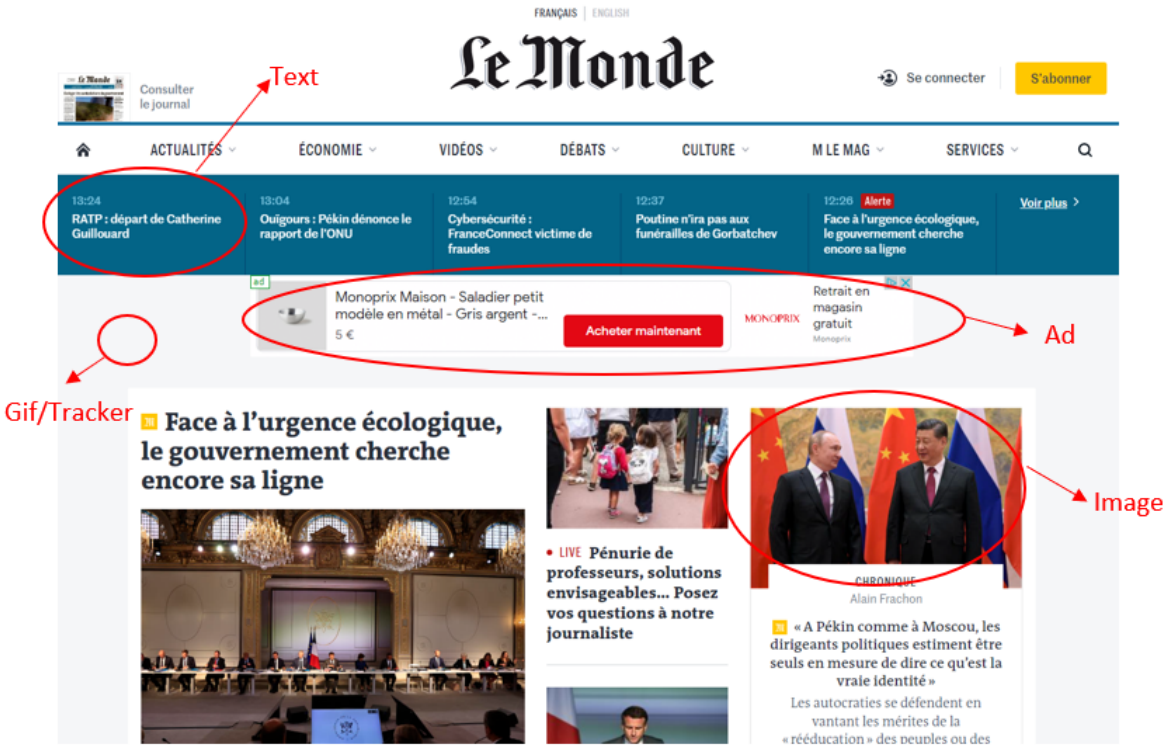

Figure 3: Pieces of data that can be found on a website

B Appendix - Websites used in the study

Table 15: List of the websites in each category

Category	Websites			
Adult	bongacams.com pornhub.com xhamsterlive.com	chaturbate.com spankbang.com youporn.com	livejasmin.com tukif.com	perfectgirls.net xhamster.com
Arts and Entertainment	1plus1tv.ru arte.tv dailymotion.com fandom.com imdb.com koreus.com ohmymag.com radio.co.ci telerama.fr wakanim.tv	6play.fr bfmtv.com deviantart.com filimo.com journaldupirate.net merojax.me primevideo.com rezka.ag telewebion.com youtube.com	adkami.com canalplus.com disneyplus.com france.tv kinogo.ge namasha.com programme-tv.net senscritique.com tf1.fr	allocine.fr crunchyroll.com doramy.club french-stream.lol kinogo.la netflix.com programme.tv spotify.com vimeo.com
Business and Consumer Services	calameo.com mondialrelay.fr societe.com	chronopost.fr packlink.com	iadfrance.fr seloger.com	logic-immo.com sncf.com
Community and Society	meetic.fr			
Computers Electronics and Technology	01net.com baidu.com clubic.com ecosia.org forumactif.com genial.ly	adobe.com bing.com commentcamarche.net ed-protect.org frandroid.com getadblock.com	apple.com boulanger.com darty.com extreme-down.live free.fr google.com	ask.com bouyguestelecom.fr duckduckgo.com facebook.com galinaturetskaya.ru google.dz

Continued on next page

Category	Websites			
	google.fr	google.ru	hp.com	ilovepdf.com
	imgur.com	instagram.com	journaldunet.com	lilo.org
	live.com	mail.ru	microsoft.com	mirrorace.org
	office.com	ok.ru	onvasortir.com	openclassrooms.com
	orange.fr	over-blog.com	oxtorrent.tv	paypal.com
	reddit.com	rutracker.org	samsung.com	savefrom.net
	sfr.fr	skype.com	smallpdf.com	speedtest.net
	stackexchange.com	stackoverflow.com	tiktok.com	tirexo.ai
	tumblr.com	uptobox.com	vk.com	wetransfer.com
	wikidot.com	wordpress.com	yandex.ru	yggtorrent.li
	yts.mx	yts.one	zoom.us	zt-protect.com
E-commerce and Shopping	aliexpress.com	aliexpress.ru	amazon.co.uk	amazon.com
	amazon.fr	auchan.fr	cdiscount.com	dealabs.com
	digikala.com	ebay.com	ebay.fr	etsy.com
	fnac.com	leboncoin.fr	rakuten.com	remisesetreductions.fr
	showroomprive.com	veepee.fr		
Finance	amazonaws.com	ameli.fr	ants.gouv.fr	axa.fr
	banquepopulaire.fr	boursorama.com	caf.fr	cic.fr
	cmb.fr	credit-agricole.fr	creditmutuel.fr	etoro.com
	labanquepostale.fr	lassuranceretraite.fr	lcl.fr	maaf.fr
	mabanque.bnpparibas	mgen.fr	smc.fr	societegenerale.fr
	tradingview.com	trustpilot.com		
Food and Drink	carrefour.com	carrefour.fr	cuisineaz.com	intermarche.com
	journaldesfemmes.fr	just-eat.fr	leclercdrive.fr	lidl.fr

Continued on next page

Category	Websites			
	mahjong.fr	marmiton.org	nounou-top.fr	
Gambling	fdj.fr	secretsdujeu.com		
Games	breakflip.com fr-org.com leagueofgraphs.com roblox.com	discordapp.com gamewave.fr liquipedia.net supersoluce.com	fextralife.com instant-gaming.com loups-garous-en-ligne.com twitch.tv	forgeofempires.com jeuxvideo.com porofessor.gg
Health	doctissimo.fr sante.fr	doctolib.fr vidal.fr	nih.gov	ooreka.fr
Heavy Industry and Engineering	edf.fr	enedis.fr	enphaseenergy.com	
Home and Garden	bricodepot.fr ikea.com	castorama.fr leroymerlin.fr	centrefrance.com manomano.fr	conforama.fr
Jobs and Career	apec.fr	indeed.com	jooble.org	urssaf.fr
Law and Government	franceconnect.gouv.fr laposte.fr service-public.fr	gouvernement.fr laposte.net yvelines.fr	impots.gouv.fr legifrance.gouv.fr	interieur.gouv.fr seine-et-marne.fr
Lifestyle	baginya.org kiabi.com zalando.fr	codesrousseau.fr laredoute.fr	galerieslafayette.com shein.com	hm.com vinted.fr
News and Media	20minutes.fr femmeactuelle.fr lefigaro.fr lesechos.fr mesopinions.com nouvelobs.com	actu.fr francetvinfo.fr lemonde.fr lexpress.fr msn.com opex360.com	dailymail.co.uk huffingtonpost.fr leparisien.fr maville.com namnak.com ouest-france.fr	donya-e-eqtesad.com lavoixdunord.fr leprogres.fr mediapart.fr newsru.com rbc.ru

Continued on next page

Category	Websites			
	rtbf.be	sputniknews.com	sudouest.fr	theguardian.com
Pets and Animals	equideow.com	zone-turf.fr		
Reference Materials	deepl.com linternaute.com multitran.com wordreference.com	lachainemeteo.com linternaute.fr pagesjaunes.fr	larousse.fr mappy.com reverso.net	linguee.fr meteofrance.com sciencedirect.com
Science and Education	ac-aix-marseille.fr ac-montpellier.fr archive.org cned.fr ecoledirecte.com etudiant.gouv.fr iledefrance.fr mit.edu overleaf.com u-picardie.fr univ-grenoble-alpes.fr univ-paris1.fr wifirst.net	ac-bordeaux.fr ac-nantes.fr auvergnerhonealpes.fr doodle.com education.fr futura-sciences.com itslearning.com mon-ent-occitanie.fr padlet.com uca.fr univ-lehavre.fr univ-pau.fr	ac-creteil.fr ac-reims.fr blackboard.com e-lyco.fr education.gouv.fr gismeteo.ru meteo60.fr monbureaunumerique.fr researchgate.net unilim.fr univ-lille.fr univ-poitiers.fr	ac-lille.fr ac-versailles.fr cairn.info eclat-bfc.fr enthdf.fr hespress.com meteociel.fr normandie-univ.fr tameteo.com unistra.fr univ-lorraine.fr univ-rouen.fr
Sports	decathlon.fr quiksilver.fr	eurosport.fr varzesh3.com	flashscore.com	lequipe.fr
Travel and Tourism	airbnb.fr paris.fr	airfrance.fr skyscanner.ru	booking.com	oui.sncf
Vehicles	caradisiac.com	lacentrale.fr	mobile.free.fr	motorsport.com

C Appendix - Variables used in the paper

Variable	Definition	Source	Year
Distance	The CEPII Gravity database gives the distance in kilometers between the capitals; the two most populated cities; between the capitals weighted by the population; between the two most populated cities weighted by the population.	CEPII	07/2022
Privacy regulation	The CNIL gives a label to each country concerning privacy regulation: Adequate country; Authority and law(s); Data protection law(s); EU or EEA Member country; Partially adequate country; No specific law.	CNIL	07/2023
Political stability	Political stability of a country is a variable built on two indexes going from 1=very low to 4=strong; indicating whether "The rules in force governing the assumption of political office by the Head of State or Government been amended to favor him/her remaining in office? Did the current Head of State or Government take office in accordance with the (potentially amended) rules in force at the time of that assumption of office (election, dynastic succession, etc.)?"	CEPII	2012
Sales tax	The tax rate per country.	Trading Economics	12/2022
IXP	The number of Internet Exchange Points in a country.	World Bank	10/2020

Table 16: List of variables