Regulating Data in the European Union and United States: Privacy, Access, Portability & APIs

Angela Mary Woodall

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2023

Abstract

Regulating Data in the European Union and United States:
Privacy, Access, Portability & APIs

Angela Mary Woodall

This dissertation examines the way that demands for more control over the collection, processing, and sharing of personal data are being managed by both government and industry leaders with strategies that appear to comply with regulations, but that fail to do so. These are "by-design" strategies used by individuals to unilaterally manage their data with automated tools.

I take a multimethod approach that combines autoethnography, reverse engineering techniques, and data analysis to assess the implementation of by-design services implemented by Facebook, Twitter, and Instagram in compliance with current European Union regulations for access and portability. I also employ archival research, discourse analysis, interviews, and participant observation.

I argue that self-led, by-design approaches do not answer the demands for more control over personal data. The regulatory and technical resources put in place for individuals to control their data are not effective because they turn over decisions about execution to an industry with no interest in sharing that data or being regulated. If policymakers continue to pursue by-design approaches, they will need to learn how to test the techniques, and the execution of the techniques, provided by industry. They will need to assess the impact on data that is made available. So that results can be evaluated, by-design tools like the ones I assessed must be accompanied by clear and detailed details about design choices and procedures. In this vein, I offer directions for critical scrutiny, including standards and measuring the impact of APIs.

I conclude that self-managed, by-design approaches are not the source of the problem. But they are a symptom of the need for critical scrutiny over the execution of tools like the ones offered by Facebook, Twitter, and Instagram. Ultimately, I found that portability and access are legally and technically fraught. However, despite the shortcomings of by-design approaches, personal data can be more effectively regulated in Europe than in the United States as the result of current regulations.

Table of Contents

List of Charts, Graphs, Illustrations

Acknowledgments

I want to express endless appreciation for the tireless oversight of this project given by my advisor, Andie Tucher. My gratitude to Richard John for the spirited, delightful conversations but also for constantly turning my attention to the past.

I owe a considerable debt to Anya Schiffrin, whose simple but critical question that helped me to find my way to the end. Her words will accompany me through the rest of my career and life. My many, many, many appreciations to Matthew Weber, who has been a delightful, generous, and thoughtful guide throughout the entire study, minus only a few months. My gratitude to Pascal Froissart for joining the committee and offering such thoughtful ideas from the beginning of my time at CELSA-Sorbonne University to the very end, during the defense. I could not have imagined a more wonderful committee. I also thank Elisheva Carlebach and Gil Hochberg, both of whose lectures continue to animate my ideas and work. I learned from them the poetry of scholarship. And, of course, John Watson, a mentor and inspiration.

I was blessed with the opportunity to begin this journey with an amazing cohort, Bernat Ivancsics, Emilie Xie, and Elena Egawhary. Thank you Lisa Bolz for helping me to continue it. I have also learned from every one of my colleagues and owe much to Sharon Ringel, Ri Pierce-Grove, and Joscelyn Jurich. Special appreciations to Cherie Henderson, Javi Sauras, and Ava Sirrah, Danielle Lee Tomson, Andi Dixon, Adelina Yankova, and Joanna Arcieri, who all shared their wonderful ideas during the iconoclastic seminars with the late Todd Gitlin. His spirit remains with us.

There are no words precise or expansive enough to capture the love and gratitude I feel for my family. For my brilliant daughters, Naima and Zoë Yi. For my mother, Linda Woodall,

who took care of them when I returned, because of her wise insistence, to school at Laney Community College in Oakland so many years ago. For the generosity of Joe and Ramona De Benedetti, and the endless kindness and humor of Rosanne, Michel, Marisa, Joe, and Stephan De Benedetti. And for Chris De Benedetti, the man who walked with me through the fire.

Dedication

I dedicate this work to my daughters and my beloved.

Introduction

Over the past 20 years, Amazon, Google, Facebook, YouTube, and Twitter have become control centers of an industry that relies on the collection and processing of personal data, and a network of constituents who want that data – from advertisers to government to social science researchers.[1] Much of this data is produced by individuals in the course of their everyday lives, yet they have had the least command over it.

Using Facebook or Twitter, for instance, means agreeing to terms and conditions that give those companies legal ownership over data.[2] As a whole, the data industry claims an exceptional level of decision-making power, entitling them to retain and repurpose user data regardless of the original intent for sharing it online.

Creating a complex legal and technical architecture has made scrutiny difficult. Facebook for example stymied scrutiny until the 2018 Cambridge Analytica data misuse scandal led to the discovery that a Facebook app was leaking the personal data of 120 million Facebook subscribers.[3] The scandal involved a Cambridge University researcher who sought permission from Facebook to gather personal information about subscribers using a questionnaire. The survey was distributed on Facebook (with the permission of the company) to subscribers, who

[1] Jef Ausloos and Pierre Dewitte, "Shattering One-Way Mirrors – Data Subject Access Rights in Practice," *International Data Privacy Law* 8, no. 1 (February 1, 2018): 4–28, https://doi.org/10.1093/idpl/ipy001.

[2] Catherine C. Marshall and Frank M. Shipman, "On the Institutional Archiving of Social Media," in *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '12 (New York, NY, USA: ACM, 2012), 1–10, https://doi.org/10.1145/2232817.2232819.

[3] Sarah Perez, "Facebook Rolls out More API Restrictions and Shutdowns," *TechCrunch* (blog), July 2, 2018, https://social.techcrunch.com/2018/07/02/facebook-rolls-out-more-api-restrictions-and-shutdowns/.

were paid to take a personality test and agreed to have their data collected for academic use. However, Facebook also allowed the collection of information about the test-takers' Facebook friends without permission. This data was also shared with political consultants at a private company (Cambridge Analytica).[4] At the time, Facebook's policies allowed only the collection of friends' data to improve user experience in the app and barred the data from being sold or used for advertising.[5]

Scandals and public pressure by civil society stakeholders intensified scrutiny about how social media networking companies like Facebook were collecting, processing, and sharing user data. Demands to know what was being held, how it was being used, and with whom it was being shared grew and continue to grow alongside the value of personal data.[6] Scrutiny led to the European Union's 2018 General Data Protection Regulation (GDPR) and, in 2022, the E.U. Digital Markets Act and Digital Services Act (DMA and DSA respectively).

These regulations are widely understood as a net positive by stakeholders across the public, private, academic, and civil society sectors. Yet there remain technical and legal questions that deserve further research so that stakeholders can reap the full benefits of the regulations. What are the best practices for ensuring privacy and security? What systems of

---

[4] Carole Cadwalladr and Emma Graham-Harrison, "Revealed: 50 Million Facebook Profiles Harvested for Cambridge Analytica in Major Data Breach," *The Guardian*, March 17, 2018, sec. News, https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election.

[5] Cadwalladr and Graham-Harrison.

[6] Helen Kennedy, "How People Feel about What Companies Do with Their Data Is Just as Important as What They Know about It," London School of Economics, *Impact of Social Sciences* (blog), March 29, 2018, https://blogs.lse.ac.uk/impactofsocialsciences/2018/03/29/how-people-feel-about-what-companies-do-with-their-data-is-just-as-important-as-what-they-know-about-it/; Lee Rainie, "Americans' Complicated Feelings about Social Media in an Era of Privacy Concerns," *Pew Research Center* (blog), accessed July 28, 2023, https://www.pewresearch.org/short-reads/2018/03/27/americans-complicated-feelings-about-social-media-in-an-era-of-privacy-concerns/; "Special Eurobarometer 359: Attitudes on Data Protection and Electronic Identity in the European Union - Data Europa EU," accessed July 23, 2023, https://data.europa.eu/data/datasets/s864_74_3_ebs359?locale=en.

governance are best suited for instituting those best practices? How might they balance the varied interests of stakeholders?[7]

This study gives attention to those open questions by looking at the way demands for more control over the collection, processing, and sharing of personal data are being managed by both government and industry leaders. I combine an assessment of current European Union regulations, primarily the GDPR and DMA, with rights established in both: the right of access to one's own data and the right to data portability. These rights – the means to know what data service providers have collected and to transfer that data between providers at will – are intended to promote data privacy, protection, and control. In this study, I look at the way that the objectives for privacy, access, and control are being pursued with regulatory and technical strategies optimized to give users more authority over their data, but that do not.[8]

Regulatory Overview

The DMA targets the potential for unfair, coercive, or deceptive practices that could obstruct competition. These measures are aimed at the data industry's largest operators including Google, Apple, Facebook, Amazon, and Microsoft. Policymakers signify them by the acronym "GAFAM" because of their entrenched and dominant control over data.[9] The DMA distinguishes

[7] For this framing I drew on Data Transfer Initiative, "The Future of Data Portability and Law: Unpublished Concept Note" (Chris Riley, 2023).

[8] Bjarki Valtysson, Rikke Frank Jorgensen, and Johan Lau Munkholm, "Co-Constitutive Complexity: Unpacking Google's Privacy Policy and Terms of Service Post-GDPR.," *NORDICOM Review: Nordic Research on Media and Communication* 42, no. 1 (July 1, 2021): 124–41, https://doi.org/10.2478/nor-2021-0033.

[9] Matthias Leistner, "The Commission's Vision for Europe's Digital Future: Proposals for the Data Governance Act, the Digital Markets Act and the Digital Services Act—a Critical Primer," *Journal of Intellectual Property Law & Practice* 16, no. 8 (August 1, 2021): 778–84, https://doi.org/10.1093/jiplp/jpab054.

companies with their kind of reach as "gatekeeper platforms." In computing, the term platform defines a core system on which other software or apps can be installed and so customized by outside users. Platform gatekeepers are so named because their technical design supports an organizational strategy where they become data bottlenecks between business users and end users. This happens when, for example, a business user like a fitness company wants to reach consumers with an exercise tracking device but, to do so, has to go through the Apple App Store and Facebook Ads. In addition to app stores and ad targeting, gatekeeper platforms offer devices, operating systems, browsers, geo-location services, social networks, tailored content, and other data-intensive products and services. Although not all platforms offer the full array of these services, the trend is toward integration.[10]

Joining them are U.S. and international efforts that call for more transparency, access, and protection.[11] Proposals for fulfilling these goals vary, but increasingly policymakers are leaning on an approach called data protection by design and default, where individuals unilaterally manage their data with automated tools.[12]

---

[10] Claudia Diaz, Omer Tene, and Seda F. Guerses, "Hero or Villain: The Data Controller in Privacy Law and Technologies," SSRN Scholarly Paper (Rochester, NY, September 5, 2013), https://papers.ssrn.com/abstract=2321480.

[11] U.S. Senator Chris Coons (D-Deleware) introduced a Platform Accountability and Transparency Act, SB 5339, in 2022 modelled on the DSA-DMA package but the bill was referred to the Committee on Health, Education, Labor, and Pensions, where it has remained since December 2022. I will discuss other legislation in subsequent chapters but overall this study is focused on European regulations because of the fragmentation of U.S. regulations. The California Consumer Privacy Act (CCPA) bill is modeled on the GDPR, but other states have local statutes, making it difficult to compare to the GDPR or DMA. Also the study is not focused on analyzing the different regulations, which would be many, between the countries although I do discuss the differences in regulatory approaches at the level of national governance in a later chapter. International efforts refer to the UNESCO, "Guidelines for Regulating Digital Platforms: A Multistakeholder Approach to Safeguarding Freedom of Expression and Access to Information" (UNESCO, February 2023), https://unesdoc.unesco.org/ark:/48223/pf0000384031.locale=en. These are voluntary guidelines that any nation can sign on to but are still in draft form.

[12] Ann Cavoukian, "Privacy by Design: The 7 Foundational Principles Implementation and Mapping of Fair Information Practices" (Information and Privacy Commissioner of Ontario, May 2010), www.ipc.on.ca.

Data Protection by Design and Default

The GDPR formulates "data protection by design and default" in Article 25 as an obligation of data controllers to integrate relevant protection of personal data in the architecture of devices.[13] Put another way, regulations should provide the necessary legal means to control how, why, and when personal data is used. By-design provides the specifications by which those regulatory objectives are translated into tools for users to control their data.[14] "Default" is defined in terms of data minimization, which means that a data controller should limit the collection of personal information to what is directly relevant and necessary to accomplish a specified purpose.[15] The means should be automated to the extent possible (a company cannot be made responsible for integrating requirements for which no technical solution has yet been developed).[16]

Proponents in government, civil society, and industry argue that access and portability give individuals more authority over their data and, in the case of portability, more options for where they keep it. It is clear from the E.U. blueprint for digital media that policymakers see the potential in access and portability for supporting normative values (autonomy, dignity, self-determination, privacy). E.U. policymakers also see the potential to create competitive strategies in an industry dominated by companies like Facebook, Twitter and Instagram.[17]

---

[13] "CCPA vs GDPR," November 30, 2020, https://www.cookiebot.com/en/ccpa-vs-gdpr-compliance-with-cookiebot-cmp/.

[14] This capacity to is referred to as information self-determination. See: Serge Gutwirth et al., *Reinventing Data Protection?*, Softcover reprint of hardcover 1st ed. 2009 édition (Springer, 2010).

[15] I have shortened the term to by-design.

[16] Mireille Hildebrandt and Laura Tielemans, "Data Protection by Design and Technology Neutral Law," *Computer Law & Security Review* 29, no. 5 (October 1, 2013): 509–21, https://doi.org/10.1016/j.clsr.2013.07.004.

[17] "A European Strategy for Data," Communication From The Commission To The European Parliament, The Council, The European Economic And Social Committee And The Committee Of The Regions (Brussels, Belgium:

The GDPR encourages self-service tools implemented by industry intended to support access and portability, by design, implying that controllers will develop resources that ease unwanted effects of mass data collection.

This study demonstrates why a closer look at by-design applications show reason for qualifying these expectations. By looking at execution, it is clear that this combination has not delivered an effective strategy for managing the collection, processing, and sharing of their personal data online.[18] And they have not eased the unwanted effects of mass data collection.

The regulatory and technical resources put in place for individuals to control their data are not effective because decisions about execution have been turned over to an industry with no interest in sharing that data or in being regulated.

I base this argument on an empirical assessment of by-design techniques made available by Facebook, Twitter, and Instagram intended to comply with the right to access personal data and the right to transfer, or "port," it. My argument is further supported by looking at the way that the approach to accommodating user rights developed and took effect, which I outline in a subsequent section.

European Commission, February 19, 2020), https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020DC0066.

[18] Users refer to individuals who produce data by using online systems – like social media. They are called data subjects in the European Union. I have adopted the term user and user rights in much of this study.

Methodology Overview

I tested compliance of access and portability designs with the GDPR and DMA by collecting data using three methods and then assessing the results.[19] The methods for collection include Archive-It, Webrecorder, and social media that I call platform downloads. Archive-It and Webrecorder are licensing services used to collect data from web- and social media sites. Platform downloads are web portals that allow subscribers to download their data from platforms themselves. Overall data was difficult to gather and the collections were incomplete. None of these services performed well at collecting social media. This was due in part to the fact that Archive-It and Webrecorder are not well designed for social media. However, Facebook blocked Webrecorder from accessing the site altogether. Twitter has since also begun restricting access to data in the same way that Facebook does. Platform downloads were the only way I was able to collect data from Facebook, Twitter, and Instagram. The obstacles I encountered led me to look more closely at the resources available for data access and portability and at the results.

I used GDPR Articles 12-25 and DMA Article 6 to assess the results. These articles address access and portability rights, and they encourage automated, by-design methods that individuals have to exercise those rights.

I identified the way that self-management, by-design methods have been implemented and accepted as a suitable strategy for exercising these rights. They have advantages such as being automated and nearly instantaneous, which was not the case before (discussed in the following section). Yet the methods have disadvantages for users, including incomplete,

[19] I assigned the name platform downloads, which do not have a consistent name. I provide a full and detailed overview in the methodology chapter.

confusing results with no recourse. My experience was like the experience many subscribers might encounter when trying to use the methods designed by Facebook, Twitter, and Instagram.[20]

In this regard, I found it necessary to understand the role of APIs (application programming interface) in the results because they are a key design feature despite sustained critical scrutiny over their unpredictable, unregulated selection and authorization methods. APIs are championed in the GDPR and DMA as effective methods for delivering policy goals without sufficient oversight.

The remainder of this chapter is structured in four main parts. In the first I provide background about how user data rights and the regulatory and technical strategies for fulfilling them developed. While I outline key policy developments between 1970 and 2022 in Chapter 1, I introduce some key concepts here because they add essential context to the way that core principles for processing personal data developed. The second section looks more closely at previous scholarship that documents the advantages and disadvantages self-management, by-design approaches – before the GDPR and leading up to the DMA. The third section outlines debates over APIs, which are important to understanding the way that government and industry are managing user rights. The fourth provides an overview and outline of the study. The methods I tested should offer more effective control over data and unwanted exposure by individuals and regulation. I conclude in the final section, that the implementation of a self-management, by-

[20] I had permission to collect data from the Facebook, Twitter, and Instagram accounts. One of the accounts belonged to an organization that I belonged to and I had the account the necessary email and password. I also used my own Facebook, Twitter, and Instagram accounts. This is also detailed in the methodology chapter and omitted here for readability.

design approach has reinforced a model in which users are responsible for exercising their rights with little or no means to challenge decision-making about how those rights will be provided.

Conceptual Frameworks

The strategy of combining industry-led solutions to data self-management has a significant history in privacy regulations in the United States and Europe dating back to the blueprint on which regulations today are based.[21] The 1973 Fair Information Practice Principles provided this blueprint.[22]

The blueprint was of American design but originated from a shared alarm over advances in the capacity of computerized data processing, which spiked in the 1960s and 1970s in the United States and Europe. Advocates inside and outside of government on both continents at the time considered the way data could be linked, shared, stored, and repurposed in databases as a danger to privacy. They responded by developing three main conceptual frameworks: data protection, access, and information self-determination. Although the frameworks follow from different intellectual sources, they revolve around the ability of individuals to control their personal information. This is what took shape as data privacy. The premise is that, in order to have data privacy, individuals need to be aware of personal data that is being collected and processed.

---

[21]Michael Veale, Reuben Binns, and Jef Ausloos, "When Data Protection by Design and Data Subject Rights Clash," *International Data Privacy Law* 8, no. 2 (May 1, 2018): 105–23, https://doi.org/10.1093/idpl/ipy002. Sometimes called "by design and default" as well. I discuss the approach at length in a subsequent chapter.

[22] U.S. Department of Health, Education and Welfare, Secretary's Advisory Committee on Automated Personal Data Systems, Records, computers, and the Rights of Citizens, *Records, Computers, and the Rights of Citizens*, OHEW Publication, (OS)73-94 (Washington, D.C.: U.S. Department of Health, Education & Welfare, 1973), https://aspe.hhs.gov/reports/records-computers-rights-citizens.

The frameworks overlapped in several ways but they split on identifying why people needed data privacy in the first place. In the first framework, control is the necessary condition for the ability to make decisions about personal data. The ability to make decisions is the condition for individual dignity, autonomy, personal liberty, and self-determination. This sociologically-based construct depends on a perception that citizens suffer a dangerous loss of control and lack of awareness when computer processing outstrips individual capacity to oversee what information is being processed. This argument is based on the principle that people can only develop freely when it is clear what others know about them. Without such knowledge, their ability to act freely is restricted. Therefore, data collecting and processing must be transparent and limited in purpose to ensure that people are aware of the information about them being managed.[23]

The rationale in the second framework extended the doctrine of due process – the right to see and contest evidence brought in criminal cases – to situations in which decisions were made based on the processing of personal data. Privacy constituted the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them would be communicated to others. On a practical level, people should be able to access their records in order to assess how a judgment about them had been made. The rationale was to give people more power vis-à-vis the entities that controlled their data and made potentially life-affecting decisions based on the data. This was in order to counterbalance the accumulation of economic and political power exercised by public and private organizations collecting and processing personal data. Only with access and control would citizens be in a position to judge

[23] René Mahieu, "The Right of Access to Personal Data: A Genealogy," *Technology and Regulation* 2021 (August 20, 2021): 62–75, https://doi.org/10.26116/techreg.2021.005.

the veracity and relevance of their data being collected and processed, and to challenge unfair decisions. This formulation was based on an argument that personal data was often collected in situations where power was unequal and the data was used to exert control over citizens in such a way that affected their opportunities. That could be a home loan or child custody. [24]

Here, proponents across juridical, government, and academic fields did not argue that computerization caused the lack of control and awareness. But, they were responding to fears that computerization (particularly data processing) was escalating both and, as a result, citizens were put at a disadvantage compared to the entities controlling their data.

These views influenced European theories about privacy, data access, and protection, from which emerged an argument that, if personal data was the center of power, then having access was a way to restore power to citizens and offer society a new way of regulating and rebalancing power structures. This critique of power structures rejected the idea that there ever existed an equality of power in relations between individuals and institutions: the relations between employer and employee, the citizen and the state, the doctor and the patient, the consumer and the producer, the holder of an electronic database and the person whose data is held. The balance of power was always skewed. Moreover, computing did not cause inequality but exacerbated it because the opaque and specialized nature of electronic data processing multiplied existing possibilities of abuse. The expert knowledge needed to use computers, and extract meaningful knowledge from databases, led to more technocratic forms of power and, in turn, meant a further loss of control for the majority of people. On the flipside, database technologies could also make access easier for people because they would no longer have to seek

---

[24] Mahieu.

out records manually. They could control the management of information about them and, ultimately, control the power based on that information.[25]

Nevertheless, access would not in itself resolve the imbalance of power. On the contrary, access could paradoxically undermine the position of individuals by functioning as a way to legitimize processing of personal data, justified by the argument of the possibility of everyone to know the information collected on themselves (i.e. how could there be a problem if everyone had the right to access their data?).

The problem was that individual pieces of information, which could be obtained through access requests, did not provide sufficient knowledge and power to the individual. The danger was that existing power structures would be reinforced. The conundrum could be addressed by giving individuals tools to access their data autonomously. But, because they alone could not be expected to reform inherently unequal social systems, collective action would be needed.[26] This might be third parties that pooled data for individuals and helped them to monitor the data controllers.

Leading up to the 1973 Fair Practices principles, these concerns centered on the flow and accessibility of information. Laws against restricting the press and speech, for example, reflect the high social value put on making information available. Conversely, too much information could give one constituent too much power over another, thus the concern over government

[25] Mahieu.

[26] Mahieu.

surveillance and individual privacy that animated attempts to rebalance power between data controllers and individuals.[27]

The Fair Practices guidelines were designed to moderate the determinantal effects of data processing systems by making the data accessible to its producers. This framework does not try to regulate the collection and processing of data. Instead, the guidelines state that individuals should know about what data has been collected about them and have the ability to correct it, or demand that it be deleted. The Fair Practices principles cemented this disclosure approach, which became the backbone of data protection, including the GDPR and U.S. privacy laws. The GDPR gave individuals legal tools for overseeing the collection, processing, and storage of their data. This includes the right of erasure, access, and portability. The law also assigns responsibility to industry for designing tools that facilitate self-management and oversight. They should be made, by design, with data protection at the forefront.

Relevant Scholarship

The by-design framework that I examine in this study is directed at developing cost-effective tools that individuals can use to manage their data with methods that industry, rather than government, are made responsible for implementing.[28] Studies before the GDPR was implemented suggest that the by-design approach could be useful because user rights were

[27] Jef Ausloos, "Chapter 2: Foundations of Data Protection Law," in *The Right to Erasure in EU Data Protection Law*, ed. Jef Ausloos (Oxford ; Cambridge, USA: Oxford University Press, 2020), 0, https://doi.org/10.1093/oso/9780198847977.003.0002; Gloria González Fuster, *The Emergence of Personal Data Protection as a Fundamental Right of the EU*, Softcover reprint of the original 1st ed. 2014 édition (Springer, 2016).
[28] Cavoukian, "Privacy by Design: The 7 Foundational Principles Implementation and Mapping of Fair Information Practices"; Veale, Binns, and Ausloos, "When Data Protection by Design and Data Subject Rights Clash."

ignored, inefficient, or underused.[29] In a 2017 empirical study of compliance with user access rights shortly before the GDPR came into force, researchers contacted 60 data controllers with requests for information about their own personal data being held by the companies.[30] They reported an often flagrant lack of awareness, organization, motivation, and consistency on the part of staff who handled their requests with company policies and existing regulations. Consequently they argued for tools that would empower individuals to access their data. Such tools could, if only marginally, insert checks and balances into the oversight of processing operations.

The researchers conceded the complexity of interests involved in balancing citizens' rights with the business interests of the companies controlling the data. For example, the right to access personal data, correct, it, or erase it could conflict with data controllers' economic freedoms. The rights of citizens could challenge trade secrecy, and the right to have sensitive data erased could conflict with various economic and property interests.

Taking account of these economic, legal, and technical contexts, the researchers evaluated possible remedies. They recommended fixes to terms of service contracts, as well as technical ones. A significant number of companies controlling the data included in the study struggled to even identify and locate the requested information. This, they argued, could be avoided by developing or reconfiguring their systems in such a way to facilitate the retrieval of relevant data in a secure and individualized way. Systems could be automated and information made machine-readable and interoperable.

---

[29] Veale, Binns, and Ausloos, "When Data Protection by Design and Data Subject Rights Clash."

[30] The GDPR came into force in 2018, a year before the study's 2017 publication.

Additionally, the controllers should provide straightforward templates to data subjects wishing to request access to their personal data. Throughout the empirical study, they noted considerable uncertainties related to the form and content of requests, which in turn led to lengthy and often unfruitful correspondence with controllers. By allowing users to build their requests following a clear and pre-determined format, the controllers could dismiss many procedural concerns and reduce the element of surprise for controllers confronted with requests from multiple sources. Data subjects would know from the start what documents to provide, which information to attach, and where to send their request.

They considered the possibility of embedding relevant features into the architecture of systems, and at every step of their processing activities. Controllers could, for example, implement machine-readable privacy policies to help people understand complex legal issues and make managing privacy preferences easier.

Execution of by-design methods has been uneven, and critical scholarship has documented problems with a narrowing focus on preventing data from being exposed (unwanted data disclosure) rather than relying on an expansive framework of rights and obligations.

The focus is understandable from companies trying to avoid being penalized for mishandling personal data. However, very often a narrow focus on risk by data controllers has come to outweigh rights. In a 2018 case study, researchers documented the way that by-design practices clashed with user rights. Data controllers failed to fulfill rights and obligations to individuals, including control and confidentiality. Engineers were designing the systems to privilege a narrow interpretation of protection over rights and they left few ways, or none at all, for individuals to be aware that these trade-offs were happening. This happened sometimes when

legal obligation to remove personal details conflicted with the right to access, erase, or object to data handling. Users couldn't erase what they couldn't find. Here rights were clashing. But problems were not disclosed or acknowledged. Rights of access, portability, and to object to processing suffered as a result. At a higher level, companies like Apple, Google, and Facebook were shifting risk onto the individuals whose data they are handling, but whose ability to manage that risk had been stripped away.[31]

These studies document the disconnect between technical and legal interpretations of key data protection notions. The disconnects have resulted in mismatches between user rights in the law and practice by data controllers. The data controllers don't acknowledge the trade-offs and disconnects and it is only by looking closely at moments when problems occurred like data breaches or obstacles to personal data that they are apparent.[32]

In one of the few studies of a platform download, information science researchers acquired social media data generated by the administration of Barack Obama during his presidency from 2009-2017. They acquired the content from the outgoing Obama administration, which had extracted the data from the platforms and packaged each set, then made the content available to scholars and groups for study including the Internet Archive through an approval process administered by the National Archives and Records Administration (NARA).

[31] Veale, Binns, and Ausloos, "When Data Protection by Design and Data Subject Rights Clash."

[32] Lee A. Bygrave, "Data Protection by Design and by Default: Deciphering the EU's Legislative Requirements," SSRN Scholarly Paper (Rochester, NY, June 20, 2017), https://papers.ssrn.com/abstract=3035164; Gerardo Con Diaz, "The Text in the Machine: American Copyright Law and the Many Natures of Software, 1974–1978," *Technology and Culture* 57, no. 4 (2016): 753–79.

The researchers sought to understand how the Obama social media data archive originated, was developed, and matured through each platform – Twitter, Vine (a now discontinued smartphone video application), and Facebook. Their comparison of the packaged collection against the version still online prompted questions about the provenance, quality, and veracity of the data extraction. Tweets and interaction data were missing, metadata was incomplete, and they questioned what else was missing. They observed that data structures and formats for each platform are unique, and, once removed from the native platform, social media data loses important context and becomes a snapshot of a moment in time. They also noted that metadata and other features of the data could change due to shifting API rules, platform redesigns, and other decisions that are not undertaken with academic research in mind but that nevertheless affect such work. They concluded that: "While the data speak to the engagement cultivated by the administration in its use of social media, the collection contains as many questions as it does answers."[33]

Similarly, portability initiatives promote the potential of privacy, access, and control by easing the move of data between service providers. Actually exercising that right has proved difficult for individuals because transfer tools are not optimized for user experience, if they exist at all. Portability can be limited because it requires users to download a structured digital file containing their data and then find a service to use the downloaded data. The decision for how users will exercise this right rests with the data controller, rather than individuals.

[33] Amelia Acker and Adam Kriesberg, "Tweets May Be Archived: Civic Engagement, Digital Preservation and Obama White House Social Media Data," *Proceedings of the Association for Information Science and Technology* 54, no. 1 (2017): 1–9, https://doi.org/10.1002/pra2.2017.14505401001.

This seemingly simple move involves numerous technical and legal constraints on all sides. The GDPR give neither individuals nor industry sufficient guidance about the coordination between hosts, or third parties operating on behalf of users' seeking to transfer their data. Data controllers face uncertainties about their responsibilities, potential liability, and the scope of data that should be made portable. And, market competition for data, and the ever-increasing accumulation for the purposes of AI applications, impedes the necessary level of coordination and standardization for effective user tools.[34]

Criticism of by-design resources as they have been implemented centers on the way they make individuals responsible for exercising their rights, but without giving them adequate, appropriate means to challenge decision-making about how those rights will be provided.[35] This includes access, disclosure, notice, and opt-out rights. Legal scholar David Solove links this model to what he calls "privacy self-management" that normalizes inadequate controls over personal data. I couple his analysis in Chapter 3 with concept of performance, which Ari Waldman uses to describe the *appearance* of effective policy.[36]

By-design options are presented as a solution capable of making otherwise inscrutable data controllers like Google and Facebook more accountable but without adequate vetting. Data controllers retain the prerogative to disable, block, or modify use of the services. They can also

[34] Chris Riley, "Data Transfer Project Use Cases," Data Transfer Initiative, accessed May 27, 2023, https://dtinit.org/use-cases.

[35] Acker and Kriesberg; Laurens Naudts, Pierre Dewitte, and Jef Ausloos, "Meaningful Transparency through Data Rights: A Multidimensional Analysis," in *Research Handbook on EU Data Protection Law* (Edward Elgar Publishing, 2022), 530–71.

[36] I discuss Solove and Waldman at length in Chapter 2.

stifle the use of these services by restrictions in their terms of service, changes to APIs, or by removing the services together. The devices and designs are deployed by data controllers because they help protect legal and business interests with a minimum of standards and oversight. The opposite model is the right to restrict the extent of data that individuals must share with data controllers and service providers to the minimum necessary for conducting a transaction.[37]

Debates about APIs

Debates over data rights and social media providers like Facebook, Twitter, and Instagram involve terms of service agreements and APIs.

APIs are like doors to data. The specifications for how these rules are written are protected by controllers as trade secrets because they are locks on those doors.

Twitter and Facebook price different levels of access to their APIs and, thus, the data. There are no- and low-cost levels that provide less data than paid ones. The most expensive levels are aimed primarily at behavioral marketing industries, which use data to target consumers based on the actions they take on websites and social media sites. These industries use the data provided by APIs to create new applications, which in turn generate more user data for both parties. Twitter and Facebook maintain the locks and the keys to this information.

[37] Diaz, Tene, and Guerses, "Hero or Villain"; Hildebrandt and Tielemans, "Data Protection by Design and Technology Neutral Law"; "Introduction: Privacy Self-Management and the Consent Dilemma," *Harvard Law Review* 126, no. 7 (2013): 1880–1904; "Privacy, Practice, and Performance," *110 CAL. L. REV. 1221*, 2022, https://doi.org/10.2139/ssrn.3784667.

Research about repurposing APIs to achieve user rights is sparse.[38] However, research has identified consistent limitations associated with the quality of APIs outputs, in particular inconsistent data, which limits reliability and reproducibility.[39]

The ability to effectively access and scrutinize social media data relies on the ability to use APIs and on an understanding of their technical constraints. This is not an expertise that many people have. By comparison, researchers, journalists, and transparency watchdogs do, and the policymakers in the United States and European Union are enlisting them in efforts to make the operations of companies like Facebook and Twitter more accountable for what happens on their platforms.

As a result, these allies are keenly aware that APIs affect data in unpredictable ways. Political science scholar Rebekah Tromble argues that the platforms and their APIs have always been proprietary black boxes never intended for scholarly use. One, they are made for commercial purposes, and, two, the type and quality of data is unclear.[40] Options are, however, shrinking. Facebook and Twitter companies don't release information about their sampling selection methods but they also offer no alternatives.

---

[38] Lorenzino Vaccari et al., "APIs for EU Governments: A Landscape Analysis on Policy Instruments, Standards, Strategies and Best Practices," *Data* 6, no. 6 (2021): 59, https://doi.org/10.3390/data6060059.

[39] Justin Chun-Ting Ho, "How Biased Is the Sample? Reverse Engineering the Ranking Algorithm of Facebook's Graph Application Programming Interface," *Big Data & Society* 7, no. 1 (January 2020), https://doi.org/10.1177/2053951720905874; Yuanbo Qiu, "The Openness of Open Application Programming Interfaces," *Information, Communication & Society* 20, no. 11 (November 2, 2017): 1720–36, https://doi.org/10.1080/1369118X.2016.1254268; Rebekah Tromble, Andreas Storz, and Daniela Stockmann, "We Don't Know What We Don't Know: When and How the Use of Twitter's Public APIs Biases Scientific Inference," SSRN Scholarly Paper (Rochester, NY, November 29, 2017), https://doi.org/10.2139/ssrn.3079927.

[40] Rebekah Tromble, "Where Have All the Data Gone? A Critical Reflection on Academic Digital Research in the Post-API Age," *Social Media + Society* 7, no. 1 (January 1, 2021): 2056305121988929, https://doi.org/10.1177/2056305121988929.

APIs are considered necessary to handling the size of data involved in the volume of data that companies like Facebook and Twitter process and store. Business users that buy data, as well as researchers, journalists, archivists, and transparency watchdogs, expect to use APIs. APIs are being integrated into government data handling and regulatory oversight. The DMA suggests APIs as a resource that will facilitate by-design approaches.

This is because APIs provide an interface necessary for two systems – Twitter and Facebook, for example – to exchange data between them (thus the name, application programming interface). APIs help different systems work together.[41]

APIs are the backbone of by-design measures but policymakers are ignoring their trade-offs and continue to be treated like neutral tools. This leads to the question of whether – given the caveats with opacity, ordering, and sampling – APIs should be used to deliver rights to personal data.[42]

Their role and drawbacks deserve more attention and caution from policymakers. But I argue that APIs are not the source of problems with the implementation of user rights. They are, however, indicative of the way both government and industry are managing their commitment to user rights with black boxes whose operations have not been accounted for.

For this reason, I raise the question of whether APIs are appropriate for the uses to which they are being repurposed and, more importantly, whether user rights can be effective without

[41] Sih Yuliana Wahyuningtyas, "Interoperability for Data Portability between Social Networking Sites (SNS): The Interplay between EC Software Copyright and Competition Law," *Queen Mary Journal of Intellectual Property* 5, no. 1 (January 1, 2015): 46–67, https://doi.org/10.4337/qmjip.2015.05.03.

[42] Sara Day Thomson and William Kilbride, "Preserving Social Media: The Problem of Access," *New Review of Information Networking* 20, no. 1 (2015): 261–75.

regulating them. Despite studies that have raised critical questions about them, by in large, APIs remain closed to outside scrutiny in much the same way that algorithms are. This study helps to open these boxes at a time when platforms are increasingly being invited to use APIs to manage developing international laws, systems, and standards.

I am not offering an argument against automation or the by-design model. The scholarship I cited above suggests that the platform downloads implemented by Facebook, Twitter, and Instagram were, in comparison, an improvement on what was available before. The platform downloads in this study include form-ready templates for requests and consent, machine-readable privacy policies, and personal data access without delays. Log-ins were password protected, the data was packaged in a machine-readable format or HTML, and available within less than an hour. Yet, I also encountered other, troubling results.

Instead, I found that the data protection by-design approach reinforces and formalizes a model in which users are responsible for exercising their rights, but are given little or no means to challenge decision-making by industry about how those rights will be provided.[43] Effective evaluation requires trying to understand complex data sets, and the systems behind them, with inadequate transparency and tools. It should be clear from my experience that this was a difficult process, which led me, as it has for other researchers, to more questions than were answered.[44] The process led me to consider access and portability rights as policy tools, raising deeper questions about what and whose objectives are in fact being pursued by government and platforms, alike.

[43] Naudts, Dewitte, and Ausloos, "Meaningful Transparency through Data Rights."

[44] Acker and Kriesberg, "Tweets May Be Archived."

This study offers insights about the development in government toward relying on private industry to facilitate the collective exercise of data rights, on behalf of citizens.[45] The risk is to normalize privacy rights that are defined and adjudicated by technical systems, including APIs, and data controllers.

Looking at by-design methods as an instance of the practice of relying on private industry to carry out policy highlights the possibilities for the effectiveness of future regulations and constraints on that efficacy. While much of the study focuses on E.U. regulations, the findings are as relevant to the United States because both share a history of the same approach to user rights.

Overview of Chapters

This is roughly divided into three sections: privacy, access, and portability regulations. Chapter 1 outlines the backdrop and context for this study's focus on access and portability and positions the study in a broader framework of policy concerned with technical and legal barriers to implementing data regulations. I compare the roots of data protection regulations in Europe and the United States between 1970 and 2020. I trace the development of data privacy protection by looking at political choices made about how emerging communication technologies that involve data should be regulated. I look at several moments in the 1970s, 1990s, and 2010s, when decisions were made about how to regulate the collection and storage of personal data. The Fair Practice principles were the result of a back-and-forth between the two continents, one influencing the other, and creating the mold for subsequent policy. While there are clear

[45] A. Giannopoulou et al., "Intermediating Data Rights Exercises: The Role of Legal Mandates," *International Data Privacy Law* 12 (November 2022), https://doi.org/10.1093/idpl/ipac017.

differences in political traditions, the United States and Europe share principles that govern data privacy. Both rely on approaches that were deployed by design to manage conflicts between the obligations to constituents, pressure from industry, and the demands of legal frameworks. This chapter helps to illustrate why I use the term data privacy protection. The term helps to gather together several constructs – privacy, control, self-determination, protection, user rights, and access – that framed the approaches to policymaking considered in this study.

Chapter 2 outlines my methodology for assessing the legal and technical tools made available for data self-management. The chapter is divided into several parts all confined to the data collection phase of the project. I include details about:

- How I developed a research strategy and why it was appropriate to the study
- How I gained access to the data
- What data I collected and how
- What my research activities were
- How I structured my analysis, and why
- How I chose to process the data, and why

The first section provides an overview of two complementary, qualitative methodologies for studying software: the first, reverse engineering and, the second, the walkthrough method.[46] I discuss how each was used in this study, their potential limitations, and how I combined the two in order to address some of those limitations. I then describe the methods I used to collect social media data from the Facebook, Twitter, and Instagram accounts of an organization that I belonged to, the Graduate Workers of Columbia. I use the last section to describe the results.

[46] Ben Light, Jean Burgess, and Stefanie Duguay, "The Walkthrough Method: An Approach to the Study of Apps:," *New Media & Society*, November 11, 2016, https://doi.org/10.1177/1461444816675438.

In Chapter 3, I summarize and analyze the results of the data collection methods according to GDPR provisions including scope, format, and privacy controls. I mainly restrict my analysis to the Twitter, Facebook, and Instagram platform downloads. These are built-in, by-design features implemented by Facebook, Twitter, and Instagram in compliance with the GDPR to provide users access to the data that controllers are collecting and processing.

Chapter 4 details my attempt to test portability rights provided in the GDPR and DMA and evaluate the results according to relevant regulatory criteria in the GDPR and DMA. I chose to test portability with automated methods developed in 2018 for making data transfers by an industry alliance led by Google and joined by Microsoft, Facebook, and Twitter. My attempt to compare techniques for portability and the data was impeded for several reasons. For this reason and for the sake of clarity, I included the portability testing methodology in this chapter rather than in the methodology chapter.

I conclude that self-led, by-design approaches are not adequate for offering control over personal data. The regulatory and technical resources put in place for individuals to control their data are not effective because they turn over decisions about execution to an industry with no interest in sharing that data or being regulated. I also conclude that by-design approaches are not the source of the problem, but are a symptom of the need for critical scrutiny over the execution of by-design tools like the ones offered by Facebook, Twitter, and Instagram. I also offer directions for critical scrutiny, including measuring the impact of APIs. Ultimately, I found that portability and access are, legally and technically, fraught. However, despite the shortcomings of by-design approaches, personal data can be more effectively regulated in Europe than in the United States because of the GDPR and DMA. Lastly, because this study involves multiple

software and computing systems, each structured in ways that affected my research, I have included a technical glossary in Appendix A.

Chapter 1: The Foundations of Data Privacy

This chapter compares the roots of data privacy protection regulations in Europe and the United States between 1970 and 2020. I trace their development by looking at political choices made about how emerging communication technologies that involve data should be regulated. I look at several moments when decisions were made about how to regulate the collection and storage of personal information in databases, i.e. personal data. While there are clear differences in political traditions, the United States and Europe share principles that govern data privacy and how they should be adjudicated. Both rely on individuals to unilaterally manage their data with automated tools. These are by-design approaches that provide individuals with methods to access their data and limited rights over it. This chapter lends a deeper perspective about the contemporary origins of a strategy that is the backbone of by-design policies in the United States and Europe.

Communications industries in Europe have been subject to stricter oversight than their U.S. counterparts.[47] Overall, even with a cultural ambivalence toward, and at times resistance to, oversight, there is an expectation that government will be involved in communications and the flow of information, including state subsidized popular media. This approach has set the tone today for regulating digital industries on both continents. However, a small number of very large online enterprises have captured the lion's share of the value generated online by data. While their business models are different, their principal product is based on personal information collected online and stored in databases. This data comes from the communication and activities of individuals on the internet.

[47] France in particular has a tradition of centralized government and regulation.

Online services that started as channels for networked communication now take the form of platforms. Architecturally, platforms like Google and Facebook are software-based systems that provide core functions on top of which outside developers can add features that extend the functions without having access to the core software and code, or control point. Platforms also describe the organizational forms that these enterprises take.[48] Their activity has become increasingly more complex, inscrutable, and unobservable as a result not only of the architecture but also because their economic activity itself is multi-faceted.[49] Platforms have become the dominant structure for online communication and information sharing. They have a global reach that supports linking geographically dispersed parties in trade and communication, as well as political, social, and cultural activity.[50] Moreover, the heterogeneity and architecture of platforms shield their economic and political activity, allowing them the flexibility to adapt to different regimes while pursuing their political-economic interests. As a result, much of the infrastructure of the internet and the rest of the digital economy is in the hands of private platform enterprises,

---

[48] José Van Dijck, "Seeing the Forest for the Trees: Visualizing Platformization and Its Governance," *New Media & Society* 23, no. 9 (September 1, 2021): 2801–19, https://doi.org/10.1177/1461444820940293.

[49] Anne Helmond, David B. Nieborg, and Fernando N. van der Vlist, "Facebook's Evolution: Development of a Platform-as-Infrastructure," *Internet Histories* 3, no. 2 (April 3, 2019): 123–46, https://doi.org/10.1080/24701475.2019.1593667.

[50] Christian Bartelheimer et al., "Systematizing the Lexicon of Platforms in Information Systems: A Data-Driven Study," *Electronic Markets* 32, no. 1 (March 1, 2022): 375–96, https://doi.org/10.1007/s12525-022-00530-6; Terry Flew and Fiona R. Martin, *Digital Platform Regulation: Global Perspectives on Internet Governance*, 1st ed. 2022 édition (Cham, Switzerland: Springer Nature Switzerland AG, 2022); Anne Helmond, "The Platformization of the Web: Making Web Data Platform Ready:," *Social Media + Society*, September 30, 2015, https://doi.org/10.1177/2056305115603080; Petro S. Korniienko et al., "Contemporary Challenges and the Rule of Law in the Digital Age," *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique* 36, no. 2 (April 1, 2023): 991–1006, https://doi.org/10.1007/s11196-022-09963-w; Amrit Tiwana, Benn Konsynski, and Ashley A. Bush, "Research Commentary — Platform Evolution: Coevolution of Platform Architecture, Governance, and Environmental Dynamics," *Information Systems Research* 21, no. 4 (December 2010): 675–87, https://doi.org/10.1287/isre.1100.0323; José van Dijck, *The Culture of Connectivity: A Critical History of Social Media* (Oxford ; New York: Oxford University Press, 2013); Shawn Walker, Dan Mercea, and Marco Bastos, "The Disinformation Landscape and the Lockdown of Social Platforms," *Information, Communication & Society* 22, no. 11 (September 19, 2019): 1531–43, https://doi.org/10.1080/1369118X.2019.1648536.

many of them American corporations that have demonstrated an intention to act like private rule makers.[51]

Having recognized the potential of data to fuel economic growth, politicians, governments, and regulators in the United States and European Union have responded with direct efforts to reign in digital giants at the center of the data industry, in particular companies like Google, Amazon, Facebook, and Twitter.[52]

Interventions take the form of competition policy like antitrust enforcement and economic regulation. Governments are also invoking the rights of individuals who use digital services as a way to justify regulating the data industry. Citing principles like a right to access, privacy, transparency, and fairness, policymakers have turned to by-design for technical features and self-management measures that allow individuals to access their data. The access enables them to know what data about them is being collected, moving the data to another service, checking its veracity, or asking for data to be deleted. The data protection by design tactic requires industry to integrate regulatory criteria into the construction of information technology. The design is, in turn, based on self-management strategies that are supposed to provide individuals with control over their data and thus privacy protection online.[53]

---

[51] Hannah Bloch-Wehba, "Global Platform Governance: Private Power in the Shadow of the State," *SMU Law Review* 72, no. 1 (January 1, 2019): 27; Korniienko et al., "Contemporary Challenges and the Rule of Law in the Digital Age."

[52] Flew and Martin, *Digital Platform Regulation*. Big tech refers to the dominant information technology companies operating today, i.e. Apple, Alphabet (Google), Amazon, Facebook (Meta), and Microsoft. They are considered in E.U. legislation to be the very large online platforms, or VLOPs. The criteria are provided elsewhere.

[53] Lee A. Bygrave, "Security by Design: Aspirations and Realities in a Regulatory Context," SSRN Scholarly Paper (Rochester, NY, May 23, 2022), https://papers.ssrn.com/abstract=4117110.

Overview

Respect for private life and the right to the protection of personal data are different but closely related. Although expressed somewhat differently, privacy in both Europe and the United States denotes a respect for private life, home and property, and personal information, as well as communications. This basic premise developed before computers. But computerized data processing over the years came to be considered as a threat to privacy because the ability to "keep oneself to oneself" was threatened by the massive amounts of information being stored in computer databases and transmitted without the subject knowing.

The right to access data was considered to provide the necessary checks and balances essential for mitigating this problem. Access is the condition for other rights, that now include the right to know what data is being collected about you, the right to verify the accuracy of data, the right to have data erased, or corrected. Also included is a right to restrict processing, move data from one provider to another, object to processing, and not to be subject to a decision based solely on automated processing.[54] I have adopted the term "user rights" to describe them. [55] In the following sections, I outline developments in the United States and Europe during the past half century that created the foundations for the approaches now in place.

The examples I have selected between 1970 and 2020 provide a context for data legislation considered in this study.[56] How user rights are conceived and legislated reflects norms about the nature and the proper role of the state, the system of political parties, the pattern of

[54] Being able to move data between providers is referred to portability. I will discuss portability and other of these rights in more detail in a later section and chapter. I also provide explanations in the glossary.

[55] Ian J. Lloyd, *Information Technology Law* (Oxford [England] ; New York : Oxford University Press, 2008), http://archive.org/details/informationtechn0000lloy_e3i8.

[56] "A European Strategy for Data: Shaping Europe's Digital Future," May 2, 2023, https://digital-strategy.ec.europa.eu/en/policies/strategy-data.

relations between economic and political interests, and the development of civil society.[57] The examples reflect how governments in the United States and Europe have managed the deepening dependence on information and communication technologies with their interest in harnessing data for economic growth and the rights of citizens.

Much of this chapter focuses on the 1970s. I also look at the 1990s and debates over user rights and regulating the internet during the Clinton Administration, when the triad of social, technical, and industry demands intensified. I then move to key provisions in the GDPR.[58]

I then outline some common explanations for differences in how the United States and Europe have managed user data rights along with an outline of relevant institutions, as well as the similarities.

Comparisons between the United States and Europe tend to project a normative idea of European governments as the guarantors of citizen rights and entitlements. Europeans do so by giving government a more prominent role in fostering social welfare and placing more limits on unfettered development of markets and technology. European governments provide, pay for, and heavily regulate essential services. In exchange, intensive government entanglement with daily life is accepted and often valued. This extends to data regulation.[59] In 1970 the German state of Hesse enacted the first data protection act directed specifically at regulating automated data processing and which applied to local government activity. The "Data Protection Act" applied to

[57] Daniel C. Hallin and Paolo Mancini, *Comparing Media Systems: Three Models of Media and Politics* (Cambridge ; New York: Cambridge University Press, 2000).

[58] "Privacy Guidelines for the National Information Infrastructure: A Review of the Proposed Principles of the Privacy Working Group."

[59] Fred H. Cate, *Privacy in the Information Age* (Washington, D.C.: Brookings Institution Press, 1997).

official files of the Hesse government and established a Data Protection Commissioner.[60] Sweden followed in 1973 with the first national statute, which extended rights to the private sector.[61] By the 1980s, France and Great Britain had all enacted some form of legislation that applied to both government and industry.

Common wisdom also holds that United States' historical circumstances, combined with a tradition of—and a persistent fear of—religious and political oppression, cultivated a suspicion of government and a relatively strong respect for markets and technology. Thus government surveillance has been seen more often than industry as a danger to privacy. By extension, the U.S. Constitution gives citizens rights against the government, but imposes few affirmative obligations on government.[62] It was in U.S. courts that the modern concept of privacy developed: individuals should be shielded from unwanted publication of embarrassing information, but not at the cost of speech.[63]

The United States is not only a strong proponent of free speech, it also sees the commercialization of the internet and data as a testament to the United States' commitment to entrepreneurship, innovation, and free markets, all of which have been drivers of U.S. economic success and technological progress. The tremendous financial and political importance to the U.S. economic growth and innovation culture of companies that deal in personal data makes

[60] "Privacy Guidelines for the National Information Infrastructure: A Review of the Proposed Principless of the Privacy Working Group."

[61] Lloyd, *Information Technology Law*.

[62] Cate, *Privacy in the Information Age*.

[63] Abraham L. Newman, *Protectors of Privacy: Regulating Personal Data in the Global Economy*, Illustrated edition (Ithaca: Cornell University Press, 2008).

American legislators and regulators less willing to challenge the business models that are behind economic success.[64] This has led to support for compliance strategies that would not be perceived as obstacles to busines. In light of their structural differences, compliance would be based on industry sectors rather than outright privacy rights. For example, the first major U.S. data privacy law, the Fair Credit Reporting Act (FCRA), targeted the credit reporting industry.[65] In contrast, Europe has codified a more top-down regulatory mode.

The normative explanation for this difference posits that the legal frameworks, and the process from which they result, reflect a European cultural, political, and social commitment to privacy as a basic human right. This is attributed to the historical surveillance endured under the Nazis during World War II and, after, the surveillance systems established by the governments of countries under Soviet control to impose political stability and avoid revolution. East Germany, one such Soviet "satellite state," created a surveillance apparatus that became notorious creating an atmosphere of suspicion and distrust toward fellow citizens and state institutions. The best known agency was the Staatssicherheitsdienst, widely known as the Stasi. The agency developed into a barely transparent security bureaucracy with manifold tasks, an enormous staff, and a network of informers, spies, and agents. [66] In response to wartime and postwar experiences,

[64] Anu Bradford, *The Brussels Effect: How the European Union Rules the World* (New York: Oxford University Press, 2020).

[65] Title VI of the Consumer Credit Protection Act.

[66] Jens Gieseke, *The History of the Stasi: East Germany's Secret Police, 1945-1990*, 1st edition (Berghahn Books, 2015); James Lay Jr., "Note by the Executive Secretary to the National Security Council on United States Policy toward the Soviet Satellites in Eastern Europe," draft policy statement (Washington, D.C, December 11, 1953), https://digitalarchive.wilsoncenter.org/document/national-security-council-nsc-174-draft-united-states-policy-toward-soviet-satellites; Andreas Lichter, Max Löffler, and Sebastian Siegloch, "The Long-Term Costs of Government Surveillance: Insights from Stasi Spying in East Germany," *Journal of the European Economic Association* 19, no. 2 (April 1, 2021): 741–89, https://doi.org/10.1093/jeea/jvaa009.

European countries developed an agenda to prevent oppressive bureaucracies capable of using record keeping for nefarious purposes.[67]

These fears became tangible as the result of a post-war proliferation of record-keeping agencies that produced increasing amounts of data about citizens. Governments and businesses in Germany and elsewhere in Europe in the 1950s and 1960s turned to automated data processing (originally developed for wartime military purposes) to handle the data they were producing. The demand generated by new, sophisticated, and ubiquitous record keeping systems to store and process data challenged conventional legal and social controls on organizations. No one wished to repeat the surveillance during World War II, by the East German government after the war or, for that matter, the Cold War spying in other Soviet satellites and throughout Europe. Combined with a mistrust that corporations would act in the public interest, the past paved the way for a robust privacy rights agendas in West Germany, in postwar European countries outside of Soviet control, and, later in E.U. policies.[68]

These narratives do not adequately explain variations in policies among European countries or their similarities with the United States, which are necessary to consider when trying to formulate effective policy today.

European policymakers actively modeled their data privacy authorities on U.S. independent regulators such as the Federal Trade Commission and the Federal Communications

[67] Cate, *Privacy in the Information Age*.

[68] Newman, *Protectors of Privacy*; William Ware, "Records, Computers and the Rights of Citizens" (U.S. Health and Human Services, June 30, 1973), https://aspe.hhs.gov/reports/records-computers-rights-citizens.

Commission.[69] The United States had its own history of surveillance They may not have equaled Europeans' experiences, but they led to several years of Congressional hearings examining the surveillance activities during the presidency of Richard Nixon and the half-century that J. Edgar Hoover oversaw U.S. domestic intelligence, which culminated in the post-Watergate support for government reform.[70]

The narratives also overlook similarities in the way that user rights have been and continue to be managed on both continents. From the earliest efforts to manage the collection of data, Europe and the United States have put the responsibility on citizens to fulfill their rights, which are based primarily on the right to request their personal records. Approaches to legislating these rights have been consistent and consistently similar for decades. By looking at the history of privacy, and user rights more broadly, it is clear that the United States and Europe influenced each other and continue to do so while still being animated by different political traditions. The principles and practices today have roots in the policies set in the 1970s.

Data Protection

The history of privacy and information surveillance predates computers. But the accumulation of personal information made possible by computer databases, escalating by the mid-1960s, created momentum for legislation that would serve as a model for the next half-century.

---

[69] Bradford, *The Brussels Effect*; Rebekah Dowd, "Digitized Data as a Political Object," in *The Birth of Digital Human Rights: Digitized Data Governance as a Human Rights Issue in the EU*, Information Technology and Global Governance (Cham: Springer International Publishing, 2022), 3–25, https://doi.org/10.1007/978-3-030-82969-8_1; Newman, *Protectors of Privacy*.

[70] "Department of Home Security: The Fair Information Practice Principles Factsheet," 2008, https://www.dhs.gov/publication/privacy-policy-guidance-memorandum-2008-01-fair-information-practice-principles.

Computers at that time mainly handled banking transactions, customer accounts of utilities, credit-card companies, large department stores, voting records, drivers' licenses, and government transactions.

This data was stored in computer databases, which, more than the collection of information alone, made combining and repurposing data possible at a scale that advocates worried would put citizens' privacy at risk. Emerging from a series of inquiries during the 1960s, advocates saw dangers in the way that data could be linked, shared, stored, and repurposed. They also worried that the scale that computers could do this work challenged existing technical, legal, and social controls on organizations, especially government ones. Researchers and policy experts also argued that these systems were highly vulnerable to sabotage and that electronically stored information could be, in the words of a 1971 appraisal, "pirated almost as easily as can files committed to paper."[71]

Committee on Automated Personal Data Systems

In 1973, then-U.S. Secretary of Health, Education, and Welfare Elliot Richardson originated the Committee on Automated Personal Data Systems, made up of nearly two-dozen people called together from academia, government, private industry, and other sectors.[72] Their task was to

---

[71] Richard L. Worsnop, "Reappraisal of Computers," in *Editorial Research Reports 1971*, CQ Researcher Online (Washington, D.C., United States: CQ Press, 1971), 345–66, http://library.cqpress.com/cqresearcher/cqresrre1971051200.

[72] This is the number of committee members who produced the report not including David Martin and several other Richardson staff who were present at the first meeting. Despite the relationship between Richardson and Nixon, which deepened in 1974 because of Watergate, I found nothing based on the records available to me to suggest direct involvement in the committee by the Nixon-Ford administration , or influence over Richardson's reason for appointing it. If there was such an involvement, the information would be included in the archives of the former presidents or, very likely, the Elliot L. Richardson Papers held by the Library of Congress, Part 1, Box I:108-173, or Part 3, Box III:48. They have not been digitized to my knowledge and are not available to me. According to his former colleague David Martin, Richardson shared some of the same concerns being considered by the Senate Judiciary Constitutional Rights subcommittee. The committee he appointed received copies of testimony Richardson

make an intensive study about the impact of computer databases on individual privacy. Richardson's special assistant to the committee, David Martin, presided over the meetings, which were recorded and transcribed for official records.[73]

Martin explained at the start of the first meeting that there were two reasons for the committee. The first was the concern over government record keeping.[74] The Department of Health, Education, and Welfare oversaw one of the largest source of government data collection at the time, including Social Security numbers. This relates to the second, narrower, reason for the committee. Richardson wanted to review the use of Social Security numbers outside of the Social Security Agency, which his department oversaw. He was prompted by a proposal to create a standard identifier for individuals to make exchanging information about them easier for computers.

The "Standard Identifier for Individuals" (as the proposed identifier was called) combined Social Security numbers with first, middle, and last names. At the time the biggest

---

before the committee, during which he told senators that the planned to appoint a group to consider the issue. I have not been able to locate the testimony. See: "Transcript of Proceedings of the Secretary's Advisory Committee on Automated Personal Data Systems (SACAPDS)" (Berkeley Law Center, April 17, 1972), https://www.law.berkeley.edu/research/bclt/research/privacy-at-bclt/archive-of-the-meetings-of-the-secretarys-advisory-committee-on-automated-personal-data-systems-sacapds/.
Otherwise, it appears from other accounts that Nixon and Ford publicly supported limits on government record keeping and less so on private industry, especially an oversight board, and were under pressure by intelligence agencies to limit oversight. For a recent interpretation, see: Dowd, "Digitized Data as a Political Object"; and Newman, *Protectors of Privacy*. Intelligence agencies also fought to limit oversight. Supporting their underlying account of political obstacles, see details about obstacles by the telecom industry and intelligence agencies available in: Douglas Metz, "Proposed Substitute for Senator Ervin's Privacy Protection Commission in S. 3418; Possible Privacy Commission Compromise," Memorandum Of Information For The File (Washington, D.C.: Domestic Council Committee On The Right Of Privacy, December 4, 1974), Box 56, folder "Privacy - Commission" of the Philip Buchen Files at the Gerald R. Ford Presidential Library.

[73] "Transcript of Proceedings of the Secretary's Advisory Committee on Automated Personal Data Systems (SACAPDS)."

[74] Ware, "Records, Computers and the Rights of Citizens."

concern was the scale at which computer databases were being used to link records across agencies, thus making anonymizing their identify that much harder.

The American National Standards Institute came up with the idea but needed government support before the identifier could be adopted.[75] Given the involvement of Social Security numbers, thus the Social Security Administration, adoption of the standard by the Health, Education, and Welfare Department was indispensable for the identifier to become an actual standard. Unused, the proposal would be ignored and left to fade away.

But the combination and cross-agency exchange of two identifiers – name and number – was exactly the kind of activity that was already worrying privacy advocates.[76] It worried Richardson and, when the U.S. Office of Management and Budget circulated a proposal for the standard and adoption, he stalled. He was not ready to say whether he would support the standard. Instead, he appointed the Committee on Automated Personal Data Systems to "develop a basis" for the decision and its possible consequences for privacy. However, Richardson decided that their study should be broader than Social Security numbers in order to effectively address privacy issues.

Richardson asked the committee to analyze and make recommendations about safeguards to protect against potentially harmful consequences that could come from using automated personal data systems and measures for remedies for harms.[77]

---

[75] The American National Standards Institute (ANSI) is a private nonprofit organization that oversees the development of voluntary consensus standards for products, services, processes, systems, and personnel in the United States.

[76] Sarah E. Igo, *The Known Citizen: A History of Privacy in Modern America* (Cambridge, Massachusetts: Harvard University Press, 2018).

[77] Caspar Weinberger succeeded Richardson in 1973.

The committee members had already become aware of the scope of data processing from a series of congressional efforts launched in the 1960 and 1970s.[78] These efforts included special committees commissioned by lawmakers or executive cabinet members, as well as the hearings attached to the 1971 House-Senate comprehensive privacy bill.[79] These commissions and hearings revealed how new recordkeeping systems were being used, the extent of data being collected, the new cadre of managers put in charge of overseeing the collection, and predicted the consequences for the privacy of citizens.[80] The details revealed in reports described massive quantities of information being collected, stored, and made available to government with little effort for unknown purposes.

The details shocked those Americans made aware for the first time of the scope and depth of information gathering practices, which we can compare to reactions over contemporary revelations about social media data sharing.

One of Richardson's committee members, a professor, said that the hearings provided a "rather astounding" portrayal of information gathering techniques used in the private sector and the linkages between information units in the private sector and the governmental sectors, including such matters as home life, drinking habits, and sexual behavior. People had to turn over these details to companies or government agencies if they wanted their services.

Computers and privacy were brought together in arguments popularized by legal scholars, among them Alan Westin and Arthur R. Miller. In commission reports and books that

---

[78] These include the 1962 Special committee on Science and Law of the New York Bar Association and1965 House Special Committee on Invasion of Privacy of the committee of Government Operations (the Gallagher Commission) in response to proposals to create government databanks centralizing information about individuals.

[79] Newman, *Protectors of Privacy*.

[80] This was led by Sen. Sam Ervin, a civil libertarian who would serve as the chairman of the Senate Watergate Committee investigating Richard Nixon. For a full account of the House-Senate privacy hearings and efforts as well as Ervin more specifically, all of which are outside the scope of my research, see Newman.

each authored, or contributed to, Westin and Miller substantiated a link between computers and the threats of computer-driven intrusion that compromised the ability of individuals to control the circulation of information about them.[81] Westin's argument originated in the idea that surveillance was happening as an accidental by-product of electronic data processing. According to Miller, once personal information has been stored, the individual to whom it refers loses the capacity to control it. Their arguments formed the foundation for an argument that the boundaries of privacy should include claiming the right to determine when, how, and to what extent personal information is communicated to others.[82]

In this formulation, privacy would not stop at an absence of information about us in the minds of others; rather, but should encompass the control we possess (or lack) over information about ourselves. That is to say, individuals, in order to have privacy, must be able to control what happens to their data. Control includes "information self-determination," which means deciding which personal data should be disclosed and how it should be used. This argument takes two forms. One is that individual control is a precondition for personal integrity, autonomy, and dignity. The advent of computer data processing risks imposing a demeaning lack of control on individuals who lack awareness about their personal data being collected and processed.

Individual autonomy follows from the control we have to consent or object to data collection and processing, such as having data about us deleted. Therefore individuals should have the right to determine when, how, and to what extent personal information is communicated to others.

---

[81] Alan F. Westin, "Science, Privacy, and Freedom: Issues and Proposals for the 1970's. Part I--The Current Impact of Surveillance on Privacy," *Columbia Law Review* 66, no. 6 (1966): 1003–50, https://doi.org/10.2307/1120997; Alan F. Westin, "Social and Political Dimensions of Privacy," *Journal of Social Issues* 59, no. 2 (July 1, 2003): 431–53, https://doi.org/10.1111/1540-4560.00072.

[82] Ausloos, "Chapter 2: Foundations of Data Protection Law"; Fuster, *The Emergence of Personal Data Protection as a Fundamental Right of the EU*.

Control is a shield against manipulation and targeting, actions that can harm, cause a loss of self-termination, which is part of what allows us to be autonomous actors in society, something related to liberty and freedom.[83] The second argument is premised on addressing imbalances of power by giving individuals the means to access and challenge information about them held by public or private entities.

These arguments have been over the decades condensed into what we now call information privacy: the ability of individuals to exercise some control over the use of information about them. Data protection describes the legal framework for how data is collected, processed, and used. Data protection does not protect individuals from all processing. Rather the goal is to prevent unlawful or disproportionate processing that deprives individuals of their rights.[84]

The committee, assembled to debate how to respond to the situation 50 years ago, recognized that their recommendations could either focus on trying to stop the technology, or recognize that computers, databases, and data-sharing were a fact of life that they could try to improve. This led to questions like: "How do you prevent access that is unauthorized? How do you give access to the individuals involved so that they can correct their own record? How do you prevent intercommunication between systems when that should be prevented?"[85]

[83] Ausloos, "Chapter 2: Foundations of Data Protection Law."

[84] Mahieu, "The Right of Access to Personal Data."

[85] "Transcript of Proceedings of the Secretary's Advisory Committee on Automated Personal Data Systems (SACAPDS)."

The committee was aware of the debates over privacy – Miller was one of the members. Privacy and control were part of their discussions and they yoked the two together. They concluded that the net effect of computerization is that even in non-governmental settings, individuals' control over the use of the personal data they gave to an organization, or that an organization obtained about him, was diminishing.[86] They also concluded that under the then-current law, a person's privacy was poorly protected against arbitrary or abusive record-keeping practices. For this reason, as well as because of the need to establish standards of record-keeping practice appropriate to the computer age, they recommended the enactment of a federal "Code of Fair Information practice" for all automated personal data systems.[87] The Code should rest on five basic principles that would be given legal effect as safeguard requirements for automated personal data systems.

Records, Computers and the Rights of Citizens

[86] William Ware, "Transmittal Letter to Secretary Honorable Caspar W. Weinberger Secretary of Health, Education, and Welfare," in *Records, Computers and the Rights of Citizens*, OHEW Publication, (OS)73-94 (Washington, D.C.: U.S. Department of Health, Education & Welfare, 1973). Ware introduced himself at the meeting as a computer specialist by profession at the Rand Corporation active in publicizing, largely within the Department of Defense, the problem of computer systems that can leak information and ways and means of providing information safeguards against such leakage; and, by extension, what computer systems that leak information can do to personal privilege or personal privacy.

[87] The word "practice" was written in the lowercase in the original.

That original set of principles was published in a 1973 report, "Records, Computers and the Rights of Citizens: Report of the Secretary's Advisory Committee on Automated Personal Data Systems." The recommendations would apply to all automated personal data systems and include penalties:

- The Code should define "fair information practice" as adherence to specified safeguard requirements
- The Code should prohibit violation of any safeguard requirement as an "unfair information practice"
- The Code should provide that an unfair information practice be subject to both civil and criminal penalties
- The Code should provide for injunctions to prevent violation of any safeguard requirement
- The Code should give individuals the right to bring suits for unfair information practices to recover actual, liquidated, and punitive damages, in individual or class actions. It should also provide for recovery of reasonable attorneys' fees and other costs of litigation incurred by individuals who bring successful suits

They also included safeguards requirements for administrative personal data systems, public notice requirements, rights for individual data subjects, guidelines for statistical reporting and research uses of administrative personal data systems. This last category in particular included trainings, precautions, punishments, and public notice requirements, and rights for individuals whose data was being used, including consent:

> Any organization maintaining a record of personal data, which it does not maintain as part of an automated personal data system used exclusively for statistical reporting or research, shall make no transfer of any such data to another organization without the prior informed consent of the individual to whom the data pertain, if, as a consequence of the transfer, such data will become part of an automated personal data system that is not subject to these I safeguard requirements or the safeguard requirements for administrative personal data systems.[88]

[88] U.S. Department of Health, Education and Welfare, Secretary's Advisory Committee on Automated Personal Data Systems, Records, computers, and the Rights of Citizens, *Records, Computers, and the Rights of Citizens*.

While the committee used the Social Security number to think through the problems and remedies for data practices, it is clear from transcripts and in the 339-page "Records, Computer, and the Rights of Citizens" report that members were thinking much more broadly. The recommendations provided the framework for general solutions and also action items to be taken both within the Health, Education, and Welfare agency and by the Federal government as a whole. These included assigning responsibility for preparing a detailed plan to carry out the action agenda to an official in Health and Human Services, which was by then being led by Caspar Weinberger in the midst of the Watergate hearings that led to the resignation of President Richard Nixon a month after the report was published.

Weinberger praised the report as deserving to be widely read and discussed but gave no indication that the recommendations should be taken seriously. He wrote in the forward to the published version that:

> The Committee obviously considers its recommendations to be a reasonable response to a difficult set of problems. The Committee has taken a firm position with which some may disagree. However we should be grateful to the Committee for speaking with such a clear voice. In doing so, it has no doubt set in motion the kind of constructive dialogue on which a free society thrives.[89]

The committee considered the principles to be the minimum set of rights that should be available to the individual. They then set out to extend those rights to citizens. An obvious mechanism was the creation of a centralized federal agency to regulate all automated personal data systems. Such an agency would be expected generally be the watchdog over all databases, public and private.

[89] "Forward," in *Records, Computers and the Rights of Citizens*, OHEW Publication, (OS)73-94 (Washington, D.C.: U.S. Department of Health, Education & Welfare, 1973), 5–7, https://aspe.hhs.gov/reports/records-computers-rights-citizens.

But because systems used by the enormous number and variety of institutions dealing with personal data vary greatly in purposes, complexity, scope and administrative context, an agency that could regulate that breadth of activity would have to be both large in scale and pervasive. The procedures for regulation or licensing would become extremely complicated, costly and might unnecessarily interfere with desirable application of computers to record keeping.

A regulatory agency with the expertise and size to manage the complexity, scope, and heterogeneity of databases spread out among public and private institutions and sectors was unlikely. For one, the political climate in Washington at the time was particularly unfavorable to government regulation that would place reporting obligations on industry. In the midst of the escalating Watergate scandal, a federal agency like this would likely have been seen as an intrusion on citizens.[90] This was evident from the unfolding hostility at the time to a Congressional proposal for an oversight board that would have been part of a comprehensive privacy bill. The idea for an enforcement agency would have lacked a constituency powerful enough to overcome that hostility. Consequently, the committee proposed a solution they considered capable of providing for the citizen strong rights, but that avoided the necessity of a regulatory body. Few of the tougher action items made it off the page. At the same time, the importance in the recommendations placed on giving individuals a way to be aware of what was being done with their data fits with the role in the new conceptualization of privacy of self-determination and control. This is still the dominant framework for managing information privacy.

[90] Sarah E. Igo, "Codes of Confidentiality and Consent," in *The Known Citizen: A History of Privacy in Modern America* (Cambridge, Massachusetts: Harvard University Press, 2018).

Managing the Balance

The report and its recommendations were condensed into a framework and called the Code of Fair Information Practices:

- There must be no personal data record-keeping systems whose very existence is secret
- There must be a way for a person to find out what information about the person is in a record and how it is used
- There must be a way for a person to prevent information about the person that was obtained for one purpose from being used or made available for other purposes without the person's consent
- There must be a way for a person to correct or amend a record of identifiable information about the person
- Any organization creating, maintaining, using, or disseminating records of identifiable personal data must assure the reliability of the data for their intended use and must take precautions to prevent misuses of the data

The Code as it was adopted was condensed and stripped of direct enforcement measures, but it provided the framework for the next major regulation, the Privacy Act of 1974.[91] Taken as a whole, the Code of Fair Practices, the Records, Computer, and the Rights of Citizens report, the

[91] "Privacy Guidelines for the National Information Infrastructure: A Review of the Proposed Principles of the Privacy Working Group,"https://archive.epic.org/privacy/internet/EPIC_NII_privacy.txt.
The Code and report were concerned about government recordkeeping more than private industry. The *proposed* Code called for two sets of safeguard requirements; one for administrative automated personal data systems and the other for automated personal data systems used exclusively for statistical reporting and research. Special safeguards were recommended for administrative personal data systems whose statistical reporting and research applications were used to influence public policy. The safeguard requirements also defined *minimum standards* of fair information practice. Under the proposed Code, violation of any safeguard requirement would constitute "unfair information practice" subject to criminal penalties and civil remedies. The Code would also provide for injunctive relief. Pending legislative enactment of such a code, the report recommends that the safeguard requirements be applied through Federal administrative action. See Cate 1997 cited here and elsewhere for more and a discussion about the development of privacy in current contexts of digital data and information technologies. For recent legislation, see Office of Science and Technology Policy, "The Blueprint for an AI Bill of Rights: Relationship to Existing Law and Policy," The White House, accessed July 13, 2023, https://www.whitehouse.gov/ostp/ai-bill-of-rights/relationship-to-existing-law-and-policy/.

Fair Credit and Reporting Act, and the 1974 Privacy Act set the framework for a subsequent wave of legislation (280 Congressional bills in four years) addressing data access and dissemination.[92]

Policy in the 1970s set the tone for managing the balance between the government's need to gather and use personal information and an individual's need to exercise control over that information in order to protect their privacy.[93] Much of the responsibility ultimately fell on individuals, who should seek out the information being collected about them, while the organizations that controlled the data were responsible for putting the procedures in place to provide access to that data. Best practices provided minimal obligations for those procedures and they were not mandatory in most cases. But they were applied in the public sector, and to the extent required in the private sector, both of which expanded the scope of data collecting again when the internet was widely introduced in the 1990s.

---

[92] Jerome J. Hanus and Harold C. Relyea, "A Policy Assessment of the Privacy Act of 1974," *American University Law Review* 25, no. 3 (1976 1975): 555–94.

[93] The Privacy Act was passed against the backdrop of building claims to "the right to know" embodied by the Freedom of Information Act, which strengthened public access to documents held by some, but not all, federal agencies: transparency activists have long been stymied by its exemption of Congress and the White House. Hanus and Relyea, "A Policy Assessment of the Privacy Act of 1974"; Michael Schudson, *The Rise of the Right to Know: Politics and the Culture of Transparency, 1945–1975* (Cambridge, Massachusetts: Belknap Press, 2015).

Global Information Infrastructure

The Fair Practices code created a baseline in the 1990s when the White House of President Bill Clinton and Vice President Al Gore set about developing the "Global Information Infrastructure" with an explicit rationale of balancing the rights of citizens with economic growth.[94]

The Clinton White House imagined the "Global Information Infrastructure" as a vast network of hardware and software, applications, activities, relationships, and information itself that was a "network of networks." At another level, the concept included the "National Information Infrastructure," with the internet at the core.[95]

Clinton and Gore saw the internet's potential for the collection, re-use, and instant transmission of information, which could – if managed correctly, they argued – drive both economic growth and lead to global transparent, meritocratic governance.[96]

---

[94] Private industry is not absent from discussions in the 1970s. But the concern was personal liberty of citizens about whom information collecting was being automated on a large scale. The record-keepers were largely government agencies. The Clinton-Gore White House explicitly set out to drive economic growth by exploiting then-new networked technologies and understood that the privacy was important to making people feel they could safely use the internet. The network that Clinton and Gore promoted included databases, images, sound recordings, library archives, or other media; standards, interfaces, and transmission codes that facilitate interoperability between networks and ensure the privacy and security of the information carried over them, as well as the security and reliability of the networks themselves. People – vendors, operators, service providers – would be creating and using the information, developing applications and services with it, and constructing facilities.

[95] National Academy of Sciences, "The Global Information Infrastructure: A White Paper Prepared for the White House Forum on the Role of Science and Technology in Promoting National Security and Global Stability" (National Academy of Sciences, March 29, 1995), https://clintonwhitehouse4.archives.gov/WH/New/Commerce/read.html; William Clinton and Albert Gore Jr., "A Framework For Global Electronic Commerce," July 1, 1997, https://clintonwhitehouse4.archives.gov/WH/EOP/OSTP/forum/html/giipaper.html.

[96] They may have seen this on their own or been convinced by industry leaders who were lobbying to privatize and commercialize development of the internet and its infrastructure, largely taking place in academic and government institutions. See: Megan Sapnar Ankerson, *Dot-Com Design: The Rise of a Usable, Social, Commercial Web* (New York: NYU Press, 2018); Tim Berners-Lee, *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web*, 1 edition (San Francisco: HarperBusiness, 2000); Shane Greenstein, *How the Internet Became Commercial: Innovation, Privatization, and the Birth of a New Network* (Princeton: Princeton University Press, 2015).

Correct management meant minimizing regulatory barriers on electronic commerce, above all by ensuring a free market for internet services unfettered by federal or state regulations and championed provisions.[97]

This is apparent in the Telecommunications Act of 1996, whose stated intent was to encourage innovation and reduce regulation among the relevant industries, thereby driving competition.[98] Likewise, Section 230 (originally included in the Communications and Decency Act), shields online businesses that provide, but aren't the original sources of, information from being treated like publishers or speakers whose communications are considered to be harmful. Section 230's protections are not absolute. The law does not protect companies that violate federal criminal law or protect companies that create illegal or harmful content. But, it provides for an extremely broad exemption from liability based on the First Amendment provisions for freedom of expression that internet-based companies and social media providers tool advantage of to protect themselves against reforms and to prevent competition.[99]

The Clinton Administration also recognized that to manage the infrastructure correctly and extract the potential economic and social benefits of the internet, personal privacy online couldn't be left entirely up to private industry. Clinton laid out how he planned to balance the "free flow of information" with the rights of individuals in 1997.[100] Americans, he argued,

---

[97] Greenstein, *How the Internet Became Commercial*.

[98] Patricia A. Aufderheide, *Communications Policy and the Public Interest: The Telecommunications Act of 1996*, 1 edition (New York: The Guilford Press, 1999).

[99] Electronic Frontier Foundation, "Section 230," *EFF Issues* (blog), n.d., https://www.eff.org/issues/cda230; Anh Nguyen, "Transatlantic Perspectives from Sciences Po Digital, Governance and Sovereignty Chair Florence G'sell and Georgetown Law Professor Anupam Chander on the Digital Services Act and Section 230," *McCourt Institute* (blog), March 13, 2023, https://mccourtinstitute.org/transatlantic-perspectives-from-professors-florence-gsell-and-anupam-chander-on-the-digital-services-act-and-section-230/.

[100] Clinton and Gore Jr., "A Framework For Global Electronic Commerce."

treasured privacy. But, at the same time, the First Amendment, a "hallmark" of U.S. democracy, protects the free flow of information. These two "competing values" – personal privacy and free flow of information – would have to be resolved in order to benefit from the commercial potentials of the infrastructure.

This was a stretch because the Constitution didn't protect the flow of information in quite the way Clinton framed it. With a sleight of hand, Clinton was trying to reframe the debate in favor of minimal government and free enterprise.

The actual conceptual framework for balancing privacy with economic growth was hammered out in a series of task force meetings and working groups whose members identified a need to adapt traditional fair information practices, which we saw in the previous section, to a broader networked communications environment.[101]

Privacy advocates generally promoted stronger protection for the public, while industry representatives argued that self-regulation was adequate and that new laws were unnecessary. For its part, the White House advocated a limited government role in legislating privacy, leaving the private sector to self-regulate how they collected, shared, and sold personal data. However, the White House repeatedly warned that if the private sector didn't effectively self-regulate, the

[101] The Privacy Working Group is one of the advisory groups created by the Information Infrastructure Task Force. The IITF was an inter-governmental organization charged with the coordination of government policy for the Information Infrastructure, chaired by the-Vice President Al Gore and Secretary of Commerce Ron Brown. There were three IITF Committees -- Information Policy, Applications, and Telecommunications Policy. The Privacy Working Group was one of three working groups within the Information Policy Committee. The other two were Information Access and Intellectual Property. The Privacy Working Group was made up of about twenty federal officials from such agencies as the Department of Justice, the National Security Agency, the Commerce Department, the Defense Department, the Office of Management and Budget, IRS, Census, the US Postal Service and other agencies. "Privacy Guidelines for the National Information Infrastructure: Review of the Proposed Principles of the Privacy Working Group," accessed May 3, 2023, https://archive.epic.org/privacy/internet/EPIC_NII_privacy.txt.

administration would face increasing pressure to play a more direct role in safeguarding consumer choice regarding privacy online.[102]

The framework for balancing privacy with economic growth that Clinton's administration produced ultimately rested on the awareness and choice of individuals: data-gatherers should provide data producers with a meaningful way to limit use and reuse of personal information. This would be done by providing the producers with details about what information they are collecting, and how they intended to use such data. Data-gatherers would give individuals reasonable means to obtain, review, correct, and limit the use of their personal information. It would be up to individuals to seek this information. The methods inspecting the personal data was largely left up to the service providers collecting and controlling it. In other words, in line with the Code of Fair Information Practices, the framework effectively shifted responsibility for privacy protection to individuals, and the controllers of the data could decide how to manage how they would provide access to the data.

Privacy as a Human Right in Europe

Like their counterparts in the United States, European regulators were also paying close attention to privacy and flows of personal data in the 1960s and 1970s.[103] They appreciated that electronic data were malleable, modular, easily searched, and even more easily reproduced. The Fair Practices committee pulled from commissions in West Germany, Sweden, Canada, and Great

[102] Clinton and Gore Jr., "A Framework For Global Electronic Commerce."
[103] "Decoding GDPR: Familiar Terms Could Cause Major Confusion When GDPR Takes Effect | Judicature," October 22, 2019, https://judicature.duke.edu/articles/decoding-gdpr-familiar-terms-could-cause-major-confusion-when-gdpr-takes-effect/.

Britain, that were producing similar studies with several identical items.[104] For example, Sweden's 1973 act included the right of access to all data about them, and, if the data are found to be incorrect, incomplete, or otherwise faulty, the information must either be corrected. Indeed, the Committee on Automated Personal Data Systems studied these and other examples. They also had similar outcomes in Europe, whose governments showed more willingness to limit their own agencies than they were private industry. The commissions convened throughout the 1970s to assess the role and impact of computerized data processing on individuals were influenced by the idea that privacy should include the ability to exercise some control over the use of information.

In contrast to the United States, by the early 1990s, nearly 20 countries in Europe had broad privacy or data protection statutes.[105] The laws and their enforcement systems varied from strong, independent commissions in France to an advisory model in Germany. But privacy and data protection had shifted from a collection of decentralized state laws to a centralized model as the European Union began to take on political authority.

Explanations for this attention to data protection point to developments in computing and communications technologies in the 1970s, which made feasible a massive expansion in multinational organizations. McDonalds opened the first store in Europe in 1971 followed by other multinationals that sought a uniformity of product and identity around the globe. Trade and information exchange at this scale required policies that could reconcile the interests of

[104] Ware, "Records, Computers and the Rights of Citizens."

[105] Cate, *Privacy in the Information Age*.

individual privacy with commercial interests. Standards needed to be harmonized or the burden on multinationals that acquired, stored, processed, and transferred data would be impossible for companies to manage.[106]

E.U. authorities responded with a host of guidelines and directives throughout the 1970s, 1980, and 1990s to manage the accumulation of data with citizens privacy.[107] In 1980, the Organization for Economic Cooperation and Development (OECD) released "Guidelines on the Protection of Privacy and Transborder Flows of Personal Data." This was followed in 1985 by the "Protection of Individuals with Regard to Automatic Processing of Personal Data."

In 1995, the European Commission introduced the Directive on the Protection of Individuals with Regard to the Processing of Personal Data on the Free Movement of Such Data, often referred to the Data Protection Directive.

Directives are not laws enacted directly by E.U. member nations. They are instead policy objectives developed at the executive level of the European Union that each nation must enact with its own laws, or face penalties imposed on each state.[108]

---

[106] Lloyd, *Information Technology Law*.

[107] Additionally, the Charter of Fundamental Rights of the European Union and the Treaty on the Functioning of the E.U. (TFEU) were introduced during this time. In addition to these the European Commission issued directives to member nations. It is not within the scope of this study to list all of the legal instruments.

[108] A "directive" is a legislative act that sets out a goal that all EU countries must achieve. However, it is up to the individual countries to devise their own laws on how to reach these goals. One recent example is the EU single-use plastics directive, which reduces the impact of certain single-use plastics on the environment, for example by reducing or even banning the use of single-use plastics such as plates, straws and cups for beverages. European Union, "Types of Legislation in the European Union," accessed May 6, 2023, https://european-union.europa.eu/institutions-law-budget/law/types-legislation_en.

Nevertheless, the Data Protection Directive (which I will refer to as the Directive) is cited as a major move toward standardizing legal frameworks on the protection of personal data across the then member nations.[109] This included the collection, use, and transfer of data.

Personal data was defined as any information relating to an identified or identifiable natural person, i.e. data subjects.[110] Processing of personal data meant any operation or set of operations which is performed upon personal data, whether or not by automatic means.[111] The Directive additionally distinguished between "processors" and "controllers." The processor was the entity that processed personal data on behalf of the controller. The controller was the entity

that determined the purposes and means of the processing of personal data.[112] Also, member nations should prohibit the transfer of data to countries outside the E.U. that did not meet the

[109] The Data Protection Directive is mentioned in each of the overviews of European and U.S. privacy law included in this study, as well as others that I found during my research but did not cite.

[110] An identifiable person is someone who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity."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 (EUR-Lex - 31995L0046 - EN)," text/html; charset=UTF-8, Official Journal L 281 , 23/11/1995 P. 0031 - 0050; (OPOCE), accessed May 6, 2023, https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX%3A31995L0046%3AEN%3AHTML.

[111] "…such as collection, recording, organization, storage, adaptation or alteration, retrieval,consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction." "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 (EUR-Lex - 31995L0046 - EN)."

[112] Definitions (d)-(f)of the act state that "Controller" shall mean the natural or legal person, public authority, agency or any other body which alone or jointly with others determines the purposes and means of the processing of personal data; where the purposes and means of processing are determined by national or Community laws or regulations, the controller or the specific criteria for his nomination may be designated by national or Community law. "Processor" shall mean a natural or legal person, public authority, agency or any other body which processes personal data on behalf of the controller. "Third party" shall mean any natural or legal person, public authority, agency or any other body other than the data subject, the controller, the processor and the persons who, under the direct authority of the controller or the processor, are authorized to process the data. "Directive 95/46/EC of the

standard of an adequate level of protection.[113] Data had to be processed fairly and lawfully and collected for specified, explicit and legitimate purposes. Further processing of data for historical, statistical or scientific purposes would be allowed, provided that member states provided appropriate safeguards in particular for personal data stored for long periods of time.[114] The Directive also provided criteria for lawful data, consent for which had to be freely given, specific, and informed.[115]Most importantly, the Directive provided "data subjects" (the individuals producing the data) with the right to access their data, in order to object to, correct, or erase it. The Directive did not specify the methods for obtaining access to the data, but it was

clear that the controllers would decide the process and that the controllers could opt for some kind of automated system.

The Directive was a response to several factors, according to legal scholars. First, the European Union's political leadership saw the negative and the positive potential of rapid

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 (EUR-Lex - 31995L0046 - EN)."

[113] Article 29 Working Party, "Working Document: Transfers of Personal Data to Third Countries: Applying Articles 25 and 26 of the EU Data Protection Directive" (European Commission, n.d.), 29, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/index_en.htm#maincontentSec20; Theresa Papademetriou, "Online Privacy Law (Part One): European Union," in *Online Privacy Laws : European Union & Select Foreign Countries*, ed. Ethan Williams, Privacy and Identity Protection: Law Library of Congress Global Legal Research Center (New York: Nova Science Publishers, Inc, 2012).

[114] "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 (EUR-Lex - 31995L0046 - EN)."

[115] "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 (EUR-Lex - 31995L0046 - EN)," text/html; charset=UTF-8, Official Journal L 281 , 23/11/1995 P. 0031 - 0050; (OPOCE), accessed May 6, 2023.

globalization and the development of technologies.[116] With the right strategic direction and political priorities, the internet could fuel an economy based on data and services (digital services). Without the right controls, Europe would cede control over the infrastructure of digital services and its economic promise to the United States and the expanding multinational big tech industry. Second, realizing the promise of the digital economy would require citizens' support, which meant responding to concerns about the possible misuse of their data. Steps needed to be taken to ensure the demand for digital services in the first place. Third, the scope of the European Union's authority was being expanded to policy, beyond the original economic objective of creating a single market toward integration, marking another step on the path toward the ongoing social-political unification of Europe.[117]

The guidelines and directives tied data protection to the goal of fully exploiting the potential of processing technologies. Data protection could be instrumental to reconciling economic interests – international trade and business – with the normative goals of privacy and personal liberties.[118] In order to reconcile the economic and normative goals, policymakers sought a common standard for data protection that was high enough to prevent objections to transferring data from one member state with particularly high standards to another with lower

---

[116] The political leadership includes commissioners from the 27 E.U. nations including a commission president; a parliament; and a council. See: "Political Leadership," accessed June 12, 2023, https://commission.europa.eu/about-european-commission/organisational-structure/how-commission-organised/political-leadership_en; "The Commissioners," accessed June 12, 2023, https://commissioners.ec.europa.eu/index_en.

[117] Papademetriou, "Online Privacy Law (Part One): European Union"; "Summary of: Treaty on the Functioning of the European Union EUR-Lex - 4301854 - EN," EUR-Lex, December 15, 2017, https://eur-lex.europa.eu/EN/legal-content/summary/treaty-on-the-functioning-of-the-european-union.html.

[118] Fuster, *The Emergence of Personal Data Protection as a Fundamental Right of the EU.*

ones. This happened for example when data authorities in France blocked data about French citizens collected by Fiat from being transferred to corporate offices in Italy.[119]

In short, like the White House, the E.U. executives recognized the economic promise of online commerce and the internet. The Directive was a balancing act between to the interests of individuals and the interests of industry. And the new political scope of the European Union gave them the authority to act.

European data protection laws share four main features. One, they typically apply to both public and private sectors. Two, they apply to a wide range of activities, including data collection, storage, use, and dissemination. Three, they are positive rights in that they provide the right to data protection and a right to privacy. Indeed, the right to data protection and privacy are recognized as fundamental human rights, meaning that they are the foundation for other rights.[120] In addition, they include a specialized agency with responsibility for monitoring data protection, which is not a feature that has been pursue in the same way in the United States.[121] Nevertheless implementation among the member nations was fragmented. For one, the framework was

[119] Newman, *Protectors of Privacy*.

[120] The right to data protection and the right to privacy are two distinct human rights recognized in the Charter of Fundamental Rights of the European Union, the Treaty on the Functioning of the E.U. (TFEU), and other laws to which all the EU Member States are parties. Theresa Papademetriou, "Online Privacy Law (Part One): European Union," in *Online Privacy Laws: European Union & Select Foreign Countries*, ed. Ethan Williams, Privacy and Identity Protection: Law Library of Congress Global Legal Research Center (New York: Nova Science Publishers, Inc, 2012).

[121] The fascinating but much-overlooked role these authorities, which are dedicated to data regulations, have played in privacy policymaking is outlined in Newman, *Protectors of Privacy*. The closest U.S. equivalent is the Federal Trade Commission, which early inspired the structure for oversight agencies in the European Union, but is quite different in ways that are outside of the scope of this chapter and study.

adopted in 1995 when the internet was still in its infancy. Moreover, the Directive wasn't a regulation directly applicable to legal systems of member states and data privacy standards still had to be established among a wide variety of individual states. The features that were developed beginning in the 1960s became part of the next wave of legislation leading to the GDPR.

Legal status of data protection

The GDPR changed the legal status of data protection that had existed in the decades since the Fair Information Practices was published. The regulation applied directly to E.U. member states and was based on an online environment that by 2010 included social networking sites (Facebook, Twitter) and search engines (Google), which were creating a more active market for data. Social networks changed the methods of sharing information and the advent of cloud computing (Amazon Web Services) allowed more data to be stored on remote servers instead of personal computers.

The GDPR was designed to synchronize data protection rules and legal certainty, remove obstacles in the flow of personal data within the European Union, and improve competition.[122] The legislation built on the principles established by the 1995 Directive by adding measures to limit data processing and make what was being done more transparent. For one, data controllers had to disclose information about processing in a concise, transparent, intelligible and easily accessible form, using clear and plain language.[123] A Right of Access was laid out in Article 15,

---

[122] Papademetriou, "Online Privacy Law (Part One): European Union."

[123] Nicholas Vollmer, "Article 12 EU General Data Protection Regulation (EU-GDPR)," text (SecureDataService, April 4, 2023), https://www.privacy-regulation.eu/en/article-12-transparent-information-communication-and-modalities-for-the-exercise-of-the-rights-of-the-data-subject-GDPR.htm.

which required the data controllers to provide a copy of the personal data they were processing. If the data producer made the request in an electronic format, and unless otherwise requested, the information should be provided in a "commonly used electronic form."

The GDPR also addressed the trade in personal data by requiring controllers to disclose the third-party recipients of their users' personal data and any personal data that they held but that didn't come from the user. Two other distinct measures included the right to portability and the right to be forgotten. The right of portability allows individuals to obtain a copy of their data from a service provider and transfer it easily to another service. The requirement that the data should be in a structured, commonly used, and machine-readable format was an established requirement because regulators understood data was kept in databases according to computer formats common across systems. This provision was intended to prevent lock-in to a service provider and make smaller, local operators more competitive by making data portability and interoperability requirements.

Access had clearly been important from the earliest days of data regulations. Data portability, which meant that data could be downloaded from one site and uploaded to another (Facebook to Twitter, for example), was now established as a technical prerequisite for access. Interoperability, the characteristic that makes portability possible, was also increasingly important. Interoperability makes it possible for different systems, devices, and applications to "connect" in order to transmit data automatically. The functions of interoperable components include data access, data transmission and cross-organizational collaboration.[124]

[124] "What Is Interoperability? | Definition from TechTarget," App Architecture, accessed May 8, 2023, https://www.techtarget.com/searchapparchitecture/definition/interoperability.

Discussion

This chapter has followed the way that data protection began in the 1970s, with legislation to control the processing of personal information by public authorities and large companies. I've tried to show moments when the Europe and the United States overlapped and diverged in their approaches to user rights.[125]

European laws offer more defined rights to citizens than in the United States. The E.U. regulations go so far as to establish data protection as a fundamental right, creating a normative foundation for transparency and access to personal data. There are also national agencies responsible for enforcing the GDPR. And E.U. policy is moving toward pre-emptive enforcement in current legislation with penalties for violations, as well as avenues to prevent them.

In the United States, the government has largely preferred sectoral approaches with narrowly drawn regulations targeting specific industries. Banking, credit, and telecommunications are among the most regulated sectors in terms of privacy.[126] But the normative commitment is to liberty and a minimal role for government.

---

[125] Europe adopted a design principle later recommended by the U.S. Federal Trade Commission to protect consumer privacy. The principle, privacy by design, builds in data processing procedures into technology from the start. Other best practices included making privacy the "default setting" for commercial data practices and giving consumers greater control over the collection and use of their personal data through simplified choices and increased transparency. See "Privacy by Design," *General Data Protection Regulation (GDPR)* (blog), accessed May 6, 2023, https://gdpr-info.eu/issues/privacy-by-design/. Also: "Protecting Consumer Privacy in an Era of Rapid Change: Recommendations For Businesses and Policymakers," Federal Trade Commission, March 1, 2012, https://www.ftc.gov/reports/protecting-consumer-privacy-era-rapid-change-recommendations-businesses-policymakers; noted in Papademetriou, "Online Privacy Law (Part One): European Union."

[126] Robert Gellman, "Review of 'Data Privacy Law: A Study of United States Data Protection.' By Paul M. Schwartz and Joel R. Reidenberg. Charlottesville, VA: Michie. 1996. 486 Pages. ISRN 1558343776.," *Government Information Quarterly* 14, no. 2 (1997): 215–17; Lloyd, *Information Technology Law*.

Although the differences between the Europe and the United States are important, they should not be overstated. First, privacy and data protections regulations came from back and forth exchanges between the Europe and the United States. Both E.U. and U.S. laws are instruments for balancing citizens' privacy interests with economic growth and political power.[127] Moreover, the regulations continue to center on users actively seeking information that is held by data processors. The by-design approaches favored on both continents require industry to build information systems that incorporate user rights into the database and computing architecture. These principles have been consistent for a half-century and shared in Europe and the United States.

Second, the criteria stop at formats and making data accessible, whose effectiveness at accommodating user rights are limited. The main question, given the data politics involved, is how users will claim and exercise the rights with the legal and technical resources at their disposal.[128] Are these policy by design mechanisms in place effective, and how are they tailored for users to take control over their data?[129] To evaluate these questions about the available resources, I pursued an empirical assessment of data rights, which suggest that, while technical frameworks are insufficient, policy frameworks are a greater source of concern. Chapter 2 outlines my methodology for these assessments.

---

[127] Bradford, *The Brussels Effect*; Julia M. Fromholz, "The European Union Data Privacy Directive," *Berkeley Technology Law Journal* 15, no. 1 (2000): 461–84; Gellman, "Review of 'Data Privacy Law: A Study of United States Data Protection.' By Paul M. Schwartz and Joel R. Reidenberg. Charlottesville, VA: Michie. 1996. 486 Pages. ISRN 1558343776."; Paul M. Schwartz, *Data Privacy Law : A Study of United States Data Protection* (Charlottesville, Va: Michie, c1996); Peter P. Swire, *None of Your Business : World Data Flows, Electronic Commerce, and the European Privacy Directive* (Washington, D.C.: Brookings Institution Press, 1998); Kamaal Zaidi, "Harmonizing U.S.-EU Online Privacy Laws: Toward a U.S. Comprehensive Regime for the Protection of Personal Data," *Michigan State University Journal of International Law* 12, no. 1 (2004 2003): 169–98.

[128] Valtysson, Jorgensen, and Munkholm, "Co-Constitutive Complexity."

[129] Valtysson, Jorgensen, and Munkholm.

Chapter 2: Methodology

Scholars have identified problems with the way that user rights are accommodated by looking at content moderation and agreements that social media subscribers have to sign in order to use services like Facebook and Twitter.[130] Others have examined researchers' access to user data.[131] This study adds an empirical test of user rights to those studies by looking at two related issues, access to personal data and the right to transfer data, referred to as portability.[132] This chapter describes the hands-on, comparative methods I used to collect, compare, and transfer data from Twitter, Facebook, and Instagram social media accounts .

This chapter is structured as follows: The first section provides an overview of two complementary, qualitative methodologies for studying software: the first, reverse engineering and, the second, the walkthrough method.[133] I discuss how each was used in this study, their potential limitations, and how I combined the two in order to address some of those limitations.

My attempt to compare techniques for portability and the data was impeded for several reasons explained in Chapter 3. For this reason and for the sake of clarity, I have included the portability testing methodology in Chapter 3 rather than here. There is good reason for this

---

[130] Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (New Haven: Yale University Press, 2018); Nancy Kim, *Wrap Contracts: Foundations and Ramifications*, *Wrap Contracts* (Oxford University Press, 2013), https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199336975.001.0001/acprof-9780199336975; Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (PublicAffairs, 2020).

[131] Axel Bruns, "After the 'APIcalypse': Social Media Platforms and Their Fight against Critical Scholarly Research," *Information, Communication & Society* 22, no. 11 (September 19, 2019): 1544–66, https://doi.org/10.1080/1369118X.2019.1637447; Jean Burgess and Axel Bruns, "Twitter Archives and the Challenges of 'Big Social Data' for Media and Communication Research," *M/C Journal* 15, no. 5 (October 11, 2012), http://journal.media-culture.org.au/index.php/mcjournal/article/view/561.

[132] Ausloos and Dewitte, "Shattering One-Way Mirrors – Data Subject Access Rights in Practice."

[133] Light, Burgess, and Duguay, "The Walkthrough Method."

decision. The present chapter focuses narrowly on collecting data from multiple sources using multiple techniques. Transferring data is not as complex because data should be sent from system to another. Facebook to Google in my case. Reverse engineering was important to me during the process and many of my observations apply in both cases. However I found it unwieldy to combine the two here and confusing for the reader and the researcher.

To test access, I collected data using three different methods and then assessed the results. The data belonged to an organization, the Graduate Workers of Columbia. Section two includes details about the technical methods I used for the collection – Archive-It, Webrecorder, and social media platform downloads.[134] Archive-It and Webrecorder are licensing services used to collect data from web- and social media sites. Platform downloads allow subscribers to collect their data from platforms directly. I refer to all three as services because they feature an interface and predetermined settings that simplify collection.[135] This is different than other methods (i.e. "scrapers") that require direct engagement with the underlying computer code and system

[134] This work took place between 2018-2020. Webrecorder's subscription operations are now, as of January 2021, largely part of a new entity called Conifer. This chapter reflects Webrecorder as it operated when I was using it and I continue to refer to Webrecorder rather than using Conifer.

[135] There is no platform download for websites. Websites or the code and objects of which they consist can be to varying degrees copied, downloaded, captured in screen shots or by web archiving services.

architecture.[136] Similarly, I excluded emails because the methods I tested are not designed for that medium.[137]

The third section describes the process of collecting data from Twitter, Facebook, and Instagram, stopping short of a detailed analysis, which is the subject of subsequent chapters. In the remainder of the third section I briefly describe the design of this study, including the techniques I employed, the cases I chose, and the nature of the data I collected, as well as the theoretical foundation of my methodology. I also found it necessary to include technical information about the services, websites, and social media in order to better understand the problems with data extraction and selective access that I encountered. I introduce these technical details in a practical way and as early as possible and discuss them at length in Appendix A.[138] The technical details are important for understanding not only how the technical systems examined here work, but also later understanding the results and assessments.

In this same vein, I include snapshots of results from the data collections. The format of the results are salient when considering the available options for data collecting. For example, Facebook blocked my attempt to use Webrecorder to collect data from the account. Without

---

[136] Scraping is a process by which data is collected from websites and saved for further research. Web scraping can be done through manual selection or it can involve the automated crawling of web pages using pre-programmed scraping applications. Unlike web archiving, web scraping is mostly used for gathering textual data. Web scraping tools also allow you to structure the data as you collect it. Instead of massive unstructured text files, scraped data can be reconfigured for spreadsheets or database formats that allow you to analyze and use it in research. It is a common method for scholarly research and journalism using social media data. See: Kent Emerson, "An Introduction to Web Scraping for Research," Research Data Services, November 7, 2019, https://researchdata.wisc.edu/news/an-introduction-to-web-scraping-for-research/.

[137] Matthew Connelly and Rohan Shah, "Here's What Data Science Tells Us about Hillary Clinton's Emails," *Washington Post*, November 2, 2016, https://www.washingtonpost.com/news/monkey-cage/wp/2016/11/02/heres-what-data-science-tells-us-about-hillary-clintons-emails/.

[138] Technical instructions and terms are also provided in detail in a separate chapter.

alternatives, the only option left is a platform download web portal offered by Facebook, which makes two formats available. One version is a visual facsimile of the original Facebook site, while the second is a computer code version. They both have their advantages based on the goal of the user. But the latter is only useful if the user has technical skills to analyze the code, skills which are necessary to understanding what data the results include or exclude.

Seeing the format of results highlights the need for regulatory guidelines that take into account the quality of the results particularly when it comes to social media platform downloads. They rely on Twitter, Facebook, and Instagram to set the standards for what is made accessible and how it is delivered, regardless of the needs of users. The following chapter outlines the limits of current regulations and platform downloads in detail based on my analysis. This chapter provides the context for understanding those findings.

Reverse Engineering and the Walkthrough

Researchers interested the social, cultural, and political effects of social media have focused critical attention on software code, algorithms, protocols, and the data aggregated in servers. The obstacles to these inquiries, documented in a variety of disciplines, range from the opacity of technical systems that are difficult to understand and deconstruct, to the opacity of the companies creating them, which keep as much as possible their operations closed to scrutiny.[139] One

[139] James Allen-Robertson, "Critically Assessing Digital Documents: Materiality and the Interpretative Role of Software," *Information, Communication & Society* 0, no. 0 (July 17, 2017): 1–15, https://doi.org/10.1080/1369118X.2017.1351575; Justin F. Brunelle et al., "The Impact of JavaScript on Archivability," *International Journal on Digital Libraries; Heidelberg* 17, no. 2 (June 2016): 95–117, http://dx.doi.org.ezproxy.cul.columbia.edu/10.1007/s00799-015-0140-8; Burgess and Bruns, "Twitter Archives and the Challenges of 'Big Social Data' for Media and Communication Research"; Kevin Driscoll and Shawn Walker, "Big Data, Big Questions| Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data," *International Journal of Communication* 8, no. 0 (June 16, 2014): 20; Sandra González-Bailón, Rafael E. Banchs, and Andreas Kaltenbrunner, "Emotions, Public Opinion, and U.S. Presidential Approval Rates: A 5-Year Analysis of Online Political Discussions," *Human Communication Research* 38, no. 2 (April 1, 2012): 121–43, https://doi.org/10.1111/j.1468-2958.2011.01423.x; Rob Kitchin, "Thinking Critically about and Researching

strategy for tackling these obstacles is reverse engineering, so named because it involves working backwards through the development or design process of a system with little or limited knowledge about the original methods used to construct it.[140] Originally developed for engineering tasks, researchers in communications, media studies, and information sciences have adapted the method to understand how a device, process, system, or software application works. Nicholas Diakopoulos, for instance, used reverse engineering to study algorithms. In one instance, he looked at how prioritization, ranking, or ordering by algorithms served to emphasize or bring attention to certain search results at the expense of others.[141] In another study, Justin Chun-Ting Ho applied the approach to a Facebook ranking algorithm in order to identify features of a post that would affect its odds of being selected in a search.[142]

Reverse engineering involves trade-offs. One in particular is that the researcher remains limited to an outsider-looking-in view with a fuzzy glimpse into how something works.[143] The second limitation is that the ability to replicate results – the standard for credible research – is limited and researchers must be cautious when making assertions about their findings. Nevertheless, reverse engineering can be a useful tool in the case when other options are not

---

Algorithms," 2017, https://doi.org/10.1080/1369118X.2016.1154087; Gary King, Jennifer Pan, and Margaret E. Roberts, "Reverse-Engineering Censorship in China: Randomized Experimentation and Participant Observation," *Science* 345, no. 6199 (2014): 1–10; Tromble, Storz, and Stockmann, "We Don't Know What We Don't Know."

[140] Meredith Broussard, *Artificial Unintelligence: How Computers Misunderstand the World*, Illustrated edition (Cambridge, Massachusetts London, England: The MIT Press, 2019); Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, 1 edition (New York: NYU Press, 2018).

[141] Nicholas Diakopoulos, "Algorithmic Accountability: On the Investigation of Black Boxes" (New York: Tow Center for Digital Journalism, Columbia University, December 3, 2014), https://www.cjr.org/tow_center_reports/algorithmic_accountability_on_the_investigation_of_black_boxes.php/; Nick Diakopoulos, "Algorithmic Accountability," *Digital Journalism* 3, no. 3 (2015): 398–415.

[142] Ho, "How Biased Is the Sample?"

[143] Kitchin, "Thinking Critically about and Researching Algorithms."

available especially if supplemented by reporting techniques that, in addition to interviews where possible, include reviews of documents – meeting notes, terms of use, developer blogs, etc.[144] For example, in the case of social media, privacy policies are essential for finding practical details about a company and serve as the basis for evaluating the practical operations of the service. The caveat is that the policies lack clarity, accessibility, and completeness, although this too can be a finding in its own right.[145] The third drawback of reverse engineering is that there is no set methodology to follow and no instructions. The walkthrough method, which I describe in the next section, gave me a device for systematizing my approach.[146]

The term walkthrough appears in reverse engineering tutorials. In software engineering or application development, a walkthrough refers to a step-by-step overview of something including a software product, a how-to, or the actual assessment of outputs. Ethnographers adapted the approach to critically analyze software application interfaces and the overlap between the two disciplines help to explain how the two methods fit together in my experience.

A walkthrough involves step-by-step observations of screens; features; and sequences of activities; as well as a close reading of documentation in order to assess the intended user and use.[147] For example, looking at user interface arrangements focuses attention on how software guides users through activities according to where buttons and menus are placed. Looking at functions and features help to see how something like pop-up windows, compulsory fields, and

---

[144] Diakopoulos, "Algorithmic Accountability," December 3, 2014.

[145] Ausloos and Dewitte, "Shattering One-Way Mirrors – Data Subject Access Rights in Practice."

[146] I found that drawing conclusions based on my observations was difficult because I could interpret the results in a variety of ways. I initially focused on archiving but contexts changed during the study and interpreting results based on regulations became a more meaningful.

[147] Light, Burgess, and Duguay, "The Walkthrough Method."

requests made by the app to link with other user accounts requires certain actions while excluding others. Some buttons like "Report" or "Share" may be smaller or harder to find than others thus leading to particular behavior. This could include Twitter's lack of an edit feature, which requires users to either delete their post or add a second post explaining the first one.

Documentation was also telling. For example, Twitter's platform download set-up informs the user that they will be provided with "the information that we believe is most relevant and useful to you." This is a technical communication about how the service works and a statement about Twitter's entitlement to withhold data that may not comply with regulations.

These details often get overlooked when code and algorithms are the main object but may be salient for interpreting the data.[148]

I could have opted for an ethnography of a coding or design team to provide detailed understanding that reverse engineering lacks, potentially increasing the accuracy of the research. But access to personnel and their work can be difficult if not impossible in companies that forbid their employees and contractors to discuss their work. In other instances, the team may not want, or feel they have the time, to talk to a researcher.[149] They may not want to be observed or have their work scrutinized.

---

[148] Light, Burgess, and Duguay.

[149] Abiola O. Fanimokun, Gary Castrogiovanni, and Mark F. Peterson, "Developing High-Tech Ventures: Entrepreneurs, Advisors, and the Use of Non-Disclosure Agreements (NDAs)," *Journal of Small Business and Entrepreneurship* 25, no. 1 (2012): 103-119,127-128; "Moderators 'not Denied Access' to Non-Disclosure Agreements, Facebook Insists," *BreakingNews.Ie*, May 20, 2021, sec. Ireland, https://www.proquest.com/docview/2529404998/citation/53CC68CCD39B4920PQ/1; Madhumita Murgia, "A Tale of Two Facebook Whistleblowers," *FT.Com*, June 22, 2022, https://www.proquest.com/docview/2692576026/citation/7E99496D23A04DE7PQ/1; Hannah Murphy and Kiran Stacey, "Facebook Libra: The inside Story of How the Company's Cryptocurrency Dream Died," *FT.Com*, March 10, 2022, https://www.proquest.com/docview/2648412117/citation/EEC3DEE5A6F64CE3PQ/1; Sarah T. Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media*, Illustrated edition (New Haven: Yale University Press, 2019).

In addition, I wanted to capture the direct experience of using the services to collect data. I drew on qualitative studies about data collecting, some specific to Archive-It, that observed practitioners.[150] But I decided that adapting reverse engineering and the walkthrough method, combined with an auto-ethnographic approach, would avoid difficult negotiations over access to workplaces and yield a detailed study grounded in first-hand practice.

I saw more utility in an auto-ethnography and targeted technical questions for my purposes. However, originally I planned on an ethnographic approach. Archive-It's director, Jefferson Bailey, agreed to my early plan to observe the Archive-It team working at the San Francisco headquarters and had given other ethnographers access as well. Later, he offered only limited access to the team's engineers for specific questions and made other administrative staff available instead. My impression from conversations was that he was trying to keep from adding to the workload of engineers. I was able to talk to them by attending meetings for Archive-It subscribers and other venues. I also interviewed former engineers and staff.

Webrecorder design and engineering staff were accessible but the latter were hesitant to have an ethnographer on site because they did not want to be distracted from their work. This was a mistake on my part because in a proposal which the group asked for I added that I would be on site and ask questions about their work in progress. It was this interruption that they

[150] Emily Maemura et al., "If These Crawls Could Talk: Studying and Documenting Web Archives Provenance," *Journal of the Association for Information Science & Technology* 69, no. 10 (October 2018): 1223–33, https://doi.org/10.1002/asi.24048; Jessica Ogden, Susan Halford, and Leslie Carr, "Observing Web Archives: The Case for an Ethnographic Study of Web Archiving," in *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17 (New York, NY, USA: ACM, 2017), 299–308, https://doi.org/10.1145/3091478.3091506; Jessica Ogden, "'Everything on the Internet Can Be Saved': Archive Team, Tumblr and the Cultural Significance of Web Archiving," *Internet Histories* 6, no. 1–2 (April 3, 2022): 113–32, https://doi.org/10.1080/24701475.2021.1985835; Ed Summers and Ricardo Punzalan, "Bots, Seeds and People: Web Archives as Infrastructure," *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, 2017, 821–34, https://doi.org/10.1145/2998181.2998345.

especially wanted to avoid but staff later told me that the founder, Elia Kramer, was reluctant and ultimately rejected my request even when I suggested I would be a silent observer.

Regarding the social media operators, I was able to speak to Twitter management staff in mid- 2019. I made the contact by attending an event in May 2019 at the Twitter offices in Manhattan, New York, for developers. The event was designed to introduce Twitter's new API. APIs, which I have said elsewhere, are used to formalize the way information is requested and exchanged between computers and are becoming the dominant method of data collecting for the commercial partners of Facebook and Twitter, as well as researchers.

The interview with Twitter came about because of my interest in reaction to the limits Twitter had recently imposed on the amount of data, and the methods for collecting it, which social scientific researchers worried would affect their work.[151] Twitter data had been easy to access compared to Facebook. But Twitter, along with Facebook, had recently changed their API to make that access harder. As a whole, the restrictions can be attributed to two main issues. One was a growing concerns over user privacy.[152] These concerns led to requirements for privacy protections in the 2018 E.U. data protection law, the General Data Protection Regulation, which Facebook and Twitter responded to by imposing more restrictions on access to data. The second was an attempt by Facebook and Twitter to better monetize user data. Researchers and archivists almost immediately began to complain about access to social media data and confusing rules about accessing and sharing that data imposed by Facebook and Twitter.[153]

---

[151] Bruns, "After the 'APIcalypse.'"

[152] Cadwalladr and Graham-Harrison, "Revealed."

[153] Bruns, "After the 'APIcalypse'"; Justin Littman, "Twitter's Developer Policies for Researchers, Archivists, and Librarians," *Medium* (blog), January 7, 2019, https://medium.com/on-archivy/twitters-developer-policies-for-researchers-archivists-and-librarians-63e9ba0433b2.

In 2019 the Twitter management staff I spoke to said they were interviewing researchers in order to understand what kind of data the research community was seeking and how Twitter might be able to respond to their needs while still protecting the privacy of their subscribers. I saw the interview as an opportunity to clarify some of the complaints by archivists and fellow researchers, especially about APIs, and to clarify my own results with collecting social media data, in particular APIs and the platform downloads.

I did not have an opportunity to talk to anyone at Facebook or Instagram. Beginning in 2019, I sent two messages through the online help center request form, none of which were answered. Since that time, Facebook has made more contact options available but, despite using the different channels, I still have received no response.

The problem with follow-up contact appears to be categorical, at least in part: Facebook provides self-help articles about privacy and, ultimately, an address for the data controller "responsible for your information." This information is listed on Facebook's privacy policy page: https://www.facebook.com/privacy/policy under a subsection, "How to contact Meta with questions." But if the option is not listed in the dropdown menu provided, there is no way to proceed to contacting anyone. The options stop there. The other option is to file a complaint with the lead supervisory authority, the Irish Data Protection Commission, or a local supervisory authority. Twitter has similar categories.

My question is about a design feature and a product, not about my data or about my privacy. Since my question does not match the pre-selected categories, there is no easy way to direct my questions.[154] It is entirely likely that my experience mirrors any other subscriber who wants to know more about these questions.

[154] Although I have since sent messages to the Irish Data Protection Commission.

I also sought technical advice from API experts and watched tutorials about using APIs, which I was already familiar with from a data science program in 2015. Reverse engineering seemed like the best and, in the case of platform downloads, possibly the only method for my research.

Using myself as the subject, I adapted reverse engineering and the walkthrough method to guide my comparison of data collecting from Archive-It, Webrecorder, and platform downloads. The walkthrough added precision to reverse engineering by guiding my attention to particular features or irregularities of both the services and the collected data. The walkthrough also gave me a way of recording the impact of the interface design and features on my experience and their significance.

I supplemented this approach with a review of technical documentation, terms of service and use documents, as well as interviews, events, meetings, and tutorials. I relied on these sources for practical instructions for how to use the services to collect data. This was necessary in the case of platform downloads particularly because so little documentation about them is available. But platform downloads appear to be a combination of interface, database, and API. The following sections details my data gathering process with details about the service providers, the services, their use, and the results.

Collection methods

Archive-It

Web archiving is the practice of collecting and preserving resources from the web (including whole websites or individual pages and media). The practice is designed to preserve the look and feel of websites. Web archiving can be done through manual selection, or it can involve the

automated extraction of web pages using pre-programmed applications. There are many applications for web archiving. The most well known design uses an automated process called "crawling" to collect pages from the web and store them on servers.[155]

Archive-It and Webrecorder are Archiving-as-a-Service (SaaS) packages used to automate the process of data collecting.[156] They are based on a licensing model in which subscribers pay for a bundle of services, including software and storage. The license gives subscribers access to the software and they can store content remotely on a server belonging to the service provider. The service provider in turn has dedicated remote servers, or has an agreement with a cloud storage providers such as Amazon or Google. Common examples are email, calendaring, and office tools (such as Microsoft Office 365).[157] Archive-It is the largest such service today and the software that it is built on, Heritrix, was one of the earliest

methods for collecting websites. Archive-It provides subscribers access to the software and storage on the organization's servers. Webrecorder works in a similar way but stores subscribers' collection remotely, on Google or Amazon cloud servers.

As I have explained elsewhere, I began collecting the Graduate Workers of Columbia Twitter, Facebook, and Instagram accounts and website in 2018. Originally, I intended to create an archival collection of GWC's social media accounts (where much of the organizing activities of the group are presently recorded), which would be donated to a repository for future research. I decided to base my doctoral research on the project. I was interested in social media archiving,

---

[155] "An Introduction to Web Archiving for Research," Research Data Services, October 15, 2019, https://researchdata.wisc.edu/news/an-introduction-to-web-archiving-for-research/.

[156] Maemura et al., "If These Crawls Could Talk."

[157] "What Is SaaS? Software as a Service | Microsoft Azure," accessed May 18, 2023, https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-saas.

which, in 2018, was a relatively new subject among communications scholars interested in documenting websites for archival and research purposes (i.e. web archiving).[158] Members of GWC agreed to the idea of an archival collection and were aware of my overlapping research interests.

I decided to use this method because I was familiar with Archive-It and had worked for the parent organization, Internet Archive, creating a collection of social media accounts. Archive-It is optimized for site-wide collecting by large archival institutions that make up the majority of the clientele, such as the Library of Congress and Columbia University. The service was developed in the early 2000s together with, and marketed to, cultural institutions and university libraries serving institutional mandates for record-keeping, as well as historians and researchers.

Archive-It was meant to make the process collecting easier especially for practitioners with little in the way of technical skills.[159] The users are actually operating a bot, called a crawler, which captures snapshots of live web content (the name widely used in media studies and web archiving to distinguish between active websites and ones that have been taken offline but can still be displayed). Crawlers are programmed to identify materials on the live web that belong in a collection, based upon the user's selection of URLs. This can be done by creating what is called a "seed list," which identifies the website addresses (URLs) that the user wants to collect. I might include https://journalism.columbia.edu/ and then create a list of pages on that

[158] Miguel Costa, Daniel Gomes, and Mário J. Silva, "The Evolution of Web Archiving," *Int. J. Digit. Libr.* 18, no. 3 (September 2017): 191–205, https://doi.org/10.1007/s00799-016-0171-9; Julien Masanès, ed., *Web Archiving*, 2006 edition (Berlin ; New York: Springer, 2006).

[159] Ian Milligan, Nick Ruest, and Jimmy Lin, "Content Selection and Curation for Web Archiving: The Gatekeepers vs. The Masses," in *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, JCDL '16 (New York, NY, USA: ACM, 2016), 107–10, https://doi.org/10.1145/2910896.2910913.

site, such as "student work" (https://journalism.columbia.edu/student-work) and then drill down from there to the type of student work categorized by media, such as the page, writing (https://journalism.columbia.edu/student-work?type=83). This is just one example, but these seed lists are important for specific collections. Another approach, which is more likely for social media, is to type in the URL, such as for Columbia Journalism School (https://twitter.com/columbiajourn) and specify how much of the site should be collected. Since Archive-It is mainly used by researchers, historians, and archivists creating collections for cultural institutions and university libraries like Columbia University, the user will have an idea of what they want based on the needs of institutional mandates. In the case where the user is not sure, they would look through the site first or simply collect as much as possible by selecting very broad settings. With time the user would begin to identify what they do or do not want. In my case, I wanted to capture everything, which is another approach but one that is not sustainable because of the time that it takes to check for missing data and because data storage can be expensive. A social media account include a variety of media including photos and videos with large file sizes. Archive-It provides data storage for subscribers, who make decisions about what to collect based on their budget.

After the URLs are identified, the user sets rules that instruct the crawler to collect snapshots of links and other objects from those pages.[160] These are packaged in a file that can be displayed using a specialized application produced by the Internet Archive, called the Wayback Machine.[161] Heritrix can be prevented from crawling by a few technologies. One is by setting

[160] Steve Schneider and Kirsten Foot, "Archiving of Internet Content," in *The International Encyclopedia of Communication*, ed. Wolfgang Donsbach (Blackwell Publishing, 2008).

[161] The Internet Archive Wayback Machine is a service that allows people to visit archived versions of websites. Visitors to the Wayback Machine can type in a URL, select a date range, and then begin surfing on an archived version of the Web. Imagine surfing circa 1999 and looking at all the Y2K hype, or revisiting an older version of

Archive-It to avoid pages that include a robots.txt file, which is a "Keep Out" sign for bots. A robots.txt file identifies which URLs a crawler can access on your site, mainly to avoid overloading a site with requests. They are a standardized convention and can be ignored. Archive-It's default setting is to ignore them.

Archive-It, crawlers, and conventions like robots.txt date back to the 1990s, when websites were made up of individual files called documents or pages, with text, photos, and other media. The documents were designed with HTML to be displayed in a web browser one at a time and they functioned fairly predictably based on established design rules, or protocols. The web was like a mini-van: a slow but reliable vehicle for moving people around to places they wanted to go.

However, websites and the contents on it are no longer static sites with documents and HTML, and they do not function predictably. Well before 2018, HTML was animated by new designs and features that changed states without triggering effects that web archiving software depended on. Now the web was like a race-car: shiny, fast, and seemingly unpredictable in the way that it worked. Archive-It developed new applications that recorded web browser activity, Umbra and Browzler, in order to adapt to the new design.

---

your favorite Web site. The Wayback Machine is also the system used to display captures of content using Archive-It. The Wayback Machine makes it possible to reconstruct the "snapshots" collected by the crawler and view them as close as possible as they originally appeared. The reconstruction depends on what the crawler collected so missing links, for example, will result in a different version than the original one. In addition, the Wayback Machine is not able to display some formats. This is a confounding and increasingly common problem that requires using Archive-It's quality assurance features to look at whether an element was not collected or cannot be displayed because of the format. The next question if collection was the problem, is why. For more about both uses of the Wayback Machine, see: "Wayback Machine General Information – Internet Archive Help Center," accessed May 26, 2023, https://help.archive.org/help/wayback-machine-general-information/; Nancy Watzman, "Wayback Machine Captures Melania Trump's Deleted Internet Bio | Internet Archive Blogs," *Wayback Machine Captures Melania Trump's Deleted Internet Bio* (blog), July 28, 2016, https://blog.archive.org/2016/07/28/wayback-machine-captures-melania-trumps-deleted-internet-bio/.

Researchers and repositories were also trying to adapt existing methods to social media. Compared to websites, social media are activity streams that may be updated after collection happens and include layers of engagement – retweets, mentions of other users, likes, un/favoriting, deletions and so on.[162] The ribbon of posts are, in the words of Stine Lomborg, emergent, editable, and undergoing a continual process of development.[163] A single tweet, for example, may contain a video, photograph, emoji, link to an article, hashtag, and @ handle signaling to another user, which provide valuable contextual information, but which also provokes thorny privacy issues. My original premise was that the design of social networking sites were incompatible with crawlers. I chose Webrecorder and platform downloads for comparison, in order to test my premise.

Webrecorder

The Webrecorder.io software developed by a former Internet Archive engineer Ilya Kreymer under the auspices of the digital art organization Rhizome, housed at the New Museum in New York.[164] As the name implies, the software records a user's web interactions by capturing browser interactions instead of through a crawler.

An alternative was a practical necessity because, as I noted before, crawlers are bots, similar to those used for search engine indexing. They are designed for sites whose architecture

[162] Acker and Kriesberg, "Tweets May Be Archived."

[163] "Researching Communicative Practice: Web Archiving in Qualitative Social Media Research," *Journal of Technology in Human Services* 30, no. 3–4 (July 1, 2012): 219–31, https://doi.org/10.1080/15228835.2012.744719.

[164] Dragan Espenschied, "Rhizome Releases First Public Version of Webrecorder," *Rhizome.Org* (blog), August 9, 2016, http://rhizome.org/editorial/2016/aug/09/rhizome-releases-first-public-version-of-webrecorder/; Rhizome, "Rhizome Awarded $600,000 by The Andrew W. Mellon Foundation to Build Webrecorder," Rhizome, January 4, 2016, http://rhizome.org/editorial/2016/jan/04/webrecorder-mellon/.

was made up of HTML, pages, and links and not for the kind of networked sites that populate the web with personalized content that changes states according to who is using it and when. These sites is often called "dynamic," to contrast them with HTML-based "static" sites.

Crawlers are designed for static sites. By comparison, Umbra, Brozzler, and Webrecorder are optimized for dynamic sites. Webrecorder was distinguished by an attention to scale and ethics.

With philanthropic funding from the Andrew Mellon Foundation, Webrecorder was launched to cater to individuals and communities pursuing what in the humanities is referred to as critical digital cultures, particularly those whose work engaged artistically with digital technology and the internet.[165] Alongside Kreymer, the team behind the software service included software developer and media artist Dragan Espenschied, Anna Perricci, an expert web archivist experienced with technically and ethically challenging collections, and lead designer, Pat Shiu, also a media artist.[166] They all had deep technical knowledge about digital media and an ethical commitment to seeking a new model of web archiving. They sought to develop a "…human-centered archival tool to create high-fidelity, interactive, contextual archives of social media and other dynamic content, such as embedded video and complex JavaScript, addressing our present and future."[167]

---

[165] Rhizome, "Rhizome Awarded $600,000 by The Andrew W. Mellon Foundation to Build Webrecorder"; "Mellon Foundation," accessed July 2, 2023, https://www.mellon.org/grant-details/webrecorder-a-web-archiving-tool:-phase-two-20444133.

[166] Sian Evans, Anna Perricci, and Amy Roberts, "'Why Archive?' And Other Important Questions Asked by Occupiers," in *Informed Agitation: Library and Information Skills in Social Justice Movements and Beyond*, ed. Melissa Morrone (Library Juice Press, 2014).

[167] Espenschied, "Rhizome Releases First Public Version of Webrecorder"; Morgan McKeehan, "Symmetrical Web Archiving with Webrecorder, a Browser-Based Tool for Digital Social Memory. An Interview with Ilya Kreymer | NDSR – NY," *National Digital Stewardship Residency (NDSR)* (blog), February 23, 2016, https://ndsr.nycdigital.org/symmetrical-web-archiving-with-webrecorder-a-browser-based-tool-for-digital-social-

High fidelity, user-control, and scale were important to preserving the "look and feel" of digital objects that register information as well as experiences. This was something that crawlers do not do well. In addition, Webrecorder sought to address issues experienced by artists working in proprietary digital environments. These artists considered the profiles they made on Instagram or other platforms to be art and they wanted the profiles to attract comments. But such works might become part of the environment that a museum collector would want to archive but with permission from the artists. Webrecorder was meant to address the need for new models responsive to the ethics of the artists and the visitors to their sites, which aligned with a critical debate over a connection between archives and racism, patriarchy, and colonialism, as well as data surveillance and privacy rights.[168] This debate was the product of both larger social and cultural movements about local knowledge, history, social consciousness, vulnerability, access, power, or control heightened by the exposure of individuals on social media.[169] The marketing material emphasized a "human scale" to collecting data that would avoid collecting methods that indiscriminately harvested data.

A key argument was that data collecting at this scale could inadvertently cause harm to people by exposing their identity and activity without their knowing.[170] Political protesters were

---

memory-an-interview-with-ilya-kreymer/; Rhizome, "Rhizome Awarded $600,000 by The Andrew W. Mellon Foundation to Build Webrecorder."

[168] Michael Connor, "Ethics and Archiving the Web" (Rhizome National Forum on Ethics and Archiving the WEb, New Museum, New York, June 27, 2018), https://vimeo.com/277335998.

[169] Arjun Appadurai and Neta Alexander, *Failure*, 1 edition (Cambridge, UK ; Medford: Polity, 2019).

[170] *Ethics and Archiving the Web: Stewardship and Usage* (New York, New York, 2018), https://vimeo.com/276935105; see also Ed Summers, Bergis Jules, and Vernon Mitchell Jr., "Documenting the Now: Ethical Considerations for Archiving Social Media Content Generated by Contemporary Social Movements: Challenges, Opportunities, and Recommendations," Documenting DocNow, July 19, 2018, https://www.docnow.io/docs/docnow-whitepaper-2018.pdf.

considered at risk especially those whose employers frown on political activity. Moreover, collecting personal information without consent would be double harm to historically marginalized groups.

From this perspective, Webrecorder was intended to offer a curation-oriented design that proponents of ethical data collecting immediately recognized as important to their efforts to create collections from historically significant events but that helped to mitigate the risk to participants.[171] The ideal use of Webrecorder would be creating high-quality captures of select objects for both individuals collecting their own work and researchers and repositories seeking alternatives to crawlers for technical and ethical reasons. Webrecorder made these issues central to its promotional outreach including a conference hosted by a New York City-based digital media art organization, Rhizome, in February 2018 called "The National Forum on the Ethics of Web Archiving."

[171] Ed Summers, "Introducing Documenting the Now," *Maryland Institute for Technology in the Humanities* (blog), February 17, 2016, https://mith.umd.edu/introducing-documenting-the-now/.

Platform downloads

Platform downloads are a useful way to back up an account before deleting it.[172] The feature also fits with the interest in keeping social media as a personal archive.[173] Individuals could use one or more of the free versions of web archiving services but the latter can be labor intensive and technically challenging for someone who wants to create a personal archive, or for a small organizations such as labor archives.[174] Platform downloads could be an option within their reach.

However, Twitter and Facebook made platform downloads available because of requirements to provide personal data to the producers formalized in the GDPR, the prevailing European Union data protection regulation at the time I was testing the download.

A key provision in the GDPR is access to personal data and portability. These provisions allow subscribers to gauge what data the platforms are collecting about them overall and move their data elsewhere. For example, if a user doesn't agree with the privacy policy of a social media provider. They want to stop using it immediately, but don't want to lose the content they have created.[175] As I outlined in the previous chapter, the GDPR, like previous regulations on

---

[172] Big Brother Watch Team, "How Can I Download a Copy of My Facebook Data? What Is Included – and What Isn't? — Big Brother Watch," How can I download a copy of my Facebook data? What is included – and what isn't? — Big Brother Watch, March 23, 2018, https://bigbrotherwatch.org.uk/2018/03/how-can-i-download-a-copy-of-my-facebook-data-what-is-included-and-what-isnt/; Jean Burgess and Axel Bruns, "Twitter Archives and the Challenges of 'Big Social Data' for Media and Communication Research," *M/C Journal* 15, no. 5 (October 11, 2012), http://journal.media-culture.org.au/index.php/mcjournal/article/view/561; Josh Constine, "Instagram Launches 'Data Download' Tool to Let You Leave," *TechCrunch*, April 24, 2018, https://social.techcrunch.com/2018/04/24/instagram-export/; Jefferson Graham, "Download Your Facebook Data: How to Do It and What You Might Find," *USA Today*, March 30, 2018, https://www.usatoday.com/story/tech/talkingtech/2018/03/30/downloaded-all-my-facebook-data-what-learned/471787002/.

[173] Jim Gemmell, Gordon Bell, and Roger Lueder, "MyLifeBits: A Personal Database for Everything," *Communications of the ACM* 49, no. 1 (January 1, 2006): 88–95, https://doi.org/10.1145/1107458.1107460.

[174] Michael Nash, ed., *How to Keep Union Records* (Chicago: Society of American Archivists, 2010).

[175] Riley, "Data Transfer Project Use Cases."

which it is based, left it up to the platforms to decide how to fulfill the requirements as long as users could download content from their accounts in a machine-readable format that would be easily transferred to another service. Twitter, Facebook, and other providers chose to comply with platform downloads.

Platform downloads are available only to subscribers who must verify their identity through a process designed to prevent unauthorized access to the data. The exact steps vary by platform, but the process itself starts with a request by the subscriber to download their data from the site. The data is filtered and made available through a simple user interface according to specific categories of data.

Downloading refers to the transmission of a file or data from one computer to another.[176] Facebook and Twitter, as well as references I read online, refer to the features as a way to download your archive, download and archive your data, access and download your data, and download a copy of your information.[177]

The term I chose, platform download, reflects the way the method is described and refers to the source, social media platforms.

---

[176] "Definition of Download," in *Merriam-Webster*, July 31, 2023, https://www.merriam-webster.com/dictionary/download.

[177] Barbara Krasnoff, "How to Download an Archive of Your Twitter Data," The Verge, November 11, 2022, https://www.theverge.com/23453703/twitter-archive-download-how-to-tweets; David Rubenking, "How to Download Your Facebook Data (and 6 Surprising Things I Found)," *PCMAG*, November 30, 2018, https://www.pcmag.com/how-to/how-to-download-your-facebook-data-and-6-surprising-things-i-found; Daniel Victor, "How to Download Your Twitter Archive," *The New York Times*, November 18, 2022, sec. Technology, https://www.nytimes.com/2022/11/18/technology/how-to-download-your-twitter-archive.html; "Accessing & Downloading Your Information | Facebook Help Center," accessed June 11, 2020, https://www.facebook.com/help/1701730696756992/?helpref=hc_fnav; "Download a Copy of Your Information on Facebook | Facebook Help Center," n.d., https://www.facebook.com/help/212802592074644; "How to Access and Download Your Twitter Data | Twitter Help," n.d., https://help.twitter.com/en/managing-your-account/accessing-your-twitter-data; "How to Download Your Twitter Archive and Tweets | Twitter Help," n.d., https://help.twitter.com/en/managing-your-account/how-to-download-your-twitter-archive.

However, the architecture and methods behind the downloads are unclear and undocumented. The platform download feature appears to be a combination of interface, database, and API. Data is requested by typing a "get" and "post" to retrieve and send data, such as tweets from July to August 2019. Requests are returned in the JSON format, a text-based format that easily converts into JavaScript. Developers use APIs to manage requests for data. The data is portable and available for extraction but can easily be made unavailable to unauthorized requests.

I was not able to confirm the use of APIs with Twitter, Facebook, or Instagram. Computer engineers I spoke to had different opinions but had not themselves tried to confirm the design of the platform downloads. But I am confident that they are indeed part of the architecture.

We do know that APIs are the primary way to get data into and out of the Twitter, Facebook, and Instagram platforms and to query data, post new stories, links etc. The services give users a way to choose which categories of data they want based on a closed menu of choices (activity across the platform, personal information, security and log-in information, friends and followers, etc.).

API technical documentation is consistent with these selections and with repeated close observations of the personal download set-up using a feature on Firefox and Chrome web browsers, which allows the user to visually observe (or, inspect) the process happening in real time. For example, I would see "get" and "post" messages on the browser inspector during the downloads – indicators that APIs are being used to get data in and out of the platform databases. Moreover, during a 2019 conversation with a Twitter researcher and manager, I noted that I was concerned that my downloads would not be comprehensive because the Twitter API are known

to limit the tweets that are returned. They accepted, albeit passively, my assertion about the use of an API in the downloads.

Another indication of APIs came from the access controls. I could not download the files myself because Facebook and Twitter require that users authenticate their credentials. This is done by providing a code, which is then sent to their phone number attached to the account. The process was consistent with authorization protocols described in API documentation and tutorials.

I could not authenticate my identity as an authorized user of the Twitter account because the phone number belonged to another union organizer. They finally had to download the files for me, which were transferred, unopened, to my computer with a flash drive. This was done in their office and, to the best of my knowledge, the process did not affect the contents of the downloads.

While a definitive lack of confirmation about the role of APIs in the research is a limitation, it is also a useful result in its own right about the limits of using platform downloads to fulfill user rights and increase transparency, which I discuss in more detail in the findings chapter. Here I provide information about the methodology I used to assess different ways of accessing data. Again, I initially expected platform downloads to be more comprehensive, but found that, one, they were not; and two, platforms were limiting other methods for extracting data and limiting user rights.

Overview of Results

Archive-It: Twitter

Initially the captured version of GWC's Twitter feed seems indistinguishable from the original (i.e. the "live web"). The text of the tweets, as well as handles, hashtags and retweets are visible along with photos, articles, videos, or other content are visible on screen. However, the crawler has only captured the first layer of content. The crawler has not captured the connected posts strung together in "threads," or comments and many of the links open to non-working pages.

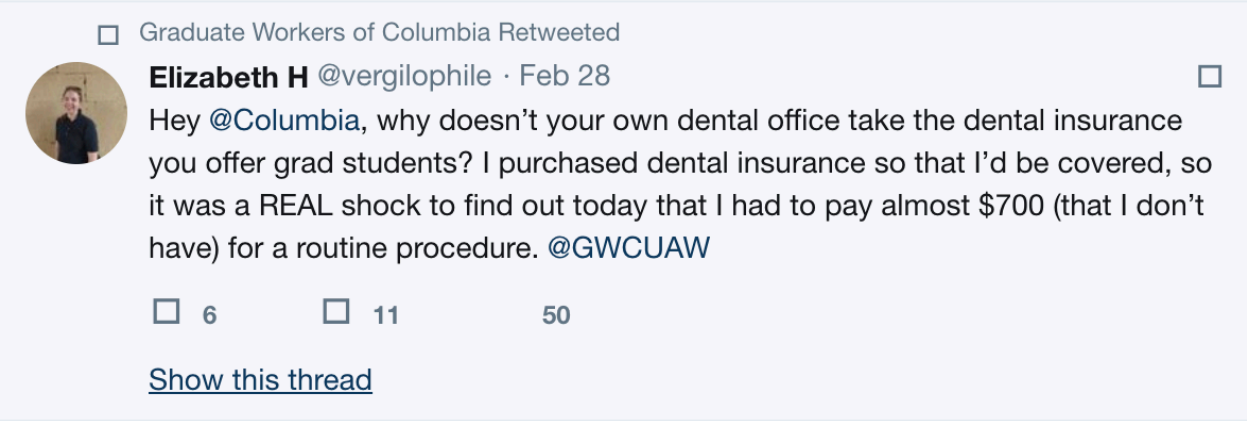

Figure 1: Archive-It did not capture the thread that should be apparent by clicking on "show this thread"

Archive-It: Facebook

The GWC Facebook crawl in the Wayback Machine displayed a replica of the GWC Facebook home page obscured by log-in page pop ups that appeared in the middle of the screen. The Facebook feed became lighter as though the site automatically applied a transparency filter of about 40 percent and a rectangular pop-up message prompted me to log in to the account to "See more of GWC-UAW Local 2110 Graduate Workers…" About a minute later, the pop-up moved

to the bottom of the screen, revealing the feed again. I could scroll down through several screens of posts but once the cursor was at the end of the screen, I couldn't access additional content and there was no way to click on the "See More" button because the bottom of the screen was blocked by the pop-up, which prevented me from being able to see if the button was there at all, indicating that more content should be accessible.

Images in some posts seemed to be split in two and filled most of the screen. A video, which did not load, floated outside of the post underneath the distorted image. From there I tried to click on several links in FB including a GWC Bargaining Committee blog that detailed negotiation sessions between the GWC and Columbia University. Instead the link opened to a page displaying the message, "The page you requested has not been archived in Archive-It." Several reasons could account for this failure, according to the message, including, most likely, that the page was outside the crawler's scope.

Under the Facebook Photos menu were 21 photos posted by other people. I clicked on the "See all" link, which returned more than 21 photos displayed in thumbnails as well as in photo albums. Facebook prevented me from accessing the albums with a log-in box that popped up in the center of the screen. The sequence was the same as the last time: The pop-up moved to the bottom of the screen and the links opened error pages identical to the previous ones except for the URL at the bottom of the message, which linked to Facebook. But even on the site I could not access the albums because of a log-in pop up that began at the bottom of the page and kept moving around the page wherever I tried to click. Reports helped to distinguish between problems with a capture (missing content) and problems with replay in the Wayback Machine. But I still could not capture a single page of the site.

The problem remained unresolved and, over time, I learned from other users that my experiences with inexplicable failures were commonplace.[178] Archivists I spoke to said they no longer tried to collect Facebook even when the content would be directly relevant to their collections.

Archive-It: Instagram

The #CUStrikeout Instagram account crawl seemed to be complete until I tried to open some of the images.[179] They were supposed to open up onto larger stand-alone versions that alongside text, a menu, icons, and comments. More than a dozen of these stand-alone images were missing as were the menus, icons, and comments.

Archive-It how-to documentation suggests using its Brozzler software rather than Heritrix for Instagram because of the platform features of dynamic elements designed to change, or adapt, according to user behavior. I used Brozzler (and Heritrix) at a variety of settings without a detectible different in the results.[180] Dynamic content did not appear to be the reason behind the

[178] These practitioners, most archivists, tended to weigh the significance of missing objects according to whether they existed elsewhere. If replicas exist, they chose not to spend their time and data resources to capture it. Many used their time to check the seed URLs before launching a crawl. An institutional archivist for a large university collection said she is satisfied with a crawl as long as she knows the record exists because it would be accessible for archival purposes, even if not immediately. But if a site was too demanding they did not include it in their collections. Facebook was one of those.

[179] In fact I analyzed two Instagram accounts: In 2018, a GWC member created an account @CUStrikeout that became the de-facto GWC Instagram site during the strike and subsequent actions. He changed the site from @CUStrikeout in May 2020 to @Columbia People's Covid Response (CPCR) a group that formed after the onset of the COVID-19 pandemic emergency measures. He maintained the GWC content but I had to repeat the same steps when the status of the accounts became clear to me.

[180] For example, changing the Brozzler settings to a data capacity to 10 Gb netted three more images that had been excluded (53 vs. 56).

omissions.[181] I repeated the Instagram crawl at different settings but improved the capture by only three images.[182]

Webrecorder: Twitter, Facebook, Instagram

I began by opening Webrecorder's home page. I could sign up for a free account on Webrecorder.io that included storage space, although I could not find information about how many gigabytes of storage space would be available to me. The other option was to download and use the Webrecorder Player desktop application to export the web archives for browsing offline. Installing and running the software on my computer was possible but it did require familiarity with GitHub, a popular online repository that makes open-source software projects available for download and repurposing to varying degrees. The practice of posting code online instead of keeping it proprietary is increasingly common because it allows companies to attract programmers interested in improving the software and adding features.

[181] Staff later suggested that the records are likely to have been captured and that playback in the Wayback Machine is the issue. But according to the how-to documentation, especially high numbers of queued URLs typically may indicate a crawler trap -- a set of web pages that create an endless number of URLs if allowed free reign, expending document/data budgets unnecessarily. Unintentional traps include calendars, which can send a crawler into an infinite loop as it seeks to capture dynamic pages that point into the future and the past. In addition to web archiving, other crawler-like bots are commonly used for marketing and spamming, which web developers will try to stop with traps because they can affect search optimization results and interfere in display of social media objects. Some traps try intentionally to catch spambots and crawlers that Heritrix crawlers resemble.

[182] Archive-It staff were at a loss for how to explain the problem with Instagram and directed me to Wayback QA (quality assurance). This was a process that entailed replaying crawls in the Wayback Machine in order to identify documents that were captured, and that were in the WARC file, but not otherwise visible. Facebook had by far the highest number of out of scope and queued URLs: a ratio of about 3-to-1 (e.g. more than 14,000 in the case of Facebook compared to Twitter's 4,500). In contrast, the number of blocked URLs was insignificant.

The download version of Webrecorder also required installing and setting up Docker, a software platform service that packages applications together in a "container" to make them easier to run.

The description on the Webrecorder GitHub site reflected a relatively high level of expertise necessary to use the downloaded version. I opted for the online account and create a collection called GWCSocialMedia. I had the option to keep my user-created collection private or to make it public (visitors to Webrecorder.io can access the content if the collection is public). But the terms of service made plain that submitting or storing content on the site gives Webrecorder "…a perpetual, irrevocable, worldwide, non-exclusive, fully paid-up, royalty-free, sublicensable, and transferable" license to that content.[183] The terms made no guarantees about longevity of the web archives on the site or of the service but it was clear that I would be relinquishing some authority over the GWC social media by capturing and storing it with Webrecorder.

For the first session, I chose to capture GWC's Twitter feed @GWCUAW and type in the URL. Webrecorder offered an auto-scroll feature which allows the app to crawl the entire Twitter feed, from the most recent at the top to the first tweet on Jan. 27, 2016 at the bottom. I hit the button and the photos and videos began scrolling down the screen. Webrecorder opened a new page that displayed a Resource Not Found (RNF) message. Like Archive-It, anything linked to in the tweet had to be captured separately, or "patched" (the missing objects can be added to the collection or segregated).

[183] "Webrecorder Terms and Policies," Webrecorder, accessed December 11, 2018, https://webrecorder.io/_policies#privacy.

I next tried to use Webrecorder on the GWC Facebook page, which required a log-in to the account. To access the account, I needed the personal log-in credentials of GWC's organizer, who set up the page several years ago, which I ultimately was able to provide.

I started the capture, with autoscroll activated, and watched the images and captions descend down the screen. The volume was massive, and the process took more than an hour. When I tried to replay the capture, the fields rendered as blank white boxes. Except for the banner at the top of the page, not a single post had been captured. Lastly, I try to capture the Instagram account, which wasn't blocked but was incomplete. I had to work backwards to identify missing content and patch it in.[184]

Platform Downloads

Twitter: Download your Twitter Archive

The platform download began with a template that guides the user through their selections. The template specified what options were available and what data could be accessed through the download. Twitter limited requests to "download your Twitter archive" to no more than every 30 days.[185] The instructions clearly state that the download would be incomplete and curated without specifying the criteria for what Twitter would include/exclude; or a measure of what would be missing. Once the request was made, I waited for Twitter to compile the data with "the information that we believe is most relevant and useful to you."

---

[184] Anna Perricci, "Webrecorder: Web Archiving for All!," https://www.slideshare.net/annaperricci/webrecorder-web-archiving-for-all.

[185] "How to Download Your Twitter Archive," accessed June 12, 2020, https://help.twitter.com/en/managing-your-account/how-to-download-your-twitter-archive.

Subscribers have the option of requesting the data in two formats. The first is HTML, which makes it easier to see a visual representation of the files using a browser similar to the live interface; and, secondly, the machine-readable format JSON, which lends itself to analyzing the content as data but not to an ordinary visual rendering. JSON requires some skill to interpret but is easily searchable, whereas the HTML version is closer to the look and feel of the live Twitter feed. I selected both formats.

Content and the accompanying metadata were delivered almost immediately in a series of folders and files.[186] The contents of the folders included the file "Your archive.html" that could be opened in a browser for an approximation of the Twitter feed as it appeared at the time the download was requested. Below are three versions of a tweet as it appears in the original (Figure 2), HTML (Figure3), and JSON(Figure4) formats. I give a brief overview of the HTML and then the JSON versions.

Figure 2: Tweet as it appears on Twitter

[186]The downloads are actually delivered in a.zip file, which is a collection of various files that have been compressed into one file to decrease the file size and make moving or sending them easier. Both Mac and Windows come with an inbuilt compression feature that enables the user to put all the files in one folder. Therefore, it is a suitable option to compile multiple items into a single file whose file size is decreased. The .zip file is opened to reveal the original files, which contain the uncompressed folders and the data they contain. As far as I know, no data is lost or affected by .zip compression. Parallel, "How to Zip a Folder on Mac," July 14, 2022.

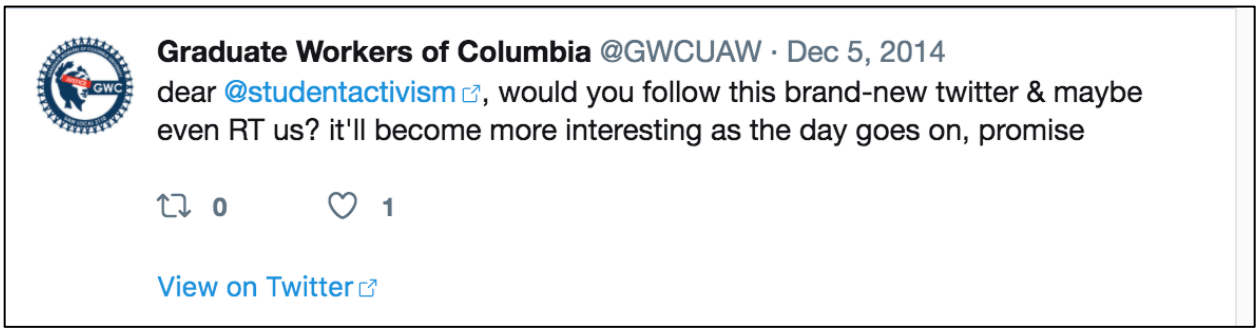

Figure 3: Tweet as it appears in "Your archive.html"

Figure 4: Abbreviated example of a single tweet and metadata delivered in JSON


```
{  "tweet" : {
    "retweeted" : false,
    "source" : "<a href=\"http://twitter.com\" rel=\"nofollow\">Twitter Web Client</a>",
    "entities" : {
      "hashtags" : [ ],
      "symbols" : [ ],
      "user_mentions" : [ {
        "name" : "Angus Johnston 😁",
        "screen_name" : "studentactivism",
        "indices" : [ "5", "21" ],
        "id_str" : "20919626",
        "id" : "20919626"
      } ],
      "urls" : [ ] },
    "display_text_range" : [ "0", "144" ],
    "favorite_count" : "1",
    "id_str" : "540875381092671488",
    "truncated" : false,
    "retweet_count" : "0",
    "id" : "540875381092671488",
    "created_at" : "Fri Dec 05 14:28:34 +0000 2014",
    "favorited" : false,
    "full_text" : "dear @studentactivism, would you follow this brand-new twitter &amp;amp; maybe even RT us? it'll become more interesting as the day goes on, promise",
    "lang" : "en"}
```

Twitter provided basic information about the contents and the way they were organized in a "read me" document that comes with the download package, while the meaning of each tag is documented on the Twitter website for application developers rather than for the subscribers.[187]

Neither explain the rationale for the file structure and categories but do help navigate the downloads. For example, "name" is the name of the person registered to the @studentactivism account, the intended recipient of the mention in the first tweet. The results also included whether the tweet was retweeted or liked. Also included were hashtags (a short keyword prefixed with the hash symbol # and a signature feature of Twitter).[188] The results appeared to be limited to engagement-type of data but were otherwise difficult to assess. For example, I cannot tell what had been deleted and the results were out of order sequentially (the first tweet in the set is dated 2020 and the last one 2019 with the first tweet from 2014 somewhere in the middle of the set).

The HTLM version helped to recreate the look and feel of the Twitter feed at the time of download but elements like buttons, date, and time were missing. And, although the HTML display of the download was searchable by oldest and newest tweet, information about the device and data were the only metadata included with the photos and videos.[189] There was no explanation for the difference between the original "live web" version of the tweet and the HTML version, or for including only certain metadata.

---

[187] "POST Statuses/Update," accessed May 26, 2023, https://developer.twitter.com/en/docs/twitter-api/v1/tweets/post-and-engage/api-reference/post-statuses-update.

[188] Axel Bruns and Jean Burgess, "The Use of Twitter Hashtags in the Formation of Ad Hoc Publics," in *Proceedings of the 6th European Consortium for Political Research (ECPR) General Conference 2011*, ed. A. Bruns and P. De Wilde (European Consortium for Political Research (ECPR) General Conference (6th), United Kingdom: The European Consortium for Political Research (ECPR), 2011), 1–9, https://eprints.qut.edu.au/46515/.

[189] Twitter strips out EXIF data captured by digital cameras that includes GPS location. Twitter uses this data to process photos but does not make it available to visitors to the site although it is not clear if the data is retained for third-party data companies. See: Twitter, "How to Post Photos or GIFs on Twitter," accessed June 2, 2020, https://help.twitter.com/en/using-twitter/tweeting-gifs-and-pictures..

Facebook: Download a copy of your data

Facebook offers users the option to download a copy of their data on a similar request template to Twitter's but is not explicit about the frequency of requests. The contents (available in JSON and HTML) loosely correspond to the original Facebook site and interface. Whereas, the live homepage was organized by posts, about, photos, videos, events, notes, and community, the native download file structure was divided by folders and subfolders that include photos, videos, wall posts, and events. For example, photos and videos arrived in a folder labeled with a uniform resource identifier (URI). The contents were then further divided into subfolders labeled posts, timeline, profile, and but also specific events (e.g Oct15thWeAreWorkersPanel_uAVODxgumw).

Figure 5: The first GWC Facebook post from Dec. 5, 2014


```
{
   "timestamp": 1417806860,
   "attachments": [
    {
     "data": [
      {
       "media": {
        "uri":
"photos_and_videos/TimelinePhotos_88VOdDYNgA/10452976_697188067046522_913521288
9245387215_o_697188067046522.jpg",
        "creation_timestamp": 1417806834,
        "media_metadata": {
         "photo_metadata": {
          "upload_ip": "209.2.225.149"
         }
        },
        "title": "Timeline Photos",
       "description": "Congratulations a majority of the 2,800 Columbia University RAs
       and TAs have formed a union and today a delegation from across campus asked
       the administration to agree to a fair, expedient process to verify majority support
       and start bargaining. \n\nToday, we delivered a letter to President Lee Bollinger
       asking the university to recognize our union. Over 200 people joined us, read
       more here: http://bit.ly/1vUvreS"}
```


The posts themselves were displayed in blocks of text stripped of their metadata as well as any photos and videos that did not originate with GWC. If a post included a .gif file posted by GWC, the image was intact. If the content shared in the post originated elsewhere but was shared by GWC, any photos and videos were missing. Some of the media was located in image folders.

Also missing were likes, comments, and direct messages, and Facebook converted URLs linked in the posts to a string of numbers. For example, in a post GWC references an initiative, UndoCU, in a link that was rendered in the platform download as:

> The Organizing Committee of Graduate Workers of Columbia-UAW stands in solidarity with @[1801737756765051:274:UndoCU] in calling for Columbia to cancel the contract with U.S. Customs and Border Protection (CBP) that started on May 29th and will continue through November 2020.

Figure 6: Rendering of a Facebook post

Searching Facebook for the ID number turned up an error message even though the UndoCU Facebook page was still active. In 2016, Facebook API downloads included a friends list in the data archive.[190] That element now appeared to be excluded in the platform downloads. There was no explanation that would help to understand the missing data or the organization of it.

Instagram

In 2012 Facebook acquired the photo and video sharing app Instagram and while there were similarities between Instagram's native download and the other platforms (dates out of order, minimal metadata, disjointed file structure), there were also several differences: most importantly, comments and direct messages were included in Instagram. Facebook does not include direct messages but both it and Twitter excludes comments.[191]

---

[190] Jed R. Brubaker and Vanessa Callison-Burch, "Legacy Contact: Designing and Implementing Post-Mortem Stewardship at Facebook," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI'16: CHI Conference on Human Factors in Computing Systems, San Jose California USA: ACM, 2016), 2908–19, https://doi.org/10.1145/2858036.2858254.

[191] Pavica Sheldon and Katherine Bryant, "Instagram," *Computers in Human Behavior* 58, no. C (May 1, 2016): 89–97, https://doi.org/10.1016/j.chb.2015.12.059.

Comments show interactions with non-members on one hand, while direct messages have the potential for insights that public-facing social media and email ordinarily would not. Some GWC members used social media more for communication than email, making direct messages one of the few traces of direct conversations among them as well as non-members. This would be important information to have for anyone creating a self-archive or to assess the data that is collected by the platform.

Instagram provided data (in JSON) about "connections" that did not seem to correspond to any feature on the site. Only by examining the file was I able to see that the file contained information about followers and the accounts followed by GWC. The process required familiarity with JSON and the text editors necessary to display it because the native download does not have a HTML version of contents that could be displayed in a browser without either. Instagram displayed the number of posts on the live version but the content that constitutes the posts were sorted into other folders, organized by date, labeled "stories," "videos," and "photos." Without a corresponding folder, it was not clear where the "tagged" data was organized. In essence, the contents were disorganized and confusing from a user perspective.

Another folder, "information about you," was empty except for "primary location" and the name of a city and state. But the GWC Instagram account was itself interesting because it was seldom used until May 1, 2020 even though the account was created earlier. Until May 1, 2020, the only content on the site was the union's logo. Site administrators began adding old content from 2018 and recent images from protests against COVID-19 pandemic emergency measures. They were all dated with the day and year they were posted instead of when they were created. Only with metadata about the date the content was created, rather than posted, would it be clear that the photos are of events unrelated to COVID-19 actions. Images lacked metadata

and, because Instagram posts are image and video based, couldn't establish a link between text and image in a JSON file the way that could be done in Twitter and Facebook. Without knowing the history of GWC or the account, the discrepancy would not be easy to detect.[192]

Discussion

Overall data was difficult to gather and the collections were incomplete. There are multiple ways to understand the reasons for this. Firstly, none of these services performed well at collecting social media.[193] My original explanation was that older techniques, made for document-based websites, weren't designed for social media. To some extent my premise was correct, evidenced by the development of new techniques (Umbra, Brozzler, and Webrecorder).

However, Facebook blocked Webrecorder from accessing the site altogether, even though I had permission to access and capture the GWC account.

Twitter has since also begun restricting access to data in the same way that Facebook does. These limits could benefit user account security in welcome ways, but also illustrates that platforms are imposing their terms on access. Platforms decide how security and user rights will be provided, which I discuss more in depth in the next chapter.

I am aware that the advantages and disadvantages of each method are relative the user and the use case. For some, the option to look at what data is being collected about them is important, while others just want to back up their data and what counts is that the data is there.

---

[192] Similar to the other sites, Instagram allows users to "request" data and sets the terms: "We've started creating a file of things you've shared on Instagram and will email a link to awwoodall@gmail.com. It may take up to 48 hours to collect this data and send it to you." The download link expires after four days.

[193] Nicholas A. John and Asaf Nissenbaum, "An Agnotological Analysis of APIs: Or, Disconnectivity and the Ideological Limits of Our Knowledge of Social Media," *The Information Society* 35, no. 1 (January 1, 2019): 1–12, https://doi.org/10.1080/01972243.2018.1542647.

Someone else like an archivist might be more interested in specific facets of a website. Context could be more important in the latter example and less so in the former. Nonetheless, the limitations on data extraction and selective access raise questions about how Facebook, Twitter, and Instagram manage access to information and suggest that current efforts to regulate platforms more effectively may fall short of goals. Having provided information about my methodology and the methods for collecting the data, I turn in Chapter 2 and 3 to assessing the data.

Chapter 3: Access in the GDPR

The right of access has been integral to privacy since the 1970s, when the first data protection instruments were written in the United States and Europe. In Europe, the right of individuals to know about information being stored about them in databases was included in early resolutions.[194] The Organisation for Economic Co-operation and Development added guidelines in a 1981 convention and the right of access to personal information was also enshrined in Article 8 of the E.U. Charter of Fundamental Rights.[195]

The right of access plays an important role in monitoring and enforcing rules designed to govern entities, public or private, that collect and process our personal information.[196] In the private sector, social media platforms like Twitter and Facebook now direct a complex, automated, and seemingly ubiquitous system for information – or, data – processing.[197] As a result, regulations are important for enhancing the transparency of personal data processing.[198]

In the European Union regulations these companies are called data controllers, entities that determine the purposes and the means of data processing, whether they are government authorities or private corporations like Twitter and Facebook.

The GDPR deals directly with data controllers, data privacy, and data protection. These principles contribute to making the operations of data controllers more transparent by giving

---

[194] "On the Protection of the Privacy of Individuals Vis-à-Vis Electronic Data Banks in the Private Sector," Pub. L. No. Res(73)22 (1973), https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680502830; "On the Protection of the Privacy of Individuals Vis-à-Vis Electronic Data Banks in the Public Sector," Pub. L. No. Res(74)29 (1974), https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=09000016804d1c51.

[195] "Article 8 - Protection of Personal Data," European Union Agency for Fundamental Rights, April 25, 2015, 8, http://fra.europa.eu/en/eu-charter/article/8-protection-personal-data.

[196] Ausloos and Dewitte, "Shattering One-Way Mirrors – Data Subject Access Rights in Practice."
[197] Ausloos and Dewitte.

[198] "Guidelines 01/2022 on Data Subject Rights: Right of Access" (European Data Protection Board, January 18, 2022).

individuals sufficient, transparent, and easily accessible information about how their data is being collected and processed. The controllers have to show what they are doing with personal data.

Articles 12-23 of GDPR encapsulate data privacy and control by establishing the right of access to personal data according to key principles including transparency and fairness, as well the overarching right of privacy. In order to realize these goals effectively in practice, the articles were written to grant individuals this right without unnecessary constraints, at reasonable intervals, and without excessive delay or expense. All this should lead to more effective enforcement of the right of access by the data subject.

In this chapter, I assess these articles and their provisions for monitoring the collection, processing, and sharing of personal data. After establishing several key terms, I provide an overview of my methodology. I mainly restrict analysis to Twitter, Facebook, and Instagram platform downloads. These are built-in features implemented by each of these "data controllers" in compliance with the GDPR to provide users access to the data that controllers are collecting and processing.

Second, I summarize the function that the GDPR ascribes to access and provide an overview of how access fits into a larger framework. Within this framework are several overarching approaches to data rights that political leaders are pursuing. Two important concepts in the GDPR are data protection by design and data protection by default, regulations that fit with the GDPR's broader risk-based approach to data protection.[199] Protection by design in the GDPR means that risks to privacy must be anticipated in the design, implementation, and function of systems, products, and business practices in order to prevent harms and provide

[199] Included in GDPR Article 25.

better, cost-effective protection for individual data privacy, and that only data which is necessary for each specific purpose of the processing should be gathered at all. A tenet of this data protection by-design, risk-based framework is providing users with self-service tools so that they can determine how their personal data is being used and whether policies are being properly enforced. Platform downloads are a manifestation of this framework.

The third section outlines the criteria I used to evaluate the results from the platform downloads according to GDPR criteria including scope, format, and privacy controls. Formats describe the form in which data is made available to users. The format should be easy to interpret, machine-readable, and portable – criteria I explain in this section. Privacy controls include ways that people are required to identify themselves in order to access their data.

I detail my findings in section four along with potential advantages and disadvantages of platform downloads. Notably, I identified the way that platform downloads have been implemented and accepted as a suitable self-service strategy for exercising access rights. Platform downloads have advantages such as being easy to download in GDPR required formats. Yet there are disadvantages for users, including incomplete, confusing results.

I found it necessary to explain the integration of APIs because, to a growing degree, they anchor data access and protection approaches, and are championed by political leaders and regulators, as well as the industry whose control over data the GDPR is intended to regulate. They are also a key feature in the architecture of platform downloads whose effect on access rights should be assessed.

I conclude in the final section that, on one hand, platform downloads offer advantages for access rights. But, on the other hand, implementation of a self-service, by-design approach has

reinforced a model in which users are responsible for exercising their rights with little or no means to challenge decision-making by controllers about how those rights will be provided.

Key terms

The GDPR defines personal data as "any information relating to an identified or identifiable natural person."[200] A broad variety of data falls within the definition of personal data, including name, address, and telephone number as well as medical findings, history of purchases, creditworthiness indicators, or the content of communications. In light of the broad scope of the definition of personal data, a restrictive assessment of that definition by the controller would lead to an erroneous classification of personal data and ultimately to a violation of the right of access.[201]

A data subject is any living person whose personal data is collected, held or processed by an organization. A data subject is granted rights under the GDPR and other instruments.
I opt for the term "user" in most cases.

Methods for Analysis and Background

This chapter evaluates how and under what conditions these rights are effective based on my own experience trying to collect data from three social media platforms, Twitter, Facebook, and Instagram using multiple methods. The methods include two software services, Archive-It and Webrecorder, and, thirdly, platform downloads.

[200] "Art. 4 GDPR – Definitions," *General Data Protection Regulation (GDPR)* (blog), 4, accessed July 31, 2023, https://gdpr-info.eu/art-4-gdpr/.

[201] "Guidelines 01/2022 on Data Subject Rights: Right of Access."

Platform downloads are a service provided by Twitter, Facebook, and Instagram for the explicit purpose of allowing users to receive a copy their data, transmitted automatically, without having to request the data from the companies and wait for a response – the essence of what it means to download something. Platform downloads were implemented in 2016 in advance of the GDPR's requirement to make data accessible to subscribers. Twitter, Facebook, and Instagram complied with the GDPR by allowing subscribers to download their data from the site (a Twitter subscriber could download their data from the Twitter site and so on) via what I call a platform download, a process that has no established name. Nor are details about their design or function published. But, as I described in Chapter 1, the design and architecture of a platform download appear to be a combination of a user interface, APIs, algorithms that sort data, and settings that limit the number of requests for data that can be made (this is called rate limiting).

By-Design, APIs, and Platform Downloads

The GDPR requires controllers to implement technical and organizational measures that meet principles of data protection. Measures should be incorporated from the start that help to fulfill criteria for data rights. The premise behind this by-design model is to build values into technology by reflecting on the potential effects on users during the design phase.

Data processing safeguards must meet the requirements of the GDPR regulations in order to protect the rights of data subjects. The requirements apply to the amount of personal data collected, the extent of their processing, the period of their storage, and their accessibility.

Providing users with self-service tools is one way to meet these expectations. Platform downloads are one version of such a tool suggested by data regulators that would allow users to download their data from the site of a data controller in a way that would comply with the

GDPR. Facebook, Twitter, and Instagram deployed their versions of this idea by repurposing their APIs and adding a user-interface to make the process easier.

I look at the advantages and disadvantages of platform downloads as an approach to fulfilling rights. Platform downloads appear to meet many of the GDPR criteria. For example, platform downloads make data available in a format that is portable, which is a GDPR requirement. APIs make automated portability possible. Portability in the GDPR allows for data subjects to receive a copy of their personal data from a controller in an electronic, commonly used, and machine-readable format, and to transmit those data to another data controller without hindrance. This means that users should be able to move their data among different applications, programs, computing environments, and cloud services. Without data portability, a person's data is accessible only through the platform on which it is created or stored.[202]

While Archive-It and Webrecorder allow subscribers to receive their data in a format that resembles the original visual representation, only platform downloads offer the option to also receive data in a portable file format. This is an advantage in terms of portability that only platform downloads provided. Platform downloads also have an advantage over the other methods in terms of privacy and consent because they prevent unauthorized access to the data. However, they had disadvantages for users in that they did not comply fully with access requirements of the GDPR.

[202] "Article 29 Data Protection Working Party" (European Data Protection Board, April 5, 2017), 29, https://edpb.europa.eu/about-edpb/more-about-edpb/article-29-working-party_en.

Privacy Self-Management and Performance

While an argument can be made for building values into technology, by-design is rooted in what Daniel Solove calls "privacy self-management." This approach by policymakers, which has remained largely unchanged since the 1970s, provides people with a set of rights to enable them to make decisions about how to manage their data. These rights consist primarily of rights to notice, access, and consent regarding the collection, use, and disclosure of personal data. The goal of this bundle of rights is to provide people with control over their personal data, and, through this control, allow them to weigh the costs and benefits of the collection, use, or disclosure of their information.

Privacy self-management, he argues, does not provide people with meaningful control over their data.

Proponents of privacy self-management envision a rational individual informed about their rights making appropriate decisions about whether to consent to various forms of collection, use, and disclosure of personal data. Empirical and social science research demonstrates that there are severe cognitive problems that undermine privacy self-management.

Returning to the characterization above of idealized rational actors, two of the most important components of privacy self-management are informing individuals about the data collected and used about them (notice) and allowing them to decide whether they accept such collection and uses (choice). European as well as U.S. lawmakers have embraced these "notice and choice" components of the Fair Information Practices framework. The lawmakers and the laws they create have normalized the practice of providing notice and choice by offering privacy notices and a choice to opt out of some of the forms of data collection and use described in the notices.

However, making decisions about how to manage personal data is not simple because of cognitive barriers like skewed decision-making and a struggle to apply knowledge to complex situations. Individuals assess familiar dangers as riskier than unfamiliar ones and are more willing to share personal data when they feel in control, regardless of whether that control is real or illusory: they are more willing to take risks, and judge those risks as less severe, when they feel in control. And people tend to make certain mistakes in judgment consistently.

In essence, if indeed people read privacy policies (and there is reason to doubt that they do), and they understand them, they often lack enough background knowledge to make an informed choice. If people read them, understand them, and can make an informed choice, their choice might be distorted.

Moreover, making privacy simple and easy to understand conflicts with fully informing people about the consequences of giving up data, which are quite complex if explained in sufficient detail for a fully-informed decision. Individuals need a deeper understanding and background to make informed choices. Many privacy notices, however, are vague about future uses of data. In addition, privacy choices are often binary. Individuals might desire more granularity in their choices, but additional granularity adds complexity and creates greater risks of confusion. Solove notes that efforts to improve education are certainly laudable, as are attempts to make privacy notices more understandable. But such efforts fail to address a deeper problem — privacy is quite complicated. This fact leads to a tradeoff between providing a meaningful notice and providing a short and simple one.

He further argues that these cognitive problems impair individuals' ability to make informed, rational choices about the costs and benefits of consenting to the collection, use, and disclosure of their personal data. Second, and more troubling, even well-informed and rational

individuals cannot appropriately self-manage their privacy due to structural problems. For one, there are too many entities collecting and using personal data to make it feasible for people to manage their privacy separately with each entity. Also, many privacy harms are the result of an aggregation of pieces of data over a period of time by different entities. It is virtually impossible for people to weigh the costs and benefits of revealing information or permitting its use transfer without an understanding of the potential downstream uses, further limiting the effectiveness of the privacy self-management framework.

In practice, automated, self-management resources like webportals and dashboards have had poor results. Reasons for the failures are complex, from technical problems like network unreliability to incorrect assumptions that users are able to exercise more control than they are capable of. To increase their success, the idea of "Data Portability As a Service" (DPaaS) have become popular among data control advocates as a way people could authorize third-party providers to exercise the right of portability in their name and have the data sent directly to another host or repository. However, even if the dashboards are user-friendly, the likelihood that individuals would take advantage of them appears to be contingent on successful outreach: Companies reported in one study that they often found it difficult to convince clients to exercise their right to data portability.[203]

There is a circular logic involved when, with each sign of failure of privacy self-management, the typical response by policy-makers, scholars, and others is to call for more, and improved, privacy self-management.[204] That is being done, increasingly, with by-design

[203] Helena Ursic, "Unfolding the New-Born Right to Data Portability: Four Gateways to Data Subject Control," *SCRIPTed: A Journal of Law, Technology & Society* 15, no. 1 (August 1, 2018): 42–69, https://doi.org/10.2966/scrip.150118.42.

[204] Solove, "Introduction."

methods. Legal scholar Ari Waldman characterizes this circular logic as a performance of privacy law compliance. One set of these "performances" include instituting internal compliance programs, implementing privacy impact assessments, translating legal requirements into technical specifications and new products, and performing internal assessments and external audits. Another set is aimed at individuals' exercise of rights. These include navigating consent toggles, opt-out buttons, cookie requests, privacy policies, and data request links. Supplementing this self-governance are practices in which regulators partner with industry to settle disputes and develop rules and where industry creates internal compliance structures for ongoing accountability. These practices constitute a roughly uniform approach to privacy law reflected in almost every recent proposal for comprehensive privacy law in the United States. The practices have roots in the Fair Information Practices code and have become entrenched, such that one country's privacy regulations resemble another.[205]

Solove and Waldman make an argument for why the legal framework that this code. and the regulation that rely on it, leave individuals with inadequate controls, notice and choice among them. Moreover, transparency leaves much to be desired when industry techniques remain opaque and veiled by secrecy.[206]

I question whether privacy self-management tactics in the GDPR like self-service platform downloads should be seen through these frameworks, or whether they offer meaningful criteria, methods, and tools for controlling how data is being processed. Does the data protection by-design approach promote access rights? Does making users responsible for exercising their

[205] *Privacy as Trust: Information Privacy for an Information Age* (Cambridge, United Kingdom ; New York, NY: Cambridge University Press, 2018); "Privacy, Practice, and Performance."

[206] Diaz, Tene, and Guerses, "Hero or Villain."

rights give them adequate, appropriate means to challenge decision-making about how those rights will be provided?[207] What kind of rights do they provide?

Key Frameworks and Criteria

Under the GDPR, the right of access consists of three components. One is confirmation of whether or not personal data is being processed. Two, access to the data. And, three, information about the processing itself. Individuals also have a right to obtain a copy of their personal data. This provision is not an additional, tacked-on right, but is a means and method for providing access to the data. Data controllers cannot impede the right of access through file formats, limits on the scope of the data returned, or boilerplate responses to requests (i.e. standardized text that is repeated without making changes to the original and does not consider the context of a request, essentially amounting to an automated response). Where data sets are complex or dense, controllers should facilitate access through tools that will help them understand the data.[208]

Scope

When making a request for access to personal data, the first thing that the data subjects need to know is whether or not the controller processes data about them. Therefore, they have a right to know if personal data about them is being processed. If processing is happening, they have a right to have access to that data and information about the processing itself.

---

[207] Acker and Kriesberg; Laurens Naudts, Pierre Dewitte, and Jef Ausloos, "Meaningful Transparency through Data Rights: A Multidimensional Analysis," in *Research Handbook on EU Data Protection Law* (Edward Elgar Publishing, 2022), 530–71,

[208] Jef Ausloos, Rene Mahieu, and Michael Veale, "Getting Data Subject Rights Right," SSRN Scholarly Paper (Rochester, NY, December 1, 2019), https://papers.ssrn.com/abstract=3544173.

Complete means the data subjects should have access to all the information that the controller processes regarding them. This also means that the controller is obliged to search for personal data throughout its IT systems and non-IT filing systems.

The GDPR explicitly states that access to personal data means access to the actual personal data, not only a general description of the data or a reference to the categories of personal data processed by the controller. Data subjects are entitled to have access to all data processed relating to them, or to parts of the data depending on the scope of their request. However, the obligation to provide access to the data does not depend on the type or source of those data. Rather, it applies to its full extent even in cases where the requesting person had initially provided the controller with the data, because the purpose is to let the data subject know about the actual processing of the data by the controller. The copy of the data must also be complete (i.e. include all personal data requested).

Controllers are not required to provide personal data that they processed in the past but that they no longer have at their disposal. For instance, the controller may have deleted personal data in accordance with its data retention policy and statutory provisions, and may thus no longer be able to provide the requested personal data.[209] However, the data must have actually been deleted from the controller's servers. This is noteworthy because controllers continue to hold and reprocess data that individuals deleted by clicking on a "delete" button.

Format

Despite the importance of formats, the GDPR is not prescriptive about them because individuals have different legitimate reasons for exercising their rights and so have different backgrounds

[209] The length of time for which the data are stored should be fixed in accordance with GDPR Article 5(1)(e).

and capabilities.[210] Secondly, the right of access may be easy and straightforward to apply in some situations, for example when a small organization holds limited information about someone. In other situations, the right of access is more complicated because the data processing is complex with regard to the number of data subjects, the categories of processed data, as well as the flow of data within and between different organizations. Considering these differences in personal data processing, the appropriate way to provide access will vary accordingly.

It is the responsibility of the controller to decide upon the appropriate form in which the personal data will be provided. But the GDPR provides users with the right to receive personal data in a structured, commonly used and machine-readable format. This makes it easier to move to different formats and also different services. Given the requirement, the format in which the data is delivered is important to consider. The most common are CSV, TXT, and JSON.

The data must be provided in such a way that makes it possible for the data subject to keep it. Thus, the requirement to provide a copy means that the information about the person who makes the request is provided in a way which allows them to retain all of the information and to refer back to it. PDFs may be one way of doing this in some cases and often are, but there are limitations to that format. The GDPR allows for PDFs. PDFs were designed originally for printing, not for data analysis. They are sometimes saved as static files, which hinders their portability, and they do not include metadata or allow for the effective re-use of the data.[211] Their usefulness is limited to older data systems, designed for shallower data collection and processing, when the number of data points about any given person were fewer than today. A

---

[210] Ausloos, Mahieu, and Veale, "Getting Data Subject Rights Right."

[211] "16 EN/WP 242 Rev. 01," 29.

print-out or summary in the past was considered sufficient to give an individual oversight about their personal data being processed, delivered in a PDF format.[212] Today there is a proliferation of tracking technologies and the growth of the behavioral marketing industry that make PDFs less attractive.[213]

Because the GDPR leaves the decision up to the data controller, the controller has to consider how formats could impact or hinder user access rights.

Readability and Compatibility

Hardware and software change often and social media platforms are known for frequent and unexpected "tweaks" and redesigns to the features on their sites. These reduce long-term readability and compatibility so that a file format or device may become obsolete and no longer readable. Controllers should consider the longevity of their file format choices to ensure long term readability and access. File formats that are more likely to be accessible (or stable) in the future have the following characteristics:

- Non-proprietary (not registered or protected as a trademark or brand name; generic)
- Open, well-documented standard (well-documented means written in inviting and clear language, that is comprehensive, detailing all aspects of the standard).[214]
- Common use by research community (more reliable results and performance; more likely to support long-term maintenance because there many stakeholders)
- Standard representation (ASCII, Unicode, both common standards. Unicode is the universal character encoding used to process, store and facilitate the interchange of text data in any language while ASCII is used for the representation of text such as symbols, letters, digits, etc. in computers. They make data legible. Exotic or arcane encoding is likely to be unreadable/discarded.)

---

[212] Portable document format.

[213] Ausloos, Mahieu, and Veale, "Getting Data Subject Rights Right"; Solove, "Introduction."

[214] Adam Scott, "The Eight Rules of Good Documentation," O'Reilly Media, April 17, 2018, https://www.oreilly.com/content/the-eight-rules-of-good-documentation/.

- Unencrypted (information or data not converted into a code to enforce authorized access but whose "keys" can be lost thereby making maintenance impossible
- Uncompressed (this means information about the data is not eliminated during processing, which can make a file unreadable) [215]

Authentication

In order to ensure the security of processing and minimize the risk of unauthorized disclosure of personal data, the controller must be able to identify the data subject, i.e. find out which data refer to the data subject, and, if there are any doubts, confirm the identity of the person. The GDPR does not impose any requirements regarding the methods for determining the identity of the data subject.

Additional information requested by the controller should not be more than the information initially needed for the verification of the data subject's identity (authentication).

If the controller has reasonable grounds for doubting the identity of the requesting person, they may request additional information to confirm identity. However, as a rule, the controller cannot request more personal data than is necessary to enable this identification, and the use of such information should be strictly limited to fulfilling the users' request. Verification cannot lead to excessive demands and/or to the collection of personal data which are not relevant or necessary to strengthen the link between the individual and the personal data requested.

Therefore, controllers should assess the proportionality of their identification measures, taking into account the type of personal data being processed (e.g. special categories of data or not), the nature of the request, and the context within which the request is being made.

[215] "File Formats for Long-Term Access | Data Management," accessed June 24, 2023, https://libraries.mit.edu/data-management/store/formats/.

Proportionality means avoiding excessive data collection while ensuring an adequate level of processing security.

The controller should implement an authentication (verification of the data subject's identity) procedure in order to be certain of the identity of the persons requesting access to their data, and ensure security of the processing throughout the process of handling an access request, including for instance a secure channel for the data subjects to provide additional information. The method used for authentication should be relevant, appropriate, proportionate and respect the data minimization principle.

The principle of "data minimization" means that a data controller should limit the collection of personal information to what is directly relevant and necessary to accomplish a specified purpose. They should also retain the data only for as long as is necessary to fulfill that purpose. In other words, data controllers should collect only the personal data they really need, and should keep it only for as long as they need it.[216]

Consequently, it is disproportionate to require a copy of an identity document if the user has already been authenticated by being logged into their accounts. Authentication is the default for platform downloads. Using a copy of an identity document as a part of the authentication process creates a risk for the security of personal data and may lead to unauthorized or unlawful processing, and as such it should be considered inappropriate in most cases.

Lastly, if controllers do not provide full access to a data subject they have to explain why within one month. The explanation has to explain the concrete circumstances and let users take action against the refusal. The explanation must include information about the possibility of

[216] "European Data Protection Supervisor," Glossary, June 20, 2023, https://edps.europa.eu/data-protection/data-protection/glossary/d_en.

lodging a complaint with a supervisory authority and seeking judicial remedy. But controllers can decide how and in what format they want to deliver these explanations.

Data protection by design

Data protection by design is part of an overall principle of how access rights should be delivered according to the GDPR. Measures should be designed to implement data-protection principles, such as data minimization, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of the regulation and protect the rights of data subjects. The controller must implement appropriate technical and organizational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility.

One way to provide the information that meets expectations about scope and does so by default is by providing the data subject with self-service tools. They could facilitate an efficient and timely handling of data subjects' requests of access and will also enable the controller to include the verification mechanism in the self-service tool.

A social media service has an automated process for handling access requests in place that enables the data subject to access their personal data from their user account. To retrieve the personal data the social media users can choose the option to "Download your personal data" when logged into their user account. This self-service option allows the users to download a file containing their personal data directly from the user account to their own computer.

Figure 7: Example from the GDPR of a self-service tool that facilitate an efficient and timely handling of data subjects' requests of access and will also enable the controller to include the verification mechanism

The use of self-service tools cannot limit the scope of personal data received. If not possible to

give all the information required in the GDPR through the self-service tool, the remaining information needs to be provided in a different manner. The controller may indeed encourage the data subject to use a self-service tool that the controller has set in place for handling access requests. However, if the controller must also handle access requests that are not sent through the established channel of communication. Further, the notion of "provide" means that "the data subject must not have to actively search for information about their rights in the GDPR amongst other information, such as privacy, or terms and conditions of use, pages of a website or app." The controller needs to take into account the quantity and complexity of the data when choosing means for providing access. But if the controller does decide on self-service tools, they must take appropriate measures to provide a concise, transparent, intelligible, and easily accessible form, using clear and plain language.[217] The requirement that providing access to the data subject has to be done in a concise and transparent form means that controllers should present the information efficiently and succinctly in order to be easily understood and captured by the data subject, especially if it is a child.

Possible limitation of the right of access

Since communicating and making available personal data to the data subject is a processing operation, the controller is always obliged to implement appropriate technical and organizational measures to ensure a level of security appropriate to the risk of the processing.[218] This applies independently of the modality in which access is provided. In case of non-electronic transmission of the data to the data subject, depending on the risks that are presented by the processing, the

[217] According to GDPR Article 12(1).

[218] in GDPR Articles 5(f), 24 and 32.

controller may consider using registered mail or, alternatively, to enable, but not oblige, the data subject to collect the file against signature directly from one of the controller's establishments. If information is provided by electronic means, the controller shall choose electronic means that comply with data security requirements. Also in case of providing a copy of the data in a commonly used electronic form, the controller must take into account data security requirements when choosing the means of how to transmit the electronic file to the data subject. This may include applying measures including encryption and password protection. In order to facilitate access to the encrypted data, the controller should also ensure that appropriate information is made available so that the data subject can access the requested information in clear ways.

Inferred data

Implementing these criteria is complex because of the sophistication and obfuscation of tracking technologies and the growth of the behavioral marketing industry.[219] Using Google translation online apps for example produces data about individuals as well as what they are trying to learn. That data is used to improve automated methods for detecting text. In many cases, scripts record keystrokes, mouse movements, and scrolling behavior, along with the entire contents of the pages a user visits, and sends them to remote servers. Unlike typical analytics services that provide aggregate statistics, these scripts are intended for the recording and playback of individual browsing sessions. The stated purpose of this data collection includes gathering insights into how users interact with websites and discovering broken or confusing pages. Text typed into forms is collected before the user submits the form, and precise mouse movements are saved, without visual indication to the user. Some companies allow publishers to explicitly link

[219] Solove, "Introduction."

to recordings of a user's real identity but even in other cases, anonymity is not assured given the scale and scope of the data being collected.[220]

In 2020, Google was accused in a proposed class action in a U.S. District Court of Northern California of illegally invading the privacy of millions of users by pervasively tracking their internet use through browsers set in "private" mode. The data could be mined to learn about users' friends, hobbies, favorite foods, shopping habits, as well as intimate and potentially embarrassing information.[221]

These are examples of inferred data. This is data derived from processing rather than directly from a person. This can be data from sensors that indirectly provide data about people, often without their direct involvement or knowledge. Companies and governments also perform analytics on direct data to infer other characteristics of the data subjects. An algorithm based on publicly available social media posts may infer that someone is not a safe driver. This analysis is not based on a single data input, such as address, but on large datasets using techniques like regression, classification, clustering, and machine learning. However, as the example of car insurance shows, they can affect individuals – as when they are denied auto insurance or required to pay higher premiums than they would otherwise pay – based on algorithmic processes beyond their knowledge. Companies have argued that inferred data is the knowledge a company

[220] Steven Englehardt, "No Boundaries: Exfiltration of Personal Data by Session-Replay Scripts," *Freedom to Tinker* (blog), November 15, 2017, https://freedom-to-tinker.com/2017/11/15/no-boundaries-exfiltration-of-personal-data-by-session-replay-scripts/.

[221] Jonathan Stempel, "Google Faces $5 Billion Lawsuit in U.S. for Tracking 'private' Internet Use," *Reuters*, June 2, 2020, sec. U.S. Legal News, https://www.reuters.com/article/us-alphabet-google-privacy-lawsuit-idUSKBN23933H.

generates from its processing activities and, therefore, is intellectual property that is owned by the creating company.[222]

California laws define inferred data as data that is derived from information otherwise considered personal information under the law; and data that used by the controller or business for the purpose of creating a profile about that consumer. The matter is unsettled in the European Union, where personal data includes opinions and inferences based on legal precedent but the GDPR stops short of extending the requirements for access to personal data to inferred data. In addition, users do not have a right to access to algorithms or other technical matters that platforms argue are trade secrets.

These processes may amplify bias and inaccuracies via positive feedback loops, which further entrench negative consequences for data subjects.[223] This is the reason for "right to be forgotten" laws that allow users to have data deleted from servers that could continue to have a negative effect on their reputation, credit score, or job and housing prospects. Thus, arguably, knowing about data that has been collected (or is the result of additional processing), and knowing about data that should have been deleted, are arguably pre-requisites to other access rights, including the right to know what data a controller has about an individual, to request to have data deleted and/or rectified. It is also necessary to effective oversight, and the principle of transparency – all of which are necessary to counterbalance the power of data controllers.[224]

---

[222] Joan Wrabetz, "What Is Inferred Data and Why Is It Important?," *Business Law Today from American Bar Association* (blog), August 22, 2022, https://businesslawtoday.org/2022/08/what-is-inferred-data-why-is-it-important/.

[223] Bart Custers, "Profiling As Inferred Data. Amplifier Effects and Positive Feedback Loops," SSRN Scholarly Paper (Rochester, NY, October 9, 2018), https://doi.org/10.2139/ssrn.3466857.

[224] Ausloos, Mahieu, and Veale, "Getting Data Subject Rights Right."

This section outlined relevant criteria for testing compliance with access provisions in the GDPR. The GDPR should make it possible, within reasonable limits, to know what data is being collected and processed in a way that helps the user to fulfill their access rights. This should be possible in a straightforward way that does not require users to share more data. In the next section, I outline my findings about how these provisions were accommodated. Do the platform downloads meet the criteria for scope and formats? Are they legible by an average subscriber (using myself as an example)? Does the format support sufficient oversight over the data processing taking place?[225] For example, is the data compatible with independent analysis tools? To what extent do the three methods fulfill the goals included in these criteria of preventing unauthorized access to an account without imposing undue burdens on legitimate access, while also minimizing data collection during the process?

In answering these questions, I look at ways in which the GDPR right of access was fulfilled by Twitter, Facebook, and Instagram. I also look at advantages and disadvantages of the different methods of collecting data as a whole as a way to assess data protection by design and default and data self-management, including such criteria as privacy and consent. Platform downloads were the only method for collecting GWC social media that provided direct access to the data. This offered a surprising measure of privacy and consent compared to the other methods but, along with the other criteria, also offers useful insights about how others will experience efforts to access their data and assess how it is being processed.

[225] Ausloos, Mahieu, and Veale.

Criteria:

- Method for downloading data and security
    a. Dedicated portal or other method
    b. Automated process
    c. API
    d. Communication
- Extent and relevance of the data (scope)
    a. Missing data (and explanations)
    b. Organization of data (and explanations)
- Portability
    a. Commonly used format (JSON, XML, CSV)
- Format
    a. Machine readable format (JSON, XML, CSV)
- Privacy
    a. Authentication

I explain the criteria below and, in the results, report issues encountered

Results

I chose format, scope, portability, and privacy according to GDPR guidelines and I highlight advantages and disadvantages of platform downloads in this context.

Service	Explanation	Format	Scope	Portability	Privacy
Facebook	Boilerplate	Yes	No	Inconclusive	Automated, Secure, API
Twitter	Boilerplate	Yes	No	Inconclusive	Automated, Secure, API
Instagram	Boilerplate	Yes	No	Inconclusive	Automated, Secure, API

Table 1: Overview of Platform Download Results

According to the GDPR, the copy of the data should be complete (i.e. include all personal data requested). I should have access to all the information that Twitter and Facebook holds regarding me. And Twitter and Facebook should search for personal data throughout its IT systems and non-IT filing systems. Moreover, the GDPR explicitly states that access to personal data means access to the actual personal data, not only a general description of the data; or a reference to the categories of personal data processed by the controller. Therefore, I should be able to use Twitter, Facebook, and Instagram's platform download features to obtain a complete copy of personal data held by each, and be able to analyze the contents and the data processing to the best of my ability.

Below is an overview of the platform download self-service pages from Twitter, Facebook, and Instagram for comparison.

Facebook

Your Activity Across Facebook: Information and activity from different areas of Facebook, such as posts you've created, photos you're tagged in, groups you belong to and more.

Personal Information: Information that you've provided when you set up your Facebook accounts and profiles.

Connections: Who and how you've connected with people on Facebook, including things like your friends and followers.

Logged Information: Information that Facebook logs about your activity, including things like your location history and search history.

Security and Login Information: Technical information and logged activity related to your account.

Apps and Websites off of Facebook: Apps you own and activity we receive from apps and websites off of Facebook.

Preferences: Actions you've taken to customize your experience on Facebook.

Ad information: Your interactions with ads and advertisers on Facebook

Twitter

Your Twitter data provides you with a snapshot of your Twitter information, including the following:

Account: If you are logged in to your Twitter account, you will see information such as your username, email addresses or phone numbers associated with your account, and your account creation details. You will also see certain information that you may have previously provided to us, such as your birthday and profile location. Whether or not you are logged in, you can also see certain information that we have inferred about your account or device such as gender and age range. You can update or correct most of this information at any time (your account creation details cannot be edited).

Account history: If you are logged in, you will also be able to see your login history, as well as the places you've been while using Twitter.

Apps and devices: You can also view the browsers and mobile devices associated with your account (if you are logged in) or current device (if logged out), and the apps you have connected to your Twitter account. If you see login activity from an app you don't recognize or that looks suspicious, you can go to the Apps tab in your settings to revoke its access to your Twitter account. The IP location shown is the approximate location of the IP address you used to access Twitter, and it may be different from your physical location.

Account activity: You will be able to see the accounts you've blocked or muted.

Interests and Ads data: You can also see interests that Twitter and our partners have inferred about your account or current device. These interests help improve your Twitter experience by, for example, showing you better content including ads, notifications, and recommended Tweets in your Home timeline and Explore. You can also view any Twitter advertisers who have included your account or current device in their tailored audiences.

You can also access additional information about your account elsewhere on Twitter while logged in, including the contacts imported from your address book, your entire Tweet history, the apps you have given access to your Twitter account, and the Twitter accounts you've muted and blocked.

Instagram

The platform download feature was difficult to find, located by clicking on a small icon at the bottom of the page. Because of Instagram' minimalist features, no identifying text was visible to identity the menu. There was no way to know how to find the platform download feature except to click through the icons. Once located, Instagram provides no information about what is included in the platform downloads. There are no other options aside from JSON and HTML formats.

Formats

The downloads were provided in two formats. One was a visual representation in HTML of the original site minus missing links, media, and features. The JSON version meets GDPR standards for providing data in a common machine-readable format that is compatible with other systems. JSON meets best practices for data formats by being non-proprietary; open and well documented; commonly used in the research community; unencrypted; uncompressed; and it can represent all of the most common printable characters, as well as the non-printable characters.[226]

JSON is not readable without additional support because it is computer code printed in a long, undifferentiated block without formatting. But it can be uploaded to a spreadsheet program like Excel or Google Sheets.

The JSON format provides the advantage of allowing anyone familiar with programming languages used for data analysis (Python or R) to import the file into an external analysis tool (such as Pandas or R-Studio)[227].

---

[226] European Union Agency for Fundamental Rights and Council of Europe, "Handbook on European Data Protection Law, 2018 Edition" (Publications Office of the European Union, April 2018).

[227] These are popular data analysis packages that can process data in order to identify pattern such as the number of times a keyword or phrase is mentioned. They are necessary to large-scale datasets that cannot be managed by close-reading.

This would provide an extra layer of analysis and scrutiny. For example, this would be one way to identify what data may be missing from the download.

Data analysis of social media data is a common strategy for researchers interested in social networks, misinformation, or other topics and who have the necessary computing skills for analyzing large-scale datasets. However, inequalities in access to computing skills have persisted for decades across geography, gender, age, and income. Thus the majority of subscribers should not be expected to use data analysis tools and methods that skilled professionals have, or be expected to in the near future.[228]

The HTML version could have advantages for users who are not familiar with data analysis tools. I could immediately see from the HTML version that the download was missing links and icons. This also meant that the HTML version was only a partial recreation of the site and I would have to search for the original media for the full representation to be restored. This was difficult because the data had no discernible order, either in JSON or HTML. In both versions, dates were out of order.

The categories of data didn't offer a logical structure for the purposes of finding data, identifying any missing data, or, generally understanding the extent of data collected about me. It is difficult to imagine how I could exercise my right to object to processing operations if I could not identify the data or the processing in the first place.

---

[228] Pernille Bjørn, Maria Menendez-Blanco, and Valeria Borsotti, *Diversity in Computer Science: Design Artefacts for Equity and Inclusion*, electronic resource, 1st ed. 2023 (Cham: Springer International Publishing : Imprint: Springer, 2023), https://doi.org/10.1007/978-3-031-13314-5; Gabriele Kaiser and Pat Rogers, eds., *Equity in Mathematics Education: Influences of Feminism and Culture* (London ; Washington, D.C: Falmer Press, 1995); Julie R. Posselt, *Equity in Science: Representation, Culture, and the Dynamics of Change in Graduate Education* (Stanford, California: Stanford University Press, 2020); Linda Skrla and James Joseph Scheurich, eds., *Educational Equity and Accountability: Paradigms, Policies, and Politics*, Studies in Education/Politics (New York: RoutledgeFalmer, 2004).

Scope

The platform downloads limit the scope of personal data received. The instructions for Twitter, Facebook, and Instagram state that the data returned will be incomplete, as well as being pre-selected and compiled by the platforms according to "the information that we believe is most relevant and useful to you." In other words, not only is the data incomplete, but is also a summary identified by categories. For example, Twitter provides a "snapshot" to users through the platform download, while according to the instructions, the full tweet history is available elsewhere. A link to the page led me to Twitter's "Help Center" page, where I found more platform download instructions. But the instructions include the disclaimer that the tweet history/platform downloads are limited to:

> The information we believe is most relevant and useful to you, including your profile information, your Tweets, your Direct Messages, your Moments, your media (images, videos, and GIFs you've attached to Tweets, Direct Messages, or Moments), a list of your followers, a list of accounts that you are following, your address book, Lists that you've created, are a member of or follow, interest and demographic information that we have inferred about you, information about ads that you've seen or engaged with on Twitter, and more.

The instructions do not include criteria for relevant and useful. But this is not information that the user should have to search for. The aim and overall structure of the right of access is to provide individuals with sufficient, transparent, and easily accessible information about the processing of their personal data so that they can be aware of, and easily verify, the lawfulness of the processing and the accuracy of the processed data. This make it easier to exercise other rights, such as having

data erased or wrong information corrected.[229] The notion of *to provide* in the GDPR means that the data subject must not have to actively search for information about their rights amongst other information, such as privacy, or terms and conditions of use, pages of a website or app. The platform download features function contrary to this criteria in the GDPR.

The platform downloads themselves are only available by searching through several layers of privacy information. These self-service instructions are concise, but not transparent. They are not in an easily accessible form, nor do they use clear and plain language.[230] Not only is the language vague and contradictory but it is also ambiguous. Under a Privacy Policy page, Facebook includes a statement that metadata about content is collected, "like the location where a photo was taken or the date a file was created," and "information about the message itself, like the type of message or the date and time it was sent." The word "like" does not provide a clear statement about the scope of the data collected and processed. Overall, there is no specific information about what data is maintained or for how long. Without a clear statement, it's not possible to know if the scope of the platform downloads fully meet GDPR standards.

In addition, Twitter and Facebook associate access to data with privacy (e.g. the downloads are available through a privacy policy page several layers into the site). Nowhere is access as a right mentioned. Rather, Facebook, by way of illustration, makes the following statement:

> We store different categories of data for different time periods, so you may not find all data from the time you joined Facebook. You won't find information or content that you deleted because we delete that content from our servers. Remember, you can access most of the content you post to a certain profile on Facebook by logging into your account and switching to that profile. Keep in mind: The categories of data we receive, collect, and save may change over time. Learn more about your Facebook data in our Privacy Policy.

[229] "Guidelines 01/2022 on Data Subject Rights: Right of Access."

[230] According to GDPR Article 12(1).

The ambiguous statement at once makes readers aware that they can access their Facebook data while also claiming the right to delete content, change categories of data that will be provided, and deny access without offering specific details.

Privacy and Consent

Platform downloads from Twitter, Facebook, and Instagram provided the most complete privacy protection and opportunities for consent. For one, they required multiple levels of authentication and validation. I was unable to access the account as a result, and had to rely on the original subscriber to request the download for me.

Nor could I collect the GWC Facebook accounts using Archive-It or Webrecorder and had limited success with Twitter and Instagram. This was apparently a function in one case of rate limiting, which refers to managing the number of requests (or "calls") for data are made on an API. This is done to limit the amount of data they can extract. The operator API will send error messages indicating "too many connections" when the limit has been exceeded. Information about the platforms' APIs and rate limits are in developer agreements. In my case, Facebook, as the API operator, limited the number of requests that could be made using Webrecorder, which appeared to be capturing the Facebook account but had not. I received no error message or indication otherwise that the collection had failed. Together with authentication procedures, the APIs and rate limiting provided automated privacy protections that would be difficult to override. Platform downloads seem to provide an ethical solution for at-scale collecting. However, platform downloads put Twitter, Facebook, and Instagram in charge of mediating privacy and consent.

Identification

The GDPR is explicit that the controller may not require the data subject to provide more information than necessary for identification. In the case of platform downloads, subscribers are identified through an authentication process identical to the process used for anyone who seeks to access subscriber data and who is not the subscriber: external "third-party" developers who create new products, researchers, journalists, and archivists. for the platform downloads, an interface simplifies the process.

I had the credentials to log-into the GWC Twitter, Facebook, and Instagram accounts but could not because of the authentication protocols linked to a mobile phone number. As I said above, this offered privacy controls. I would have to log into the account, and thus follow the same authentication protocols, to ask questions about the data and the platform downloads. This is because

Twitter, Facebook, and Instagram provide users with a form on the site that requires a log-in and password authentication. I did not realize this until more than a year after my data collection. In addition to written correspondence, I decided to pose general questions by logging into my own account to use the platform download feature for my data and send questions to the platform through the form on the site.

Follow-up questions about the platform downloads required me to provide a government-issued identification, i.e. drivers' license even though I was logged into my account. This practice is not only prohibited in the GDPR, but was also prohibited in previous E.U. directives because the practice is contradictory to the principle of data minimalization. Moreover, the inquiries included in the form were preselected. Twitter provided a comment box. But I was

warned in boilerplate language that questions would be ignored if they didn't match preselected criteria, such as "I am requesting Twitter account information."

Data portability

The platform downloads offer an advantage in terms of portability because they are available in a machine-readable format (JSON). But the HTML (Hypertext Markup Language) version could also be portable because it is a text-based approach to describing the structure of content contained in an HTML file. The HTML markup tells a web browser how to display text, images and other forms of multimedia on a webpage. The contents would not be displayed visually like the original without additional elements, including CSS and JavaScript, which are not as portable, if at all. Moreover, Archive-It and Webrecorder may not meet the portability standards because they store data in a specialized format, WARC. This format is shared by web archiving services alone so does not offer the kind of wide-spread compatibility that the GDPR portability provisions appears to include.

However, the platform downloads may offer only limited portability because of the way that data is organized according to categories specific to the platform.[231] In addition, Facebook, Twitter, and Instagram provide separate features for users who want to transfer their data to another service. The GDPR includes nothing to stop platforms from intervening, and the transfer may be complicated for the average user. But it could put users at a disadvantage by diminishing their right to transfer their data independently to another service. To be transparent and simpler to assess, standards about these transfers, and alternatives, should be made available.

[231] "Interoperability," European Data Protection Supervisor, accessed June 24, 2023, https://edps.europa.eu/data-protection/our-work/subjects/interoperability_en.

Discussion

The platform downloads formally met some GDPR standards, such as formatting and portability, and offer advantages. But, in practice, they present limitations to each of the criteria and they fail overall in terms of access and user rights. Generic "read me" explainers that accompanied the downloads clearly stated that the results would be a sample of the data on account holders' sites, and also curated. Unclear is how the platforms decide what is most relevant and useful to the person requesting the download, and what is included and what is left out. Given that the results were fractured, incomplete, and difficult to interpret, they needed an explanation to specify how the data was sampled and categorized. In fact, none of the platforms identified the method they use to compile the downloads. Without more disclosure about how the data is compiled, their information value is inherently limited and the results must be interpreted accordingly. This may be somewhat beside the point because the results were not in line with GDPR standards in the first place.

The examples here show that Facebook, Twitter, and Instagram impose their criteria on what data counts as interesting and relevant to their users. They also demonstrated their ability and willingness to prevent users from employing other methods (Archive-It, Webrecorder) to collect data. This was an effective privacy measure but is also means that private companies are now adjudicating privacy with automated methods that are poorly documented and opaque to users and regulators.

Policymakers leave fulfilling requirements up to private industry, thereby opening the door to self-service approaches like platform downloads and APIs. While the design of both offer a data protection by design, they do not support transparency or fairness. Moreover, rights can only be exercised through a proprietary API and in the current state the ability of users to effectively

interact with "their" data effectively depends on their ability to use an API and on their understanding of its technical constraints.[232] The way Twitter, Facebook, and Instagram describe platform downloads normalizes their intervention. Account holders must "request" their data from the platform, which then "prepares" the download. The platforms call personal downloads archives and copies, implying that the results will replicate the live site and be complete. Yet, the results are a snapshot of their account rather than a rendering of it. Ultimately, platform downloads reinforce control by platforms.[233]

While the GDPR appears to offer coherent enforcement, political and corporate interests hinder their effectiveness at fulfilling user rights. This was illustrated by the way that Twitter, Facebook, and Instagram chose to comply with provisions in the GDPR.

Twitter, Facebook, and Instagram did not canvass users about their preferences or needs with regard to accessing their data. Instead the companies implemented platform downloads by repurposing an automated system they use to make data available to commercial brokers and application developers that retains control over the data. As a result, access rights involve complex data formats, algorithms, and APIs. Formats are intelligible only with additional technical skills and software. And portability is limited.

Do platform downloads offer a reliable method of accommodating user rights to access? My original premise was that the technical design of platform downloads prevented access rights from being sufficiently accommodated. In particular, I identified platform downloads and, specifically APIs, as an impediment to effective data protection. Indeed, the downloads do not

---

[232] Cornelius Puschmann and Jean Burgess, "The Politics of Twitter Data," SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, January 23, 2013), https://doi.org/10.2139/ssrn.2206225.

[233] Chinmayi.

meet standards of fair processing, data minimization, data quality, or protection. And APIs are problematic for reasons summarized in this and other chapters. But the failure of effective legislation lies with political leaders who are outsourcing the role of oversight to users and allowing the industry they are supposed to regulate to define the way that user rights will be accommodated.

This inquiry is a step toward establishing a conceptual and methodological framework for assessing platform regulations by assessing platform downloads. Data rights are considered necessary for individuals to better understand how and for what purposes their data are being processed. They are also intended to help make digital infrastructures more transparent and contribute the means for empowering individuals and society, as well as realigning information-driven power asymmetries.[234] Looking at user rights through an empirical assessment as this study does provides the opportunity to reconsider letting an industry premised on the accumulation of data regulate itself using automated systems that enforce control of data.

Platform downloads were implemented by social media platforms in response to demands for more transparency about the companies' data collection and sharing practices. They do not meet the criteria by which stakeholders have sought to regulate platforms. Instead, platforms have sought ways that appear to comply with regulations but do so without ceding in meaningful ways to the demands for accountability and transparency. Personal downloads illustrate the way that platforms structure data as well as decide what will be provided to subscribers, who will be challenged to identify the data collected by the platforms, how it was shared, and with whom.

At the same time, the assessments demonstrate that APIs and rate limits may have advantages for privacy and consent because they are effective at blocking unauthorized access.

[234] Naudts, Dewitte, and Ausloos, "Meaningful Transparency through Data Rights."

Nonetheless, they show APIs and rate limits to be inadequate tools for privacy and consent because they are black boxes whose operations are inconsistent and because they provide no way for the subscriber to inquire about, or contest the results.

Making users responsible for exercising their rights does not give them adequate, appropriate means to challenge decision-making about how those rights will be provided.[235] Platform download are repurposed applications designed to control access to data stored on the company servers. Facebook, Twitter, and Instagram repurposed platform downloads for fulfilling a requirement in the GDPR to access and thus control personal data.

The reasoning of that requirement is quite expansive and normative, including dignity and autonomy, and the GDPR includes steps for individuals to exercise authority over their personal data. Those steps are based on a contentious self-management, by-design premise. Nevertheless, the findings in this chapter show that Facebook, Twitter, and Instagram are operating with a view limited to avoiding data breaches and other problems for which E.U. regulators could penalize them with fines and subscribers could penalize them by ending their use of the service, thereby withholding the data that the services depend on. The platform downloads are repurposed in a way that limits user rights to a narrow interpretation by Facebook, Twitter, and Instagram. And, they are doing so by repurposing APIs, algorithms, and rate limits. Although research about personal downloads is scarce, researchers have raised consistent concerns about the way in which APIs and algorithms manage access to data as well as affect the quality of the data that can be accessed.

[235] Acker and Kriesberg; Laurens Naudts, Pierre Dewitte, and Jef Ausloos, "Meaningful Transparency through Data Rights: A Multidimensional Analysis," in *Research Handbook on EU Data Protection Law* (Edward Elgar Publishing, 2022), 530–71,

Putting data protection onto the shoulders of individuals requires rights and resources if they are to determine what information is disclosed, and to whom. Yet, platforms implemented these methods without public debate or assessment by regulators. Regulators, relying on by-design principles, left the execution up to the companies to decide. In this case, platform downloads met narrow criteria for offering automated options to individuals to make it easier access their data. But regulators missed the opportunity for policymaking that was relevant and met the needs of the constituents, whose options remain limited.

Consequently, while the rights and principles are clear, our ability to exert the right over our personal data is unsettled at best, requiring a level of sophistication and technical know-how that is not widespread.[236] Platform downloads are akin to gestures that make it appear that Facebook, Twitter, and Instagram are complying with the GDPR by giving access to their subscribers. Similar to Waldman's social practice of law, these gestures on the part of platforms include writing policies and making tweaks to their sites, interpreting the policies and tweaks in blogs as a service to consumers (business and individuals), consulting with regulators, translating legal requirements into technical specifications and new products, and conducting internal assessments and external audits. They perform the management of privacy and access but they do not adequately meet the standards for either. Taken as a whole, by-design tactics in the GDPR like self-management platform downloads are inadequate. In the next chapter I look at whether the inadequacies I found are addressed in the Digital Markets Act, new European Union regulations being put into place.

[236] Christian Katzenbach and João Carlos Magalhães, "Platform Governance Archive," n.d.

Chapter 4: Portability in the DMA

In the previous chapter I assessed user right measures in the GDPR linked to privacy and the protection of personal data, including the right to access and transfer personal data, called portability. I compared criteria in the GDPR for access and portability with platform downloads, a built-in tool for downloading data provided by Facebook, Twitter, and Instagram from their sites. In this chapter I assess key regulations in the E.U. Digital Markets Act (DMA) for portability.

The DMA was introduced in 2022 along with the Digital Services Act (DSA). Together they are considered to be the most comprehensive and toughest attempt to regulate large companies that dominate digital services, including Facebook, Twitter, and Instagram[237]. The DSA and DMA contribute to existing regulations with two main goals:

1. To create a safer online environment in which the fundamental rights of all users of digital services are protected especially against misinformation and the effects of hate speech
2. To establish a level playing field to foster innovation, growth, and competitiveness, both in the European Single Market and globally[238]

The DSA is directed at the first goal while the DMA is directed at the second. I look at DMA regulations for portability that are intended to make it easier to transfer data between online companies.

[237] Regulations apply to respective holding companies, such as Meta.

[238] "The Digital Services Act Package | Shaping Europe's Digital Future," May 5, 2023, https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package.

The GDPR was explicitly a data rights policy tool, which sought to give individuals more control vis-à-vis the companies that control their data. Those controllers have since become larger and have become dominant firms with anticompetitive business models and tactics. Various solutions have been identified, from more aggressive enforcement of existing competition laws, to anti-trust actions to control the acquisition of emerging competitors. The promotion or mandating of data portability are often included as a key part of digital competition policy reform agendas.[239]

Portability is closely related to the right of individuals to access their personal data. Both portability and access are E.U. rights whose value is linked to data control and privacy. Access is a right to know about data that has collected, processed, and stored. Data portability as a right enables demanding, receiving, and transferring data elsewhere, notably between different applications, programs, and service providers. Individuals can switch between providers or keep data in two places at once (called, respectively, "switching" and "multi-homing"), making it as easy as possible for users to try new services.[240]

An important addition in the DMA is the requirement for gatekeepers targeted by the regulation to provide tools that facilitate data portability not only for individual end users but also business users, in real time and continuously.

A business user accesses core platform services for the purpose of, or in the course of, providing goods or services to end users – subscribers, consumers, data subjects, or, what I have

[239] OECD, "Data Portability, Interoperability and Digital Platform Competition," OECD Competition Committee Discussion Paper, 2021, http://oe.cd/dpic.

[240] "The Digital Markets Act: Ensuring Fair and Open Digital Markets," accessed July 31, 2023, https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_en.

chosen to call in this chapter, users or individuals. They are the ones producing personal data, which is at the heart of markets that the DMA is seeking to make more competitive through portability and, at the other end, regulate with audits, investigations, penalties, fines, and court reviews.

Gatekeepers should give these in-between services continuous and real-time access to data with technical measures including APIs or an integrated tool for small volume business users. They should so this without intentional obstructions. The data access will not be complete for these business users. By contrast, individuals should have access to all the data they create on the sites and they should be able to decide what to do with it. Facilitating switching and letting individuals keep their data in multiple places should lead, in turn, to an increased choice and an incentive for gatekeepers and business users to innovate.[241] The rationale for portability is clear from this explanation in the DMA:

> Gatekeepers benefit from access to vast amounts of data that they collect while providing the core platform services, as well as other digital services. To ensure that gatekeepers do not undermine the contestability of core platform services, or the innovation potential of the dynamic digital sector, by restricting switching or multi-homing, end users, as well as third parties authorised by an end user, should be granted effective and immediate access to the data they provided or that was generated through their activity on the relevant core platform services of the gatekeepers…Facilitating switching from service to service, or having multi-homing should lead, in turn, to an increased choice for end users and acts as an incentive for gatekeepers and business users to innovate.[242]

[241] "Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on Contestable and Fair Markets in the Digital Sector (Digital Markets Act)," § 59-60 (n.d.).

[242] DMA Recital 59. (A recital is a statement that explains the rationale for a contract or a law and provides context information.)

In essence, the expectation is for portability to put individual users in a position to claim privacy, control, and access by threatening to move their data to another platform.[243]

While the rule may have the potential to facilitate the former objectives, portability at its core is being used in the DMA as a policy tool to promote competition without restraining the production of data. Again, data is a commodity whose value governments are seeking to maximize. The tension is in protecting individuals enough to encourage them to continue producing data. Portability gives regulators a tool to reduce the market power of the gatekeepers while increasing consumer choice without hampering data production, while also promoting as much entry and expansion of new operators as feasible.[244]

Despite the significant theoretical attractiveness of data portability, experience with applying these measures to digital platforms remains limited, or are only in the early stages of implementation.[245] Deep and unresolved legal and technical barriers remain, requiring more research. Even as momentum for portability grows, open questions remain concerning how various laws and legal frameworks should be coordinated. Data controllers will have to work with third parties hosts but are responsible for maintaining security and privacy of user data during the transfer. The controllers thus face uncertainty about their liability and the scope of data that should be made portable. It is equally unclear whether regulators have devised a measure for assessing the success of data portability requirements. Data portability measures

---

[243] While I discuss the DMA, similar arguments and policies are being put forth in U.S. anti-trust and other bills. See: Practical Law, "House Antitrust Subcommittee Unveils Five Big Tech Antitrust Bills," *Reuters*, June 14, 2021, sec. Legal Industry, https://www.reuters.com/legal/legalindustry/house-antitrust-subcommittee-unveils-five-big-tech-antitrust-bills-2021-06-14/. The bills do not mention the technical means that will be required but portability and interoperability are posed as solutions in policy and industry discussions.

[244] Fiona Scott Morton et al., "Equitable Interoperability: The 'Super Tool' of Digital Platform Governance," in *Policy Discussion Paper No. 4* (Digital Regulation Proect, Yale University: Yale Tobin Center for Economic Policy, 2021), 32.

[245] Ursic, "Unfolding the New-Born Right to Data Portability."

may need to include the ability for users to give consent. But what does meaningful consent look like for data portability?[246]

An outstanding issue involves the mechanisms through which data is provided to users. The DMA requires the gatekeepers to provide individual and business users with tools that facilitate continuous and real time data portability. The criteria for these tools, and how the goals of the portability and protection measures should be realized, remain an open question. In fact, little is known about the ways in which individuals have been able to make use of portability rights using tools developed by data portability initiatives.

In this chapter, I look at the portability obligations in current E.U. regulations and at the way in which Facebook, Twitter, and Instagram are meeting those obligations.

I find reason to support the premise of portability and the obligations specified in Article 6 of the DMA, which I outline in the first section. I then summarize the goals, efforts, and obstacles for portability in previous rules, most notably in the GDPR, which made portability a right connected to data control and privacy. I include details about how the technical details of the policy were formulated by E.U. advisors and how they were then formalized in 2018 by an association, the Data Transfer Project. The members of this project include several of the largest gatekeepers identified in the DMA – Google, Microsoft, Facebook, and Twitter. This group developed early methods for portability in advance of the GDPR.

I assess existing options that two of the original members, Facebook and Twitter, made available to users for controlling their data by making it transferrable. I also add Instagram and two new competitors, Threads and Mastodon.[247] Threads was released in 2022 by Facebook

[246] Riley, "Data Transfer Project Use Cases."

[247] I excluded the original members Google and Microsoft because this study is about social media platforms.

founder Mark Zuckerberg. Mastodon is an independent company popularized after Twitter's acquisition in 2022 by Elon Musk. I found that Facebook and Mastodon do little to offer effective portability. Twitter offers no dedicated option for transferring personal data. Instead Twitter provides a platform download, which complies with the letter of the law if not the spirit. I was unable to test Threads altogether, details about which I discuss in a subsequent section.

Along the way I review key terms and concepts. In addition to portability this includes interoperability. Portability allows users to move their data between different applications, programs, computing environments, and cloud services. Interoperability describes the ability of computer systems or software to exchange and make use of that data.[248] Effective data portability involving digital platforms require some degree of interoperability, so that data transfers will be sufficiently useful and dynamic to achieve their competition objectives. APIs are expected to facilitate these goals by creating the conditions for both.[249]

Portability and interoperability go hand-in-hand technically and the DMA has elevated their importance with expectations that there will be more room for competition by smaller operators, more options and control for consumers, and, generally, more options online. My assessment of portability tools and techniques shows that, despite the overall need for portability, the expectations in the DMA for competition and user control are unrealistic. In the case of portability, my findings suggest the opposite: that platforms use these options to give the appearance of compliance.

---

[248] The entirety of Article 7 (eight sections) in the DMA is devoted to interoperability including eight sections, but this is intended for messaging, video-conferencing and e-mail services (the "number-independent interpersonal communications services").

[249] Scott Morton et al., "Equitable Interoperability: The 'Super Tool' of Digital Platform Governance."

Overview of Portability and Interoperability Rights

Since portability allows the direct transmission of personal data from one source to another, the right is intended to be an important tool that will support the free flow of personal data as well as market competition by preventing users from being "locked-in" to one company's service.

Data portability is outlined in Article 6 of the DMA. When exercising Article 6, gatekeepers must provide end users, and third parties authorized by an end user to access data on their behalf, at their request and free of charge, with effective portability of data. This data will have been provided by the end user, or generated through the activity of the end user in the context of the use of the relevant core platform service. Access to this data must be "continuous and real-time."[250] The gatekeeper must offer no-cost ("free-of-charge") tools that facilitate the effective exercise of such data portability.[251]

Portability in the DMA allows for users to receive a copy of their personal data from one service provider in an electronic, commonly used, and machine-readable format, and to transmit that data to another provider (and without any intentional impediments). For example, I might be more willing to switch from Twitter to a competitor if I can transfer my data and contacts (the subscribers I follow and who follow me).

[250] "Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on Contestable and Fair Markets in the Digital Sector (Digital Markets Act)," § Article 8, Compliance (n.d.).

[251] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector (Digital Markets Act).

(9)	The gatekeeper shall provide end users, and third parties authorized by an end user, at their request and free of charge, with effective portability of data provided by the end user or generated through the activity of the end user in the context of the use of the relevant core platform service, including by providing, free of charge, tools to facilitate the effective exercise of such data portability, and including by the provision of continuous and real-time access to such data.[252]
(11)	The gatekeeper shall provide business users and third parties authorised by a business user, at their request, free of charge, with effective, high-quality, continuous and real-time access to, and use of, aggregated and non-aggregated data, including personal data, that is provided for or generated in the context of the use of the relevant core platform services or services provided together with, or in support of, the relevant core platform services by those business users and the end users engaging with the products or services provided by those business users. With regard to personal data, the gatekeeper shall provide for such access to, and use of, personal data only where the data are directly connected with the use effectuated by the end users in respect of the products or services offered by the relevant business user through the relevant core platform service, and when the end users opt into such sharing by giving their consent.[253]

Table 2: Portability-relevant sections in Chapter III Article 6 - Obligations for gatekeepers susceptible of being further specified and recitals

Additionally, gatekeepers should also ensure, by means of appropriate and high quality technical measures, such as APIs, that end users, or third parties authorized by end users, can freely transfer the data continuously and in real time.

Data Portability before the DMA

Data portability has been included in legislation since 1995.[254] The GDPR made it a right enforceable across the European Union. Additional details came from an advisory group – the

[252] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector (Digital Markets Act).

[253] "Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on Contestable and Fair Markets in the Digital Sector (Digital Markets Act)," § 11 (n.d.).

[254] "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 (EUR-Lex - 31995L0046 - EN)."

Article 29 Data Protection Working Party, or A29WP – tasked with providing guidance on interpreting and implementing the provisions.

The guidelines described what data should be included in response to a portability request, or scope. The working party interpreted the term "provided" in Article 20 broadly to include data actively and knowingly provided by the data subject (the person whose personal data is collected) and data gathered by data controllers.[255] As long as it was technically feasible, users had the right to directly transfer their data from one controller to another, without hindrance from the former (where the data resides).

The working group expected the scope to include all personal data about a data subject and which that person provided to a data controller or any "observed" data provided by virtue of using a service or device. For example, search history, traffic data, and location data.[256]

The GDPR provisions require that the data be provided in a "structured, commonly used and machine-readable format." Machine-readable refers to a file format structured so that software applications can easily identify, recognize and extract specific data, including individual statements of fact, and their internal structure. Machine-readable formats could be open or proprietary; they could be formal standards or not. But documents encoded in a file format that limits automatic processing would not meet the standard of machine-readable because the data cannot, or cannot easily, be extracted.[257]

---

[255] Janis Wong and Tristan Henderson, "How Portable Is Portable? Exercising the GDPR's Right to Data Portability," in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, UbiComp '18 (New York, NY, USA: Association for Computing Machinery, 2018), 911–20, https://doi.org/10.1145/3267305.3274152.

[256] "16 EN/WP 242 Rev. 01."

[257] European Parliament, Council of the European Union, "Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 Amending Directive 2003/98/EC on the Re-Use of Public Sector Information Text with EEA Relevance," June 26, 2013, http://data.europa.eu/eli/dir/2013/37/oj/eng.

The formats should be interoperable, defined in the GDPR as "the ability of disparate and diverse organisations to interact towards mutually beneficial and agreed common goals, involving the sharing of information and knowledge between the organisations, through the business processes they support, by means of the exchange of data between their respective…systems." No specific format was required but suggestions included XML, JSON, and CSV. These formats make data portability easier because they structure data in a way that makes it easy for other systems to process. The format and structure matter. Imagine trying to upload a PDF document or one formatted for Microsoft Word into a spreadsheet formatted for Excel. By contrast, a CSV format can be used in Excel and other spreadsheet applications like Google Sheets.

Portability and interoperability are both relevant to making data transfers technically feasible, and the advisory group distinguished between the two. According to the group, the right to receive portable personal data in common, machine-readable formats is not the same as making data interoperable across different platforms.[258] The former is a minimal standard that facilitates interoperability of the data format. Machine readable formatting is a specification for the means of interoperability, which is the desired outcome.[259] They suggested APIs as a way of achieving these outcomes and lessen the potential burden resulting from repetitive requests.[260]

Making it easier to get around obstacles set up by gatekeepers is one of the ways the DMA fills gaps left open by earlier legislation. Other gaps include technical solutions to avoid

[258] Wong and Henderson, "How Portable Is Portable?"

[259] "16 EN/WP 242 Rev. 01."

[260] "16 EN/WP 242 Rev. 01." defined an API as "the interfaces of applications or web services made available by data controllers so that other systems or applications can link and work with their systems."

hindrances like a lack of portability, interoperability, APIs, or formats and compliance. Gatekeepers now have to demonstrate compliance with the DMA and that they are trying to achieve the objectives outlined in the act and other E.U. regulations. The DMA also combines technical solutions and compliance: Gatekeepers should ensure the compliance with the DMA by design. This means building compliance measures into technological design. This might mean making default settings easy to change, avenues for communication that are not automated, or designing for interoperability. Gatekeepers will be able to consult with the European Commission and technical staff as well as stakeholders on these measures.[261]

At the same time, E.U. authorities included portability in Article 6, "Obligations for gatekeepers susceptible of being further specified." This means that stakeholders on both sides will be ironing out the details in the months ahead.

Industry Response to Data Portability in the GDPR

Portability and interoperability measures are not new to competition policy, nor are they confined to digital platforms. Telephone number portability measures in the 1990s were intended to encourage competition and interoperability concerns regarding the Windows operating system were being considered by antitrust enforcers around the same time. Even though not identified as such, the original Fair Practices principles included the working equivalent of portability. Data portability as a concept originated with Internet users' need to transfer data they had been accumulating, such as e-mail, friends' lists or address books from one service to another service. Later on, the primary aim was to enable users to easily move their data although it was also

[261] "Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on Contestable and Fair Markets in the Digital Sector (Digital Markets Act)," § Recital 65 (n.d.).

regarded as a mechanism necessary to prevent unfair competition and make data protection of individuals effective.[262] In 2007, the "Bill of Rights for Users of the Social Web" claimed a core set of rights including the freedom to grant persistent access to their personal information to trusted external sites. This meant, by extension, portability.[263] Although the declaration had no legal force, it was a first step to a right for data portability and, a few months later, the Data Portability Project was founded. The founders' ambition was to discuss and work on solutions to remove constraints to data portability.[264] There were regulatory initiatives as well: the Electronic Frontier Foundation, a privacy rights organization, produced "A Bill of Privacy Rights for Social Network Users," that claimed the concept of data portability was fundamental to promoting competition and ensuring that users could truly maintain control over their information.[265] That same year, the White House launched the "My Data" initiative with the intent of "empowering all Americans" with personal data access, interoperability, and portability.[266]

The group started developing a common set of interoperable standards and formats to support portability as early as 2012, when the first draft of portability requirements in the GDPR were released. In 2016 – the same year the GDPR became law – Google, Microsoft, and Facebook

[262] Ursic, "Unfolding the New-Born Right to Data Portability."

[263] Steve O'Hear, "A Bill of Rights for Users of the Social Web," *ZDNET* (blog), September 6, 2007, https://www.zdnet.com/home-and-office/networking/a-bill-of-rights-for-users-of-the-social-web/.

[264] Barbara Van der Auwermeulen, "How to Attribute the Right to Data Portability in Europe: A Comparative Analysis of Legislations," *Computer Law & Security Review* 33, no. 1 (February 1, 2017): 57–72, https://doi.org/10.1016/j.clsr.2016.11.012.

[265] Kurt Opsahl, "A Bill of Privacy Rights for Social Network Users," Electronic Frontier Foundation, May 19, 2010, https://www.eff.org/deeplinks/2010/05/bill-privacy-rights-social-network-users.

[266] Kristen Honey, Phaedra Chrousos, and Tom Black, "My Data: Empowering All Americans with Personal Data Access," whitehouse.gov, March 15, 2016, https://obamawhitehouse.archives.gov/blog/2016/03/15/my-data-empowering-all-americans-personal-data-access.

launched an industry association called the Data Transfer Project.[267] Twitter and Apple joined the project soon after.[268] Google and other members of the Data Transfer Project published similar announcements listing a set of principles for the initiative:

- Guarantee privacy and security. Ensuring these values is crucial to fostering trust among users that their data will be protected. User trust is a prerequisite for widespread adoption of data portability tools.
- Make data portability reciprocal between importers and exporters. Absent reciprocity, the flow of information would be one-sided, and conditions of lock-in could resurface after users switch to a new service.
- Focus portability on personal data. The scope of data made portable should extend to all the information a user has provided or generated while interacting with a service.[269]

A white paper published online in July 2018 reflects the complexity of developing portability techniques and tools to fulfill these aspirations. There is no one single way of handling portability issues, which involve data formats and structures, as well as the heterogeneity of human language syntax and semantics. There are also policy facets. Issues of identity, security, and privacy must be addressed by both the sending service and the receiving one. A transfer between two providers has to allow each to maintain control over the security of their service and the data being accessed. Everything has to be maintained consistently and be flexible enough for

[267] Chris Riley, "Data Transfer Initiative," accessed July 24, 2023, https://dtinit.org/.

[268] Damien Kieran, "Putting People First on Data Portability," July 20, 2018, https://blog.twitter.com/en_us/topics/company/2018/putting-people-first-on-data-portability. The last update to the Twitter blog where company announcements are made is dated April 25, 2023. I found no updates about portability after this 2018 post. Google had in fact begun a portability project in 2012 and the A29WP recommended trade associations like the Data Transfer Project.

[269] Craig Shank, "Microsoft, Facebook, Google and Twitter Introduce the Data Transfer Project: An Open Source Initiative for Consumer Data Portability," EU Policy Blog, July 20, 2018, https://blogs.microsoft.com/eupolicy/2018/07/20/microsoft-facebook-google-and-twitter-introduce-the-data-transfer-project-an-open-source-initiative-for-consumer-data-portability/.

new formats and use cases brought about by future innovation. Most of all, companies have to be motivated in the first place to build both export and import functionality into their services.[270]

Data Transfer Project members produced Google Takeout, Facebook's Transfer your Information, Apple's Data and Privacy Page, and Microsoft's Privacy Dashboard.[271] They made reciprocal two-way portability available so that subscribers could transfer their data automatically between the services, thereby creating, at least at this level, interoperable systems. The Data Transfer Project and the members' efforts reflected the best practices outlined by the A29WP advisors. The GDPR gave them an added incentive.[272]

Portability in Practice

Neither the effects of the GDPR or the Data Transfer Project materialized as anticipated. A 2018 study, focused on file formats, tested the portability right by sending requests by email to 230 controllers. The study found widescale confusion about the rights, obligations, and formats.[273] Researchers looking at provisions for portability among Internet of Things device makers (where personal information was collected by physical sensors were installed in traditionally private

[270] Chris Riley, "Data Transfer Initiative: Overview," accessed July 24, 2023, https://dtinit.org/overview.

[271] Julie Brill, "Putting People First on Data Portability," *Regulatory Affairs at Microsoft* (blog), May 21, 2018, https://blog.twitter.com/en_us/topics/company/2018/putting-people-first-on-data-portability; "Introducing Data Transfer Project: An Open Source Platform Promoting Universal Data Portability," *Google Open Source Blog* (blog), July 20, 2018, https://opensource.googleblog.com/2018/07/introducing-data-transfer-project.html.

[272] For formats, the Data Transfer Project adopted CSV, JSON, and XML, which were, or were fast becoming, industry standards. They also developed a limited approach that converts a range of proprietary formats into one of several data models useful for transferring data.

[273] Wong and Henderson, "How Portable Is Portable?"

settings, referred to by the acronym IoT) found little support for portability rights and data silos.[274]

In an economic and legal analysis of data portability in the context of in-car IoT systems, Daniel Gill and Jakob Metzger argued that if data protection expectations remained unfulfilled and the framework remained inefficient, data subjects still have only very limited control over their data because they are not able to gain a fair share of the value created from the data that they have generated through their activities. They argued for obligatory standardized formats; "open" (as opposed to proprietary) standards; the mandatory introduction of standardized technical solutions such as APIs; and data transfer features that are designed to be highly visible and convenient. Lastly, they suggested obligatory reciprocal data transfers between data controllers of similar size and capacity and expanding the criteria of data required to be made available to inferred data.[275]

An assessment of Facebook's portability tool found distinct deficiencies. Some data was available through a Facebook API, other data was available only through the platform downloads, and still other data wasn't available at all. The researchers concluded that the only way for someone to transfer their personal data would be to use both methods, but that neither

[274] Daniel Gill and Jakob Metzger, "Data Access through Data Portability – Economic and Legal Analysis of the Applicability of Art. 20 GDPR to the Data Access Problem in the Ecosystem of Connected Cars," SSRN Scholarly Paper (Rochester, NY, May 5, 2022), https://papers.ssrn.com/abstract=4107677; Sarah Turner et al., "The Exercisability of the Right to Data Portability in the Emerging Internet of Things (IoT) Environment," *New Media & Society* 23, no. 10 (October 1, 2021): 2861–81, https://doi.org/10.1177/1461444820934033; Lachlan Urquhart, Neelima Sailaja, and Derek McAuley, "Realising the Right to Data Portability for the Domestic Internet of Things," *Personal and Ubiquitous Computing* 22, no. 2 (April 1, 2018): 317–32, https://doi.org/10.1007/s00779-017-1069-2.
[275] "Data Access through Data Portability – Economic and Legal Analysis of the Applicability of Art. 20 GDPR to the Data Access Problem in the Ecosystem of Connected Cars."

were sufficient. They also found that Facebook couldn't import data exported from its own site.[276]

The studies provide valuable evaluations of portability measures by platforms and provisions in the DMA. I add to them by assessing the available procedures, tools, and techniques for exercising the right of data portability. I look at the way that data portability has been implemented, the challenges, and the extent to which the Digital Markets Act addresses those challenges.

Methodology and criteria

In this section, I look at the way portability has been implemented by Facebook, Twitter, and Instagram, and Mastodon against 4 main criteria:

- Data portability available
- Method for requests
    a. Dedicated portal or other method
    b. Automated process
- Method for transfers
    a. Available procedures, tools, and techniques
    b. Data format
    c. Real-time and continuous transfer
- Extent and relevance of the data (scope)
    a. Missing data
    b. Organization of data (and explanations)

[276] John Musser and Adam DuVander, "How Facebook Makes It Nearly Impossible For You To Quit," ProgrammableWeb: API University, n.d., accessed August 6, 2020. The Programmable Web series has been discontinued and the content of the site is no longer online. See: Lane Kin, "ProgrammableWeb Is Shutting Down," API Evangelist, October 15, 2022, https://apievangelist.com/2022/10/15/programmableweb-is-shutting-down/. I used a PDF version of the Musser and DuVander series that I downloaded from the website in 2020.

I also provide details about any issues I encountered.[277] I report the results according to the presence or absence of features either required or suggested in E.U. policy. Machine-readable formats are not required but are a best practice, while real-time and continuous transfer is included in the DMA. I also compare the results against the principles enumerated by the Data Transfer Project. Given that these were developed by some of the same companies that I assessed, the two should align where technically feasible. On this basis, several criteria demand attention: Facebook, Twitter, and Instagram should offer portability by equipping users with robust tools that bolster their control of personal data. These tools should be easy to use and accessible via interoperable interfaces. In addition, control over personal data is not only a technical feature but also a qualitative one. This means that I can move the data between platforms but also that the data is complete and not provided only in part according to decisions made by the platforms.

One limitation in this comparison is that I analyzed only a small set of social media platforms, some of which were not part of the Data Transfer Project industry alliance. One of those is Threads, which I did not originally intend to include. In designing the assessment I started with testing the transfer of my personal data from Twitter to Mastodon and to Facebook, and vice-versa. Those options were not possible as I explain below. Facebook does however offer an option to transfer data to Threads, a sister company to Facebook and Instagram, all founded by Mark Zuckerberg. Facebook, Instagram, and Threads are part of his holding company, Meta. Zuckerberg released Threads in July 2023. The app became as a text-based

[277] The criteria is adapted from Turner et al., "The Exercisability of the Right to Data Portability in the Emerging Internet of Things (IoT) Environment."

social network alternative and competitor to Twitter, which was sold in 2022. However, I was not able to test the transfer because Threads is not available in the European Union. The release in Europe has been delayed in order to ensure that Threads' policies are in line with the GDPR and DMA rules for gatekeepers, in particular the potentially anti-competitive combination of user data across services outlined forcefully in Article 5 of the DMA.[278]

Twitter was sold by its founders in October 2022 to Elon Musk, whose oversight has been criticized in news reports. These reports as well as blogs and posts on social media sites have described subscribers shutting down their Twitter accounts or at a minimum seeking alternatives. One of those alternatives is Mastadon, which provides an opportunity to consider expectations in the DMA that portability will support switching and avoid lock-in.[279] However, ultimately, I was unable to assess the portability criteria in all but one case. Subsequent sections outline the results and findings, which are valid despite the limitations because those limitations reflect the effectiveness of DMA regulations.

[278] "Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on Contestable and Fair Markets in the Digital Sector (Digital Markets Act)," § 5 (n.d.).

[279] Kalley Huang, "What Is Mastodon and Why Are People Leaving Twitter for It?," *The New York Times*, November 7, 2022, sec. Technology, https://www.nytimes.com/2022/11/07/technology/mastodon-twitter-elon-musk.html.

Service-to-Service	Transfer possible	Data received	Scope of data	Format	Method
Facebook-Google	Yes	Yes	Incomplete	No	Automated, Secure, Dedicated; API - inconclusive
Twitter-Mastadon	No	Incomplete	N/A	N/A	N/A
Instagram/ Facebook-Threads	Limited (available to U.S. subscribers only)	N/A	N/A	N/A	N/A

Table 3: Overview of Portability Results

Results

Twitter

Twitter does not provide a portability tool and, to my knowledge, there is only one mention of portability on the website. Consulting Twitter's Help Center led to Twitter's Privacy Policy and a section called, "How Can I Control My Data." I found a reference in the next page under the heading, "5.1 Access, Correction, Portability," which led to a link to the following: "You can download a copy of your information, such as your Tweets, by following the instructions here."

The embedded link led to a platform download page titled, "How to Download Your Twitter Archive."[280]

Facebook

Facebook offers a dedicated portability tool accessible from a Facebook Help Center page. The transfer tool is available by clicking on a link, "Transfer Your Information to a Service Off of Facebook." Facebook grants the destination service permission to access Facebook and the user account; and the destination service grants Facebook the reciprocal access permissions.[281] However, portability is limited. Available destinations are limited and divided according to the type of data:

- Facebook posts and notes: Google Docs, Daybook, Blogger, Wordpress.com/Jetpack
- Facebook photos and videos: Google Photos, Dropbox, Koofr, Photobucket, Backblaze
- Facebook events: Google Calendar

I chose to transfer my Facebook data and media to Google Docs and Google Photos, respectively. The transfer between Facebook and Google was finished within the hour and posts/media from Nov. 2008 to July 23, 2023 were delivered to my Google Drive account, each post in a separate Google doc file, compiled in folders. Media included in posts were in separate folders. The first post, from 2008, appears to be my first post to Facebook, reminding me that Facebook originally began with a prompt on the screen ("What's on Your Mind"), which was replaced by status updates, which most often in my case early on began with "is..."[282]

---

[280] https://help.twitter.com/en/managing-your-account/how-to-download-your-twitter-archive.

[281] Facebook also provides a second option, "Download Your Information," which provides a copy of Facebook data to "keep or transfer to another service." This is a platform download.

[282] David Smith, "Analysis of Facebook Status Updates," *Revolutions* (blog), December 29, 2010, https://blog.revolutionanalytics.com/2010/12/analysis-of-facebook-status-updates.html.

Absent in the transfer file on Google is everything but the time stamp and text of the post:

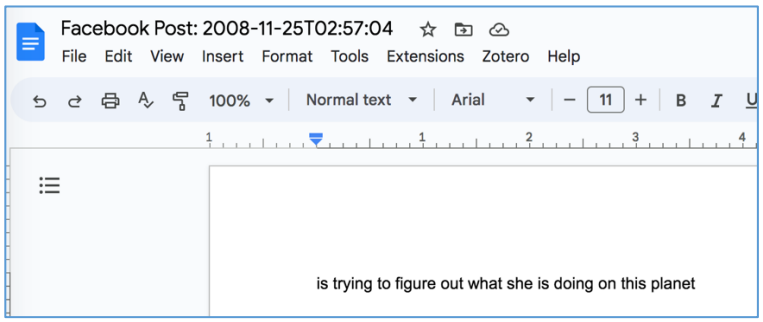

Figure 8: Facebook post as it appears in the Facebook-Google transfer

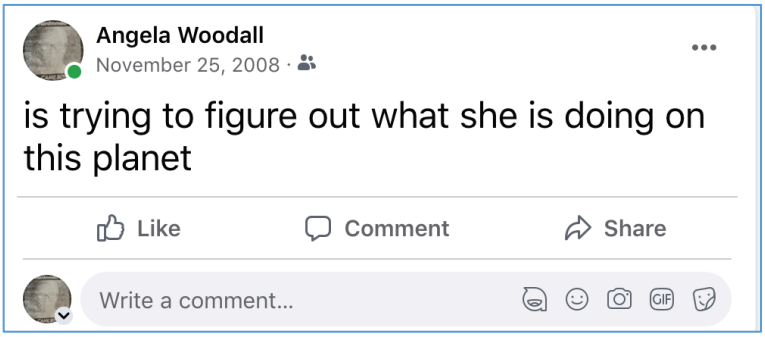

Figure 9 : Facebook post as it appears on Facebook as of July 27, 2023

Some links in original posts are excluded. Entire posts are excluded as well as photos, which are incomplete. Facebook withholds data according to several guidelines.

Subscribers can transfer **c**opies of posts, photos, and videos that they created and can see on their profile, including:

- Status updates
- Posts with a photo, photo album, gif or video that you've uploaded to Facebook
- Posts with a link to another website
- If you posted a video from another website, the video won't transfer. If you included a link to the video in your post, the link will transfer.

There is a long list of content that cannot be transferred that includes (but is not limited to) posts that:

- were posted on a friend's profile.
- you didn't create, even if you are tagged in them.
- you've added to your archive or trash.
- that include a life event.
- that are automatically created when you change your profile picture.

After reviewing the entire list which takes up more than a (printed) page, I was unable to discern which reason applied to the omitted posts. [283]

Threads

Facebook, Instagram, and Threads are owned by Facebook.[284] Threads was released in the United States and Europe in 2022 and requires subscribers to sign up by using their Instagram account username and password. This design links the two applications together and Threads and Instagram share data. Deleting the Threads app will delete subscribers' Instagram account, which could be a form of lock-in.[285]

I was not able to transfer data from any service to Threads because the app is not available for use in the European Union.[286] I could download the Threads application but not open it. Among

[283] For a full list, see: "Transfer Your Information to a Service off of Facebook | Facebook Help Center," accessed July 27, 2023, https://www.facebook.com/help/230304858213063.

[284] Threads is listed under Instagram Inc. on the company website.

[285] Clothilde Goujard, Edith Hancock, and Pieter Haeck, "Why Europeans Don't Have Threads yet," *POLITICO*, July 6, 2023, https://www.politico.eu/article/why-europeans-dont-have-threads-yet-twitter-meta/; Armani Syed, "Why E.U. Users May Not Get to Use Threads," *Time*, July 6, 2023, https://time.com/6292586/privacy-concerns-threads-meta/.

[286] Threads may be available to devices registered to a non-E.U. internet service provider address. The address could be masked with a virtual private network (VPN), which is a service that encrypts the connection of a device

the relevant articles in the DMA, Article 5 obligates gatekeepers to do any of the following: (a) process, for the purpose of providing online advertising services, personal data of end users using services of third parties that make use of core platform services of the gatekeeper; (b) combine personal data from the relevant core platform service with personal data from any further core platform services or from any other services provided by the gatekeeper or with personal data from third-party services; (c) cross-use personal data from the relevant core platform service in other services provided separately by the gatekeeper, including other core platform services, and vice versa; and (d) sign in end users to other services of the gatekeeper in order to combine personal data. There are other provisions that could apply in the case of the Threads-Instagram data sharing.

Mastodon

Mastodon is an independent *social networking text-based service* with microblogging features similar to Twitter. Data can be exported from Mastodon in a CSV format using a dedicated portal on the site available via Settings > Export, which has a pushbutton feature that is easy to find and use.

The CSV file is an "archive" of current followed accounts, currently created lists, currently blocked accounts, currently muted accounts, and currently blocked domains. Posts and media can be downloaded in a format called ActivityPub JSON.

JSON is a commonly used and machine-readable format considered as a best practice option.[287] ActivityPub is a niche standard developed to support independent blogging networks

---

and can make it appear to be in a different location. I have decided not to test this option, although it may be useful to future research.

[287] I have described JSON at length in the glossary and elsewhere.

like Mastodon. This is not a proprietary standard (ActivityPub is open source, meaning that anyone can use and modify the code). But ActivityPub is not standard or commonly-used, and, together, ActivityPub JSON appears to be software limited to ActivityPub-formatted documents.[288] Mastodon currently does not support importing posts or media due to technical limitations that are explained in technical terms that have generated a series of articles online intended to help users understand how to use the site.[289] Twitter data cannot be transferred to Mastodon.

Discussion

Before platform downloads and portals like the one offered by Facebook, users had to email or write to the service providers for their data, which was delivered in a variety of formats, or not at all. The E.U. regulations advance a by-design self-management strategy for portability that individuals can adopt unilaterally with the right techniques and tools, like those designed as part of the Data Transfer Project.[290] Automated data portability tools can be an advantage to users if they make an otherwise potentially difficult process easier and they appear to be necessary given the DMA requirement for real-time and continuous transfers.

Some previously common formats, like PDFs, are obstacles to transferring data to another service. The data is not meant to be downloaded to store or even read through. Portable

[288] ActivityPub is a new format to me and this is my understanding of ActivityPub JSON, which I have not tried to reverse engineer for this study.

[289] Amanda Silberling, "A Beginner's Guide to Mastodon, the Open Source Twitter Alternative," *TechCrunch* (blog), July 24, 2023, https://techcrunch.com/2023/07/24/what-is-mastodon/; Oladipo Tamilore, "A Beginner's Guide to Mastodon," Buffer Resources, November 16, 2022, https://buffer.com/resources/mastodon-social/.

[290] Veale, Binns, and Ausloos, "When Data Protection by Design and Data Subject Rights Clash."

data is meant to be transferred from one system to another, which means the format and structure must be configured for those systems, including APIs. That is why best practices call for JSON, XML, and CSV. They are compatible with most systems.

PDFs do not meet these requirements. Neither does the format and structure of the documents transferred from Facebook to Google. The Facebook transfer to Google provided stand-alone text in individual Google docs. The format does not meet existing standards and would be difficult and time-consuming to consolidate in order to assess whether the data was complete. Alternatively, JSON, XML, and, to a degree CSV, are user-*unfriendly* (JSON and XML especially can seem like a blob of data) and require specific skills that may (or may not) be within reach to many subscribers who want to transfer their data.[291]

Method and Scope

Portability portals, where they exist, appear to be similar to platform downloads in their design and operation: a simple user interface with APIs providing the technical features necessary to transfer data between services. The use of APIs, whose use will become more integrated in order to facilitate portability, calls attention to the need for attention to the scope of the data. APIs as I have discussed in this study are known to provide incomplete results with minimal transparency.[292]

Standards for scope (extent and relevance of the data) have to be more coherent and include consequences for the providers. When, for example, Facebook withholds certain

[291] "Digital Services Act," accessed March 6, 2023, https://digitalservicesact.cc/. Article 44(b).

[292] Amelia Acker, "Social Media Data Preservation in an API-Driven World"; Ho, "How Biased Is the Sample?"; Tromble, Storz, and Stockmann, "We Don't Know What We Don't Know."

categories of data, the user must evaluate whether the criteria is legitimate. Deciding what is legitimate can be difficult, especially when the files are delivered individually, as the Facebook-Google set was. More explanations pointing to relevant criteria and regulation could be helpful or, just as plausible, more confusing. However, the technical feasibility appears to be more important as a feature than the quality of the data provided in the language of the GDPR and DMA.[293]

Discussion

Twitter does not appear to offer a dedicated portability feature and only offers a platform download on the site.[294] As I noted earlier, there is a relationship between the right to data portability and access. Access is a right of individuals to know what personal data that concerns them is being collected, processed, and held. Portability is intended in the DMA to give individuals not only the right to demand their data and receive the data, but to also have it transferred elsewhere, notably between different applications, programs, and services, which is spelled out in Article 6.[295] This leads to the question of whether a platform download intended for one purpose (access) should be repurposed to fill a portability right.

---

[293] Clarence Smith, "Portability and Digital Markets Act," July 25, 2023. I have changed all names of correspondents to pseudonyms for confidentiality. This is a summarized comment by a portability expert who is familiar with the DMA and manages a company that facilitate data portability.

[294] I could find no new mention of data portability on Twitter, now named *X*, as of July 25, 2023. I tried several different searches on Google and the Twitter site. Going through the Help Center to the Privacy Policy and "How Can I Control My Data" I found 5.1 "Access, Correction, Portability," which led to a link to the following: "You can download a copy of your information, such as your Tweets, by following the instructions here" and ultimately to the platform download page "How to Download Your Twitter Archive." https://help.twitter.com/en/managing-your-account/how-to-download-your-twitter-archive

[295] "16 EN/WP 242 Rev. 01."

Nevertheless, the Twitter platform download, while not a best practice, appears to comply with existing regulations.

Twitter may have been engaged in a more questionable tactic. Several accounts I found online reported that Twitter was blocking transfers of subscribers' followers/friends to Mastodon. This was happening to subscribers using dedicated external services that acted on their behalf to make the transfers. The services were prevented from using the Twitter API necessary for the transfers even with the authorization of the user, who initiates the process.[296] The services were detected because they access the subscriber's account by logging in with the subscriber's credentials (email and password). Twitter detects this and blocks the third-party sign-in. Automated portability requests work on the same principle in that a service (i.e. Facebook), must sign into another service, Google. This is a valuable feature for subscriber security raises questions about whether Twitter is exceeding the terms of the DMA regarding business-users. Twitter's blocking log-ins from services that facilitate portability, such as in the case of Twitter and Mastodon, could present an unfair disadvantage for Twitter against not only Mastodon's ability to enter a market where there is a demand by virtue of subscribers, but also those subscribers trying to obtain their data. This scenario is one at the heart of DMA gatekeeper obligations and provisions for portability in Article 6. The end effect is that, from a user perspective, Mastadon and Twitter presented the most confusing experiences with trying to transfer personal data.

[296] Karissa Bell, "Twitter Is Shutting down Its Free API, Here's What's Going to Break," *Engadget*, February 8, 2023, https://www.engadget.com/twitter-new-developer-terms-ban-third-party-clients-211247096.html; Karissa Bell, "Twitter's New Developer Terms Ban Third-Party Clients," *Engadget*, January 19, 2023, https://www.engadget.com/twitter-new-developer-terms-ban-third-party-clients-211247096.html; Huang, "What Is Mastodon and Why Are People Leaving Twitter for It?"

E.U. regulations and the automated systems also help industry by clarifying rules, guidelines, and standards, as well as saving their staff from fulfilling the requests, and training them about how to do so. When I encountered unexpected obstacles transferring my personal data from Facebook, Twitter, and Instagram, as well as difficulty trying to understand the source of the obstacles, I was directed to forms with limited options and so many fields that that I gave up trying out of frustration. The frustration is compounded by flawed how-to explanation pages for completing transfers (or platform downloads) that were out-of-date, unclear, and incomplete in their instructions. The links to making requests must be more prominent on the sites rather than mislabeled and only available through multiple levels in the privacy section of the sites.

The DSA includes the requirement that providers of intermediary services designate a single point of contact for recipients of services, enabling rapid, direct and efficient communication in particular by easily accessible means such as telephone numbers, email addresses, electronic contact forms, chatbots or instant messaging (communications with chatbots must be made apparent). Further, providers of intermediary services should allow recipients of services to choose means of direct and efficient communication which do not solely rely on automated tools. Providers should make all reasonable efforts to guarantee that sufficient human and financial resources are allocated to ensure that this communication is performed in a timely and efficient manner. This would also help in the other instances as well.

The DMA gives regulators tools to take back some of platforms' autonomy over their operations in so far as size, competitiveness, and fairness are concerned. Portability could be useful to strategies for increasing transparency, fairness, and market competition if consumers can automatically and easily transfer their data from one service to another. This assessment shows

that portability is not an effective self-management strategy that individuals can adopt unilaterally.[297] My experience and assessment illustrate how fraught and complex portability is. But the problems were not entirely due to technical limitations. The example of Twitter demonstrated that a new owner can impose significant and sometimes confusing changes in policy and philosophy. Musk blocked services deployed on behalf of subscribers who were seeking to take control of their data. Platform downloads were their only alternative. This method meets the barest of obligations in the DMA – if at all.

The Twitter example also shows the way that APIs can be equally useful for preventing portability as they can for providing the conditions for interoperable systems, thus demonstrating the need for close inspection of APIs. Moreover, the example shows the limitations of relying on by-design, self-service measures, an observation carried over from the previous chapter about platform downloads. Without a holistic view of portability specifically, but by-design approaches more generally, it is unclear how well-served users who want to unlock their data will be. Even murkier is how the opportunity to use portability to make platform operations more transparent and competitive will be successful.

[297] Veale, Binns, and Ausloos, "When Data Protection by Design and Data Subject Rights Clash."

Conclusion

The prominence and importance of digital platforms in multiple domains have mushroomed during the past decade, leading to demanding social, economic, and policy questions. The questions arise from the complicated interactions between the constituents on these platforms – individuals producing data, as well as nations, states, researchers, journalists, advertisers, and the platform companies themselves. In response, these groups have called on governments to intervene with regulation that sets the terms for interactions between platforms and constituents.

The most urgent claims are for transparency, privacy, and access to data and the protection of it. These wants clearly occupy an important role in monitoring and enforcing rules designed to govern entities, public or private, that collect and process our personal information.[298] In the private sector, social media platforms like Twitter and Facebook now occupy an important role directing a complex, automated, and seemingly ubiquitous system for data collecting, processing, and storing.[299] As a result, regulations that respond to these claims are important for enhancing the transparency of these activities.

In this dissertation I looked at one set of constituents at the center of these regulations and the debates that accompany them: the individuals who produce the data which fuels an entire industry, but who have the least control over their it. Their claims to more control over the

[298] Ausloos and Dewitte, "Shattering One-Way Mirrors – Data Subject Access Rights in Practice."

[299] Ausloos and Dewitte.

collection, processing, and sharing of personal data are being managed by both government and industry leaders with data self-management, by-design approaches.

I asked whether the data protection by-design approach will have the desired effects that individuals are seeking and policymakers are pursuing. I approached this question with an empirical assessment of platform downloads and portability tools provided by Facebook, Twitter, and Instagram. They are, or are becoming, the only ways that individuals can download their data or move it. I compared the results to criteria in the GDPR and Digital Markets Act.

I showed that these approaches do not answer the demands for more control over their data. By-design approaches that I tested appear to comply with regulations but they fail to offer individuals with effective resources.[300] Based on the evaluations, I argued that the regulatory and technical resources put in place for individuals to control their data are not effective because they turn over decisions about execution to an industry with no interest in sharing that data or being regulated.

Next Steps

Provisions for access and portability in the DMA are designed to encourage platform users to produce data, which gatekeepers must make available to some degree to both producers and commercial users in order to drive economic benefits. Consequently, data protection is all the more important. However, by-design models cannot assure data protection in the way that regulators seem to believe possible and may undermine their efforts instead.

[300] Valtysson, Jorgensen, and Munkholm, "Co-Constitutive Complexity."

It is true that data protection measures have made it possible to deny data-gathering cookies on websites. So-called consent requests, combined with the obligation of transparency, were instituted to give back control to individuals over the use of their personal data. Websites must offer information that helps visitors reasonably understand why their information is being collected, the purpose of collection and who will access it, and the amount of the information exchanged. Website designers tie this obligation to messages about accepting cookies before being allowed to access the site. They make the process of opting out so tedious that policymakers now speak of consent fatigue: it's easier to accept the data gathering than not. This in turn speaks to the need for an on-the-ground, detailed understanding of how people use social media. But opting out of cookies should not be the end-all of data protection that it has become.[301]

In this respect, I see two problems with the by-design model, one technical and the other social. From the first standpoint, the by-design model relies on platforms' compliance without adequate oversight of the technical systems being implemented. This became clear in the instance of access and portability. The by-design systems like platform downloads themselves need to be tested, well documented, and monitored, which is presently not the case. Enforcement by regulators will also clearly be needed for the implementation of portability and interoperability measures. The example in Chapter 4 demonstrated that these measures will need

[301] Luis Montezuma and Tara Bassirian, "How to Avoid Consent Fatigue," *IAPP: The Privacy Advisor* (blog), January 29, 2019, https://iapp.org/news/a/how-to-avoid-consent-fatigue/; U.K. Information Commissioner's Office, "When Is Consent Appropriate?" (ICO, May 19, 2023), https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/consent/when-is-consent-appropriate/.

to include conditions for access in order to distinguish legitimate limitations (e.g. relating to security and data protection) from anticompetitive strategies.[302]

Self-managed, by-design approaches are not the source of the problem. But they are a symptom of the need for critical scrutiny over the execution of tools like the ones offered by Facebook, Twitter, and Instagram. If policymakers continue to pursue by-design approaches, they will need to learn how to test the techniques, and the execution of the techniques, provided by industry. They will need to assess the impact on data that is made available. So that results can be evaluated, by-design tools like the ones I assessed must be accompanied by clear and detailed details about design choices and procedures.

In this regard, policymakers will have to create effective criteria for measuring the impact of APIs. Instead they are calling for more APIs despite reasons why they should not. These reasons are not a secret. Scholars, archivists, and journalists outside the industry have documented the destabilizing impact of APIs on the quality of their datasets and thus reproducibility, the gold standard for good documentation. Industry insiders who rely on APIs for data have observed the same drawbacks. Both the insiders and outsiders have documented APIs' negative impact on their access to data. This is worth noting because APIs are, or are intended to become, the engines of all data access and portability. The evaluation of platform downloads show why this should be of concern to all stakeholders. Enough research calling attention to the quality and opacity of APIs exist to motivate caution and inquiries when it comes to user rights. The results in this study lay the foundation for developing the standards and criteria for the necessary work.

---

[302] OECD, "Data Portability, Interoperability and Digital Platform Competition," Competition Committee Discussion Paper, 2021, http://oe.cd/dpic.

In addition to the European Data Protection Supervisor, the DMA and DSA include the provision for two new oversight bodies. In the DMA, a high-level group of supervisors is being drawn from various E.U. oversight agencies.[303] In the DSA, a new Digital Services Coordinator has means to investigate and fine companies.[304] Oversight and enforcement will be complicated by the fact that APIs are not well-documented (it was difficult to even determine whether an API was part of the platform download system).

APIs are also sorely in need of standards based on clear criteria. Private industry will continue to fight disclosure, but their refusals offer policymakers the opportunity to show willingness by the E.U. to enforce data protection measures that exceed cookies. Otherwise, policymakers risk undermining the authority of the laws and institutions they represent.

Secondly, the movement toward self-management, by-design measures is a way of automating oversight, which in turn normalizes certain expectations about the way that resources should be allocated and the capacity to fulfill imagined needs based on institutional control and expert systems.[305] Policy statements and industry narratives alike share a vision for a new era of seamless interactions and data sharing taking place across a digital landscape of uninterrupted engagement.[306] Automated by-design models are meant to eliminate friction in order to achieve

---

[303] The group will be composed of 30 representatives nominated from the Body of the European Regulators for Electronic Communications (BEREC), the European Data Protection Supervisor (EDPS) and European Data Protection Board, the European Competition Network (ECN), the Consumer Protection Cooperation Network (CPC Network), and the European Regulatory Group of Audiovisual Media Regulators (ERGA). The High Level Group can provide the Commission with advice and expertise with the aim of ensuring that the DMA and other sectoral regulations applicable to gatekeepers are implemented in a coherent and complementary manner.

[304] "Article 41 - Powers of Digital Services Coordinators," https://digitalservicesact.cc/dsa/art41.html.

[305] Angela Woodall and Sharon Ringel, "Blockchain Archival Discourse: Trust and the Imaginaries of Digital Preservation," *New Media & Society*, November 22, 2019, https://doi.org/10.1177/1461444819888756.

[306] "A European Strategy for Data: Shaping Europe's Digital Future," May 2, 2023, https://digital-strategy.ec.europa.eu/en/policies/strategy-data; Meta Utopia, "Effortless Communication — Stay Connected with My World," *Medium* (blog), October 5, 2023, https://medium.com/@MW_MetaUtopia/effortless-communication-stay-connected-with-my-world-a74899de635a.

these imagined scenarios. As I detailed in Chapter 3, Twitter provides seamless self-service portals for communication, but the company uses them like a one-way mirror. Twitter gave me no way to question, let alone contest, their access policies and instruments. My experience with their one-way mirrors was shared by the U.S. Department of Justice, whose lawyers (seeking tweets from the account of former president Donald Trump) were unable to reach Twitter when forced to use the same online self-help portal. Yet the portal is exactly the kind of tool that fits with the self-management, by-design model.

This study showed that regulation can be effective but not always in the way that we might expect. This happened in the case of Facebook and the complicated data sharing between Instagram and the new social networking app Threads. Each of the three services is owned by Mark Zuckerberg, who delayed the release of the Threads app in Europe. He did this in order to assess whether the data sharing between Threads and Instagram violated E.U. restrictions clearly outlined in Article 5 of the DMA. The app is available in the United States but not to residents of Europe, where I am. This meant could I not access Threads. It also mean that I could not transfer my data to from Twitter to Threads. The example demonstrated the effect that E.U. regulations can have, which I would have overlooked otherwise. The example also, as I said in the introduction, illustrated how fraught portability and access are legal and technically and that policy does not always have effects intended by policymakers.

This is a reminder of the critical need for empirical research. But the takeaway from this double-sided story is this: Even as new offices are being established, the automated self-led, by-design model risks crowding out the essential expertise necessary to realize the stated goals of regulations like the DMA, creating less oversight and setting terms for user participation according to fixed technical protocols of systems like APIs. The history of data protection

demonstrates that friction is necessary for adequate oversight. That friction has come from national data authorities, who have defied not only the industry leaders but also policymakers by demanding concrete data protection actions on behalf of citizens. Without them, we will be left with systems that provide seamless interaction, but with little recourse when we encounter failure. The risk is that by replacing friction with a version of automated authority handed down to us from platforms and policymakers, we lose our most important allies. That story has yet to be told.

Bibliography

"A European Strategy for Data." Communication From The Commission To The European Parliament, The Council, The European Economic And Social Committee And The Committee Of The Regions. Brussels, Belgium: European Commission, February 19, 2020. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020DC0066.

"A European Strategy for Data: Shaping Europe's Digital Future," May 2, 2023. https://digital-strategy.ec.europa.eu/en/policies/strategy-data.

"Accessing & Downloading Your Information | Facebook Help Center." Accessed June 11, 2020. https://www.facebook.com/help/1701730696756992/?helpref=hc_fnav.

Acker, Amelia. "Social Media Data Preservation in an API-Driven World." Presented at the Society of American Archivists, 2019.

Acker, Amelia, and Adam Kriesberg. "Tweets May Be Archived: Civic Engagement, Digital Preservation and Obama White House Social Media Data." *Proceedings of the Association for Information Science and Technology* 54, no. 1 (2017): 1–9. https://doi.org/10.1002/pra2.2017.14505401001.

Allen-Robertson, James. "Critically Assessing Digital Documents: Materiality and the Interpretative Role of Software." *Information, Communication & Society* 0, no. 0 (July 17, 2017): 1–15. https://doi.org/10.1080/1369118X.2017.1351575.

AltexSoft. "REST API: Key Concepts, Best Practices, and Benefits." Accessed June 8, 2023. https://www.altexsoft.com/blog/rest-api-design/.

Amazon Web Services, Inc. "What Is an API? - Application Programming Interfaces Explained - AWS." Accessed June 15, 2023. https://aws.amazon.com/what-is/api/.

Ankerson, Megan Sapnar. *Dot-Com Design: The Rise of a Usable, Social, Commercial Web*. New York: NYU Press, 2018.

Apify Blog. "Facebook Data Mining with Web Scraping," March 17, 2023. https://blog.apify.com/facebook-data-mining-with-web-scraping/.

App Architecture. "What Is Interoperability? | Definition from TechTarget." Accessed May 8, 2023. https://www.techtarget.com/searchapparchitecture/definition/interoperability.

Appadurai, Arjun, and Neta Alexander. *Failure*. 1 edition. Cambridge, UK ; Medford: Polity, 2019.

Arellano, Kelly. "SOAP vs REST (vs JSON): Web API Services [2021] | RapidAPI." Rapid Blog, February 13, 2019. https://rapidapi.com/blog/soap-vs-rest-api/.

"Article 29 Data Protection Working Party." European Data Protection Board, April 5, 2017. https://edpb.europa.eu/about-edpb/more-about-edpb/article-29-working-party_en.

Article 29 Working Party. "Working Document: Transfers of Personal Data to Third Countries: Applying Articles 25 and 26 of the EU Data Protection Directive." European Commission, n.d. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/index_en.htm#maincontentSec20.

Aufderheide, Patricia A. *Communications Policy and the Public Interest: The Telecommunications Act of 1996*. 1 edition. New York: The Guilford Press, 1999.

Ausloos, Jef. “Chapter 2: Foundations of Data Protection Law.” In *The Right to Erasure in EU Data Protection Law*, edited by Jef Ausloos, 0. Oxford ; Cambridge, USA: Oxford University Press, 2020. https://doi.org/10.1093/oso/9780198847977.003.0002.

Ausloos, Jef, and Pierre Dewitte. “Shattering One-Way Mirrors – Data Subject Access Rights in Practice.” *International Data Privacy Law* 8, no. 1 (February 1, 2018): 4–28. https://doi.org/10.1093/idpl/ipy001.

Ausloos, Jef, Rene Mahieu, and Michael Veale. “Getting Data Subject Rights Right.” SSRN Scholarly Paper. Rochester, NY, December 1, 2019. https://papers.ssrn.com/abstract=3544173.

Bartelheimer, Christian, Philipp zur Heiden, Hedda Lüttenberg, and Daniel Beverungen. “Systematizing the Lexicon of Platforms in Information Systems: A Data-Driven Study.” *Electronic Markets* 32, no. 1 (March 1, 2022): 375–96. https://doi.org/10.1007/s12525-022-00530-6.

Bell, Karissa. “Twitter Is Shutting down Its Free API, Here’s What’s Going to Break.” *Engadget*, February 8, 2023. https://www.engadget.com/twitter-new-developer-terms-ban-third-party-clients-211247096.html.

———. “Twitter’s New Developer Terms Ban Third-Party Clients.” *Engadget*, January 19, 2023. https://www.engadget.com/twitter-new-developer-terms-ban-third-party-clients-211247096.html.

Berners-Lee, Tim. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web*. 1 edition. San Francisco: HarperBusiness, 2000.

BIG BROTHER WATCH TEAM. “How Can I Download a Copy of My Facebook Data? What Is Included – and What Isn’t? — Big Brother Watch.” How can I download a copy of my Facebook data? What is included – and what isn’t? — Big Brother Watch, March 23, 2018. https://bigbrotherwatch.org.uk/2018/03/how-can-i-download-a-copy-of-my-facebook-data-what-is-included-and-what-isnt/.

Bjørn, Pernille, Maria Menendez-Blanco, and Valeria Borsotti. *Diversity in Computer Science: Design Artefacts for Equity and Inclusion*. Electronic resource. 1st ed. 2023. Cham: Springer International Publishing : Imprint: Springer, 2023. https://doi.org/10.1007/978-3-031-13314-5.

Bloch-Wehba, Hannah. “Global Platform Governance: Private Power in the Shadow of the State.” *SMU Law Review* 72, no. 1 (January 1, 2019): 27.

Bradford, Anu. *The Brussels Effect: How the European Union Rules the World*. New York: Oxford University Press, 2020.

Braun, Joshua. “Transparent Intermediaries: Building the Infrastructures of Connected Viewing.” In *Connected Viewing*. Routledge, 2013.

Braunstein, Mark. “Healthcare in the Age of Interoperability: Part 2.” *IEEE Pulse* (blog), November 15, 2018. https://www.embs.org/pulse/articles/healthcare-in-the-age-of-interoperability-part-2/.

BreakingNews.ie. “Moderators ‘not Denied Access’ to Non-Disclosure Agreements, Facebook Insists.” May 20, 2021, sec. Ireland. https://www.proquest.com/docview/2529404998/citation/53CC68CCD39B4920PQ/1.

Brian, Marshall. “How Web Pages Work.” HowStuffWorks, September 5, 2000. https://computer.howstuffworks.com/web-page.htm.

———. “How Web Servers Work.” HowStuffWorks, April 1, 2000. https://computer.howstuffworks.com/web-server.htm.

Brill, Julie. “Putting People First on Data Portability.” *Regulatory Affairs at Microsoft* (blog), May 21, 2018. https://blog.twitter.com/en_us/topics/company/2018/putting-people-first-on-data-portability.

Broussard, Meredith. *Artificial Unintelligence: How Computers Misunderstand the World*. Illustrated edition. Cambridge, Massachusetts London, England: The MIT Press, 2019.

Brubaker, Jed R., and Vanessa Callison-Burch. “Legacy Contact: Designing and Implementing Post-Mortem Stewardship at Facebook.” In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2908–19. San Jose California USA: ACM, 2016. https://doi.org/10.1145/2858036.2858254.

Brunelle, Justin F., Mat Kelly, Michele C. Weigle, and Michael L. Nelson. “The Impact of JavaScript on Archivability.” *International Journal on Digital Libraries; Heidelberg* 17, no. 2 (June 2016): 95–117. http://dx.doi.org.ezproxy.cul.columbia.edu/10.1007/s00799-015-0140-8.

Bruns, Axel. “After the ‘APIcalypse’: Social Media Platforms and Their Fight against Critical Scholarly Research.” *Information, Communication & Society* 22, no. 11 (September 19, 2019): 1544–66. https://doi.org/10.1080/1369118X.2019.1637447.

Bruns, Axel, and Jean Burgess. “The Use of Twitter Hashtags in the Formation of Ad Hoc Publics.” In *Proceedings of the 6th European Consortium for Political Research (ECPR) General Conference 2011*, edited by A. Bruns and P. De Wilde, 1–9. United Kingdom: The European Consortium for Political Research (ECPR), 2011. https://eprints.qut.edu.au/46515/.

Burgess, Jean, and Axel Bruns. “Twitter Archives and the Challenges of ‘Big Social Data’ for Media and Communication Research.” *M/C Journal* 15, no. 5 (October 11, 2012). http://journal.media-culture.org.au/index.php/mcjournal/article/view/561.

Bygrave, Lee A. “Data Protection by Design and by Default: Deciphering the EU’s Legislative Requirements.” SSRN Scholarly Paper. Rochester, NY, June 20, 2017. https://papers.ssrn.com/abstract=3035164.

———. “Security by Design: Aspirations and Realities in a Regulatory Context.” SSRN Scholarly Paper. Rochester, NY, May 23, 2022. https://papers.ssrn.com/abstract=4117110.

Cadwalladr, Carole, and Emma Graham-Harrison. “Revealed: 50 Million Facebook Profiles Harvested for Cambridge Analytica in Major Data Breach.” *The Guardian*, March 17, 2018, sec. News. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election.

Cate, Fred H. *Privacy in the Information Age*. Washington, D.C.: Brookings Institution Press, 1997.

Cavoukian, Ann. “Privacy by Design: The 7 Foundational Principles Implementation and Mapping of Fair Information Practices.” Information and Privacy Commissioner of Ontario, May 2010. www.ipc.on.ca.

Clinton, William, and Albert Gore Jr. “A Framework For Global Electronic Commerce,” July 1, 1997. https://clintonwhitehouse4.archives.gov/WH/EOP/OSTP/forum/html/giipaper.html.

Connelly, Matthew, and Rohan Shah. “Here’s What Data Science Tells Us about Hillary Clinton’s Emails.” *Washington Post*, November 2, 2016. https://www.washingtonpost.com/news/monkey-cage/wp/2016/11/02/heres-what-data-science-tells-us-about-hillary-clintons-emails/.

Connor, Michael. “Ethics and Archiving the Web.” Presented at the Rhizome National Forum on Ethics and Archiving the WEb, New Museum, New York, June 27, 2018. https://vimeo.com/277335998.
Constine, Josh. “Instagram Launches ‘Data Download’ Tool to Let You Leave.” *TechCrunch*, April 24, 2018. https://social.techcrunch.com/2018/04/24/instagram-export/.
Cookiebot. “CCPA vs GDPR,” November 30, 2020. https://www.cookiebot.com/en/ccpa-vs-gdpr-compliance-with-cookiebot-cmp/.
Costa, Miguel, Daniel Gomes, and Mário J. Silva. “The Evolution of Web Archiving.” *Int. J. Digit. Libr.* 18, no. 3 (September 2017): 191–205. https://doi.org/10.1007/s00799-016-0171-9.
Crawford, Stephanie. “How HTML5 Works.” HowStuffWorks, August 24, 2011. https://computer.howstuffworks.com/html-five.htm.
Custers, Bart. “Profiling As Inferred Data. Amplifier Effects and Positive Feedback Loops.” SSRN Scholarly Paper. Rochester, NY, October 9, 2018. https://doi.org/10.2139/ssrn.3466857.
Data Transfer Initiative. “The Future of Data Portability and Law: Unpublished Concept Note.” Chris Riley, 2023.
“Decoding GDPR: Familiar Terms Could Cause Major Confusion When GDPR Takes Effect | Judicature,” October 22, 2019. https://judicature.duke.edu/articles/decoding-gdpr-familiar-terms-could-cause-major-confusion-when-gdpr-takes-effect/.
“Definition of Download.” In *Merriam-Webster*, July 31, 2023. https://www.merriam-webster.com/dictionary/download.
“Department of Home Security: The Fair Information Practice Principles Factsheet,” 2008. https://www.dhs.gov/publication/privacy-policy-guidance-memorandum-2008-01-fair-information-practice-principles.
Diakopoulos, Nicholas. “Algorithmic Accountability: On the Investigation of Black Boxes.” New York: Tow Center for Digital Journalism, Columbia University, December 3, 2014. https://www.cjr.org/tow_center_reports/algorithmic_accountability_on_the_investigation_of_black_boxes.php/.
Diakopoulos, Nick. “Algorithmic Accountability.” *Digital Journalism* 3, no. 3 (2015): 398–415.
Diaz, Claudia, Omer Tene, and Seda F. Guerses. “Hero or Villain: The Data Controller in Privacy Law and Technologies.” SSRN Scholarly Paper. Rochester, NY, September 5, 2013. https://papers.ssrn.com/abstract=2321480.
Diaz, Gerardo Con. “The Text in the Machine: American Copyright Law and the Many Natures of Software, 1974–1978.” *Technology and Culture* 57, no. 4 (2016): 753–79.
“Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms - Nicolas Suzor, 2018.” Accessed February 11, 2022. https://journals.sagepub.com/doi/10.1177/2056305118787812.
“Digital Services Act.” Accessed March 6, 2023. https://digitalservicesact.cc/.
Dijck, José van. *The Culture of Connectivity: A Critical History of Social Media*. Oxford ; New York: Oxford University Press, 2013.
Dowd, Rebekah. “Digitized Data as a Political Object.” In *The Birth of Digital Human Rights: Digitized Data Governance as a Human Rights Issue in the EU*, 3–25. Information Technology and Global Governance. Cham: Springer International Publishing, 2022. https://doi.org/10.1007/978-3-030-82969-8_1.

"Download a Copy of Your Information on Facebook | Facebook Help Center," n.d. https://www.facebook.com/help/212802592074644.

Driscoll, Kevin, and Shawn Walker. "Big Data, Big Questions| Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data." *International Journal of Communication* 8, no. 0 (June 16, 2014): 20.

Electronic Frontier Foundation. "Section 230." *EFF Issues* (blog), n.d. https://www.eff.org/issues/cda230.

Emerson, Kent. "An Introduction to Web Scraping for Research." Research Data Services, November 7, 2019. https://researchdata.wisc.edu/news/an-introduction-to-web-scraping-for-research/.

Englehardt, Steven. "No Boundaries: Exfiltration of Personal Data by Session-Replay Scripts." *Freedom to Tinker* (blog), November 15, 2017. https://freedom-to-tinker.com/2017/11/15/no-boundaries-exfiltration-of-personal-data-by-session-replay-scripts/.

Espenschied, Dragan. "Rhizome Releases First Public Version of Webrecorder." *Rhizome.Org* (blog), August 9, 2016. http://rhizome.org/editorial/2016/aug/09/rhizome-releases-first-public-version-of-webrecorder/.

Ethics and Archiving the Web: Stewardship and Usage. New York, New York, 2018. https://vimeo.com/276935105.

EUR-Lex. "Summary of: Treaty on the Functioning of the European Union EUR-Lex - 4301854 - EN," December 15, 2017. https://eur-lex.europa.eu/EN/legal-content/summary/treaty-on-the-functioning-of-the-european-union.html.

European Data Protection Supervisor. "Interoperability." Accessed June 24, 2023. https://edps.europa.eu/data-protection/our-work/subjects/interoperability_en.

European Parliament, Council of the European Union. "Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 Amending Directive 2003/98/EC on the Re-Use of Public Sector Information Text with EEA Relevance," June 26, 2013. http://data.europa.eu/eli/dir/2013/37/oj/eng.

European Union. "Types of Legislation in the European Union." Accessed May 6, 2023. https://european-union.europa.eu/institutions-law-budget/law/types-legislation_en.

European Union Agency for Fundamental Rights. "Article 8 - Protection of Personal Data," April 25, 2015. http://fra.europa.eu/en/eu-charter/article/8-protection-personal-data.

Evans, Sian, Anna Perricci, and Amy Roberts. "'Why Archive?' And Other Important Questions Asked by Occupiers." In *Informed Agitation: Library and Information Skills in Social Justice Movements and Beyond*, edited by Melissa Morrone. Library Juice Press, 2014.

Fanimokun, Abiola O., Gary Castrogiovanni, and Mark F. Peterson. "Developing High-Tech Ventures: Entrepreneurs, Advisors, and the Use of Non-Disclosure Agreements (NDAs)." *Journal of Small Business and Entrepreneurship* 25, no. 1 (2012): 103-119,127-128.

Federal Trade Commission. "Protecting Consumer Privacy in an Era of Rapid Change: Recommendations For Businesses and Policymakers." Federal Trade Commission, March 1, 2012. https://www.ftc.gov/reports/protecting-consumer-privacy-era-rapid-change-recommendations-businesses-policymakers.

Ferrari, Elena. "Access Control." In *Encyclopedia of Database Systems*, edited by Ling Liu and M. Tamer Özsu, 9–14. New York, NY: Springer, 2018. https://doi.org/10.1007/978-1-4614-8265-9_6.

"File Formats for Long-Term Access | Data Management." Accessed June 24, 2023. https://libraries.mit.edu/data-management/store/formats/.

Flew, Terry, and Fiona R. Martin. *Digital Platform Regulation: Global Perspectives on Internet Governance*. 1st ed. 2022 édition. Cham, Switzerland: Springer Nature Switzerland AG, 2022.

"Forward." In *Records, Computers and the Rights of Citizens*, 5–7. OHEW Publication, (OS)73-94. Washington, D.C.: U.S. Department of Health, Education & Welfare, 1973. https://aspe.hhs.gov/reports/records-computers-rights-citizens.

Fromholz, Julia M. "The European Union Data Privacy Directive." *Berkeley Technology Law Journal* 15, no. 1 (2000): 461–84.

Fuster, Gloria González. *The Emergence of Personal Data Protection as a Fundamental Right of the EU*. Softcover reprint of the original 1st ed. 2014 édition. Springer, 2016.

Garrett, Jesse James. "Ajax: A New Approach to Web Applications." *Adaptive Path: Ideas* (blog), February 18, 2005. http://www.adaptivepath.com/ideas/essays/archives/000385.php.

Gellman, Robert. "Review of 'Data Privacy Law: A Study of United States Data Protection.' By Paul M. Schwartz and Joel R. Reidenberg. Charlottesville, VA: Michie. 1996. 486 Pages. ISRN 1558343776." *Government Information Quarterly* 14, no. 2 (1997): 215–17.

Gemmell, Jim, Gordon Bell, and Roger Lueder. "MyLifeBits: A Personal Database for Everything." *Communications of the ACM* 49, no. 1 (January 1, 2006): 88–95. https://doi.org/10.1145/1107458.1107460.

General Data Protection Regulation (GDPR). "Art. 4 GDPR – Definitions." Accessed July 31, 2023. https://gdpr-info.eu/art-4-gdpr/.

General Data Protection Regulation (GDPR). "Privacy by Design." Accessed May 6, 2023. https://gdpr-info.eu/issues/privacy-by-design/.

Giannopoulou, A., J. Ausloos, S. Delacroix, and H. Janssen. "Intermediating Data Rights Exercises: The Role of Legal Mandates." *International Data Privacy Law* 12 (November 2022). https://doi.org/10.1093/idpl/ipac017.

Gieseke, Jens. *The History of the Stasi: East Germany's Secret Police, 1945-1990*. 1st edition. Berghahn Books, 2015.

Gill, Daniel, and Jakob Metzger. "Data Access through Data Portability – Economic and Legal Analysis of the Applicability of Art. 20 GDPR to the Data Access Problem in the Ecosystem of Connected Cars." SSRN Scholarly Paper. Rochester, NY, May 5, 2022. https://papers.ssrn.com/abstract=4107677.

Gillespie, Tarleton. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven: Yale University Press, 2018.

Glossary. "European Data Protection Supervisor," June 20, 2023. https://edps.europa.eu/data-protection/data-protection/glossary/d_en.

González-Bailón, Sandra, Rafael E. Banchs, and Andreas Kaltenbrunner. "Emotions, Public Opinion, and U.S. Presidential Approval Rates: A 5-Year Analysis of Online Political Discussions." *Human Communication Research* 38, no. 2 (April 1, 2012): 121–43. https://doi.org/10.1111/j.1468-2958.2011.01423.x.

Google Open Source Blog. "Introducing Data Transfer Project: An Open Source Platform Promoting Universal Data Portability," July 20, 2018. https://opensource.googleblog.com/2018/07/introducing-data-transfer-project.html.

Goujard, Clothilde, Edith Hancock, and Pieter Haeck. “Why Europeans Don’t Have Threads yet.” *POLITICO*, July 6, 2023. https://www.politico.eu/article/why-europeans-dont-have-threads-yet-twitter-meta/.

Graham, Jefferson. “Download Your Facebook Data: How to Do It and What You Might Find.” *USA Today*, March 30, 2018. https://www.usatoday.com/story/tech/talkingtech/2018/03/30/downloaded-all-my-facebook-data-what-learned/471787002/.

Greenstein, Shane. *How the Internet Became Commercial: Innovation, Privatization, and the Birth of a New Network*. Princeton: Princeton University Press, 2015.

“Guidelines 01/2022 on Data Subject Rights: Right of Access.” European Data Protection Board, January 18, 2022.

Gutwirth, Serge, Yves Poullet, Paul de Hert, Cécile de Terwangne, and Sjaak Nouwt. *Reinventing Data Protection?* Softcover reprint of hardcover 1st ed. 2009 édition. Springer, 2010.

Hallin, Daniel C., and Paolo Mancini. *Comparing Media Systems: Three Models of Media and Politics*. Cambridge ; New York: Cambridge University Press, 2000.

Hanus, Jerome J., and Harold C. Relyea. “A Policy Assessment of the Privacy Act of 1974.” *American University Law Review* 25, no. 3 (1976 1975): 555–94.

Helmond, Anne. “The Platformization of the Web: Making Web Data Platform Ready:” *Social Media + Society*, September 30, 2015. https://doi.org/10.1177/2056305115603080.

Helmond, Anne, David B. Nieborg, and Fernando N. van der Vlist. “Facebook’s Evolution: Development of a Platform-as-Infrastructure.” *Internet Histories* 3, no. 2 (April 3, 2019): 123–46. https://doi.org/10.1080/24701475.2019.1593667.

Hildebrandt, Mireille, and Laura Tielemans. “Data Protection by Design and Technology Neutral Law.” *Computer Law & Security Review* 29, no. 5 (October 1, 2013): 509–21. https://doi.org/10.1016/j.clsr.2013.07.004.

Ho, Justin Chun-Ting. “How Biased Is the Sample? Reverse Engineering the Ranking Algorithm of Facebook’s Graph Application Programming Interface.” *Big Data & Society* 7, no. 1 (January 2020). https://doi.org/10.1177/2053951720905874.

Honey, Kristen, Phaedra Chrousos, and Tom Black. “My Data: Empowering All Americans with Personal Data Access.” whitehouse.gov, March 15, 2016. https://obamawhitehouse.archives.gov/blog/2016/03/15/my-data-empowering-all-americans-personal-data-access.

“How to Access and Download Your Twitter Data | Twitter Help,” n.d. https://help.twitter.com/en/managing-your-account/accessing-your-twitter-data.

“How to Download Your Twitter Archive.” Accessed June 12, 2020. https://help.twitter.com/en/managing-your-account/how-to-download-your-twitter-archive.

“How to Download Your Twitter Archive and Tweets | Twitter Help,” n.d. https://help.twitter.com/en/managing-your-account/how-to-download-your-twitter-archive.

Huang, Kalley. “What Is Mastodon and Why Are People Leaving Twitter for It?” *The New York Times*, November 7, 2022, sec. Technology. https://www.nytimes.com/2022/11/07/technology/mastodon-twitter-elon-musk.html.

IEEE Xplore. “Terminology: API.” Accessed July 5, 2023. https://developer.ieee.org/Terminology.

Igo, Sarah E. “Codes of Confidentiality and Consent.” In *The Known Citizen: A History of Privacy in Modern America*. Cambridge, Massachusetts: Harvard University Press, 2018.
———. *The Known Citizen: A History of Privacy in Modern America*. Cambridge, Massachusetts: Harvard University Press, 2018.
John, Nicholas A., and Asaf Nissenbaum. “An Agnotological Analysis of APIs: Or, Disconnectivity and the Ideological Limits of Our Knowledge of Social Media.” *The Information Society* 35, no. 1 (January 1, 2019): 1–12. https://doi.org/10.1080/01972243.2018.1542647.
“JSON vs XML | AppMaster.” Accessed July 3, 2023. https://appmaster.io/blog/json-vs-xml.
Kaiser, Gabriele, and Pat Rogers, eds. *Equity in Mathematics Education: Influences of Feminism and Culture*. London ; Washington, D.C: Falmer Press, 1995.
Kennedy, Helen. “How People Feel about What Companies Do with Their Data Is Just as Important as What They Know about It.” London School of Economics. *Impact of Social Sciences* (blog), March 29, 2018. https://blogs.lse.ac.uk/impactofsocialsciences/2018/03/29/how-people-feel-about-what-companies-do-with-their-data-is-just-as-important-as-what-they-know-about-it/.
Kieran, Damien. “Putting People First on Data Portability,” July 20, 2018. https://blog.twitter.com/en_us/topics/company/2018/putting-people-first-on-data-portability.
Kim, Nancy. *Wrap Contracts: Foundations and Ramifications*. *Wrap Contracts*. Oxford University Press, 2013. https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199336975.001.0001/acprof-9780199336975.
Kin, Lane. “ProgrammableWeb Is Shutting Down.” API Evangelist, October 15, 2022. https://apievangelist.com/2022/10/15/programmableweb-is-shutting-down/.
King, Gary, Jennifer Pan, and Margaret E. Roberts. “Reverse-Engineering Censorship in China: Randomized Experimentation and Participant Observation.” *Science* 345, no. 6199 (2014): 1–10.
Kitchin, Rob. “Thinking Critically about and Researching Algorithms,” 2017. https://doi.org/10.1080/1369118X.2016.1154087.
Korniienko, Petro S., Oleh V. Plakhotnik, Hanna O. Blinova, Zhanna O. Dzeiko, and Gennadii O. Dubov. “Contemporary Challenges and the Rule of Law in the Digital Age.” *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique* 36, no. 2 (April 1, 2023): 991–1006. https://doi.org/10.1007/s11196-022-09963-w.
Krasnoff, Barbara. “How to Download an Archive of Your Twitter Data.” The Verge, November 11, 2022. https://www.theverge.com/23453703/twitter-archive-download-how-to-tweets.
Kritika, Garg. “Twitter Was Already Difficult To Archive, Now It’s Worse!” *Old Dominion University Web Science and Digital Libraries Research Group* (blog), July 15, 2020. https://ws-dl.blogspot.com/2020/07/2020-07-15-twitter-was-already.html.
Lahey, Michael. “Invisible Actors: Web Application Programming Interfaces, Television, and Social Media.” *Convergence* 22, no. 4 (August 1, 2016): 426–39. https://doi.org/10.1177/1354856516641915.
Lakshmi. “SOAP vs REST vs JSON - Differences and How They Are Useful,” September 2, 2021. https://seagence.com/blog/soap-vs-rest-vs-json/.

Law, Practical. “House Antitrust Subcommittee Unveils Five Big Tech Antitrust Bills.” *Reuters*, June 14, 2021, sec. Legal Industry. https://www.reuters.com/legal/legalindustry/house-antitrust-subcommittee-unveils-five-big-tech-antitrust-bills-2021-06-14/.

Lay Jr., James. “Note by the Executive Secretary to the National Security Council on United States Policy toward the Soviet Satellites in Eastern Europe.” Draft policy statement. Washington, D.C, December 11, 1953. https://digitalarchive.wilsoncenter.org/document/national-security-council-nsc-174-draft-united-states-policy-toward-soviet-satellites.

Learning Center. “What Is Rate Limiting | Types & Algorithms | Imperva.” Accessed June 15, 2023. https://www.imperva.com/learn/application-security/rate-limiting/.

Leistner, Matthias. “The Commission’s Vision for Europe’s Digital Future: Proposals for the Data Governance Act, the Digital Markets Act and the Digital Services Act—a Critical Primer.” *Journal of Intellectual Property Law & Practice* 16, no. 8 (August 1, 2021): 778–84. https://doi.org/10.1093/jiplp/jpab054.

Lichter, Andreas, Max Löffler, and Sebastian Siegloch. “The Long-Term Costs of Government Surveillance: Insights from Stasi Spying in East Germany.” *Journal of the European Economic Association* 19, no. 2 (April 1, 2021): 741–89. https://doi.org/10.1093/jeea/jvaa009.

Light, Ben, Jean Burgess, and Stefanie Duguay. “The Walkthrough Method: An Approach to the Study of Apps:” *New Media & Society*, November 11, 2016. https://doi.org/10.1177/1461444816675438.

Littman, Justin. “Twitter’s Developer Policies for Researchers, Archivists, and Librarians.” *Medium* (blog), January 7, 2019. https://medium.com/on-archivy/twitters-developer-policies-for-researchers-archivists-and-librarians-63e9ba0433b2.

Lloyd, Ian J. *Information Technology Law*. Oxford [England] ; New York : Oxford University Press, 2008. http://archive.org/details/informationtechn0000lloy_e3i8.

Lomborg, Stine. “Researching Communicative Practice: Web Archiving in Qualitative Social Media Research.” *Journal of Technology in Human Services* 30, no. 3–4 (July 1, 2012): 219–31. https://doi.org/10.1080/15228835.2012.744719.

Lomborg, Stine, and Anja Bechmann. “Using APIs for Data Collection on Social Media.” *The Information Society* 30, no. 4 (August 8, 2014): 256–65. https://doi.org/10.1080/01972243.2014.915276.

Maemura, Emily, Nicholas Worby, Ian Milligan, and Christoph Becker. “If These Crawls Could Talk: Studying and Documenting Web Archives Provenance.” *Journal of the Association for Information Science & Technology* 69, no. 10 (October 2018): 1223–33. https://doi.org/10.1002/asi.24048.

Mahieu, René. “The Right of Access to Personal Data: A Genealogy.” *Technology and Regulation* 2021 (August 20, 2021): 62–75. https://doi.org/10.26116/techreg.2021.005.

Marshall, Catherine C., and Frank M. Shipman. “On the Institutional Archiving of Social Media.” In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, 1–10. JCDL ’12. New York, NY, USA: ACM, 2012. https://doi.org/10.1145/2232817.2232819.

Masanès, Julien, ed. *Web Archiving*. 2006 edition. Berlin ; New York: Springer, 2006.

McKeehan, Morgan. “Symmetrical Web Archiving with Webrecorder, a Browser-Based Tool for Digital Social Memory. An Interview with Ilya Kreymer | NDSR – NY.” *National Digital Stewardship Residency (NDSR)* (blog), February 23, 2016.

https://ndsr.nycdigital.org/symmetrical-web-archiving-with-webrecorder-a-browser-based-tool-for-digital-social-memory-an-interview-with-ilya-kreymer/.
"Mellon Foundation." Accessed July 2, 2023. https://www.mellon.org/grant-details/webrecorder-a-web-archiving-tool:-phase-two-20444133.
Metz, Douglas. "Proposed Substitute for Senator Ervin's Privacy Protection Commission in S. 3418; Possible Privacy Commission Compromise." Memorandum Of Information For The File. Washington, D.C.: Domestic Council Committee On The Right Of Privacy, December 4, 1974. Box 56, folder "Privacy - Commission" of the Philip Buchen Files at the Gerald R. Ford Presidential Library.
Milligan, Ian, Nick Ruest, and Jimmy Lin. "Content Selection and Curation for Web Archiving: The Gatekeepers vs. The Masses." In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, 107–10. JCDL '16. New York, NY, USA: ACM, 2016. https://doi.org/10.1145/2910896.2910913.
Murgia, Madhumita. "A Tale of Two Facebook Whistleblowers." *FT.Com*, June 22, 2022. https://www.proquest.com/docview/2692576026/citation/7E99496D23A04DE7PQ/1.
Murphy, Hannah, and Kiran Stacey. "Facebook Libra: The inside Story of How the Company's Cryptocurrency Dream Died." *FT.Com*, March 10, 2022. https://www.proquest.com/docview/2648412117/citation/EEC3DEE5A6F64CE3PQ/1.
Musser, John, and Adam DuVander. "How Facebook Makes It Nearly Impossible For You To Quit." ProgrammableWeb: API University, n.d. Accessed August 6, 2020.
Nash, Michael, ed. *How to Keep Union Records*. Chicago: Society of American Archivists, 2010.
National Academy of Sciences. "The Global Information Infrastructure: A White Paper Prepared for the White House Forum on the Role of Science and Technology in Promoting National Security and Global Stability." National Academy of Sciences, March 29, 1995. https://clintonwhitehouse4.archives.gov/WH/New/Commerce/read.html.
Naudts, Laurens, Pierre Dewitte, and Jef Ausloos. "Meaningful Transparency through Data Rights: A Multidimensional Analysis." In *Research Handbook on EU Data Protection Law*, 530–71. Edward Elgar Publishing, 2022. https://www.elgaronline.com/display/edcoll/9781800371675/9781800371675.00030.xml.
Network Centric Operations Industry Council Interoperability - NCOIC. "What Is Interoperability?," July 14, 2014. https://web.archive.org/web/20140714143139/http://www.ncoic.org/what-is-interoperability.
Newman, Abraham L. *Protectors of Privacy: Regulating Personal Data in the Global Economy*. Illustrated edition. Ithaca: Cornell University Press, 2008.
Nguyen, Anh. "Transatlantic Perspectives from Sciences Po Digital, Governance and Sovereignty Chair Florence G'sell and Georgetown Law Professor Anupam Chander on the Digital Services Act and Section 230." *McCourt Institute* (blog), March 13, 2023. https://mccourtinstitute.org/transatlantic-perspectives-from-professors-florence-gsell-and-anupam-chander-on-the-digital-services-act-and-section-230/.
Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. 1 edition. New York: NYU Press, 2018.
OECD. "Data Portability, Interoperability and Digital Platform Competition." OECD Competition Committee Discussion Paper, 2021. http://oe.cd/dpic.

Office of Science and Technology Policy. “The Blueprint for an AI Bill of Rights: Relationship to Existing Law and Policy.” The White House. Accessed July 13, 2023. https://www.whitehouse.gov/ostp/ai-bill-of-rights/relationship-to-existing-law-and-policy/.

Official Journal L 281 , 23/11/1995 P. 0031 - 0050; “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 (EUR-Lex - 31995L0046 - EN).” Text/html; charset=UTF-8. OPOCE. Accessed May 6, 2023. https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX%3A31995L0046%3AEN%3AHTML.

Ogden, Jessica. “‘Everything on the Internet Can Be Saved’: Archive Team, Tumblr and the Cultural Significance of Web Archiving.” *Internet Histories* 6, no. 1–2 (April 3, 2022): 113–32. https://doi.org/10.1080/24701475.2021.1985835.

Ogden, Jessica, Susan Halford, and Leslie Carr. “Observing Web Archives: The Case for an Ethnographic Study of Web Archiving.” In *Proceedings of the 2017 ACM on Web Science Conference*, 299–308. WebSci ’17. New York, NY, USA: ACM, 2017. https://doi.org/10.1145/3091478.3091506.

O’Hear, Steve. “A Bill of Rights for Users of the Social Web.” *ZDNET* (blog), September 6, 2007. https://www.zdnet.com/home-and-office/networking/a-bill-of-rights-for-users-of-the-social-web/.

On the protection of the privacy of individuals vis-à-vis electronic data banks in the private sector, Pub. L. No. Res(73)22 (1973). https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680502830.

On the protection of the privacy of individuals vis-à-vis electronic data banks in the public sector, Pub. L. No. Res(74)29 (1974). https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=09000016804d1c51.

Opsahl, Kurt. “A Bill of Privacy Rights for Social Network Users.” Electronic Frontier Foundation, May 19, 2010. https://www.eff.org/deeplinks/2010/05/bill-privacy-rights-social-network-users.

Papademetriou, Theresa. “Online Privacy Law (Part One): European Union.” In *Online Privacy Laws : European Union & Select Foreign Countries*, edited by Ethan Williams. Privacy and Identity Protection: Law Library of Congress Global Legal Research Center. New York: Nova Science Publishers, Inc, 2012.

Parallels. “How to Zip a Folder on Mac,” July 14, 2022. https://www.parallels.com/tips/zip-unzip/mac/zip/folder/.

Perez, Sarah. “Facebook Rolls out More API Restrictions and Shutdowns.” *TechCrunch* (blog), July 2, 2018. https://social.techcrunch.com/2018/07/02/facebook-rolls-out-more-api-restrictions-and-shutdowns/.

Perricci, Anna. “Webrecorder: Web Archiving for All!” Internet presented at the ARLIS/NA 2017, March 28, 2018. https://www.slideshare.net/annaperricci/webrecorder-web-archiving-for-all.

“Political Leadership.” Accessed June 12, 2023. https://commission.europa.eu/about-european-commission/organisational-structure/how-commission-organised/political-leadership_en.

Posselt, Julie R. *Equity in Science: Representation, Culture, and the Dynamics of Change in Graduate Education*. Stanford, California: Stanford University Press, 2020.

"POST Statuses/Update." Accessed May 26, 2023. https://developer.twitter.com/en/docs/twitter-api/v1/tweets/post-and-engage/api-reference/post-statuses-update.
"Privacy Guidelines for the National Information Infrastructure: A Review of the Proposed Principless of the Privacy Working Group." Accessed May 3, 2023. https://archive.epic.org/privacy/internet/EPIC_NII_privacy.txt.
Puschmann, Cornelius, and Jean Burgess. "The Politics of Twitter Data." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, January 23, 2013. https://doi.org/10.2139/ssrn.2206225.
Qiu, Yuanbo. "The Openness of Open Application Programming Interfaces." *Information, Communication & Society* 20, no. 11 (November 2, 2017): 1720–36. https://doi.org/10.1080/1369118X.2016.1254268.
Rainie, Lee. "Americans' Complicated Feelings about Social Media in an Era of Privacy Concerns." *Pew Research Center* (blog). Accessed July 28, 2023. https://www.pewresearch.org/short-reads/2018/03/27/americans-complicated-feelings-about-social-media-in-an-era-of-privacy-concerns/.
Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector (Digital Markets Act), § 59-60 (n.d.).
Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector (Digital Markets Act), § Article 8, Compliance (n.d.).
Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector (Digital Markets Act), § Recital 65 (n.d.).
Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector (Digital Markets Act), § 5 (n.d.).
Research Data Services. "An Introduction to Web Archiving for Research," October 15, 2019. https://researchdata.wisc.edu/news/an-introduction-to-web-archiving-for-research/.
Rhizome. "Rhizome Awarded $600,000 by The Andrew W. Mellon Foundation to Build Webrecorder." Rhizome, January 4, 2016. http://rhizome.org/editorial/2016/jan/04/webrecorder-mellon/.
Riley, Chris. "Data Transfer Initiative." Accessed July 24, 2023. https://dtinit.org/.
———. "Data Transfer Initiative: Overview." Accessed July 24, 2023. https://dtinit.org/overview.
———. "Data Transfer Project Use Cases." Data Transfer Initiative. Accessed May 27, 2023. https://dtinit.org/use-cases.
Roberts, Sarah T. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Illustrated edition. New Haven: Yale University Press, 2019.
Rubenking, David. "How to Download Your Facebook Data (and 6 Surprising Things I Found)." *PCMAG*, November 30, 2018. https://www.pcmag.com/how-to/how-to-download-your-facebook-data-and-6-surprising-things-i-found.
"Sample XML File (Books.Xml)," October 27, 2016. https://learn.microsoft.com/en-us/previous-versions/windows/desktop/ms762271(v=vs.85).

Schneider, Steve, and Kirsten Foot. “Archiving of Internet Content.” In *The International Encyclopedia of Communication*, edited by Wolfgang Donsbach. Blackwell Publishing, 2008.
Schudson, Michael. *The Rise of the Right to Know: Politics and the Culture of Transparency, 1945–1975*. Cambridge, Massachusetts: Belknap Press: An Imprint of Harvard University Press, 2015.
Schwartz, Paul M. *Data Privacy Law : A Study of United States Data Protection*. Charlottesville, Va: Michie, c1996.
Scott, Adam. “The Eight Rules of Good Documentation.” O’Reilly Media, April 17, 2018. https://www.oreilly.com/content/the-eight-rules-of-good-documentation/.
Scott Morton, Fiona, Gregory Crawford, Jacques Crémer, David Dinielli, Amelia Fletcher, Paul Heidhues, Monika Schnitzer, and Katja Seim. “Equitable Interoperability: The ‘Super Tool’ of Digital Platform Governance.” In *Policy Discussion Paper No. 4*, 32. Yale University: Yale Tobin Center for Economic Policy, 2021.
Shank, Craig. “Microsoft, Facebook, Google and Twitter Introduce the Data Transfer Project: An Open Source Initiative for Consumer Data Portability.” EU Policy Blog, July 20, 2018. https://blogs.microsoft.com/eupolicy/2018/07/20/microsoft-facebook-google-and-twitter-introduce-the-data-transfer-project-an-open-source-initiative-for-consumer-data-portability/.
Sheldon, Pavica, and Katherine Bryant. “Instagram.” *Computers in Human Behavior* 58, no. C (May 1, 2016): 89–97. https://doi.org/10.1016/j.chb.2015.12.059.
Silberling, Amanda. “A Beginner’s Guide to Mastodon, the Open Source Twitter Alternative.” *TechCrunch* (blog), July 24, 2023. https://techcrunch.com/2023/07/24/what-is-mastodon/.
Skrla, Linda, and James Joseph Scheurich, eds. *Educational Equity and Accountability: Paradigms, Policies, and Politics*. Studies in Education/Politics. New York: RoutledgeFalmer, 2004.
Smith, Clarence. “Portability and Digital Markets Act,” July 25, 2023.
Smith, David. “Analysis of Facebook Status Updates.” *Revolutions* (blog), December 29, 2010. https://blog.revolutionanalytics.com/2010/12/analysis-of-facebook-status-updates.html.
Solove, Daniel. “Introduction: Privacy Self-Management and the Consent Dilemma.” *Harvard Law Review* 126, no. 7 (2013): 1880–1904.
“Special Eurobarometer 359: Attitudes on Data Protection and Electronic Identity in the European Union - Data Europa EU.” Accessed July 23, 2023. https://data.europa.eu/data/datasets/s864_74_3_ebs359?locale=en.
Stempel, Jonathan. “Google Faces $5 Billion Lawsuit in U.S. for Tracking ‘private’ Internet Use.” *Reuters*, June 2, 2020, sec. U.S. Legal News. https://www.reuters.com/article/us-alphabet-google-privacy-lawsuit-idUSKBN23933H.
Summers, Ed. “Introducing Documenting the Now.” *Maryland Institute for Technology in the Humanities* (blog), February 17, 2016. https://mith.umd.edu/introducing-documenting-the-now/.
Summers, Ed, Bergis Jules, and Vernon Mitchell Jr. “Documenting the Now: Ethical Considerationsfor Archiving Social Media Content Generated by Contemporary Social Movements: Challenges, Opportunities, and Recommendations.” Documenting DocNow, July 19, 2018. https://www.docnow.io/docs/docnow-whitepaper-2018.pdf.
Summers, Ed, and Ricardo Punzalan. “Bots, Seeds and People: Web Archives as Infrastructure.” *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work*

and Social Computing - CSCW '17, 2017, 821–34. https://doi.org/10.1145/2998181.2998345.
Swire, Peter P. *None of Your Business : World Data Flows, Electronic Commerce, and the European Privacy Directive*. Washington, D.C.: Brookings Institution Press, 1998.
Syed, Armani. "Why E.U. Users May Not Get to Use Threads." *Time*, July 6, 2023. https://time.com/6292586/privacy-concerns-threads-meta/.
Tamilore, Oladipo. "A Beginner's Guide to Mastodon." Buffer Resources, November 16, 2022. https://buffer.com/resources/mastodon-social/.
"The Commissioners." Accessed June 12, 2023. https://commissioners.ec.europa.eu/index_en.
"The Digital Markets Act: Ensuring Fair and Open Digital Markets." Accessed July 31, 2023. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_en.
"The Digital Services Act Package | Shaping Europe's Digital Future," May 5, 2023. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package.
Thomson, Sara Day, and William Kilbride. "Preserving Social Media: The Problem of Access." *New Review of Information Networking* 20, no. 1 (2015): 261–75.
Tiwana, Amrit, Benn Konsynski, and Ashley A. Bush. "Research Commentary — Platform Evolution: Coevolution of Platform Architecture, Governance, and Environmental Dynamics." *Information Systems Research* 21, no. 4 (December 2010): 675–87. https://doi.org/10.1287/isre.1100.0323.
"Transcript of Proceedings of the Secretary's Advisory Committee on Automated Personal Data Systems (SACAPDS)." Berkeley Law Center, April 17, 1972. https://www.law.berkeley.edu/research/bclt/research/privacy-at-bclt/archive-of-the-meetings-of-the-secretarys-advisory-committee-on-automated-personal-data-systems-sacapds/.
"Transfer Your Information to a Service off of Facebook | Facebook Help Center." Accessed July 27, 2023. https://www.facebook.com/help/230304858213063.
Tromble, Rebekah. "Where Have All the Data Gone? A Critical Reflection on Academic Digital Research in the Post-API Age." *Social Media + Society* 7, no. 1 (January 1, 2021): 2056305121988929. https://doi.org/10.1177/2056305121988929.
Tromble, Rebekah, Andreas Storz, and Daniela Stockmann. "We Don't Know What We Don't Know: When and How the Use of Twitter's Public APIs Biases Scientific Inference." SSRN Scholarly Paper. Rochester, NY, November 29, 2017. https://doi.org/10.2139/ssrn.3079927.
Turner, Sarah, July Galindo Quintero, Simon Turner, Jessica Lis, and Leonie Maria Tanczer. "The Exercisability of the Right to Data Portability in the Emerging Internet of Things (IoT) Environment." *New Media & Society* 23, no. 10 (October 1, 2021): 2861–81. https://doi.org/10.1177/1461444820934033.
Twitter. "How to Post Photos or GIFs on Twitter." Accessed June 2, 2020. https://help.twitter.com/en/using-twitter/tweeting-gifs-and-pictures.
UNESCO. "Guidelines for Regulating Digital Platforms: A Multistakeholder Approach to Safeguarding Freedom of Expression and Access to Information." UNESCO, February 2023. https://unesdoc.unesco.org/ark:/48223/pf0000384031.locale=en.
Urquhart, Lachlan, Neelima Sailaja, and Derek McAuley. "Realising the Right to Data Portability for the Domestic Internet of Things." *Personal and Ubiquitous Computing* 22, no. 2 (April 1, 2018): 317–32. https://doi.org/10.1007/s00779-017-1069-2.

Ursic, Helena. “Unfolding the New-Born Right to Data Portability: Four Gateways to Data Subject Control.” *SCRIPTed: A Journal of Law, Technology & Society* 15, no. 1 (August 1, 2018): 42–69. https://doi.org/10.2966/scrip.150118.42.

U.S. Department of Health, Education and Welfare, Secretary’s Advisory Committee on Automated Personal Data Systems, Records, computers, and the Rights of Citizens. *Records, Computers, and the Rights of Citizens*. OHEW Publication, (OS)73-94. Washington, D.C.: U.S. Department of Health, Education & Welfare, 1973. https://aspe.hhs.gov/reports/records-computers-rights-citizens.

Vaccari, Lorenzino, this link will open in a new window Link to external site, Monica Posada, this link will open in a new window Link to external site, Mark Boyd, Mattia Santoro, and this link will open in a new window Link to external site. “APIs for EU Governments: A Landscape Analysis on Policy Instruments, Standards, Strategies and Best Practices.” *Data* 6, no. 6 (2021): 59. https://doi.org/10.3390/data6060059.

Valtysson, Bjarki, Rikke Frank Jorgensen, and Johan Lau Munkholm. “Co-Constitutive Complexity: Unpacking Google’s Privacy Policy and Terms of Service Post-GDPR.” *NORDICOM Review: Nordic Research on Media and Communication* 42, no. 1 (July 1, 2021): 124–41. https://doi.org/10.2478/nor-2021-0033.

Van der Auwermeulen, Barbara. “How to Attribute the Right to Data Portability in Europe: A Comparative Analysis of Legislations.” *Computer Law & Security Review* 33, no. 1 (February 1, 2017): 57–72. https://doi.org/10.1016/j.clsr.2016.11.012.

Van Dijck, José. “Seeing the Forest for the Trees: Visualizing Platformization and Its Governance.” *New Media & Society* 23, no. 9 (September 1, 2021): 2801–19. https://doi.org/10.1177/1461444820940293.

Veale, Michael, Reuben Binns, and Jef Ausloos. “When Data Protection by Design and Data Subject Rights Clash.” *International Data Privacy Law* 8, no. 2 (May 1, 2018): 105–23. https://doi.org/10.1093/idpl/ipy002.

Victor, Daniel. “How to Download Your Twitter Archive.” *The New York Times*, November 18, 2022, sec. Technology. https://www.nytimes.com/2022/11/18/technology/how-to-download-your-twitter-archive.html.

Vollmer, Nicholas. “Article 12 EU General Data Protection Regulation (EU-GDPR).” Text. SecureDataService, April 4, 2023. https://www.privacy-regulation.eu/en/article-12-transparent-information-communication-and-modalities-for-the-exercise-of-the-rights-of-the-data-subject-GDPR.htm.

Wahyuningtyas, Sih Yuliana. “Interoperability for Data Portability between Social Networking Sites (SNS): The Interplay between EC Software Copyright and Competition Law.” *Queen Mary Journal of Intellectual Property* 5, no. 1 (January 1, 2015): 46–67. https://doi.org/10.4337/qmjip.2015.05.03.

Waldman, Ari Ezra. *Privacy as Trust: Information Privacy for an Information Age*. Cambridge, United Kingdom ; New York, NY: Cambridge University Press, 2018.

———. “Privacy, Practice, and Performance.” *110 CAL. L. REV. 1221*, 2022. https://doi.org/10.2139/ssrn.3784667.

Walker, Shawn, Dan Mercea, and Marco Bastos. “The Disinformation Landscape and the Lockdown of Social Platforms.” *Information, Communication & Society* 22, no. 11 (September 19, 2019): 1531–43. https://doi.org/10.1080/1369118X.2019.1648536.

Ware, William. “Records, Computers and the Rights of Citizens.” U.S. Health and Human Services, June 30, 1973. https://aspe.hhs.gov/reports/records-computers-rights-citizens.

———. “Transmittal Letter to Secretary Honorable Caspar W. Weinberger Secretary of Health, Education, and Welfare.” In *Records, Computers and the Rights of Citizens*. OHEW Publication, (OS)73-94. Washington, D.C.: U.S. Department of Health, Education & Welfare, 1973.

Watzman, Nancy. “Wayback Machine Captures Melania Trump’s Deleted Internet Bio | Internet Archive Blogs.” *Wayback Machine Captures Melania Trump’s Deleted Internet Bio* (blog), July 28, 2016. https://blog.archive.org/2016/07/28/wayback-machine-captures-melania-trumps-deleted-internet-bio/.

“Wayback Machine General Information – Internet Archive Help Center.” Accessed May 26, 2023. https://help.archive.org/help/wayback-machine-general-information/.

Webrecorder. “Webrecorder Terms and Policies.” Accessed December 11, 2018. https://webrecorder.io/_policies#privacy.

Westin, Alan F. “Science, Privacy, and Freedom: Issues and Proposals for the 1970’s. Part I--The Current Impact of Surveillance on Privacy.” *Columbia Law Review* 66, no. 6 (1966): 1003–50. https://doi.org/10.2307/1120997.

———. “Social and Political Dimensions of Privacy.” *Journal of Social Issues* 59, no. 2 (July 1, 2003): 431–53. https://doi.org/10.1111/1540-4560.00072.

“What Is SaaS? Software as a Service | Microsoft Azure.” Accessed May 18, 2023. https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-saas.

Wong, Janis, and Tristan Henderson. “How Portable Is Portable? Exercising the GDPR’s Right to Data Portability.” In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, 911–20. UbiComp ’18. New York, NY, USA: Association for Computing Machinery, 2018. https://doi.org/10.1145/3267305.3274152.

Worsnop, Richard L. “Reappraisal of Computers.” In *Editorial Research Reports 1971*, 345–66. CQ Researcher Online. Washington, D.C., United States: CQ Press, 1971. http://library.cqpress.com/cqresearcher/cqresrre1971051200.

Wrabetz, Joan. “What Is Inferred Data and Why Is It Important?” *Business Law Today from American Bar Association* (blog), August 22, 2022. https://businesslawtoday.org/2022/08/what-is-inferred-data-why-is-it-important/.

Zaidi, Kamaal. “Harmonizing U.S.-EU Online Privacy Laws: Toward a U.S. Comprehensive Regime for the Protection of Personal Data.” *Michigan State University Journal of International Law* 12, no. 1 (2004 2003): 169–98.

Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs, 2020.

Appendix A : Technical Glossary

This chapter describes the basic features of data collecting. The chapter introduces technical terms encountered during this study about data collecting and that are necessary to understanding the technology and regulations involved in the practice.

This includes, first, a summary of the architecture and formats that lay the foundation for any data collecting. Software must be designed according to the architecture and formats that structure the internet and websites. They must also be compatible with the formats and architecture of data and databases. Thus design and compatibility are important to understanding the methods for data collecting used in this study and how they were adapted to new web features and in particular social media. These include methods used in this study: Archive-It, Webrecorder, and platform downloads.

The first two are software specially designed for collecting websites and data. A platform download is an assembly of different elements including an interface, algorithms, and APIs (application programming interface). The overview of technical formats and features of the web is provided to facilitate a common technical understanding of how data is collected and the issues with that collection discussed in the remainder of this study.

Digital: media, formats, extensions, and files

The internet is an information system that links together computers, allowing the movement and exchange of data across these connections, or networks.[307] It is often called a network of

[307] Barry Leiner et al., "Brief History of the Internet," *Internet Society* (blog), 1997, https://www.internetsociety.org/internet/history-internet/brief-history-internet/.

networks. The web – originally called the World Wide Web, or www – is an interconnected system of "pages" accessible via the internet.[308] One way to distinguish the two is that the web is one of many applications built "on top of" the internet.[309] Several main components make the web operational:[310]

- HTTP (hypertext transfer protocol) provides the instructions to transfer data between two computers.
- URLs (uniform resource location) provide the information that allows pages to be linked together and navigable (through hyperlinks), e.g. www.someurl.com.[311] They identify where something is located on the web.
- HTML (hypertext markup language) created by Tim Berners-Lee in 1989 to encode electronic documents, focusing on presenting text and URLs.[312] HTML is necessary to provide instructions for formatting and displaying a webpage.[313]
- JavaScript is a programming language used to add interactive features to a website.
- CSS is a markup language used to enhance the appearance of HTML, such as color and alignment.

---

[308] "World Wide Web - MDN Web Docs Glossary: Definitions of Web-Related Terms | MDN," accessed July 13, 2021, https://developer.mozilla.org/en-US/docs/Glossary/World_Wide_Web.

[309] "World Wide Web - MDN Web Docs Glossary."

[310] "World Wide Web - MDN Web Docs Glossary."

[311] Or URI (uniform resource identifier).

[312] Stephanie Crawford, "How HTML5 Works," HowStuffWorks, August 24, 2011, https://computer.howstuffworks.com/html-five.htm.

[313] Crawford.

- Web pages are simple text files that contain text plus HTML, CSS, and JavaScript to display on a screen.[314]
- Web browsers are computer programs used to navigate to a location on the internet where web pages are stored and pull the page through the network and into the user's machine. Web browsers also interpret the set of HTML tags within the page in order to display the page on the user's screen.[315]
- Web servers are computer software that can respond to a browser's request for a page, and deliver the page to the web browser through the internet.[316]
- XML (extensible markup language) is a markup language related to HTML. Whereas HTML displays data and describes the structure of a webpage, XML is used to transfer data between computers, applications, etc.[317]
- JSON (JavaScript object notation) is, like XML, a format for exchanging data. But JSON is more flexible than XML because it can be used with any other programming language, and the underlying data structure of JSON is platform-independent. The language-independent nature of JSON makes it ideal for use in web development, where data may need to be exchanged with other programing languages like JavaScript.[318]

---

[314] Crawford.

[315] Crawford.

[316] Marshall Brian, "How Web Pages Work," HowStuffWorks, September 5, 2000, https://computer.howstuffworks.com/web-page.htm.

[317] "Sample XML File (Books.Xml)," October 27, 2016, https://learn.microsoft.com/en-us/previous-versions/windows/desktop/ms762271(v=vs.85).

[318] "JSON vs XML | AppMaster," accessed July 3, 2023, https://appmaster.io/blog/json-vs-xml.

- Interoperability is a term from the information technology industry that defines an approach to allowing computers and other electronic devices to "relate" to each other through standards and formats. The simplest form of interoperability is the ability to transport data from one system to another regardless of the purpose, content, and format of the data. Interoperability requires sufficient shared standards to enable a receiving system to parse, store, and use data from a sending system as though the data had originated in the receiving system. This is difficult not only because the data has to be reliably transferred in a specific language and a structured format.[319]
- A basic example of interoperability is a USB device (thumb drive or memory stick) that is interoperable with any modern computer terminal's corresponding USB port (the slot where the stick gets plugged in). You plug the USB device into the computer terminal and you can access the contents (on the USB device) because there is an information exchange. This information is transportable to secondary computers, printers and other peripheral devices because the USB storage device was designed in accordance with interoperability standards by an assortment of organizations and government agencies. Conversely, a VHS cassette is totally inoperable with a Beta playback machine. This means that no information exchange between the VHS cassette and the Beta playback equipment occurred because they were designed with incompatible standards.[320]

[319] Mark Braunstein, "Healthcare in the Age of Interoperability: Part 2," *IEEE Pulse* (blog), November 15, 2018, https://www.embs.org/pulse/articles/healthcare-in-the-age-of-interoperability-part-2/.

[320] "What Is Interoperability?," Network Centric Operations Industry Council Interoperability - NCOIC, July 14, 2014, https://web.archive.org/web/20140714143139/http://www.ncoic.org/what-is-interoperability.

- The internet and web were also originally defined by limited formats and compatibility. Activity online drove the need for interoperability. Today anyone with an email address or access to the web is a member of the world's largest and most interoperable network.[321]
- API (application programming interface) is a method for sharing information, which supports interoperability by specifying what information can be requested, how requests for the information should be formatted, and what information will be returned (i.e. the data that is requested or an error message). APIs provide interoperability by defining the rules (protocols) for how different systems will connect to each other.[322]

Formats specify *how* the information should be encoded in a file (a file is a container for data organized in some way).[323] Computers need to know the format of a file, which programmers accomplish by adding a suffix to a filename, i.e. *somewebfile.html* (it makes sense why the suffix is called a filename extension but note the period between the filename and the extension).

A file's name "somewebfile" is mostly arbitrary, while the suffix ".HTML" is specific to the format. Software, in most cases, can't open a file without a filename extension. Web applications don't know what to do with "somewebfile" without the .HTML and we will also get an error message when we try to open this file type with software that is not designed for it.[324]

---

[321] "What Is Interoperability?"

[322] "Terminology: API," IEEE Xplore, accessed July 5, 2023, https://developer.ieee.org/Terminology; Wahyuningtyas, "Interoperability for Data Portability between Social Networking Sites (SNS)."

[323] "Computer File," in *Wikipedia*, July 7, 2021.

[324] Ryan and Sampson, The No-Nonsense Guide to Born Digital Content.

Files contain information that informs a computer about the file's format and embeds information about the objects, such as an image: its size and color , and information that comprises the images so that it can be represented not as a series of numbers (0 and 1) but as a flower, a loved one, or a blue sky.

HTML structures a document with "tags" that browsers can interpret and render. The browser is designed to interpret the structure of a document through HTML, such as paragraphs, quotes, lists. CSS (cascading style sheets) adds styles to the content. JavaScript adds features, or events, like infinite scrolls on social media sites, drop-down menus, dynamic displays. Web browsers communicate via HTTP (or HTTPS) with a server, which displays content that is structured, formatted, and designed with a combination of HTML, CSS, and JavaScript.

JavaScript changed the way that information is exchanged between computers so that much of the activity takes place on the user's side, instead of sending a request to a server that then delivers the page to the web browser through the internet.[325] JavaScript now allows the user's interaction to happen asynchronously, that is, independent of communication with a server.[326]

HTML and HTTP are still the backbone of webpages, and connecting resources through hyperlinks (URLs) continues to be a defining concept of the web.[327] But increasingly since 1996, when the web began to be commercially developed, webpages are no longer static, but are dynamic – a word used to mean they change state and are modular, made up of a variety of

---

[325] Marshall Brian, "How Web Servers Work," HowStuffWorks, April 1, 2000, https://computer.howstuffworks.com/web-server.htm.

[326] Jesse James Garrett, "Ajax: A New Approach to Web Applications," *Adaptive Path: Ideas* (blog), February 18, 2005, http://www.adaptivepath.com/ideas/essays/archives/000385.php.

[327] "World Wide Web - MDN Web Docs Glossary."

objects with specific functions. The continued adoption of new web technologies has made the pages personalized and more interactive. These details are important to understanding my comparison of data collecting methods: how and why methods are increasingly ineffective and restricted.

One of the methods I used to collect social media, Archive-It, was designed to retrieve and store pages of websites. The web pages (associated images, style sheets, JavaScript, and other objects) are stored in files that can be displayed using specialized software. Although this process involves a number of different formats, the core format is HTML, which is intended to be rendered by a web browser.

By contrast, collecting social media involves retrieving data from a site like Twitter or Facebook using an API. The advantages of collecting from the API include efficiency, more metadata, and formats that return data pre-structured in a way that facilitates computation techniques for analyzing it. They allow data to be transferred from one online service to another. Hypothetically, an API could be used to download data from Twitter in order to transfer it to another social media service, like Mastodon. But, access to data through APIs must be obtained from the service provider (e.g. Twitter). Increasingly platforms including web services like Archive-It are making data inaccessible to methods that the platforms do not first authorizeThis control has commercial and security advantages, which I explain elsewhere.

What is an API?

APIs are software features that mediate interactions between applications, data, and devices.[328] More specifically, they are protocols, which grant managed access between devices.[329] API is a generic term but most often refers to a web API that is used to standardize data exchanged over the web.[330] Generally speaking, this means that two applications can be used to automatically exchange information (send, receive, and respond to a request). If we post a message to Facebook or check the weather on an iPhone app, an API connects to the Internet and sends a request for data to a server. The server then retrieves that requested data, interprets it, performs the necessary actions and sends it back to your device.[331] The application interprets that data and presents the requested information in a readable way, the basic approach to interoperability.

A web API is a small but crucial part of the process of communicating between databases that opens doorways to data, but only to specific sets of data that the provider wants to be seen.[332]

---

[328] MuleSoft Videos, *What Is an API?*, accessed July 15, 2021, https://www.youtube.com/watch?v=s7wmiS2mSXY.

[329] Alexander R. Galloway, *The Interface Effect*, 1st edition (Cambridge, UK ; Malden, MA: Polity, 2012); Alexander R. Galloway, *Protocol: How Control Exists after Decentralization* (Cambridge, Mass.: MIT Press, 2004) in Jose van Dijck, *The Culture of Connectivity: A Critical History of Social Media* (Oxford ; New York: Oxford University Press, 2013).

[330] Michael Lahey, "Invisible Actors: Web Application Programming Interfaces, Television, and Social Media," *Convergence* 22, no. 4 (August 1, 2016): 426–39, https://doi.org/10.1177/1354856516641915.

[331] "What Is an API? (Application Programming Interface)," MuleSoft, accessed July 15, 2021, https://www.mulesoft.com/resources/api/what-is-an-api.

[332] Lahey, "Invisible Actors."

The technical explanation of a web API-enabled transaction begins with a computer program that includes a "GET" or "POST" function that is executed when triggered.[333] These "GET" and "POST" functions transfer data over HTTP, which is the protocol that defines how messages are formatted and transmitted across the web.

Explanation about what an API is and does often rely on comparisons. A common analogy involves ordering take-out from an online restaurant app: we select options from a menu then communicate the selections, formatted in a specific way, through the app. The instructions are relayed to a restaurant, whose staff compile and then distribute the order to the kitchen staff. The staff slice, chop, and mix the ingredients that make up your order, which, once completed, is conveyed to a delivery service and then to your doorstep. Similarly, APIs receive requests for data that are then compiled and distributed to the computer server from which the request came. The data (stored in a database) is sliced and chopped and mixed according to the request.

Now imagine what would have happened if the restaurant on the other end could not interpret the order because it was in the wrong format or language. That is where the web API steps in and eases the translation from one format (or, computer language) to another, which called interoperability. The API does this by translating all information into a "data-interchange format" that most computer programming languages can read, such as Extensible Markup Language (XML) or JavaScript Object Notation (JSON).

APIs are becoming the prevailing method for accessing and collecting social media data regardless of the purpose because they are technologically agnostic.[334] They also allow application developers to connect new add-ons to an existing service, like Twitter and Facebook

[333]For this explanation I rely on Lahey, "Invisible Actors."

[334] Lahey.

(e.g. the dating app Tindr). APIs are also an interface for researchers to collect data from a given social media service for empirical analysis. Private companies, from startups to enterprises, have been using APIs for several years now, and recently the public sector have become interested in APIs.[335]

Providers like Twitter publish instructions on their sites about what data is available and how to request it through their APIs. They include these instructions in documentation. This documentation should describe what the API does; the type of data available; requirements for using the API; the fees; the authorization steps; examples for how to format a request correctly and the format of data that will be returned.

Types of APIs

APIs can be private, available only to operators in a company. Ones that are available to use online are called "public," or "open" APIs. These are used to make data available to third parties without exposing more data than the provider wants to make available.

Returning to the example of online food orders, we can think of APIs as a web service that provides machine-to-machine information exchanges over a network. Users can only access the front-end interface of an application or website. All the data is stored on remote servers and gets transmitted to the users' device through the API. For the data transfer from the server to the users' device, APIs use different architectures or protocols.[336] These fall into two categories:

---

[335] Stine Lomborg and Anja Bechmann, "Using APIs for Data Collection on Social Media," *The Information Society* 30, no. 4 (August 8, 2014): 256–65, https://doi.org/10.1080/01972243.2014.915276; Vaccari et al., "APIs for EU Governments."

[336] Kelly Arellano, "SOAP vs REST (vs JSON): Web API Services [2021] | RapidAPI," Rapid Blog, February 13, 2019, https://rapidapi.com/blog/soap-vs-rest-api/; Lakshmi, "SOAP vs REST vs JSON - Differences and How They Are Useful," September 2, 2021, https://seagence.com/blog/soap-vs-rest-vs-json/.

REST APIs and SOAP APIs. They are the two most common approaches for transferring data over a network using API requests.[337]

REST is considered less secure than SOAP. But some of the web's largest online services like Yahoo, Amazon, eBay, and Google offer REST APIs for their most popular features. The trade-off in security is not so great as to exceed the benefits of interoperability and ease of use: Learning to write queries is easier than SOAP. And REST APIs also take up less data, making this an attractive choice in terms of profitability.

A company's decision to opt for SOAP or REST will be based on a variety of considerations. But, taken as a whole, these APIs tie together the various computing systems, sites, and apps found online into a cohesive whole.[338] These terms are good to know because they are often used in representations of APIs, such as those in Diagram 1 and 2 below.

The most common API operations are GET, POST, PUT, PATCH, and DELETE.[339] Facebook, however, developed its own version, called GraphAPI, which is less linear in the exchange between devices and the server, and provides more tailored requests than the other versions. However, they share most of the same functions. This information is necessary to

[337]SOAP (Simple Object Access Protocol) and REST (Representational State Transfer) offer different methods to interact with a web service. The significance of the differences between the two are beyond the scope of the discussion here. But, a reason to select one standard over another includes security. SOAP can be used as an architecture for APIs to transfer complex and highly confidential files involving financial transactions, telecommunications, and identity management services. Paypal's payment API is one of the well-known SOAP APIs. Facebook also uses SOAP. REST is a common choice for building complex public APIs because its architecture is flexible and lightweight. Many social media platforms use REST APIs so that the developers can integrate their applications or websites into the platform to create additional services. A well-known public REST API is Twitter's APIs. See: Lakshmi, "SOAP vs REST vs JSON - Differences and How They Are Useful."

[338] Arellano, "SOAP vs REST (vs JSON)"; Lakshmi, "SOAP vs REST vs JSON - Differences and How They Are Useful."

[339] "REST API: Key Concepts, Best Practices, and Benefits," *AltexSoft* (blog), accessed June 8, 2023, https://www.altexsoft.com/blog/rest-api-design/.

identifying that an API is part of the architecture of a site, what kind of API it is, how to structure requests for data, and for troubleshooting.

This is a basic example of an API request (or, call) to retrieve (GET) information:

http://hapi.fhir.org/baseDstu3/Condition?code=http://snomed.info/sct|73211009

The first part (blue) specifies the server where the information is stored; the second part (green) specifies the type of resource that is desired; the third part (brown) provides sufficient information for the server to retrieve the correct resources.[340] Diagrams 1 and 2 shows the process that happens when a request is sent.

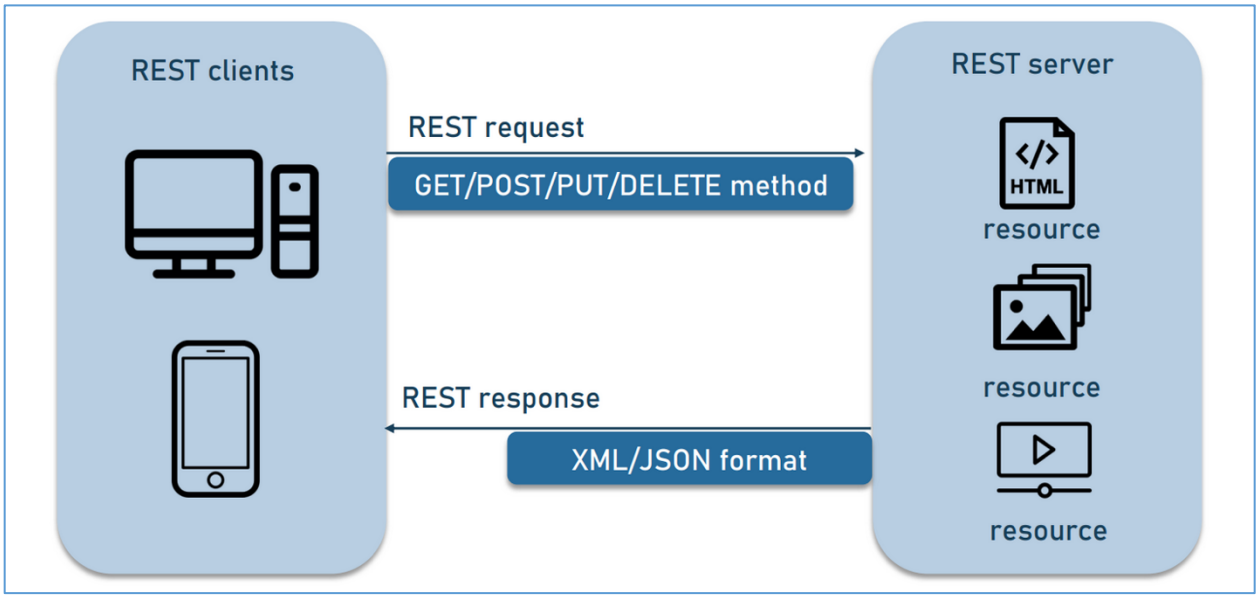

Diagram 1: Representation of an API request from a device to a server[341]

[340] Braunstein, "Healthcare in the Age of Interoperability."

[341] Elena Ferrari, "Access Control," in *Encyclopedia of Database Systems*, ed. Ling Liu and M. Tamer Özsu (New York, NY: Springer, 2018), 9–14, https://doi.org/10.1007/978-1-4614-8265-9_6.

Different coding languages like JSON are intelligible to each other. This means a range of different programs and computer languages can have access to your data through these common formats, thereby mitigating certain problems of portability and interoperability. These are qualities that make it easy to automatically download and transfer data from one service to another. This quality makes APIs attractive to regulators, who have added provisions in privacy and data protection regulations like the Digital Services Act in the European Union that data be formatted for portability and interoperability.[342]

Security

APIs are also an access control mechanism programmed to decide whether an access request can be authorized or should be denied. Authorizations are stored into a system and are then used to verify whether an access request can be authorized or not based on the credentials.[343] This verification is handled largely by tokens and keys. These are used to authorize users to make an API request, or "call." Authentication tokens automatically check that the users are who they claim to be and that they have access rights for that particular API call. For example, when you log in to your email server, your email client uses authentication tokens for secure access. API keys automatically verify the program or application making the API call. They identify the application and ensure the user has the access rights required to make the particular API call. They allow API monitoring in order to gather data on usage.[344]

---

[342] Ferrari.

[343] Ferrari.

[344] "What Is an API? - Application Programming Interfaces Explained - AWS," Amazon Web Services, Inc., accessed June 15, 2023, https://aws.amazon.com/what-is/api/.

In the case of social media, the token must be supplied to the API for validation every time a request is made for user data. These access tokens are only valid for a relatively short period of time, so a method is provided for an application to supply an old access token in exchange for a new one. This means that if the user were to revoke permission to access their data, the request for a new access token will be denied, and the application will no longer have access to that user's data.[345]

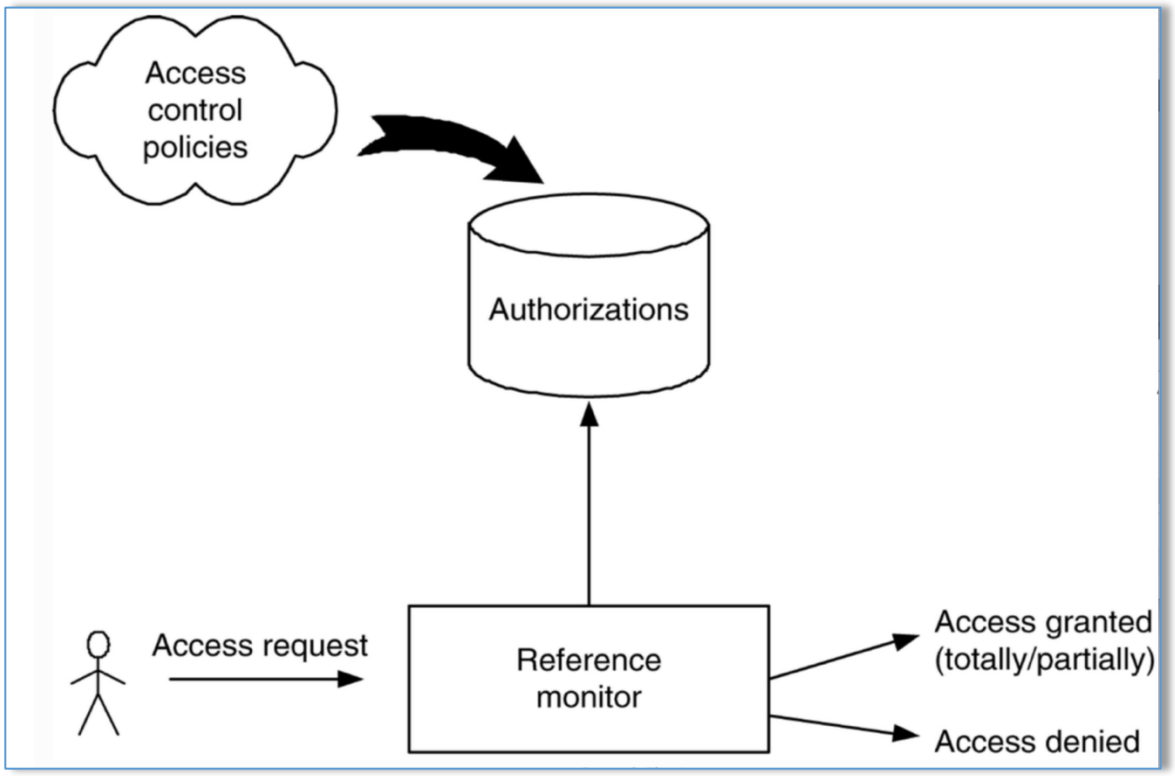

Diagram 2: **Representation of an API access authorization process**[346]

[345] "Facebook Data Mining with Web Scraping," Apify Blog, March 17, 2023, https://blog.apify.com/facebook-data-mining-with-web-scraping/.
[346] Ferrari, "Access Control."

Rate Limits

Rate limiting is another access control mechanism. Rate limiting is a technique to limit traffic across online networks. The purpose is to prevent users from requesting more data than servers can handle at one time. For example, rate limiting makes it harder for malicious actors to overburden a system with attacks, like Denial of Service (DoS). These involve attackers flooding a target system with requests and consuming too much network capacity, storage, and memory. A server will crash, making services inaccessible. APIs that use rate limiting can throttle or temporarily block any client that tries to make too many API calls. This might slow down a throttled user's requests for a specified time or deny them altogether.

Rate limiting works within applications, not in the web server. Rate limiting typically involves tracking the IP addresses where requests originate and identifying the time lapsed between requests. IP addresses are the application's main way to identify who has made each request.

Rate limiting functions by measuring the elapsed time between every request from a given IP address and tracking the number of requests made in a set time frame. If an IP address makes too many requests within the specified timeframe, the rate-limiting mechanism throttles the IP address and doesn't fulfill its requests until a certain amount of time passes. Rate-limited applications can tell individual users to slow down if they make requests too frequently.

Limits can be based on the user's IP address or the location where the request originated. There are also types of rate-limiting algorithms. For example, "fixed-window" rate-limiting algorithms restrict the number of requests allowed during a given time frame (i.e. time window). A server's rate-limiting component might implement an algorithm that accepts up to 200 API

requests per minute. There is a fixed window starting from a specified time—the server will serve no more than 200 requests between 9:00 and 9:01, but the window will reset at 9:01, allowing another 200 requests until 9:02.

Despite the differences, the premise is that rate limiting ensures that legitimate requests can reach the system server and access information without affecting an application's overall performance. One explanation compares rate limiting to police officers pulling over drivers who exceed the speed limit or parents telling their children not to eat too much sugar in a short period.[347]

However, unlike speed limit enforcement by police or sugar limits imposed by parents, rate limiting can happen without explanation, notice, or recourse. In one instance, Twitter launched a widescale redesign to its website in July 2019. Archivists trying to access an otherwise accessible account of Donald Trump using different data collecting tools received error messages but no explanation for why the content was unavailable. Ultimately they determined that rate limits were preventing them from collecting the data, or even accessing the content.[348] Additionally, the rate-limiting algorithms that control the specific types of limiting that happens are opaque, if they are disclosed at all. I found none that are. APIs also cause interoperability and portability problems.

This chapter provides an overview of operations and terms that are necessary to understanding the remainder of this study. For example, regulators have identified access, transparency, and

[347] "What Is Rate Limiting | Types & Algorithms | Imperva," *Learning Center* (blog), accessed June 15, 2023, https://www.imperva.com/learn/application-security/rate-limiting/.

[348] Garg Kritika, "Twitter Was Already Difficult to Archive, Now It's Worse!," *Old Dominion University Web Science and Digital Libraries Research Group* (blog), July 15, 2020, https://ws-dl.blogspot.com/2020/07/2020-07-15-twitter-was-already.html.

portability as key problems that need to be addressed in user rights legislation. They see promise APIs as a solution to providing these features. But protocols, API operations, rate limits, algorithms, and components like HTTP and URLs are important in their own right because they are constituent features of our information and communications infrastructure.[349] It was only by tracing each of the features outlined in this chapter that I was able to discern that platform downloads include APIs and algorithms, and thus effectively assess their role. I began to understand why other methods were unsuccessful.

We engage with APIs, rate limits, algorithms, protocols every day and they affect the way they and we understand access, transparency, and sharing. Given their place in our information sharing infrastructures and personal data protection regulations, understanding and assessing of these features and their operations is necessary.

[349] Joshua Braun, "Transparent Intermediaries: Building the Infrastructures of Connected Viewing," in *Connected Viewing* (Routledge, 2013).

Appendix B

Appendices go here, after the text and references.