Challenges in archiving the personalized web

Erwan Le Merrer, Camilla Penzo, Gilles Tredan, Lucas Verney

Challenges in archiving the personalized web

Erwan Le Merrer, Camilla Penzo, Gilles Tredan, Lucas Verney

Abstract: The decision-making algorithms embedded within online platforms are determining content shown to users. This personalization steers the dissemination of information, in contrast with the idea of a universal World Wide Web. Personalization thus generates a combinatorial explosion of different versions of the web, rendering each user's experience distinct. This raises critical questions: what elements of a personalized web should be archived? How can the collected user journeys capture a representative picture of our times? Navigating personalization is essential to capture the contemporary web experience, yet it presents methodological and technical challenges. In this chapter, we identify key challenges in performing a representative sampling of personalization within online platforms.

Keywords: personalization, archival, YouTube, 2022 French presidential election.

1. Introduction

The web has evolved from its static origins to a dynamic landscape where each user encounters an ever-changing and algorithmically tailored version. A few years ago, studies showed that people were generally unaware of the existence of algorithmic personalization (Eslami et al. 2015; Powers 2017). More recent studies (Schmidt et al. 2019; Eg, Demirkol Tønnesen, and Tennfjord 2023), however, suggest that users may grasp the notion that online content is filtered or that recommendations are based on their profiles, even if they are not necessarily familiar with algorithmic processes. Nevertheless, details regarding the algorithmic personalization of each platform remain undisclosed to both users and regulators. Recommendation algorithms, despite their critical role in selecting and ranking information, can inadvertently reinforce popularity as self-fulfilling prophecies (Salganik and Watts 2008). Furthermore, these algorithms often overlook the verification of information sources, potentially leading to the propagation of disinformation and the creation of filter bubbles.

Given that personalization inherently renders each user experience unique, collecting the entirety of the internet might offer limited insights into user experiences and journeys on online platforms, see e.g. "The Christchurch Call to Action to Eliminate Terrorist and Violent Extremist Content Online". Consequently, archiving user journeys amid algorithmic decisions becomes essential to understand individual and group dynamics, addressing a salient need in multiple contexts, such as e-commerce, web search, and social media (Schafer, Truc, and Badouard 2019).

While some recent approaches proposed means for users to collect their personal web experience (Kiesel et al. 2018), this chapter focuses on the challenges arising from the need of global and systematic archival means, with a specific focus on a concrete use case, the

Erwan Le Merrer, CNRS, France, erwan.le-merrer@inria.fr, 0000-0001-8344-2135
Camilla Penzo, PEReN, France, camilla.penzo@finances.gouv.fr
Gilles Tredan, CNRS, France, gtredan@laas.fr, 0000-0003-4473-4332
Lucas Verney, PEReN, France, lucas.verney@finances.gouv.fr, 0000-0002-1361-1703

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Erwan Le Merrer, Camilla Penzo, Gilles Tredan, Lucas Verney, *Challenges in archiving the personalized web*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.10, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 79-94, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

YouTube recommender (Covington, Adams, and Sargin 2016). We highlight and contemplate the complex interplay of methodological and technical decisions required to collect a personalized web. Emphasizing the combinatorial explosion of different web versions—each tailored to a specific user profile—we underscore the unobservable nature of these variations. Dealing with the personalization of the web is necessary to accurately capture the user experience surfing the contemporary web, but it also raises several methodological and technical challenges.

1.1 A computer scientist take on archiving personalization

This chapter reflects our position as researchers actively engaged in the technical aspects of auditing online platforms. This nascent research field is at the crossroads of several computer science fields, such as information retrieval, data science and security by certain aspects. As such, our position inherently carries a technical bias that we humbly endeavor to overcome in the development of this chapter. We believe that the outcomes of our (technical) experience, navigating the intricacies of personalization layers omnipresent on major platforms, have implications that reach beyond the realms of auditing and our technical expertise.

Defining platform personalization is a straightforward task; it involves tailoring the content suggested to users based on their past behavior and (estimated) preferences. However, it is important to recognize that personalization encompasses various practices.

We can distinguish between coarse and fine-grained personalization. An example of coarse-grained personalization is the automatic selection of the user interface language based on their inferred location (e.g. displaying an interface in French to users with a French IP address). Coarse-grained personalization is a broad approach that uniformly impacts large sets of users. Primarily, such personalization influences how contents are displayed on the interface rather than the selection of displayed contents. In contrast, fine-grain personalization aims to predict which content will likely appeal to each user. An illustration of this is Twitter's algorithmic Timeline (Bandy and Diakopoulos 2021). Implementing this type of personalization requires a sophisticated platform mechanism. Firstly, the platform observes a user's reactions to specific content, such as monitoring where the user's mouse hovers or tracking which videos were watched entirely versus those that were quickly discarded. These observations are then stored and the platform transforms them into criteria for selecting the most relevant content to present next. Throughout this chapter, we will use the term *user profile* to denote the information the platform possesses about a given user.

A closely related, yet distinct concept is *contextual recommendation*. Contextual recommendation selects items to present to a user based on the item currently being 'consumed', rather than depending on the user's past item consumption. The conceptual difference is fundamental, and aligns with a valuable mathematical abstraction, as contextual recommendation adheres to a Markovian model where the future is independent of the past, given the present. In practice, however, observing this distinction proves challenging. Modern platforms typically generate a set of 'hybrid' recommendations that rely on both the context (the current item) and the user's past history (Le Merrer and Tredan 2018). The thin line separating the two becomes even more blurred when

considering that contextual recommendations are (often) computed using techniques like collaborative filtering (exemplified by phrases such as "users watching X also watch Y"), which rely on users' watch history to assess content similarity. Consequently, in platforms using hybrid recommendations, a user's history may contain items originating from both past contextual and personalized recommendations, establishing a mutual induction between the two recommendation types.

1.2 A combinatorial explosion of versions of the web

Consider a hypothetical platform offering 100 items. To collect contextual recommendations for each item by visiting its page, one would need 100 visits. Now consider the platform incorporating personalization, where visitors receive recommendations based on the two last visited items. In this scenario, observing recommendations associated with all items, one would need to make a staggering 10,000 visits. If the platform uses the last five item visits to compute recommendations for a given item, an exhaustive observation would require 10 billion visits. What was once a modest website transforms into an intricate personalization labyrinth.

This rough estimation underscores two fundamental and technical challenges inherent in archiving a personalized web. The first challenge is evident: the exponential number of visits required for exhaustive exploration renders such thoroughness *practically unattainable*. The second challenge, more nuanced, involves the need for *certain assumptions* about the internals of the recommendation system to conduct such analysis (e.g. the number of previously visited items influencing the user recommendation for the currently visited item).

While these challenges are technical in origin, we contend that their resolution cannot be solely technical. Archiving, and especially web archiving, grapples with the difficult questions of archive curation and selection (Milligan, Ruest, and Lin 2016).

1.3 Data collection setup and terminology

We consider the conventional operational framework for web archiving, wherein robots (hereafter *bots*) gather data from the public web pages of the targeted website. The term *platform* denotes an online website hosting one or more *algorithms* or models with which our bots interact. An example of such models is YouTube's recommendation algorithm, responsible for personalizing video content for users. In our scenario, we assume a lack of agreement with the observed platform, implying that no application programming interface (API) is accessible for data collection, nor for collecting users profiles or the recommendations they receive. While we will discuss alternative approaches, throughout this chapter we will refer to the use of bots for the extraction of data from online platforms. These bots are programs in the form of scripts designed to automate specific data extraction tasks, such as emulating a user on a platform to access and extract the personalization proposed to that user.

1.4 When personalization becomes profiling

While emphasizing the need to archive personalization in today's web, it becomes

imperative to discuss how personalization has now evolved into a much more intrusive practice referred to as *user profiling*. The ubiquity and popularity of mobile versions of online platforms have become increasingly pronounced in our daily lives. The vast majority of mobile users install apps aligned with their interests, needs, and daily routines, facilitating a highly refined personalization process. Companies are now capitalizing on their ability to accurately profile mobile users, see e.g. (Farseev et al. 2020), asserting that they enhance user experiences or make lifestyle improvements, when in reality, their primary objective is finely tailored advertisements. User profiling through mobile applications involves a chain of processes, starting with the analysis of user data collected through the application. This analysis exploits correlations between the application's usage patterns and the user's personality traits, reaching a point where the platform or mobile app producer can predict the user's most personal characteristics (Gustarini et al. 2016; Xu et al. 2016). Effective profiling can transcend demographics, personal interests and lifestyles can be inferred, delving into personality traits and psychological states (Zhao et al. 2019) as well. With psychological profiling data, influencing user behavior, whether through product sales or other actions, becomes very effective.

To complete the profiling paradigm, data is now considered the new gold, bought and sold by entities with novel business models (data brokers), e.g. (Andrés, Azcoitia, and Laoutaris 2022). These entities connect heterogeneous data from multiple sources to maximize their predictive power and, consequently, their economic value. Clearly, this emerging trend raises concerns about user privacy, but has for the moment only prompted a limited response from civil society (Exodus Privacy; NOYB European Center for Digital Rights 2023). We argue that, in the pursuit of archiving the contemporary web experience, both personalization and profiling should find a space in the records for historians.

2. Motivation: Example on YouTube

We now turn our attention to the study of YouTube's personalization, driven by the fact that 71% of U.S. teenagers reportedly consult YouTube daily ("Teens, Social Media and Technology", 2023).

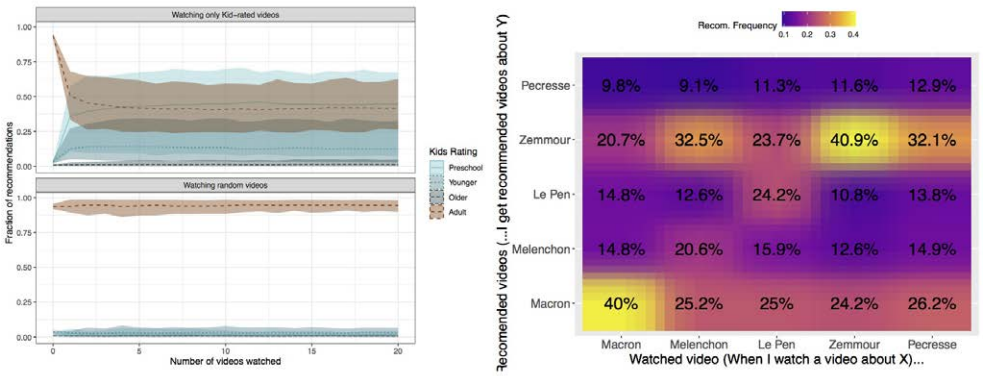
2.1 Measuring personalization using YouTube Kids

"YouTube is one of the largest scale and most sophisticated industrial recommendation systems" (Covington, Adams, and Sargin 2016), a system that has recently been at the center of recent controversies (Ledwich and Zaitsev 2020) due to its potential societal impact. We conducted a bot-driven study of personalization on YouTube (Le Merrer, Tredan, and Yesilkanat 2023), specifically focusing on measuring its consequences, with an application in the context of children recommendations. Notably, we discovered that the video identifiers were consistent across YouTube and YouTube Kids, the latter being a platform tailored for young users under the age of 13. We could thus automatically identify videos labeled as "for kids" on YouTube, i.e. those videos that appear on the platform YouTube Kids, providing a quantitative characterization of the effects of personalization. We used bots with two distinct behaviors: 'Control' bots that start with no profile and watch random videos from YouTube's personalized homepage, and 'kid' bots, that also

start with no history but exclusively watch ‘kid’ videos.

Figure 1 (left panel) presents the evolution of recommendations collected by each bot based on its behavior on the platform, as a function of the number of previously watched videos. Control bots consistently encounter a vast majority of adult (i.e. ‘non-kid’) videos (approximately 97%), whereas kid bots quickly trigger personalization, causing a shift in the recommended content mix towards a majority of kid videos. This significant change occurs within the first three watched videos and stabilizes after the fifth video.

Figure 1. (left) Composition of video recommendations (by age type) based on the number of previously watched videos, starting from empty profiles, for two video consumption profiles. (right) Candidate recommendation matrix for the 5 main candidates during the French 2022 presidential election.



In an archival context, we believe this experience carries two takeaways. Firstly, it is possible to generate and observe personalization using bots. Secondly, ‘control’ bots and ‘kid’ bots exhibit profoundly different encounters with the same YouTube platform after watching a few (< 5) videos. The contrasting experiences of kids and adults could be seen as a stark example of a filter bubble, a concept introduced by Eli Pariser (Pariser 2011). However, these filter bubbles likely lead every user into his or her own subjective journey on YouTube.

2.2 Collecting personalization during the French presidential campaign

Using the same setup as in (Le Merrer, Tredan, and Yesilkanat 2023), we gathered personalization data related to the French 2022 presidential election. While a comprehensive analysis of the collected material goes beyond scope of this chapter, we aim to present a perspective that sheds light on the methodological challenges posed by personalization.

Figure 1 (right panel) illustrates the response of personalization to bots with no prior history, and that randomly watch videos of the ‘French news’ YouTube page (from February 1 to April 10 during the first election round). Approximately 180 times a day, a bot with no profile watches five random videos consecutively from YouTube recommendations on the news page. We observe the titles of all the watched and subsequently suggested videos, considering a video to be about a specific candidate if their name or the name of their political party appears in the title. Titles lacking any mention of a candidate are disregarded. The figure demonstrates what happens when a bot inadvertently watches a video about candidate X: which videos are then recommended to the bot? To exemplify: when I watch a video about candidate Mélenchon, 32.5% of the recommendations (related to any considered candidate) are about candidate Zemmour.

We believe Figure 1 (right panel) vividly illustrates the challenges posed by personalization: after watching a single video about candidate Macron, users receive approximately twice as many recommendations about Macron compared to Zemmour, and vice-versa. Consequently, these users encounter different personalized recommendations, leading to a divergent perspective of the French political landscape on the platform.

Before presenting the conclusions drawn from this observation, it is important to acknowledge the limitations of our approach. Notably, we assign videos to a politician in a straightforward manner (based on name presence in the title) and our quantitative analysis does not delve into the semantics within each mention of a candidate. Hence, a video criticizing a candidate is treated equivalently as a video endorsing them, despite the potential substantial differences in their impact on personalization. Producing such an aggregate figure necessitates compressing millions of recommendations—referring to complex media objects— collected over more than two months into a 5×5 color matrix: it is necessarily partial (incomplete) and potentially biased. Moreover, one may argue that our bot behaviors are overly simplistic and fail to represent actual user experiences (long watch histories and diverse interests, etc.). This objection represents a central challenge that we will address in the subsequent discussion.

In the context of the French election, a mandate regulates the equal division of speaking times among the candidates in the traditional broadcasting media, starting 15 days before the election. This rule, overseen by an independent institution (Arcom), aims to foster fair competition among candidates, implementing equality at the producer level. However, transposing this rule to a personalized platform such as YouTube presents two significant challenges. Firstly, while media operate within well-defined categories requiring licenses, anyone can be a content producer on YouTube. Consequently, binding every content producer to a national rule appears difficult. The second challenge arises when aiming for equality at the receiver level. Since personalization tailors the experience on YouTube for each individual user, assessing an “average" speaking time is nearly impossible.

Elections hold great significance in the political life of democratic countries and arguably possess considerable historical value. However, a clear rule like ‘equal speaking time’ becomes nebulous when applied to personalized platforms. We contend that the same complexity applies to archival policies during elections: selecting content for archiving to provide an accurate retrospective view for historians in our contemporary times requires handling the personalization layer through which we observe online platforms.

3. Challenges

In this section, we present a structured overview of the main challenges encountered in the archiving of journeys on a personalized platform. At a broad level, three classes of challenges emerge: technical challenges arising from the algorithmic nature of the media platform, methodological challenges pertaining to the archivist’s selection of methods for constructing the archival fonds, and usability challenges focused on strategies enabling effective exploration of the archived fonds by future users. In essence, to archive personalization successfully, three fundamental questions must be addressed: how to collect personalization (technical), which aspects of personalization to collect (methodological), and how to present the collected personalization (usability). These questions are interdependent and mutually influence each other. For instance, technical limitations in observing all individual personalizations necessitate methodological choices, and these choices subsequently impact how the archive is then presented to users (usability). Given the interdependence of these issues, we advocate for the integration and juxtaposition of diverse disciplinary perspectives, such as computer science, history, and usability, to construct coherent solutions that facilitate accurate future analyses of our contemporary personalized experiences.

3.1 Technical Challenges

3.1.1 Platform opacity

Modern recommenders leverage a multitude of features, often numbering in the hundreds, to personalize users' experiences (Covington, Adams, and Sargin 2016). These include user-related data, such as demographics and consumption habits. While general techniques for implementing recommenders are publicly available (Gupta et al. 2020), the specific features employed by a corporate recommender in production are typically kept secret. Consequently, understanding the bot-simulated features that influence

personalization becomes a speculative endeavor for the programmer/archivist.

The inability to know and interact with every possible user feature used in the recommendation algorithm places the archivist in a *black-box* interaction scenario with the algorithm. Judging the impact of a particular feature on the resulting personalization necessitates tedious trial and error as illustrated in the previous section.

Despite the technical impossibility of representing every detail of a real user, we believe that the coarse-grain traits of simulated user profiles, precisely defined in the subsequent sections, can yield valuable insights. Simulated user profiles aim to represent user profiles of interest in a coarse manner, rather than ultra specific ones as exemplified by Mozilla's approach to highlight the existence of online personalization (Mozilla 2020): with 'TheirTube', they showcases ultra-coarse profiles such as 'liberal' or 'climate denier', asserting that their watch history encapsulates these personas and thus large user categories.

In order to craft synthetic yet more relevant profiles, discussions with sociologists and statisticians become crucial in crafting representative sets of personas, which can then be presented to algorithms through bots. Identified classes of people experiencing discrimination are also vital to extract personalization for further research into potential bias.

We note that, following the *Digital Services Act, 2022*, research endeavors have begun questioning the possibility of inferring which features impact algorithmic decisions (Rastegarpanah, Gummadi, and Crovella 2021), with the aim of exposing objectionable behaviors.

3.1.2 Frugality in load-responsible and non-interfering data extraction

When crawling a static website for archival purposes, the resulting server load is proportional to the disk space required to implement the website. The scenario shifts significantly, however, when dealing with dynamic websites using personalization, as they are designed to generate, filter, or sort vast amounts of content tailored to users with the aim of prolonging their stay on the website (Covington, Adams, and Sargin 2016). Consequently, crawls and data collections can be virtually endless, allowing bots to navigate through an intricate maze of personalized content. For this reason, frugality becomes a crucial practical consideration in the collection process, ensuring not only a respectful interaction with the platform infrastructure, but also for extracting a manageable amount of data for archival purposes.

Drawing a parallel with the general principle of minimal interaction with an object of study *in vivo*, we emphasize the necessity for frugal extraction.

Avoid loading platform infrastructures

Personalization on platforms has evolved to rely predominantly on complex machine learning models (Covington, Adams, and Sargin 2016). Consequently, engaging with these platforms entails compute-intensive processes in comparison to their static counterparts. When using bots for measurements and data extractions, it is imperative to consider the resulting load on the platforms to ensure responsible operation. Specifically, interactions should not disrupt the platform's service by employing overly heavy machinery to achieve collection objectives.

To avoid such disruptions, platforms commonly adopt defensive measures as rate-limiting mechanisms (Cloudfare 2023). Data extraction must accordingly account for these considerations by estimating what is tolerable for the platform.

Avoid bias in observed recommenders

Data extraction should ideally be conducted without interfering significantly with the recommender. Unlike platforms serving static websites, modern algorithms and models continuously track and adapt for up-to-date personalization, introducing the likelihood that bot actions become integrated into the functioning of the recommender, through such mechanisms as re-training (fine-tuning) based on user activity logs.

The degree of bias introduced is directly proportional to the similarity between bot actions and user actions. To exemplify the point, and at the other extreme, offensive bots may engage in *poisoning* attacks (Fang, Gong, and Liu 2020), interacting with specific items, to prompt the recommender to promote them to a larger audience. It is worth noting that this philosophy of 'just enough' interaction aligns with legal considerations, such as in the European legal system, where the data collection infringement (breaching terms of service for instance) by an auditor to collect evidence must be proportionate to support a given claim (Le Merrer, Pons, and Tredan 2023).

Avoid being sand-boxed

In a tactic infamously illustrated by the 'dieselgate' scandal, certain operators may be inclined to detect and create specific favorable versions of their systems during regulatory audits, a practice known as 'deceptive manipulation' (Siano et al. 2017). This behavior could extend to archival initiatives.

While it is essential for bots to behave in a manner indistinguishable from legitimate human-operated accounts (Cresci 2020) to avoid being detected, the archival context introduces unique challenges. Simulating a user with a bot requires obfuscation to effectively trigger and collect accurate personalization. Consequently, and depending on the targeted platform, bot actions may extend beyond mere metadata collection. For example, on platforms like YouTube, bots might emulate video visualization to conceal their true nature. Although this incurs significant traffic generation, it may be deemed unavoidable to achieve effective personalization thus the data extraction goal.

3.2 Methodological Challenges

3.2.1 Realism and representativity

Data collection from users and associated limitations

Common practices for collecting data on how online platforms personalize the user's experience include data acquisitions (Hosseinmardi et al. 2021) and data donations (Ohme and Araujo 2022), where users willingly share or sell data related to their personalized experiences on specific platforms. This can happen through the use of a dedicated plugin in their web browser, see for example the 2017 ProPublica article. While this approach provides valuable information for archivists, it unfortunately introduces significant

problems.

The first challenge arises from the widespread (and still growing) use of mobile applications to access platforms, replacing the conventional web browser access. These applications, tightly controlled by platform providers, conveniently prevent data extraction, and mobile operating systems do not support the use of plugins. Additionally, the scale of reaching and persuading a large audience to participate in a common archival objective proves complex and often costly. Consequently, the data obtained may not be sufficiently representative for upstream analysis by researchers, leading to potential biases since those willing to install plugins are likely tech enthusiasts, representing only a specific subset of society.

Personalization relies on platform algorithms applied to user profiles, containing the history of user actions. However, gathering data from users does not ensure the completeness of data in this intricate relationship presenting challenges akin to any data collection in a vast array of possibilities. This completeness is essential for performing unbiased and meaningful analyses of collected data donations.

Lastly, as personalization exposes users' tastes and habits, raising concerns about privacy, compliance with legal requirements becomes a critical consideration. For a detailed discussion on the impact of the nature of the collected data on legal obligations, please refer to Le Merrer, Pons, and Tredan (2023).

Personas from simulated users

Personas, in the context of simulated users, refer to users simulated by bots with a well-defined agenda: persona x might simulate on YouTube a video game enthusiast residing in the USA, while persona y might simulate a French individual using YouTube as a news source. Scripting allows these bots to exhibit various behaviors, employ geographically distributed IP addresses, and interact with the platform incorporating daily habits, for example. The art of crafting advanced bots lies in constructing the most realistic interactions to convincingly impersonate specific user types (Cresci 2020). Control conditions can be established to ensure that programmers accurately trigger personalization with their bots (Le Merrer, Tredan, and Yesilkanat 2023).

These bots address some of the challenges associated with obtaining personalization data from real users. They offer full control over the actions taken, directly linked to systematically collected personalization. Bias is minimized, as programmers control the history of actions and metadata associated with all their bots.

Conveniently, bots are more straightforward to set up than recruiting real users. Bringing the analysis to a larger scale relates only to the cost of hosting these scripts and the data they generate. Furthermore, there are no legal issues concerning personal data (at least in the European Union, as exposed by Le Merrer, Pons, and Tredan (2023), as the personalization of the collected data does not involve real individuals. However, a drawback is the inability for programmers/archivists to ascertain whether their bots have been detected by the platform. This introduces the possibility that the platform might willingly treat these bots differently, potentially offering similar personalization as real users with comparable profiles or occasionally biasing their personalization as it sees fit.

In the following section, we delve into the challenges associated with these personas,

and their role in extracting personalization.

Data collection from simple and specific actions

In an alternative scenario, an archivist may find the need to extract personalization data from profiles characterized by clear histories and routine actions. These actions are clearly not aimed at approaching user behavior, but rather focus on extracting consistent data across time. Consider the recommendations made to a profile diligently visiting YouTube's news page every day at noon, or those made in response to a profile limited to entering a set of predefined words of interest in the search bar. Despite the simplicity of these scenarios, they allow for precise tracking of the recommender system's evolution on the platform.

In this simplified case, the archivist might opt for a blank user, i.e. a user profile devoid of any history or prior interactions with the platform. The recommendation algorithm would not be influenced by previous choices, with the aim of having recommendations from the platform in the most neutral as possible scenario.

Another synthetic data extraction approach involves a one-shot, yet potentially comprehensive, gathering of personalization data in response to well-defined sequential actions on the platform. This approach can serve as a basis for auditing a specific aspect of the recommender at a given point in time.

3.2.2 Mainstream vs. fringe profiles

Personalization can be envisioned as a vast space, like a country, where each potential user profile corresponds to an address. Given the impossibility of exhaustively exploring this space, a deliberate selection, or sampling, must be made: where should the focus of observation lie? While virtually any focusing strategy is possible, we briefly introduce two paradigmatic ones.

The first, which we term mainstream, involves focusing the observation of personalization on the most prevalent profiles, those with the most common tastes and behaviors among the user population. In our metaphorical country of personalization, this corresponds to directing sampling towards densely populated areas, such as the capital city. The primary advantage of this strategy lies in its efficiency: each personalization is likely to capture the experience of a substantial user base. Randomly sampling inhabitants of Greece would yield roughly one third residing in the Athens region. Likewise, programming bots to watch videos suggested to an empty profile at random would likely result in mainstream tastes.

The clear drawback of this strategy is its potential to overlook what is not mainstream and which could hereafter be referred to as 'fringe' personalization. This pertains to how personalization influences users who are not representative of the overall user population. For example, in the Facebook-Cambridge Analytical data scandal (Insider 2019), Cambridge Analytical targeted highly specific profiles that diverged from the mainstream. Similarly, the rabbit-hole phenomenon, often studied as a fringe personalization regime, focuses on particular sub-populations, such as anti-vaccine advocates, conspiracy theorists, and far-right movements. A mainstream-only archival strategy would not support such studies, even though they hold value for archiving.

3.3 Usability Challenges

Once technical and methodological solutions have been developed, a final challenge lies in determining the appropriate methods for exploring the collected personalized data (as highlighted in Kelly et al. 2018, and put in relation with the Wayback Machine). A general approach would be to target the most accurate browsing experience, allowing future archive users to closely experience the mechanics of contemporary systems. However, implementing such a system would require significant efforts in emulating the logic of each target website. For example, TikTok and YouTube obey different browsing mechanisms and each requires recreation. Moreover, the archive is destined to be an imperfect copy of the original platform, capturing only a fraction of the website realistically, and unable to reproduce the complete dynamics of these social networks.

An opposing approach could aim for a unified presentation enabling future archive users to compare media platforms on an identical basis, with an implicit emphasis on content rather than presentation.

A central ergonomic challenge lies in navigating personalization in itself. While the Wayback Machine provides a suitable slider for exploring the (continuous though discretely sampled) temporal dimension of a web archive, envisioning an interface for exploring personalization poses a unique question. For archives based on synthetic profiles, TheirTube[1] prompts visitors to select one of the personas used for collection. However, no such solution exists for archives based on (real) data collections directly from users. This distinction illustrates how a usability approach is contingent on the technical and methodological decisions that shaped the personalization data collection.

4. Conclusions and open questions

In this chapter, we assert, from our technical perspectives, that personalization poses a challenge to traditional web archiving methods. We demonstrate how personalization impacts data collection on YouTube and the technical challenges associated with its analysis. Our data collection on YouTube emphasizes that the notion of a universal web no longer holds. There is no singular version of YouTube, as each user is presented with content tailored to their past actions or user profile characteristics. We contend that an effective archival strategy must include the archiving of contemporary personalization, in a consistent manner, and in addition with proposals to leverage ‘emergency’ and focused archiving of some platforms during important events for instance (Schafer, Truc, and Badouard 2019). Addressing this challenge raises several technical, methodological, and usability issues, such as how to manipulate personalization, which personalized versions to archive, and how to present this personalization to future archive users. The interconnected nature of these problems underscores the conclusion of this chapter: the need to integrate and reconcile diverse disciplinary perspectives (computer science, history, usability) to construct coherent solutions facilitating accurate future analyses of our contemporary, personalized times. Although all possible approaches may come with advantages and

[1] https://www.their.tube

drawbacks, we believe it is crucial to define a set of good practices facing the necessary archiving of personalized content.

While our primary focus has been on web archiving, personalization also impacts the information disseminated to users through mobile applications (apps). The technical opacity inherent in mobile apps compared to web pages adds an extra layer of complexity. Due to the increased collection of personal information through these mobile apps, personalization becomes extremely efficient and is referred to as user profiling.

We argue that, with the goal of archiving the user experience in today's web interactions, both personalization and profiling should be integral to the archival process, for they are deeply embedded in our digital lives and should be preserved for the benefit of historians.

Finally, while our current focus lies in understanding how personalization impacts our methods for documenting the history of the web, we believe that the very process of personalization (along with profiling) ought to be studied as a historical phenomenon, thereby recognizing its central role as a contemporary opinion-maker.

References

Azcoitia, Santiago Andrés, and Nikolaos Laoutaris. 2022. “A Survey of Data Marketplaces and Their Business Models.” *arXiv*. https://doi.org/10.48550/ARXIV.2201.04561.

Bandy, Jack, and Nicholas Diakopoulos. 2021. “Curating Quality? How Twitter’s Timeline Algorithm Treats Different Types of News.” *Social Media + Society* 7 (3).

Cloudfare. 2023. “What is rate limiting? | Rate limiting and bots.” https://web.archive.org/web/20240424000000*/https://www.cloudflare.com/learning/bots/what-is-rate-limiting/.

Covington, Paul, Jay Adams, and Emre Sargin. 2016. “Deep Neural Networks for Youtube Recommendations.” In *Proceedings of the 10th Acm Conference on Recommender Systems*, 191–98.

Cresci, Stefano. 2020. “A Decade of Social Bot Detection.” *Commun. ACM* 63 (10): 72–83.

Eg, Ragnhild, Özlem Demirkol Tønnesen, and Merete Kolberg Tennfjord. 2023. “A Scoping Review of Personalized User Experiences on Social Media: The Interplay Between Algorithms and Human Factors.” *Computers in Human Behavior Reports* 9: 100253.

Eslami, Motahhare, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. ““I Always Assumed That I Wasn’t Really That Close to [Her]’: Reasoning About Invisible Algorithms in News Feeds.” In *Proceedings of the 33rd Annual Acm Conference on Human Factors in Computing Systems*, 153–62. CHI ’15. New York, NY, USA: Association for Computing Machinery.

NOYB European Center for Digital Rights. 2023. “How Mobile Apps Illigally Share Your Personal Data.” https://web.archive.org/web/20240424000000*/https://noyb.eu/en/how-mobile-apps-illegally-share-your-personal-data.

Fang, Minghong, Neil Zhenqiang Gong, and Jia Liu. 2020. “Influence Function Based Data Poisoning Attacks to Top-N Recommender Systems.” In *Proceedings of the Web Conference 2020*, 3019–25.

Farseev, Aleksandr, Qi Yang, Andrey Filchenkov, Kirill Lepikhin, Yu-Yi Chu-Farseeva, and Daron-Benjamin Loo. 2020. “SoMin.ai: Personality-Driven Content Generation Platform.” *arXiv E-Prints*, November, arXiv: 2011.14615.

Gupta, Udit, Carole-Jean Wu, Xiaodong Wang, Maxim Naumov, Brandon Reagen, David Brooks, Bradford Cottel, et al. 2020. “The Architectural Implications of Facebook’s Dnn-Based Personalized Recommendation.” In *2020 Ieee International Symposium on*

High Performance Computer Architecture (Hpca), 488–501. IEEE.
Gustarini, Mattia, Marcello Paolo Scipioni, Marios Fanourakis, and Katarzyna Wac. 2016. "Differences in Smartphone Usage: Validating, Evaluating, and Predicting Mobile User Intimacy." *Pervasive and Mobile Computing* 33: 50–72.
Hosseinmardi, Homa, Amir Ghasemian, Aaron Clauset, Markus Mobius, David M Rothschild, and Duncan J Watts. 2021. "Examining the Consumption of Radical Content on Youtube." *Proceedings of the National Academy of Sciences* 118 (32): e2101967118.
Business Insider. 2019. "The Cambridge Analytica Whistleblower Explains How the Firm Used Facebook Data to Sway Elections." https://web.archive.org/web/20240424000000*/https://www.businessinsider.com/cambridge-analytica-whistleblower-christopher-wylie-facebook-data-2019-10?r=US&IR=T.
Kelly, Mat, Justin F Brunelle, Michele C Weigle, and Michael L Nelson. 2013. "A Method for Identifying Personalized Representations in Web Archives." *D-Lib Magazine* 19 (11–12).
Kiesel, Johannes, Arjen P de Vries, Matthias Hagen, Benno Stein, and Martin Potthast. 2018. "WASP: Web Archiving and Search Personalized." https://ceur-ws.org/Vol-2167/paper6.pdf
Ledwich, Mark, and Anna Zaitsev. 2020. "Algorithmic Extremism: Examining Youtube's Rabbit Hole of Radicalization." First Monday.
Le Merrer, Erwan, Ronan Pons, and Gilles Tredan. 2023. "Algorithmic Audits of Algorithms, and the Law." *AI and Ethics*, 1–11.
Le Merrer, Erwan, and Gilles Tredan. 2018. "The Topological Face of Recommendation." In Complex Networks & Their Applications Vi: *Proceedings of Complex Networks 2017 (the Sixth International Conference on Complex Networks and Their Applications),* 897–908. Springer.
Le Merrer, Erwan, Gilles Tredan, and Ali Yesilkanat. 2023. "Modeling Rabbit-Holes on Youtube." *Social Network Analysis and Mining* 13 (1): 100.
Milligan, Ian, Nick Ruest, and Jimmy Lin. 2016. "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.
Mozilla. 2020. "Political Advertisements from Facebook." https://web.archive.org/web/20240424000000*/https://foundation.mozilla.org/en/blog/step-inside-someone-elses-youtube-bubble.
Ohme, Jakob, and Theo Araujo. 2022. "Digital Data Donations: A Quest for Best Practices." *Patterns* 3 (4).
Pariser, Eli. 2012. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin Books.
Powers, Elia. 2017. "My News Feed Is Filtered?" *Digital Journalism* 5 (10): 1315–35.
Exodus Privacy. "Exodus Privacy Analyzes Privacy Concerns in Android Applications." https://web.archive.org/web/20240424000000*/http://https://exodus-privacy.eu.org/.
ProPublica. 2017. "Political Advertisements from Facebook." https://web.archive.org/web/20240424000000*/https://www.propublica.org/article/help-us-monitor-political-ads-online.
Rastegarpanah, Bashir, Krishna Gummadi, and Mark Crovella. 2021. "Auditing Black-Box Prediction Models for Data Minimization Compliance." *Advances in Neural Information Processing Systems* 34: 20621–32. https://proceedings.neurips.cc/paper_files/paper/2021/file/ac6b3cce8c74b2e23688c3e45532e2a7-Paper.pdf
Digital Services Act. 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending

Directive 2000/31/EC (Text with EEA Relevance). OJ L. https://web.archive.org/web/20240424000000*/http://data.europa.eu/eli/reg/2022/2065/oj/eng.

Salganik, Matthew J., and Duncan J. Watts. 2008. “Leading the Herd Astray: An Experimental Study of Self-Fulfilling Prophecies in an Artificial Cultural Market.” *Social Psychology Quarterly* 71 (4): 338–55.

Schafer, Valérie, Gérôme Truc, Romain Badouard, Lucien Castex, and Francesca Musiani. 2019. “Paris and Nice Terrorist Attacks: Exploring Twitter and Web Archives.” *Media, War & Conflict* 12 (2): 153–70.

Schmidt, Jan-Hinrik, Lisa Merten, Uwe Hasebrink, Isabelle Petrich, and Amelie Rolfs. 2019. “How Do Intermediaries Shape News-Related Media Repertoires and Practices? Findings from a Qualitative Study.” *International Journal of Communication* 13 (0). https://web.archive.org/web/20240424000000*/https://ijoc.org/index.php/ijoc/article/view/9080.

Siano, Alfonso, Agostino Vollero, Francesca Conte, and Sara Amabile. 2017. “‘More Than Words’: Expanding the Taxonomy of Greenwashing After the Volkswagen Scandal.” *Journal of Business Research* 71: 27–37.

“Teens, Social Media and Technology.” 2023. Pew Research Center. https://web.archive.org/web/20240424000000*/https://www.pewresearch.org/internet/2023/12/11/teens-social-media-and-technology-2023/.

“The Christchurch Call to Action to Eliminate Terrorist and Violent Extremist Content Online.” n.d. https://web.archive.org/web/20240424000000*/https://www.christchurchcall.com/assets/Documents/Christchurch-Call-full-text-English.pdf.

Xu, Runhua, Remo Manuel Frey, Elgar Fleisch, and Alexander Ilic. 2016. “Understanding the Impact of Personality Traits on Mobile App Adoption – Insights from a Large-Scale Field Study.” *Computers in Human Behavior* 62: 244–56.

Zhao, Sha, Shijian Li, Julian Ramos, Zhiling Luo, Ziwen Jiang, Anind K. Dey, and Gang Pan. 2019. “User Profiling from Their Use of Smartphone Applications: A Survey.” *Pervasive and Mobile Computing* 59: 101052.