International Journal of Communication 20(2026), 993–1012 1932–8036/20260005

The Quotidian Web and the Accidental Archive

ETHAN ZUCKERMAN
RYAN MCGRADY
University of Massachusetts Amherst, USA

Video hosting sites like YouTube are commonly understood through their professional creators and viral content, but they are also rich, global repositories of cultural memory and everyday life. However, their opaque, algorithmically optimized, commercial structure poses challenges to research. We argue that such platforms function as "accidental archives" that capture details of quotidian life that often escape curatorial intention. To consider the unique insights into daily life that such archives preserve, we examine two other accidental archives: a collection of late 20th-century photos and the preserved ruins of Pompeii. We present a four-part mixed-methods approach to researching quotidian video: solving the technical problem of random sampling, conducting metadata analysis, using additional computational means to augment data, and qualitatively analyzing video content.

Keywords: YouTube, research methods, social video, archives

YouTube and other platforms that host online video are not just central nodes in the participatory Internet. We propose considering them as "accidental archives," incomplete records of everyday life that could provide an irreplaceable resource for scholarship on the evolution of culture and human behavior. While accidental archives are valuable, they are also complicated to study, hard to navigate, and fraught with ethical concerns.

We begin by examining two accidental archives—a collection of amateur photos held by the OSA Archivum in Budapest, Hungary, and the preserved Roman city of Pompeii—to understand the value of these unique resources. We then consider YouTube as an accidental archive, commonly misunderstood due to consumer-side bias, generalizations from popular content, and cultural myopia, but newly accessible for archive research through our random sampling technique.

Notably, we believe a random sample offers the ability to study quotidian video, which provides both diverse, global portraits of everyday life and a perspective on how the advent of normalized, everyday video depicts and transforms societies. We argue that, while happenstance has made these collections

Ethan Zuckerman: ethanz@umass.edu
Ryan McGrady: rmcgrady@umass.edu
Date submitted: 2025-07-29

accessible, we should work to preserve—and archive in a more traditional sense—representative samples of YouTube, TikTok, and other platforms as resources for scholarship now and in the future. To do so, we describe the four components of our mixed-methods approach to research: producing a random sample; analyzing the sample's metadata to search for patterns and identify areas for further investigation; augmenting the data with language detection, topic modeling, and other tools; and collaborating with culturally knowledgeable scholars to undertake qualitative research.

"I Only Want the Good Ones"

The Goldberger House is a handsome, if unflashy, edifice in Budapest built to house the corporate offices of the Goldberger textile factory. It survived World War II largely unscathed and was nationalized by the Hungarian government in 1948. In 1981, it became the Konsumex Dollar Store, selling goods otherwise inaccessible in Hungary that could be purchased only with hard currency (The Goldberger House, 2011). The Goldberger's huge glass roof, spanning an inner courtyard, now covers the Vera and Donald Blinken Open Society Archives (often abbreviated as the OSA Archivum), an extraordinary collection of post-World War II media documenting the history of the Eastern Bloc, propaganda, samizdat, and the 1956 Hungarian revolution.

The Archivum contains documents of obvious historical value, including the archives of Radio Free Europe and Radio Liberty, of the Index on Censorship, and the personal papers of Hungarian historical figures, but the most extraordinary part of the collection may be a set of a quarter million amateur photographs that depict life in Hungary between World War I and 1988 ("HU OSA 320-2," n.d.). Some of these images were donated by individuals sharing family photo albums, but most arrived via a more roundabout route. State-owned photo developer Főfotó Vállalat maintained massive rolls of photos that had been flagged by technicians as having flaws ("HU OSA 320-2," n.d.). Archivum archivist Zsuzsa Zadori explains:

> When an individual walked into [Főfotó], she/he was to check a box on the request form: "I want ALL the images," or "I WANT ONLY THE GOOD ONES." Those opting for the second never had the chance to see their bad/faulty/out of focus/overexposed etc. images in print. They were discarded by the photoshop machine operators. Some of these "bad" images were later rescued from the garbage bins by a group of media/visual studies researchers. (Z. Zadori, personal communication, January 18, 2024)

One hundred thousand of these faulty images from 1985 to 1988 exist on massive, uncut rolls of photo papers, while a smaller subset have been cut into individual images. An even smaller set is available online, blurred to remove identifying facial features, as none of the subjects of the photos gave permission for their images to be archived or could have imagined their discarded photos would become a historical archive ("Rontott fotók," n.d.).

Zadori describes the collection as "a silent archive," problematic for researchers because there are no historical metadata for any of the images, no permission from the creators of the photos or their subjects, and thorny copyright issues surrounding each photo. There are endless questions a researcher might ask of

https://doi.org/10.65476/ppzen716

the people in the photos, but they cannot be asked, for reasons including practicality and privacy. Despite those challenges for research, the photos collectively represent a picture of life in Hungary just before the 1989 Rendszerváltás ("system change") with detail and comprehensiveness that an intentionally curated archive would have a hard time achieving.

A scholar of Hungary's late-socialist period could find countless narratives among these images. Even a casual perusal reveals contrasts—plain, featureless, functional exteriors in public, but private spaces that are richly patterned and colored, apparent despite the years degrading the photos' cyan dye. A woman in a polka-dot blouse and flowered apron stands in front of Toile de Jouy curtains and a modernist striped bedspread. Other interiors feature faux tiger rugs, a wall-sized photo of a waterfall, a bar with a tiled backsplash and upholstered vinyl. To us, it reads as domestic individualism in a society that often required uniformity, but a Hungarian who lived through the period might read it entirely differently or find narratives in the Skodas and Ladas on the streets or the kerchiefs children wore as part of their school uniform.

The Főfotó archive is not only silent, but accidental. T. R. Schellenberg (1956) defined archives as "those records of any public or private institution which are ajudged worthy of permanent preservation for reference and research purposes and which have been deposited or have been selected for deposit in an archival institution" (p. 16). The Főfotó collection, however, is an archive whose contributors did not realize they were submitting documentation of everyday life for examination decades or centuries hence (neither the photographers nor the technicians who developed and discarded them). The collection took place without curation, and its preservation preceded any judgment about reference or research worth. Ironically, the photos were preserved because they had been discarded.

As an accidental archive, it achieves something that eludes most archives: It provides a picture not of a curated vision of reality, but of the quotidian—of the extraordinary richness of everyday life. It is a valuable resource for understanding the lived Hungarian experience and for teaching young archival scholars how to unpack histories from quotidian objects while navigating the complex ethical issues of unpermissioned data. It is also a useful inspiration that encourages us to ask what we can learn from other accidental collections of quotidian objects.

Fast Food in Pompeii

Ancient Rome is one of the most thoroughly documented eras, from contemporary histories written by Tacitus or Livy to document triumphs and defeats to later epics like Gibbon's (1776–1788/1994) *The History of the Decline and Fall of the Roman Empire*, a cautionary tale about the dangers of losing civic virtues. While these accounts immortalize generations of great (and often greatly problematic) men, they have great lacunae as concerns the experience of ordinary Romans. One of the chief sources of knowledge about the lived experience in the Roman Empire comes not from these careful accounts, but from history's most famous accidental archive.

The small Roman city of Pompeii has arguably been frozen in time at least thrice. In 79 AD, an eruption of Mt. Vesuvius buried Pompeii and nearby Herculaneum under 4–6 meters of volcanic ash. For the few who remained in the city after the eruption began, about 1,500–2,000 of the city's estimated 20,000

residents, death was likely instantaneous as pyroclastic gases swept through the city, felling people in the middle of their homes and workplaces. It was first excavated by survivors, their descendants, and opportunistic treasure hunters who dug through the ash to extract statues and other valuables (Beard, 2008). Eruptions in the 5th and 6th centuries buried the city further (Mastrolorenzo et al., 2002), erasing it from maps until a rediscovery in the late 16th century (de Divitiis et al., 2004).

In the 1870s, archeologist Giuseppe Fiorelli discovered he could fill the cavities in the ash, left by decomposed bodies, with liquid plaster, reconstructing those who perished cowering under their cots, sometimes clutching one another (Dwyer, 2010). However, Fiorelli and his colleagues could not resist the urge to curate an idealized version of reality, manipulating human remains to create casts that better evoke the victims' "spirit and character" (Lazer et al., 2021, p. 104).

The explorers who unearthed much of Pompeii and Herculaneum in the 18th and 19th centuries used methods we would condemn today, tunneling through layers of artifacts to free the mosaics and frescoes that showed the artistic achievements of the time. Often, these explorers would pry works off walls and floors to transport them to museums and private collections. Those left behind were often exposed to the elements and have degraded and faded (de Caro, 2015). Rubble from Allied bombings during World War II buried part of the city a third time, destroying some structures while preserving others.

When Pompeii was a key part of the Grand Tour—the trip gentlemen took through Europe with a learned tutor as an educational rite of passage—the focus of travel to Pompeii was on the masterpieces of art and architecture, which necessarily meant a focus on the grand public spaces and the most opulent private residences. For many decades, archeologists followed suit, but for at least the last half century, they have been more focused on the quotidian, from the contents of storage jars to the writing on the walls.

While formal Latin has been preserved and taught in universities and the Church for decades, less was known about "vulgar" Latin spoken by everyday citizens. At least 11,000 inscriptions have been found on the walls of Pompeii, an extraordinary set of texts that has shed light on matters civic (electioneering was a common reason to write graffiti) and personal. A famous set of inscriptions discovered on a tavern wall depicts taunts between a lovelorn admirer of Iris, the bartender, and a rival for her affections (Benefiel, 2010). Wall writing was common in Pompeii, perhaps because other inexpensive writing materials were uncommon, and some of the most informative writings have been the most ordinary. A study of five lists of expenses, inscribed on different walls across the city, offers insights into how Pompeiians ate, what daily expenses looked like, and the relationships between free men and their slaves, who ate cheaper bread (Bowes, 2021).

Some of the most exciting recent discoveries come from examining Pompeiian ceramics. One major recent site is a thermopolium—basically a fast-food counter where many ordinary citizens would have taken their meals (Cascone, 2021). Ordinary homes in Pompeii were cramped and did not have kitchens, so lower-class citizens ate meals outside their homes, while the wealthy had kitchens and dining rooms. Based on the remnants of food in amphorae (large jars), we know a good deal about what citizens ate and how far their food traveled to reach them.

https://doi.org/10.65476/ppzen716

After years of studying the extraordinary, archeologists now find insight in the most humble of remains: remnants of dates, apples, pulses, hazelnuts, and walnuts, found preserved in sewers, documenting an enormous range of foodstuffs available to Pompeiians (Meyer, 1980; Rowan, 2017). Accounts scratched into walls meanwhile remind us that many citizens survived most days on bread, cheese, oil, and wine. Frescoes and mosaics may have documented the life of the elite, but quotidian remains—like the sewage and graffiti that shed light on economics, trade, social structure, and the health of citizens—are what we need to understand the experience of ordinary Pompeiians.

Buried Under Data

The concept of an accidental archive is not well studied. Amy Tector (2006) gave a detailed account of a decision by the Canadian national archives to catalog works by author Magdalena Eggleston. Though her degree of success did not rise to the archives' typical standards for importance, she had a key advantage over other authors: The archive happened to receive her manuscripts in a collection of works by her famous spouse. Tector (2006) calls it an "almost accidental" archive, using the experience to emphasize the constructedness of literary canons. Stefanie Schulte Strathaus and Vinzenz Hediger published an edited volume in 2023 about accidental film archives, which formed as contingencies, byproducts, omissions, and rescues, presenting them as important not just to preserve the past for its own sake, but to broaden the film canon and influence the future (Strathaus & Hediger, 2023).

Tector's (2006) "almost accidental" archive is like our other (entirely) accidental archives in that its potential was realized subsequent to its possession, but different in that it exists within a structure set up for preservation. Lorena Ramírez-López (2015) focused on this vulnerability of accidental archives, framing them as similar to defunct archives, but without the archival intent. Valuable materials were instead collected while pursuing other goals, typically business, and become imperiled when those other goals fail.

The popularization of personal computers, the Internet, and cloud computing all spurred waves of new interest in archiving, even as the quantity of materials to be archived radically expanded (e.g., Kahle, 1997). In the early days of the Internet, it was common for journalists and scholars to relate the volume of data online to physical media—how many reams of paper to print Usenet message boards or how many historical libraries to hold all the information on the Web (e.g., O'Neill et al., 2003). By the mid-2000s, such comparisons had gone out of fashion. The Internet was not just too big, but too dynamic, with websites moving from static documents to database-generated content. The rise of YouTube in 2005 made clear that analogies to paper libraries were obsolete.

Today, YouTube is the second most popular website in the world, and its visitors spend more time with it than any of the other sites in the top 20 ("Youtube.com Website Analysis for May 2025," 2025). It has become the default video-hosting platform for the Internet, serving an unfathomably wide range of purposes. It has, in addition, become an important source of not just entertainment but also education, gave rise to a large "creator economy," remediated traditional media, and is one of the foremost disseminators of music, news, television, propaganda, and advertising (Burgess & Green, 2018; Cayari, 2011; Snickars & Vonderau, 2009). Through its accumulation of a vast quantity of audiovisual materials by

and about humans, it has taken on another role: YouTube videos are—potentially, contingently—a massive accidental archive.

Moving-image archives have always been among the most difficult to access (Prelinger, 2007), as their materials require specialized and expensive equipment to reproduce and distribute. YouTube, with the help of substantial public and private infrastructural investment, practically erased such limitations—suddenly, videos were accessible anywhere. Already in 2009, archivists like Rick Prelinger (2009) observed a potentially harmful conflict between the public's perception of the platform as an always accessible archive and YouTube's goals, which prioritize accumulation and attention over preservation.

When used by nations and peoples, selective archives function as "a recording of history from a particular perspective [that] cannot provide transparent access to the events themselves" (Manoff, 2004, p. 14). Some parts of a culture are preserved at the expense of others, often to remember a specific representation of a people. David Greetham (1999) argues these "conservational decisions are contingent, temporary, and culturally self-referential, even self-laudatory: we want to preserve the best of ourselves for those who follow" (p. 9). This is not fundamentally a bad thing. Selective archiving allows vulnerable communities to choose how they are represented, rather than being subject to distortions or stereotypes (Jo & Gebru, 2020), and renowned Canadian archivist Terry Cook (2013) saw 21st-century technologies ushering in a new archival paradigm centered on community.

In the Főfotó collection within the Archivum, unusually, key conservational decisions were made not by trained archivists, but nameless photo technicians, concerned that their printers delivered too much magenta or blurred prints. The few that Archivum staff have since digitized and shared online represent curatorial decisions about what might be interesting to a contemporary audience, what might illustrate recent historical change, or perhaps what struck a curator as especially moving or beautiful, but the Főfotó collection as a whole was curated by chance.

The whole video archive of YouTube is accidental in that it accumulated without archival intent, but the way it is organized and accessed is decidedly curated. However, curation on YouTube, which does not view itself as having archival duties, functions not to remember, represent, or valorize, nor to showcase subjective beauty, but to maximize attention according to the rationalities of platform capitalism. YouTube may be vast, but most of what we know about it is based on a small portion of popular videos: the professional creators, viral hits, and other attention-optimized content prioritized by the YouTube recommendation system, which drives most of its traffic. YouTube's default mode of curation provides an impoverished view of the people who use and rely on it.

As a result, the most popular tip of the YouTube iceberg is better understood than the great mass of unpopular content. Some of these reasons relate to an overall consumer-side bias in research: We ascribe importance to what people watch rather than what people upload. Where YouTube is entertainment, people write about the most successful programs; where YouTube is news, people pay attention to the most influential sources; where YouTube is education, view count might be a proxy for quality.

https://doi.org/10.65476/ppzen716

Problems arise from a view of YouTube that prioritizes the video consumer's perspective in general and the most popular content in particular, beyond the simple reality that such a skewed sample necessarily distorts perceptions of the whole. One pitfall is generalizing from popular content, assuming unpopular content follows the same basic forms and logics of popular content, but simply fails to achieve a wide audience. Everyone is presumed to be participating in the creator economy, uploading content with the intention of going viral or building a large audience. As Lauren Berliner (2024) put it, the longtime YouTube tagline "*broadcast yourself* has evolved into a feedback loop that encourages *yourself* to conform to norms" (p. 168). In other words, the rationalities of YouTube's algorithmically organized creator culture have a flattening effect (Chayka, 2024). Regardless, there is also an extraordinary amount of content on YouTube that is not trying to go viral, which is invisible without a broader view.

Another pitfall is cultural myopia—the assumption that the ways people I know use YouTube are how the whole world uses YouTube. If I use YouTube for old movies, childhood cartoons, and occasional home repair advice, I might miss common uses outside of my lived experience, like watching livestreams of religious services. It is likely that anyone researching how people use YouTube will have difficulty thinking outside their personal experience within a culture, and highly likely that they overlook culturally specific uses, like the video-based good morning messages prevalent among South Asian users (Gupta et al., 2019).

There is another reason we understand YouTube through the popular: It is simply easier to study. Popular content is more visible, analyzed by researchers, shared on social media, discussed among friends, and, crucially, recommended by YouTube. As a for-profit company that makes money by maximizing attention, YouTube has a financial motivation to recommend videos that viewers are likely to click. If there are 1,000 equally relevant videos on the same subject, but five have millions of views, the wisdom of the crowd suggests that those five are more likely to engage viewers and keep them on YouTube, racking up ad views. Almost three-quarters of traffic on YouTube is driven by platform recommendations (Kiros, 2022).

Researchers who want a clearer picture of the whole accidental archive, including not just the popular and professional, but the less popular and quotidian, face challenging sampling problems. Scientific claims about a large population made from small samples rely on those samples being representative of the whole, but there is no straightforward way to create such a sample on YouTube. As a result, studies are more often based on opportunistic samples, beginning with a set of known videos or channels and building a sample from there. Properly understanding the everyday uses of YouTube is a technical challenge, not just a methodological one.

The neglect of unpopular content and technical barriers to its study are not unique to YouTube. It characterizes most of the large digital platforms—the small number of large companies that own, influence, and unintentionally archive so much communication and expression. Much as it is easier to study elite Romans—their legacy preserved in written accounts, statues, and palaces—than ordinary citizens, it is easiest to understand the traces of Internet royalty. YouTube archeologists will have little difficulty unearthing MrBeast and Joe Rogan, but for all the importance of these influencers, there is a wealth of

information buried underneath—how everyone else on the planet who uploaded slices of their lives over the last two decades lived, worked, and played.

Understanding YouTube requires different tools than plaster casts used in Pompeii. Media archaeology offers one possible alternative, focusing on forgotten or neglected media forms and examining them to better understand the present. As framed by Erkki Huhtamo and Jussi Parikka (2011), media archaeologists use the language of archaeology, such as "'excavating' media-cultural phenomena," to explain the way they "rummage textual, visual, and auditory archives as well as collections of artifacts, emphasizing both the discursive and the material manifestations of culture" (p. 3). Media archaeologists often examine overlooked media forms rather than overlooked uses of currently popular media, but the approach is useful for two reasons: First, we appreciate its dogged commitment to interdisciplinarity in the study of media; second, it highlights the importance of the media technology itself.

Wolfgang Ernst (2012) does not deal with YouTube directly in *Digital Memory and the Archive*, but he provides a useful perspective on what is necessary to study a platform like YouTube as a form of archive. YouTube is not simply a tool—even calling it a "platform" risks mischaracterizing its real place in the world (Gillespie, 2010)—but a technical system. It is an archive, but before we can treat it as such, we must appreciate the extent to which digital media have transformed what an archive is—dynamic, changing from instant to instant, and "generative itself in algorithmically ruled processuality" (Parikka, 2012, p. 29). Ernst (2012) relies heavily on Foucault's reframing of archives as not neutral storehouses, but productive discursive systems that condition what it is possible to say in each historical context. Archives, however, are increasingly technological, so Ernst encourages us to focus not just on linguistic rules but also on the rules of codes and protocols—archives shaped less by people creating documents and more by what can be technically stored, accessed, and retrieved.

To fully understand such an archive, or the infrastructure of digital memory, technical work is required. A researcher must take it apart, reverse engineer it, and run experiments to figure out enough of its structure to reveal some of the secrets it hides. If a Vesuvian moment freezes YouTube under ash, what are future archeologists likely to find online, spinning up those ancient hard drives? Fiorelli poured plaster of Paris to make visible the traces of those who built Pompeii. We have our own tool of the trade: the random sample.

Studying the Quotidian Web

The quotidian Web—the ways in which people use platforms like YouTube in their everyday lives, before their content is screened, prioritized, recommended, and measured by the platform's algorithms—is understudied. This is in no small part because of technical and practical challenges, but also because of lack of attention. For studies of influence and, of course, of popularity itself, it makes sense to focus on samples with the most views, and popular content is often the default starting point for Web research in general.

That the quotidian Web is understudied does not mean it is unstudied, however. YouTube, in particular, has attracted several scholars interested in peering into the glut of content that languishes in obscurity. Lauren Berliner (2022) has written several pieces on the subject, proposing a "methodology of

https://doi.org/10.65476/ppzen716

unwatched digital media" (p. 1) that challenges metrics-based value judgments of importance, urging researchers to focus more on the "why" of uploads than who watches it. The term she uses to describe "online user-produced audiovisual media that has been uploaded online but not viewed or circulated beyond the maker and their immediate, already-invested audience" is "digital obscura," and it provides insight not just into the culture, behavior, and trends but also the inner workings of how digital media infrastructure operates (Berliner, 2024, p. 165). She draws attention to some other projects aiming to showcase neglected videos, but highlights the difficulty of finding these videos as platforms reduce the ability to avoid algorithmic filtering.

Our multidisciplinary approach to studying the quotidian Web has four levels: First, build a random sample. This typically involves some computational savvy to figure out, as the process will be different for each platform. Second, analyze the sample's metadata to make representative claims about the platform as a whole and to look for patterns for closer study. Third, augment representative data using language detection and transcription, topic modeling and clustering, and other technical tools. Fourth, watch many random videos to appreciate the qualitative dimension of how people use the platform. Apart from beginning with a representative sample, the process is nonlinear. For example, data produced through additional tools will demand reanalysis of patterns in the data and provide new ways to focus qualitative work.

Random Sampling

In 2021, we set out to study harmful speech on YouTube. We obtained a sample of channels that had been flagged for extremism and began training models to classify video transcripts. Though we had access to some of the most problematic content on the platform, the kind of language we were looking for was still difficult to find. If harmful speech is not common in a data set of extremist YouTube channels, how common could it be on YouTube as a whole?

Such questions about prevalence of any content on a platform like YouTube are difficult to answer because the denominators (Zuckerman, 2021) in the equation are typically missing (*n* harmful videos out of *N* total videos). In other words, we could not calculate prevalence because we had no good estimate for the number of videos on YouTube. We were shocked—how was it possible that one of the most important websites in the world, used and relied on by so many people, did not disclose the most basic facts about itself?

Once we asked, "How many videos are on YouTube?" other simple, but unanswered, questions leapt to mind. How many YouTube videos are in English versus how many are in Korean? How many are games, and how many are news? How many views do videos get on average? Are they mostly short or mostly long? We refocused on randomly sampling YouTube to provide these missing figures.

When YouTube's creators began storing videos, they prepared to host an absurd number, indexing each video with a random number between zero and 18.4 quintillion (2^{64}). It would be convenient for us if YouTube simply started numbering videos from zero and incrementing by one for each upload, but as a for-

profit company, obscuring such numbers avoids leaking data that could be used by competitors or regulators.

Instead, platforms often create unique identifiers through hashing, transforming data into a fixed-size string using an algorithm that is difficult to reverse engineer. Some methods incorporate both hashing and non-random data. Twitter, for example, invented a schema called Snowflake to create identifiers that are both partly random and roughly chronologically sortable (Baumgartner, 2019). YouTube's schema precedes Snowflake and appears to be purely random, with no chronological information. This means it is much harder to estimate how many videos are posted to YouTube: Guess a number between one and 2^{64}, check for a video, and repeat until you find enough videos to make claims about representativeness without too much uncertainty. If that sounds easy, remember there are 18,446,744,073,709,551,616 possibilities. For that space to be just 10% full, every human on earth would have to upload about 230 million videos.

It took us more than a billion guesses for every video we could add to our sample. Fortunately, we found some shortcuts, like the ability to search for uppercase and lowercase letters simultaneously, but it still took our university-grade computing cluster months to put together a sample of about 10,000 videos (McGrady et al., 2023). To make life easier for future researchers, we used our random sample to verify another method that was proposed in 2011, but relied on some odd mechanics, calling its randomness into question (Zhou et al., 2011). While it can only return results that include a hyphen in the identifier, it is much faster, and we found it was similar enough to our sample that we recommend it to researchers.

Having solved random sampling for YouTube, we could undertake multiple levels of analysis, described below: finding patterns in metadata, qualitatively analyzing random videos, and training machine learning models to process videos at scale. We are also turning to other platforms, namely TikTok and Sharechat. The way each organizes its content is different, so we must begin the process of analyzing URLs and video identifiers anew. For example, using knowledge that the first half of a TikTok identifier is a timestamp and that the second half includes some non-random data like a data type (video, user, etc.), we were able to produce a representative sample by iterating guesses combining random identifiers and random seconds in time.

Patterns in Data

Armed with a representative random sample, our lab used the sample size, number of guesses it took to find them, and total size of the search space to produce a defensible estimate of YouTube's size: about 15 billion public videos as of mid-2024 (Zheng et al., 2024). With a mean duration of 558 seconds, there are more than 2.3 billion hours of watchable video (over 265,000 years). We received an unusual confirmation via YouTube's press release celebrating its 20th anniversary, announcing a count for the first time: "over 20 billion uploaded videos" (Mahesh, 2025, para. 1). We ran our estimation scripts again the next day, producing a new estimate of 19.4 billion. As YouTube's statement leaves ambiguity—How much more than 20? Does "uploaded videos" include private or removed videos?—and our estimates include only public videos, the total including private and deleted videos is likely well over 20 billion, rounded down to

https://doi.org/10.65476/ppzen716

align with the emerald anniversary. Our year-by-year calculations showed that YouTube is growing at an accelerating pace, especially since 2020, when COVID-19 lockdowns likely led to more people hanging around at home, creating and uploading videos.

Having a random sample allows us to make a range of claims about YouTube as a whole—average video length, view count distribution, frequency of videos over an hour or under 15 seconds, and so on—solving for many missing denominators and distributions (McGrady et al., 2023).

What quickly became clear was just how unrepresentative the experience of an average YouTube viewer is. For all the drama in comment sections, most videos have no comments at all. The median YouTube video has only 41 views (Zheng et al., 2024). Billions have fewer than 10 views. Four percent of YouTube has never been seen by even one person (we suspect this may be because of third-party mobile apps that make it easy to upload low-effort videos or even photos). Chances are, most of the videos you watch have more than 10,000 views; that pool of videos constitutes only 4% of the whole (McGrady et al., 2023). We know the tours of Pompeii focused on its wealthy citizens, and masterpieces do not adequately represent the lives of most of its citizens, so why do we accept a view of YouTube through its own privileged class?

A large enough representative sample can be further subdivided into other random samples—by year, category, and view count—to look for patterns in how people use YouTube and how those uses change over time. Some of our most interesting findings, however, required augmenting the metadata YouTube provides with additional information like the language spoken.

Augmenting Representative Data

YouTube provides metadata, but they did not answer all our questions. An obvious question is which languages and geographies are represented within the corpus. It is tempting to assume, based on the popularity of English on the Web or the company's U.S. origins, that YouTube would be mostly English. It is furthermore easy to assume, mistakenly, that the ways English speakers use YouTube are generalizable to other linguistic communities—a persistent problem in platform research. Indeed, cross-language comparisons are valuable, but rare (Matassi & Boczkowski, 2023).

YouTube's captioning system provides an indication of a video's language, but more than half of videos do not have caption data. We set up a language detection pipeline, ultimately using OpenAI's Whisper (OpenAI, n.d.), which proved effective at transcribing videos (this made sense when it was reported that Whisper was built to ingest YouTube videos as training data; Metz et al., 2024). We work with speakers of various languages to calibrate Whisper so we know which confidence thresholds give us sufficient accuracy.

Augmenting our data set with language, we can see English represents a plurality (28%), but not a majority of YouTube. Spanish, Portuguese, and Hindi, for example, are well represented, and there are about 20 languages that represent at least 1% of the corpus each (200 million videos; Zheng et al., 2024). Segmenting by language, we compared metadata among populations. Comparing English, Hindi, Spanish,

and Russian samples, we found several patterns and anomalies: fewer videos categorized as news in English, for example, and longer videos in Spanish. However, the starkest difference was in just how new, rapidly growing, and short Hindi videos were (McGrady et al., 2025), something we attribute in part to the popularity and subsequent ban of TikTok and that we continue to investigate through qualitative methods, described below.

Not all videos can be accurately transcribed—YouTube video is challenging for transcription software, with cross-talk, background music, and poor sound quality—but our ability to transcribe videos is increasing. With transcripts, we can study content and employ topic-clustering natural language processing methods. Our early experiments demonstrate that we can automatically create identifiable clusters for a range of topics like online gaming and religion. We plan to build topical collections in multiple languages for scholars who can add the necessary linguistic and cultural knowledge and context. We are also exploring topic modeling using other machine learning signals. Image-based clustering, for instance, would help us identify common media forms like video game livestreams or memes.

Representative samples, augmented by transcripts and classifiers, enable a wide range of possible studies. For example, mis-/disinformation researchers seeking to understand the influence of vaccine misinformation understandably focus attention on influencers who reach large audiences, but studying the information's spread requires either consumer-side methods like polling or following the spread of narratives through lower-attention videos. More broadly, with random samples, the ability to parse corpora by time and language, and the ability to identify topics, we can study the spread of ideas, memes, or trends across the Internet over the course of months and years.

Into the Realm of Randomness

Studying the quotidian Web is necessarily multidisciplinary. Obtaining a random sample requires computational creativity, and analyzing the sample requires data science. We see great opportunity in studying the transcripts of videos using machine learning, but even when transcripts are available for qualitative study, they are imperfect and omit context like inflection, multiple speakers, and, of course, visuals. We run the risk of limiting our understanding of the videos through logocentric assumptions about how to study the corpus.

To properly understand the content of the videos and the many ways people use YouTube as uploaders, it is necessary to watch random videos. Our random samples are statistically different from the opportunistic collections of videos scholars typically study, but statistics do not properly communicate just how different the experience of watching random YouTube is. To help potential collaborators understand the weirdness of the space, we compiled a demo reel, selecting a few seconds from a dozen videos at random and editing them together:

- Overexposed living room window, faint music, a voice announcing falling snow
- South Asian religious leader answering on-screen questions in his native language
- Homemade advertisement for a used BMW
- Woman singing a hymn in English in her church

https://doi.org/10.65476/ppzen716

- Minecraft gameplay
- Young girl dancing to Mexican music
- Religious ceremony in an Indian home

Faced with this diversity of language and content, our first step as researchers was simply to try to identify patterns. We drafted a code book, but it quickly became apparent that several of our assumptions were mistaken. Questions about monetization and editing choices, for example, were mostly irrelevant, as little professionally produced content appeared in our samples.

We did find an enormous number of video games, a surprising number of religious services, and countless uses we had not anticipated. It became evident that uploaders of many of the videos were not interested in the creator economy; they were not trying to become the next MrBeast. Many challenged the assumption that uploading to YouTube meant you were trying to "broadcast yourself" to a wide audience at all. Our sample included classified ads, condo board meetings, homework assignments, family birthdays, selfies set to music, and messages to friends. If one person watches a classified ad for a car and buys the car or if a meeting is uploaded for the benefit of those who could not attend and it receives 10 views, the video was a success. If your classified ad or meeting went viral, it is likely that something went very wrong.

Because of language barriers and a lack of cultural context, our dominant experience was one not understanding what we were watching. Thankfully, our metadata analyses provided clues for areas to examine more closely. We formed teams in our lab to study patterns we found in Korean and Hindi. Korean YouTube had more videos categorized as news than other languages we analyzed, and our collaborator saw some evidence of right-leaning podcast-style news commentary, leading to a research project about the relationship between professional and amateur newscasters on YouTube.

From our metadata findings that Hindi videos were newer/shorter and exhibited an unusual pattern of liking relative to views, we are exploring a hypothesis that Hindi users often make public videos for friends and family, not for general audiences, resulting in large numbers of "accidental vlogs" (McGrady & Snehi, 2025). Not only were videos more like private video postcards, but comments also used terms of endearment that indicated spouses or relatives. Another collaborator is conducting interviews to explore a hypothesis about the use of short video with WhatsApp for small group communication in South Asia. Sharing short videos with close friends and family allows easy communication between people with low levels of literacy or who find texting in non-ASCII languages awkward.

Our Hindi project thus involves technical problem solving to create a random sample, data analysis to find patterns in metadata, augmentation of data using language models, forming teams capable of qualitative work to investigate patterns in data, content analysis, and finally ethnography. In this sense, the challenge of understanding quotidian video resembles the challenges of evaluating diets in Pompeii based on fossilized food scraps in a preserved thermopolium. The Pompeii researchers built teams that included archeologists, nutritionists, language experts, biologists, and others to understand what they were seeing. We may need similarly diverse teams to understand the dynamics of livestreaming religious revival events in Bangladesh, one of our future areas of study.

The Ethics of Accidental Archives

Archeologists in Pompeii have a luxury: Their subjects are long dead, and claims made about them are unlikely to lead to a lawsuit. We believe researchers of the quotidian Web have a responsibility more akin to the curators of the "silent" Főfotó archive, who are careful with personally identifiable information about the people in the photos—people who are likely still alive and never consented to being part of the archive. Our tools for sampling YouTube do not retrieve videos that are set to private, but we need to consider that uploaders do not necessarily understand that these videos are public by default or may presume they are too obscure to find (Selinger & Hartzog, 2018).

The assumption that all uploaders want a large audience is baked into the names of projects like the Lonely Web (Veix, 2016). Art projects including IMG_0001 (Walz, n.d.) and Petit Tube (van der Cruyssen, 2011) have made quotidian videos visible through simple algorithmic means: They search for strings like "IMG_" or "MOV_" that commonly appear in file names of videos carelessly uploaded to YouTube. While these services offer glimpses of the strangeness we see in the random videos, users should understand they may be watching something uploaders may not have wanted strangers to see.

Long before the Internet, Erving Goffman (1959) described how people alter the way they present themselves depending on their audience, but doing so has become more complicated as a small number of large platforms have concentrated communication. danah boyd (2008) used the term "collapsed contexts" to explain the effects of Facebook users being limited to only a single, unalterable identity, forcing teenagers to talk with their friends the way they speak with parents or grandparents. Such context collapse can have profound consequences.

To ensure we do not harm someone by exposing their video out of context, we are developing privacy agreements to control access to our samples, meaning we can share lists of videos with researchers who agree not to further disseminate them or connect them with personally identifying information. Nonetheless, we face a basic, but important, aesthetic problem: How can we give audiences for our work a feel for these videos and their content? In presentations of our work, we have used compilations of short clips, but removed identifiers and declined to publish them online. When the BBC wrote an article about our work (Germain, 2025), it contacted video authors and embedded only videos where a creator gave permission. This may be a best practice, but many YouTube uploaders are "silent," unidentifiable and unreachable through YouTube. Relying on permission may also introduce language biases—it may be harder for English-speaking researchers to obtain permission to use Hindi videos, for instance. We seek dialogue with curators of more conventional archives about how to navigate these thorny privacy issues.

Preserving Everyday Life

The potential value of the accidental archive argues for permanent preservation. Our aspiration might be to preserve everything hosted by YouTube or TikTok, but that would require the resources of a company like Google to accomplish and conflicts with the imperatives of profit-driven accidental archives.

https://doi.org/10.65476/ppzen716

Prioritizing the preservation of random samples ensures that, at minimum, representative snapshots persist in the historical record.

Digital files do not degrade in the way the Archivum's photographs have reddened with time, but hard drives do get corrupted or overwritten, and platforms become inaccessible. Consider GeoCities, an early host of Internet homepages that closed abruptly, leading to an extraordinary archiving effort in 2009 (Bril, 2023). Society's most ambitious Internet archiving enterprise, the Internet Archive, is profoundly limited in what it can preserve (Thelwall & Vaughan, 2004). It does wonderful opportunistic work, like archiving Russian TV during the Ukraine war, but cannot cover everything. We need a global effort to create and preserve representative samples of YouTube, TikTok, and other video-hosting sites. This involves developing technical tools to create, augment, and analyze samples, as well as systems to host and distribute samples to researchers who agree to use them ethically. There is also a need for better legal frameworks for ethical public interest research that may technically violate platforms' terms of use and institutional support structures to support and advocate for such research in the absence of new laws.

Numerous archives and museums hold collections of snapshots from the advent of consumer photography in the United States. If we had the means to search across all those collections, we could understand not only details of lives lived in parallel in different parts of the world but also the history of how photography, as a lived practice, was shaped by the millions of people who adopted it and, in turn, how photography reshaped their lives. This archive exists today for video. As videography becomes pervasive and normalized, we can ask how people choose to share, broadcast, and preserve videos of their experience. This is not something we can understand by just preserving the most prominent examples any more than we can understand what was eaten in ancient Rome by studying the frescoes in the wealthiest homes.

The accidental archive—the preservation of Pompeii's poorest homes and residents under ash, the record of Hungary just as it emerged from socialism into a market economy—offers an unprecedented opportunity to understand the evolution of culture. An unsurpassed archive of quotidian video exists today: We need to learn to use it, to preserve it, and to learn from it.

References

Baumgartner, J. M. (2019). *Reconstructing Twitter's firehose*. Retrieved June 1, 2025, from https://docs.google.com/document/d/1xVrPoNutyqTdQ04DXBEZW4ZW4A5RAQW2he7qIpTmG-M/edit?tab=t.0

Beard, M. (2008). *Pompeii: The life of a Roman town*. London, UK: Profile Books.

Benefiel, R. R. (2010). Dialogues of ancient graffiti in the house of Maius Castricius in Pompeii. *American Journal of Archaeology, 114*(1), 59–101. https://doi.org/10.3764/aja.114.1.59

Berliner, L. (2022). Towards a methodology of unwatched digital media. *Feminist Media Histories, 8*(2), 219–230. https://doi.org/10.1525/fmh.2022.8.2.219

Berliner, L. (2024). . . . Like no one is watching: Taking digital obscura seriously. *Journal of Cinema and Media Studies, 63*(3), 164–169. https://doi.org/10.1353/cj.2024.a927692

Bowes, K. (2021). Tracking consumption at Pompeii: The graffiti lists. *Journal of Roman Archaeology, 34*(2), 552–584. https://doi.org/10.1017/S104775942100060X

boyd, d. (2008). *Taken out of context: American teen sociality in networked publics* [Doctoral dissertation, University of California Berkeley]. Social Science Research Network. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1344756

Bril, M. (2023). Performatively archiving the early Web: One terabyte of kilobyte age. *VIEW Journal of European Television History and Culture, 12*(23), 69–85. https://doi.org/10.18146/view.293

Burgess, J., & Green, J. (2018). *YouTube: Online video and participatory culture* (2nd ed.). Cambridge, UK: Polity Press.

Cascone, S. (2021, August 13). *An ancient fast food restaurant in Pompeii that served honey-roasted rodents is now open to the public*. Artnet. https://news.artnet.com/art-world/pompeii-opens-recently-discovered-ancient-fast-food-restaurant-1998265

Cayari, C. (2011, July 8). The YouTube effect: How YouTube has provided new ways to consume, create, and share music. *International Journal of Education & the Arts, 12*(6), 1–30. http://www.ijea.org/v12n6

Chayka, K. (2024). *Filterworld: How algorithms flattened culture*. London, UK: Heligo Books.

Cook, T. (2013). Evidence, memory, identity, and community: Four shifting archival paradigms. *Archival Science, 13*, 95–120. https://doi.org/10.1007/s10502-012-9180-7

de Caro, S. (2015). Excavation and conservation at Pompeii: A conflicted history. *The Journal of Fasti Online: Archaeological Conservation Series, 3*, 1–31.

de Divitiis, E., Cappabianca, P., Esposito, F., & Cavallo, L. M. (2004). The legacy of Pompeii and its volcano. *Neurosurgery, 55*(4), 989–1006. https://doi.org/10.1227/01.neu.0000142521.78944.3a

Dwyer, E. J. (2010). *Pompeii's living statues: Ancient Roman lives stolen from death*. Ann Arbor: University of Michigan Press.

Ernst, W. (2012). *Digital memory and the archive*. Minneapolis: University of Minnesota Press.

https://doi.org/10.65476/ppzen716

Germain, T. (2025, April 23). *The hidden world beneath the shadows of YouTube's algorithm*. BBC. https://www.bbc.com/future/article/20250306-inside-youtubes-hidden-world-of-forgotten-videos

Gibbon, E. (1994). *The history of the decline and fall of the Roman Empire* (D. Womersley, Ed., Vols. 1–6). London, UK: Penguin Press. (Original work published 1776–1788)

Gillespie, T. (2010). The politics of "platforms." *New Media & Society, 12*(3), 347–364. https://doi.org/10.1177/1461444809342738

Goffman, E. (1959). *The presentation of self in everyday life*. New York, NY: Anchor.

Greetham, D. (1999). "Who's in, who's out": The cultural poetics of archival exclusion. *Studies in the Literary Imagination, 32*(1), 1–28.

Gupta, A., Singh, S. K., Ahuja, K., & Gupta, A. (2019). Good morning turning to spam morning. In V. Gunjan, V. Garcia Diaz, M. Cardona, V. Solanki, & K. Sunitha (Eds.), *ICICCT 2019—System reliability, quality control, safety, maintenance and management* (pp. 1–11). Singapore: Springer. https://doi.org/10.1007/978-981-13-8461-5_1

HU OSA 320-2 amateur photographs. (n.d.). Blinken OSA Archivum. https://catalog.archivum.org/catalog/jDe85pb2

Huhtamo, E., & Parikka, J. (2011). *Media archaeology: Approaches, applications, and implications*. Berkeley: University of California Press.

Jo, E. S., & Gebru, T. (2020, January). Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *FAT '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 306–316). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372829

Kahle, B. (1997, March 1). Preserving the Internet. *Scientific American*. https://web.archive.org/web/19971011050140/http://www.archive.org/sciam_article.html

Kiros, H. (2022, September 20). Hated that video? YouTube's algorithm might push you another just like it. *MIT Technology Review*. https://www.technologyreview.com/2022/09/20/1059709/youtube-algorithm-recommendations/

Lazer, E., Welch, K., Vu, D., Vu, M., Middleton, A., Canigliula, R., Luyck, S., Babino, G., & Osanna, M. (2021). Inside the casts of the Pompeian victims: Results from the first season of the Pompeii Cast Project in 2015. *Papers of the British School at Rome, 8*, 101–136. https://doi.org/10.1017/S0068246220000264

Mahesh, V. (2025, April 23). 20 ways we're celebrating two decades of YouTube. *YouTube Official Blog*. https://blog.youtube/news-and-events/happy-birthday-youtube-20/

Manoff, M. (2004). Theories of the archive from across the disciplines. *Portal: Libraries and the Academy, 4*(1), 9–25. https://doi.org/10.1353/pla.2004.0015

Mastrolorenzo, G., Palladino, D. M., Vecchio, G., & Taddeucci, J. (2002). The 472 AD Pollena eruption of Somma-Vesuvius (Italy) and its environmental impact at the end of the Roman Empire. *Journal of Volcanology and Geothermal Research, 113*(1–2), 19–36. https://doi.org/10.1016/S0377-0273(01)00248-7

Matassi, M., & Boczkowski, P. J. (2023). *To know is to compare: Studying social media across nations, media, and platforms*. Cambridge, MA: MIT Press.

McGrady, R., & Snehi, H. (2025). The ethics of accidental vlogs. *M/C Journal, 28*(4). https://doi.org/10.5204/mcj.3201

McGrady, R., Zheng, K., Curran, R., Baumgartner, J., & Zuckerman, E. (2023). Dialing for videos: A random sample of YouTube. *Journal of Quantitative Description: Digital Media, 3*, 1–85. https://doi.org/10.51685/jqd.2023.022

McGrady, R., Zheng, K., & Zuckerman, E. (2025). One platform, four languages: Comparing English, Spanish, Hindi, and Russian YouTube. *Social Media & Society, 11*(3), 1–21. https://doi.org/10.1177/20563051251363216

Metz, C., Kang, C., Frenkel, S., Thompson, S. A., & Grant, N. (2024, April 8). How tech giants cut corners to harvest data for A.I. *The New York Times*. https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html

Meyer, F. G. (1980). Carbonized food plants of Pompeii, Herculaneum, and the Villa at Torre Annunziata. *Economic Botany, 34*, 401–437. https://doi.org/10.1007/BF02858317

O'Neill, E. T., Lavoie, B. F., & Bennett, R. (2003). Trends in the evolution of the public Web. *D-Lib Magazine, 9*(4). https://www.dlib.org/dlib/april03/lavoie/04lavoie.html

OpenAI. (n.d.). *Whisper* [Computer software]. https://github.com/openai/whisper

Parikka, J. (2012). Archival media theory: An introduction to Wolfgang Ernst's media archaeology. In W. Ernst (Ed.), *Digital memory and the archive* (pp. 1–22). Minneapolis: University of Minnesota Press.

Prelinger, R. (2007). Archives and access in the 21st century. *Cinema Journal, 46*(3), 114–118.

https://doi.org/10.65476/ppzen716

Prelinger, R. (2009). The appearance of archives. In P. Snickars & P. Vonderau (Eds.), *The YouTube reader* (pp. 268–274). Stockholm, Sweden: National Library of Sweden.

Ramírez-López, L. (2015). *An accidental archive: A case study of IndieCollect* [Master's thesis, New York University]. Moving Image Archiving and Preservation MA. https://miapnyu.org/program/student_work/2015spring/15s_3490_ramirezlopez_thesis_final_y.pdf

Rontott fotók (családi és amatőr fotók 1989-ből). (n.d.). Blinken OSA Archivum. https://1989.archivum.org/rontott-fotok

Rowan, E. (2017). Sewers, archaeobotany, and diet at Pompeii and Herculaneum. In M. Flohr & A. Wilson (Eds.), *The economy of Pompeii* (pp. 111–134). Oxford, UK: Oxford University Press.

Schellenberg, T. R. (1956). *Modern archives: Principles and techniques*. Chicago, IL: University of Chicago Press.

Selinger, E., & Hartzog, W. (2018). Obscurity and privacy. In J. C. Pitt & A. Shew (Eds.), *Spaces for the future: A companion to philosophy of technology* (pp. 119–129). New York, NY: Routledge.

Snickars, P., & Vonderau, P. (Eds.). (2009). *The YouTube reader*. Stockholm, Sweden: National Library of Sweden.

Strathaus, S. S., & Hediger, V. (2023). Capillary, migratory, projective: Inventing 'cinema's past so that it may have a future. An introduction. In S. S. Strathaus & V. Hediger (Eds.), *Accidental archivism* (pp. 47–61). Lüneburg, Germany: Meson Press.

Tector, A. (2006). The almost accidental archive and its impact on literary subjects and canonicity. *Journal of Canadian Studies, 40*(2), 96–108. https://doi.org/10.1353/jcs.2007.0024

The Goldberger House. (2011). *Blinken OSA Archivum*. https://archivum.org/about-us/the-goldberger-house

Thelwall, M., & Vaughan, L. (2004). A fair history of the Web? Examining country balance in the Internet Archive. *Library & Information Science Research, 26*(2), 162–176. https://doi.org/10.1016/j.lisr.2003.12.009

van der Cruyssen, Y. (2011). *Petit Tube*. https://petittube.com/

Veix, J. (2016, January 6). *How the weird, unfiltered Internet became a media goldmine*. Fusion. https://www.jezebel.com/how-the-weird-unfiltered-internet-became-a-media-goldm-1793853905

Walz, R. (n.d.). *IMG_0001.* https://walzr.com/IMG_0001/

Youtube.com Website Analysis for May 2025. (2025, May). SimilarWeb. https://www.similarweb.com/website/youtube.com/#overview

Zheng, K., McGrady, R., & Zuckerman, E. (2024, June 17). TubeStats. https://tubestats.org/

Zhou, J., Li, Y., Adhikari, V. K., & Zhang, Z. (2011). Counting YouTube videos via random prefix sampling. In *IMC '11: Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement* (pp. 371–380). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2068816.2068851

Zuckerman, E. (2021, November 2). *Facebook has a misinformation problem, and is blocking access to data about how much there is and who is affected*. The Conversation. https://theconversation.com/facebook-has-a-misinformation-problem-and-is-blocking-access-to-data-about-how-much-there-is-and-who-is-affected-164838

https://doi.org/10.65476/ppzen716