Communications of the Association for Information Systems

Volume 56 Paper 9

3-24-2025

Training Vision AI Models with Public Data: Privacy and Availability Concerns

Abdulkareem Alsudais
Prince Sattam bin Abdulaziz University

Follow this and additional works at: https://aisel.aisnet.org/cais

Recommended Citation
Alsudais, A. (2025). Training Vision AI Models with Public Data: Privacy and Availability Concerns. Communications of the Association for Information Systems, 56, 209-229. https://doi.org/10.17705/1CAIS.05609

This material is brought to you by the AIS Journals at AIS Electronic Library (AISeL). It has been accepted for inclusion in Communications of the Association for Information Systems by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Training Vision AI Models with Public Data: Privacy and Availability Concerns

Cover Page Footnote

This manuscript underwent peer review. It was received 07/15/2024 and was with the authors for three months for two revisions. Babak Abedin served as Associate Editor.

This article is available in Communications of the Association for Information Systems: https://aisel.aisnet.org/cais/vol56/iss1/1

Communications of the Association for Information Systems

Research Article DOI: 10.17705/1CAIS.05609 ISSN: 1529-3181

Training Vision AI Models with Public Data: Privacy and Availability Concerns

Abdulkareem Alsudais

Information Systems Department
Prince Sattam bin Abdulaziz University
Al-Kharj 11942, Saudi Arabia
0000-0002-9961-6889

Abstract:

This paper contributes to research on the ethics of utilizing publicly available images and videos in training AI models by analyzing five prominent open research datasets containing images and videos collected from web user-generated content. This study investigates the current unavailability of these images and videos to understand the extent to which users remove or limit the visibility of their content. This could indicate their opposition to the perpetual use of their images or videos in open datasets, current AI models, or the training of future models. The findings reveal that all five datasets have a substantial number of items that are no longer accessible via their original URLs. Further, a longitudinal analysis over two and a half years reveals a statistically significant increase in this unavailability. The study identifies and categorizes the factors driving this unavailability, including account termination, content being made private by users, and items removed by platforms due to policy violations. This study shows that a significant portion of users may eventually choose to remove their content from the web. This adds valuable insights to AI ethics research, highlighting privacy and the users' right to be forgotten in the context of publicly shared images and videos.

Keywords: Artificial Intelligence, Ethics, Open Data, Data Privacy, Right to be Forgotten.

This manuscript underwent peer review. It was received 07/15/2024 and was with the authors for three months for two revisions. Babak Abedin served as Associate Editor.

1 Introduction

Recent efforts by Information Systems (IS) scholars have increasingly focused on examining various aspects of Artificial Intelligence (AI) from an IS perspective, with a growing body of work addressing issues related to AI ethics (Ashok et al., 2022; Kaur et al., 2024; Kordzadeh & Ghasemaghaei, 2022; Mirbabaie et al., 2022). At the same time, advancements in AI have led to the widespread adoption of commercial tools such as ChatGPT and Dall-E that allow users to engage via websites and apps with advanced AI models. Users can utilize these models to complete tasks traditionally viewed as challenging such as text generation, image captioning, and text summarization. Many of these AI models are oriented towards vision and language related tasks, predominantly involving the generation or modification of videos and images from user inputs. Notable applications include ones focusing on the creation of videos from textual prompts, the generation of text captions for provided images, and the alteration of images based on a combination of an image and a textual prompt. Despite their significant impact, these models have attracted criticism for several reasons, such as the presence of embedded biases (Cheong et al., 2023; Luccioni et al., 2024) and their potential inclusion of copyrighted materials (Benhamou & Andrijevic, 2022; Y. Wang et al., 2023). A primary contributing factor to these issues and others is the training data used to develop these models (Agerfalk et al., 2022). Although these commercial tools may not disclose detailed information about their training data, advances in AI and deep learning, in general, can be attributed to specific open datasets widely used in research. Thus, studying these datasets and their images and videos may provide insights into these models and the underlying data upon which they depend.

One common source of training data for large AI models is user-generated online content. This includes social media posts and comments from platforms like Reddit and X, images from photo-sharing sites such as Flickr and Instagram, and videos from platforms like YouTube and Snap. Prior research suggests that users are often unaware that their content is being used in publicly available research datasets (D. Wang et al., 2015). Furthermore, users typically are not informed or contacted when their data are used to train Large Language Models (LLMs) or multimodal AI models. The inclusion of this data in AI training raises several ethical questions, including whether users should have the option to opt-in before their data are used and how they can retract their data from training datasets and developed machine learning models. These types of concerns have contributed to the emergence of a field known as machine unlearning (Bourtoule et al., 2021; Cao & Yang, 2015), which explores topics related to the methods and techniques for removing data from trained machine learning models. This is necessary for various reasons, such as identifying and removing problematic or illegal content, or complying with “right to be forgotten” laws that mandate the removal of personal data upon request (Villaronga et al., 2018).

All of this raises concerns about the inclusion of data in current AI models after uploaders have removed them from public web locations, as well as about the future use of this content in training AI models. In a recent study, Carlini et al. (2023) examined image diffusion models such as DALL-E 2 and Stable Diffusion, revealing that these models can “memorize individual images from their training data and emit them at generation time.” Furthermore, the authors were able to “extract over a thousand training examples from state-of-the-art models, ranging from photographs of individual people to trademarked company logos." Along the same lines, an initial motivation for this research was the observation that many images and videos from large AI research datasets are no longer accessible online via their original URLs. This situation raises an important question: What should happen to models trained on web user-generated content when users delete or restrict public access to these images and videos? Moreover, given that developers and researchers are likely to retain copies of these items, should their use continue indefinitely in research and as input data for training future AI models?

The primary objective of this paper is to analyze the issue of using web user-generated content in the training of commercial video and image AI models. Specifically, it examines the extent to which users may wish to opt-out of having their posted videos and images remain public, which may indicate a preference against the use of their data in other contexts. Since it is infeasible to examine the training data of specific commercial video and image AI models, this study instead focuses on five influential open video and image datasets frequently employed by the scientific community. The rationale is to use these five datasets as a case study to argue that findings on the use of user-generated images and videos could be generalized to video and image AI models that depend on leveraging such content in their training and development processes, often without the uploaders' awareness or consent. The five datasets are

ActivityNet (Caba Heilbron et al., 2015), AVSpeech (Ephrat et al., 2018), ImageNet (Deng et al., 2009; Yang et al., 2020), Open Images (Kuznetsova et al., 2020), and YT_boundingboxes (Real et al., 2017).

Utilizing these five datasets, this study makes several key contributions toward enhancing our understanding of the extent to which users who upload their images and videos online might later choose to delete them or alter their visibility settings. First, each video and image in these datasets is associated with a URL, which is visited to assess the current unavailability within the dataset. Second, this unavailability is reanalyzed at four distinct times: May 2021, November 2021, October 2022, and December 2023. Statistical tests are then applied to identify any statistically significant trends in unavailability. Third, as these datasets primarily contain URLs from Flickr and YouTube, and these platforms specify reasons for content unavailability, these reasons are systematically categorized and quantified. Specifically, the paper addresses the following research questions:

RQ1: What is the current unavailability of URLs linking to user-generated videos and images in the five datasets?

RQ2: By rerunning the same analysis four separate times between May 2021 and December 2023, what significant trends are identified regarding the unavailability of user-generated videos and images?

RQ3: What are the primary reasons for the unavailability of URLs linking to user-generated videos and images in the five datasets?

2 Related Work

2.1 Information Systems Research on AI and Data

Information Systems (IS) scholars have examined Artificial Intelligence (AI) broadly, as well as specific ethical concerns related to AI. In one paper, Mirbabaie et al. (2022) analyzed previous papers on AI and ethics within the IS research context, offering recommendations for scholars interested in this area. They concluded their paper with a table outlining potential research questions for IS scholars, organized by ethical themes, AI ethics principles, and research dimensions. Examples of ethics principles relevant to this study include “obtain informed consent” and “prevent harm to humans.” In a related systematic literature review, the authors focused on AI and employee privacy, identifying several research gaps for IS scholars (Kaur et al., 2024). The authors further stressed the importance of transparency in how organizations collect and use data. They also recommended the integration of ethical AI practices that prioritize and protect employees' privacy. Another literature review examined explainable AI in the context of IS (Brasse et al., 2023). In another systematic literature review specifically focusing on the negative aspects of people analytics, the authors identified several areas of concern when organizations use AI and analytics to manage employees, highlighting privacy and algorithmic bias as two primary issues (Giermindl et al., 2022). Finally, Varsha (2023) identified similar factors in their systematic literature review, with data privacy highlighted as a major area of concern.

In addition to these literature reviews, various studies have examined AI and data usage in organizational contexts (Kotlarsky et al., 2024; Krasikov & Legner, 2023; Lee et al., 2023). In one such study, the authors found that “strategies must strike a balance between safeguarding data privacy and security while facilitating the process of capturing, using, and reusing data within and across the organization(s)” (D. Xu et al., 2024). Additionally, in relation to this study’s focus on image and video datasets, relevant research in IS also exists in this area. In one paper, the authors analyzed thumbnails for videos from an online video-sharing service and investigated the features users highlight in the thumbnails of their uploaded content (Dedema & Herring, 2023). In another study, the author examined educational YouTube videos and analyzed how learners engage with videos in informal learning settings (Shen, 2023). Finally, IS researchers have also utilized editorials to reflect on AI’s challenges, opportunities, and potential negative impacts (Abbasi et al., 2024; Berente et al., 2021; Mikalef et al., 2022). In one editorial, the authors highlighted several issues of concern, noting that in some jurisdictions “organizations must obtain valid consent for the collection and use of personal data, and must provide individuals with access to their personal data, the right to rectify inaccurate data, and the right to be forgotten” (Arora et al., 2023).

2.2 Open Data

Since this research utilizes five open research datasets, its findings are potentially valuable for scholars exploring open science and datasets. Open data refers to data collected and freely available from governments, organizations, researchers, or individuals (Sadiq & Indulska, 2017), and has supported diverse scientific inquiries (Alsudais et al., 2022; Cantador et al., 2020). Research on open data has covered topics such as data sharing by researchers (Kim & Nah, 2018; Liu & Wei, 2023), issues related to open government data and international open data initiatives (Jetzek et al., 2019; Lnenicka et al., 2022), and challenges in open data for scientific research (Raffaghelli & Manca, 2023; X. Wang et al., 2021). Several authors have studied how to improve open research in general (AlNoamany & Borghi, 2018; Besançon et al., 2021). Additionally, efforts to refine open research have led to the creation of guidelines such as the FAIR principles—Findability, Accessibility, Interoperability, and Reusability—to enhance data reusability (Wilkinson et al., 2016).

In recent years, open data has emerged as a key topic, influencing scientific research in two primary ways. First, there has been a push to encourage, or sometimes require, researchers to publicly provide their research data to enhance reproducibility and verify findings. This practice, known as data sharing, helps other researchers to quickly access and build upon previous studies (Fecher et al., 2015; Park & Wolfram, 2017). Second, a significant focus has been placed on conducting studies that primarily involve the collection and creation of large new datasets relevant to one or more disciplines. These datasets are essential in establishing a foundation for research within a field. For example, in AI research, the introduction of new datasets like MSR-VTT, which includes videos and their descriptions (J. Xu et al., 2016), allows researchers to develop video captioning models and benchmark their performance against others. Rankings based on accepted metrics are often published to highlight comparative model performance. Additionally, open datasets often include user-generated content, which one author defines as “any kind of text, data or action performed by online digital systems users, published and disseminated by the same user through independent channels, that incur an expressive or communicative effect either on an individual manner or combined with other contributions from the same or other sources” (Santos, 2022). The datasets utilized in this study are open datasets of user-generated images and videos, specifically selected for their relevance to AI research. These datasets are used to analyze issues related to the unavailability of images and videos and the perpetual use of copies of these data in AI model training.

2.3 AI Ethics and Datasets Issues

The recent surge in AI research, tools, and open-source models has prompted increased efforts to address privacy and ethical concerns surrounding AI in general and the data used to train AI models (Leschanowsky et al., 2024; Mirbabaie et al., 2022). Previous studies have also explored various challenges inherent in open datasets for machine learning, including biases affecting model accuracy (Mehrabi et al., 2022; Stock & Cisse, 2018), unfair representation of certain groups (Shankar et al., 2017; Suresh & Guttag, 2021), and issues with incorrect or missing data (Alsudais, 2021b; Gaffney & Matias, 2018). One form of bias, representation bias, may lead to datasets that “lack the diversity of the population, with missing subgroups and other anomalies” (Mehrabi et al., 2022).

An example of this representation bias is the geographical bias in datasets like ImageNet and Open Images, where the majority of images originate from the United States, Great Britain, and European countries (Shankar et al., 2017). Additionally, early versions of ImageNet included potentially problematic categories under the main “person” category, such as “developer” and “programmer,” with the latter primarily featuring images of white, male, and European individuals, thus underrepresenting other racial, ethnic, and gender groups (Yang et al., 2020). Moreover, the category labeled “Iraqi” was filled with war-related images, which could lead algorithms to misclassify images from conflict zones as “Iraqi” (Alsudais, 2022). To address these issues, the creators of ImageNet have released an updated version of the dataset that corrects these and other issues identified in the earlier release (Yang et al., 2020).

Recent studies have explored ethical and legal considerations related to AI models and datasets. Several studies have proposed new techniques to mitigate existing biases in large vision datasets (Georgopoulos et al., 2021; Rajabi et al., 2023; A. Wang et al., 2022). Recent efforts have also examined “right to be forgotten” laws, studying their impacts and complexity. In one significant study, the authors examined this ruling, which was enacted in Europe and allows individuals to request the removal of URLs appearing in search results linked to their names (Bertram et al., 2019). Over five years, this resulted in 3.2 million URLs being requested for removal from Google Search results, highlighting the substantial impact of this

ruling. In another paper, the author investigated legal challenges associated with machine unlearning in the context of the "right to be forgotten" laws (Juliussen et al., 2023). Ultimately, these related studies highlight the research community's active engagement in exploring ethical and legal issues surrounding data, machine learning, machine unlearning, and AI, with a particular focus on image and video datasets and models. This study approaches the topic from a different perspective, examining and quantifying the availability of images and videos from large vision AI datasets over time, and potentially offering insights relevant to these areas of research.

3 Research Methods

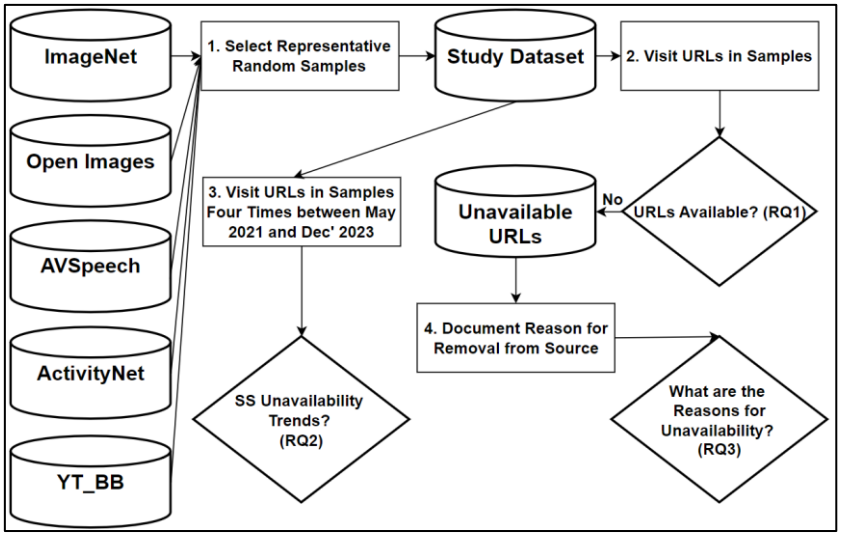

Figure 1. Conceptual Framework for this Study

In this section, the methods used to investigate the research questions are outlined. Subsection 3.1 describes the datasets employed to explore these questions. This is followed by a detailed explanation of the analytical approach, including the tests utilized in this study and how each question is answered. This section concludes by summarizing the research questions and the steps followed to address them. Figure 1 presents a conceptual framework of the research, outlining the primary steps and the research questions. In summary, the study began by selecting five open research datasets containing images or videos. Representative samples from each dataset were then selected, and the unavailability of images and videos was quantified (RQ1). This quantification process was repeated four times to identify any significant trends (RQ2). Finally, the reasons for unavailability were analyzed and quantified (RQ3).

3.1 Datasets

Ideally, investigating the use of user-generated content in training large commercial video or image AI models would involve examining the current training data from such models. However, since these models typically do not disclose their underlying training data, this paper aims to utilize datasets that are likely similar to the ones employed in their training. For instance, if a model is designed for video generation and relies on data from YouTube, there are several video datasets introduced by the scientific community that are often used to benchmark and evaluate the performance of various related models. By focusing on these datasets as case studies, this paper seeks to contribute to the growing body of knowledge that aims to understand, analyze, and enhance our comprehension of the issues associated with the use of user-generated videos and images in AI model training. Additionally, by analyzing these open research datasets, the findings can be extended to the science of science research that aims to study issues and concerns about open research datasets in general.

Accordingly, many video and image datasets commonly used in AI research were reviewed, and five were selected to investigate the questions posed in this paper. Before the selection of the five datasets for this study, specific criteria were established for their inclusion. These criteria emphasized selecting only datasets that had been introduced and detailed in published, peer-reviewed papers, were extensively used by the scientific community, and remained available for download. The primary focus of this study is

on: 1) the examination of user-generated content (videos or images) through the URLs where the content is hosted, and 2) the quantification of how many videos and images have been removed from platforms like YouTube or Flickr, along with categorizing the reasons for removal. Consequently, another criterion was that only video and image datasets that included URLs linking directly to the original items and their uploaders were selected. Based on these general criteria, The five datasets chosen were ActivityNet (Caba Heilbron et al., 2015), AVSpeech (Ephrat et al., 2018), ImageNet (Deng et al., 2009; Yang et al., 2020), Open Images (Kuznetsova et al., 2020), and YT_boundingboxes (Real et al., 2017). The citations refer to the original papers where these datasets were introduced.

Table 1. The Five Datasets Analyzed in this Study

Dataset	Content type	Size	Additional description
ActivityNet	YouTube videos	4,819 videos	The videos belong to one of 203 activity classes, such as “cycling,” “shaving,” and “playing piano.”
AVSpeech	YouTube videos	271,613 videos	The videos are divided into segments totaling 4,700 hours, each featuring a speaker.
ImageNet	Images from websites	14,197,122 images	The images are organized into several high-level categories, which include multiple subcategories structured based on the WordNet hierarchy of objects.
Open Images	Images from Flickr	1,743,042 images	The images belong to 600 object classes, with images containing bounding boxes that identify the objects within them.
YT_BB	YouTube videos	253,569 videos	From these videos, approximately 380,000 segments are extracted, each ranging from 15 to 20 seconds in length.

Table 1 lists the five datasets and their sizes (number of images or videos) as well as additional information about their content. ActivityNet includes videos collected from YouTube of physical human activities such as “dancing,” “playing volleyball,” and “kayaking.” The dataset includes the URLs where these videos were originally posted. AVSpeech comprises YouTube videos featuring individuals speaking on various topics, with videos in languages such as English, Spanish, and Russian, and also includes the original YouTube URLs. The original videos were then segmented into short snippets, each capturing a brief segment extracted from the video of a speaker speaking.

ImageNet has millions of images sourced from various websites, with over 50% coming from Flickr, a platform for posting and sharing images. A previous study indicated that, of the images on ImageNet, those collected from Flickr were more likely to still be available online than those from other sources (Alsudais, 2019). The Fall 2011 release of this dataset, which contains over 14M images, is used in this study. Similarly, Open Images is a dataset that has millions of images from Flickr. As of the writing of this paper, the dataset has seven different releases; In this study, the sixth release is used. Finally, YT_boundingboxes (YT_BB) features YouTube videos with annotated bounding boxes around single objects. In summary, three of the five datasets consist of URLs of videos uploaded to YouTube, while the other two comprise URLs of images uploaded to Flickr or, in the case of ImageNet, Flickr, and other websites.

To demonstrate the influence of these datasets on the research community, a search was completed in April 2024 to determine how many times the papers where these datasets were first introduced have been cited in Web of Science (WoS) and Google Scholar (GS). Although ImageNet is the most referenced, with over 66,098 citations in GS, the remaining datasets have each been cited at least 650 times in GS and 348 times in WoS. Additionally, ActivityNet and Open Images have received 2,571 and 2,406 citations in GS, respectively.

3.2 Unavailability Determination (RQ1)

The first research question of this study focuses on determining the unavailability of images and videos in the five selected datasets by visiting URLs and quantifying how many are no longer accessible online. To achieve this, several methods were initially considered, including web scraping and APIs. However, the chosen strategy involved generating representative random samples for the five datasets. These samples were then used to determine the availability of URLs. For each dataset, a 4% margin of error and a 99% confidence level were established, determining sample sizes of 1,033 for AVSpeech, 853 for ActivityNet, 1,037 for ImageNet, 1,036 for Open Images Dataset, and 1,033 for YT_BB.

To assess the online availability of images and videos in the samples, each item was embedded into an online spreadsheet. An “unavailability” column was added, where a cell value of “1” indicates availability and “0” indicates unavailability. If an image or video did not appear in the sheet when embedded, likely indicating unavailability, the URL associated with the content was further checked to verify its online status, as some videos and images might still be accessible online even if not available as embedded content. Figure 2 includes two screenshots, each representing a sample of tested URLs from two datasets. The first screenshot is from Open Images, showing image availability, while the second is from ActivityNet, displaying a sample of videos and their availability. For Open Images, the screenshot reflects the fourth availability check completed in December 2023, with the “unavailability reason” column showing the specific reasons for unavailability identified at that time. For ActivityNet, the screenshot shows examples of unavailability reasons provided by YouTube such as “this video isn’t available anymore” and “this video is private.”

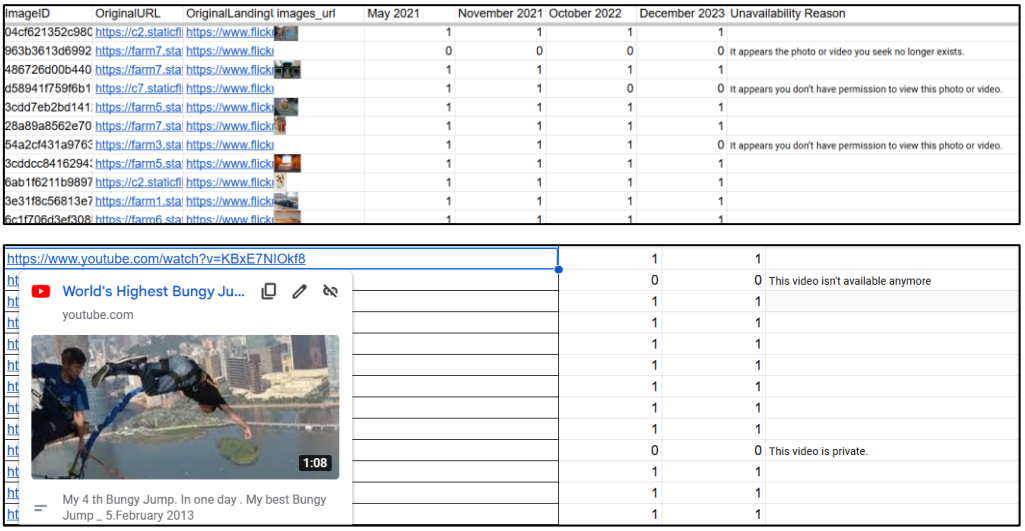

Figure 2. Screenshots of Online Sheets used to Verify Availability

To verify the reliability of using randomly generated samples from these datasets to address the research questions, three separate samples were generated for ImageNet, Open Images, and YT_BB, resulting in a total of nine samples. The same test was not conducted for the remaining two datasets, as they also consisted of YouTube videos, similar to YT_BB. These samples were analyzed to quantify the number of unavailable images or videos, and consistency was assessed by comparing the results across the three samples per dataset. The analysis showed no statistically significant differences in unavailability between the samples. In other words, the number of unavailable images or videos for each set of samples fell within the specified margins of error. Further details about this process are provided in the results section. Ultimately, the use of samples proved to be a reliable approach, supporting the study's methodology for answering the research questions.

Before adopting the sampling method, other alternatives, such as web scraping and Application Programming Interfaces (APIs), were considered but ultimately not used due to inherent limitations. Web scraping is prone to errors that can result in the collection of inaccurate data, and certain websites employ techniques to mislead scrapers (Landers et al., 2016), which compromises the reliability of results. Furthermore, YouTube's policies explicitly prohibit scraping their video content (YouTube, 2024). Using APIs for processing URLs posed additional challenges, such as API rate limits that restrict data collection frequency. Given the large size of the datasets, even YouTube's academic API was not a viable option, and Google’s API policies classify data collection based on a script iterating over a list as “scraping,” which is also prohibited. Additionally, since ImageNet contains data collected from various sources, no single API solution could accommodate this, and verifying the web scraping policies for each individual webpage is impractical. Most importantly, despite these challenges associated with scraping or APIs, the random sampling approach based on established methodologies offered a reliable and accurate solution and was therefore implemented to investigate the availability of URLs.

3.3 Trends Investigation (RQ2)

The second research question in this study examines whether there are any statistically significant decreasing trends in unavailability over time. To investigate this, images and videos from the five samples were tested at four intervals: May 2021, November 2021, October 2022, and December 2023. A limitation of this study is the non-uniform intervals between these testing points. The results from each testing session were recorded and quantified, and the unavailability data for each sample at each interval were compiled into a list. The Pearson correlation coefficient was used to determine whether statistically significant trends exist, using the four data points for each dataset and the corresponding time intervals. P-values were calculated for each sample to assess the statistical significance of the trends, with a significance level set at 0.05. In summary, a dataset is considered to be experiencing a statistically significant trend in unavailability if the p-value falls below this threshold. The results section presents these p-values along with Pearson's r values for each dataset.

3.4 Reasons for Removal (RQ3)

An important aspect of this research is to quantify the reasons behind the unavailability of images and videos in the five datasets. Given that three of the datasets consist of YouTube videos, and YouTube provides reasons for video unavailability, it is feasible to process all unavailable videos and document the reasons as stated by YouTube. Initial analysis revealed categories of reasons including copyright infringement and users making their public videos private. A prior study by Kurdi et al. (2020) processed a dataset of YouTube videos to determine the percentage that became unavailable a week after their initial upload. In a subsequent study, Kurdi et al. (2021) identified reasons for removal and quantified the number of inaccessible videos for each category, which were similar to those discovered in this study.

To quantify the reasons for unavailability in ActivityNet, AVSpeech, and YT_BB, the reason for each unavailable video was copied from YouTube and recorded in the verification sheets. These reasons were then categorized into six groups, and the number of unavailable videos attributed to each reason was counted for each dataset. For ImageNet, which comprises images from various websites around the web, it is not possible to identify uniform reasons for removal as there is no single service hosting all these images. For Open Images, only two types of messages are displayed for unavailable content on Flickr: 1) "It appears the photo or video you seek no longer exists," and 2) "It appears you don't have permission to view this photo or video."

In the end, this question was addressed by analyzing data from YouTube and Flickr for unavailable content. For YouTube, the six categories of unavailability were quantified, while for Flickr, the number of images corresponding to each of the two types of unavailability messages was counted. This analysis was conducted twice for YouTube data (November 2021 and December 2023) and once for Flickr data (December 2023). A limitation of this study is that this analysis was not run concurrently with the unavailability check that has been run four times.

3.5 Summary of Research Questions

In summary, the initial step focused on confirming the reliability of the sampling strategy in addressing the research questions. This was achieved by analyzing three randomly generated samples from three of the datasets to ensure that unavailability numbers fell within the specified margins of error. Following this, RQ1 was addressed by assessing the unavailability of images and videos in five randomly generated samples from the five datasets. Then, RQ2 was examined by investigating the presence of any statistically significant decreasing trends over time across these datasets by testing the unavailability at four distinct intervals. For RQ3, the reasons for unavailability were quantified across four of the five datasets. Table 2 provides a summary of the research steps, the datasets or subsets utilized, and the main objective of each step. In the table, the initial step is unnumbered to ensure alignment between the remaining steps and their corresponding subsection numbers in this section.

Table 2. Summary of Research Steps, Datasets, and Objectives

	Step	Dataset	Objective
-	Test the reliability of the sampling strategy	Three separate samples were generated for ImageNet, Open Images, and YT_BB (a total of nine samples)	Assess the unavailability of images and videos in the samples, confirming that the number of unavailable items in the three samples for each dataset falls within the specified margins of error, providing evidence of the sampling strategy's reliability.
1	Generate random samples	All five datasets in their entirety	Generate representative random samples from the five datasets for use in subsequent analyses. These five samples serve as the subsets utilized in the remaining steps.
2	Quantify unavailability (RQ1)	The five random samples generated from Step #1	View all images and videos in the samples to verify and quantify their current online unavailability.
3	Investigate statistically significant trends (RQ2)	The five random samples generated from Step #1	The samples are tested at four different points in time: May 2021, November 2021, October 2022, and December 2023, to determine whether statistically significant trends in unavailability exist for each dataset
4	Identify the reasons for unavailability (RQ3)	Four of the five random samples generated from Step #1	Document and categorize the reasons for unavailability as provided by YouTube and Flickr, summarizing the primary causes of content unavailability.

4 Results

4.1 Current Unavailability of Images and Videos (RQ1)

Table 3. Unavailability Results for the Five Datasets

Dataset	Sample size	Unavailable images or videos	Percentage	Estimation	
				From	To
ActivityNet	853	168	19.7%	757 videos	1,142 videos
AVSpeech	1033	221	21.4%	47,261 videos	68,990 videos
ImageNet	1037	537	51.79%	6,784,805 images	7,920,575 images
Open Images	1036	133	12.84%	154,048 images	293,491 images
YT_BB	1033	108	10.46%	16,381 videos	36,667 videos

The five samples were analyzed to determine the current unavailability of images and videos within these subsets. The findings reveal a significant portion of content is now unavailable online across all datasets. Specifically, the unavailability percentages were 19.7% for ActivityNet, 21.4% for AVSpeech, 51.79% for ImageNet, 12.84% for Open Images, and 10.46% for YT_BB. Table 3 presents the counts and percentages of content unavailable as of the last URL check in December 2023. Notably, ImageNet exhibits the highest rate of unavailability, with over half of its images no longer accessible. This is likely due to its age and the diversity of its website sources, unlike the other datasets which primarily use Flickr or YouTube. The table also provides an estimated count of unavailable videos or images, derived by aggregating the unavailability percentages. These estimates were calculated using the defined margin of error, which was set at 4% when determining the appropriate sample size. For example, since 21.4% of the videos for AVSpeech were not available, the actual videos that were unavailable are approximated to be between 17.4% (21.4% - 4%) and 25.4% (21.4% + 4%), which equals to 47,261 and 68,990 videos.

To ensure the reliability of these findings, specifically the ability of the sampling strategy to generate reliable results, three random samples from ImageNet, three from Open Images, and three from YT_BB were generated and analyzed for unavailability. The unavailability check was completed in May 2021 for the first two datasets and October 2024 for YT_BB. The unavailability percentages for ImageNet samples were 47.55%, 45.42%, and 46.20%. For Open Images, they were 9.17%, 8.02%, and 7.63%, and for YT_BB, they were 11.14%, 8.23%, and 10.75% for YT_BB (Figure 3). The maximum difference between the highest and lowest percentages was 2.13 for ImageNet, 1.54 for Open Images, and 2.9 for YT_BB.

The consistency of these variances within the margin of error set at 4% and a confidence level of 99% confirms the validity of using randomly generated samples for this research.

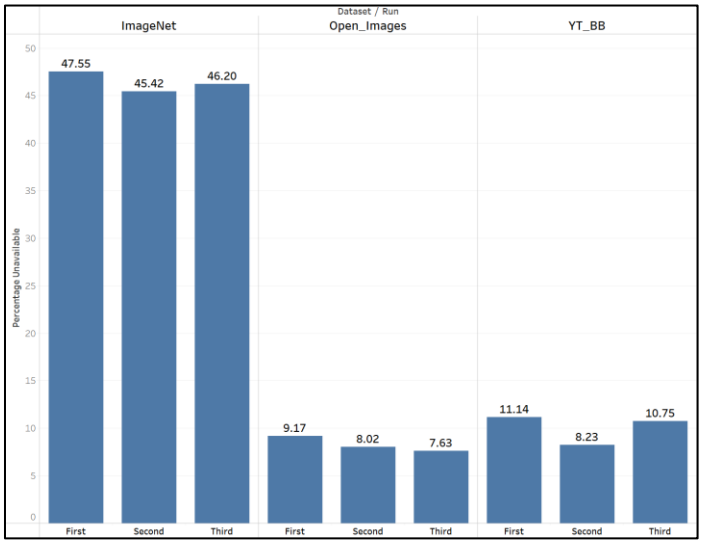

Figure 3. Results of using Randomly Generated Samples for ImageNet, Open Images, and YT_BB

4.2 Trends Determination (RQ2):

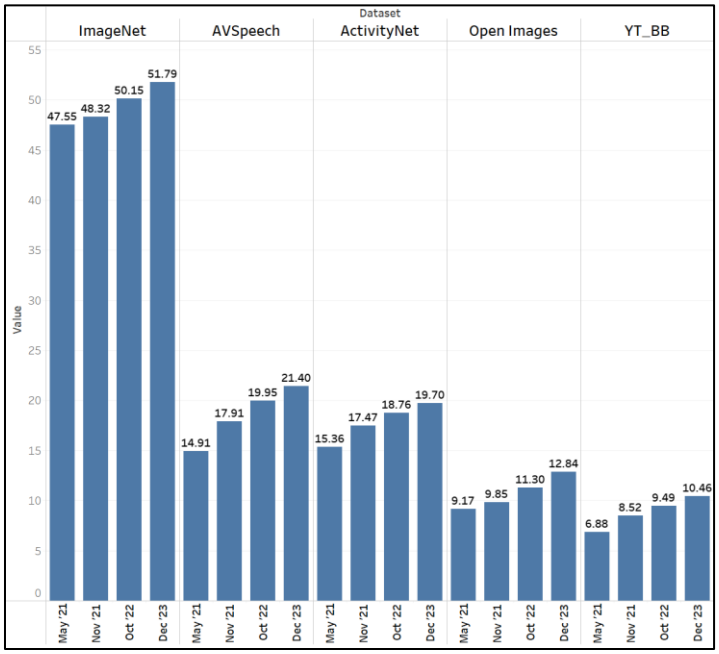

Figure 4. Unavailability Percentages the Four Times the Tests were Completed

Table 4. Determination of the Presence of Significant Trends

Dataset	Pearson's r	p-value
ActivityNet	0.9831	0.0168
AVSpeech	0.9871	0.0128
ImageNet	0.9876	0.0123
Open Images	0.9866	0.0133
YT_BB	0.9901	0.0098

The second research question investigates whether there are statistically significant trends in the unavailability of images and videos, based on checks conducted on four dates for the five datasets. The same images and videos were revisited in May 2021, November 2021, October 2022, and December 2023. Figure 4 shows the results for each dataset, revealing an overall increase in unavailability across all five datasets. The most substantial increases from May 2021 to December 2023 were observed in AVSpeech (6.49%) and ActivityNet (4.34%), followed by ImageNet (4.24%). The smallest increases were noted for Open Images (3.67%) and YT_BB (3.58%). These values represent the differences in unavailability between two dates, calculated by subtracting the initial value from the later one, rather than reflecting percentage changes. To assess the statistical significance of these trends, the Pearson correlation coefficient test was conducted using the unavailability counts. The results confirmed that the trends for each dataset are statistically significant. Table 4 provides the p-values and Pearson's r values from these tests. Ultimately, the findings show that all datasets experienced a statistically significant increase in unavailability.

4.3 Reasons for Removal (RQ3)

The third research question investigates the reasons for unavailability. For the three video datasets, this was addressed by revisiting the URLs of unavailable videos and recording the reasons for unavailability as specified by YouTube. This process was conducted twice: first in November 2021 and again in December 2023. Five reasons were identified through this analysis. The first is simply that the video is no longer available, which provides little insight into the specific cause of unavailability compared to the other reasons. The second reason is that the user made the video private. The third is the termination of the user's account. The fourth is removal due to a copyright claim, and the fifth is a violation of YouTube policies, which could include "violent or graphic content" or breaches of "community guidelines" or terms of service.

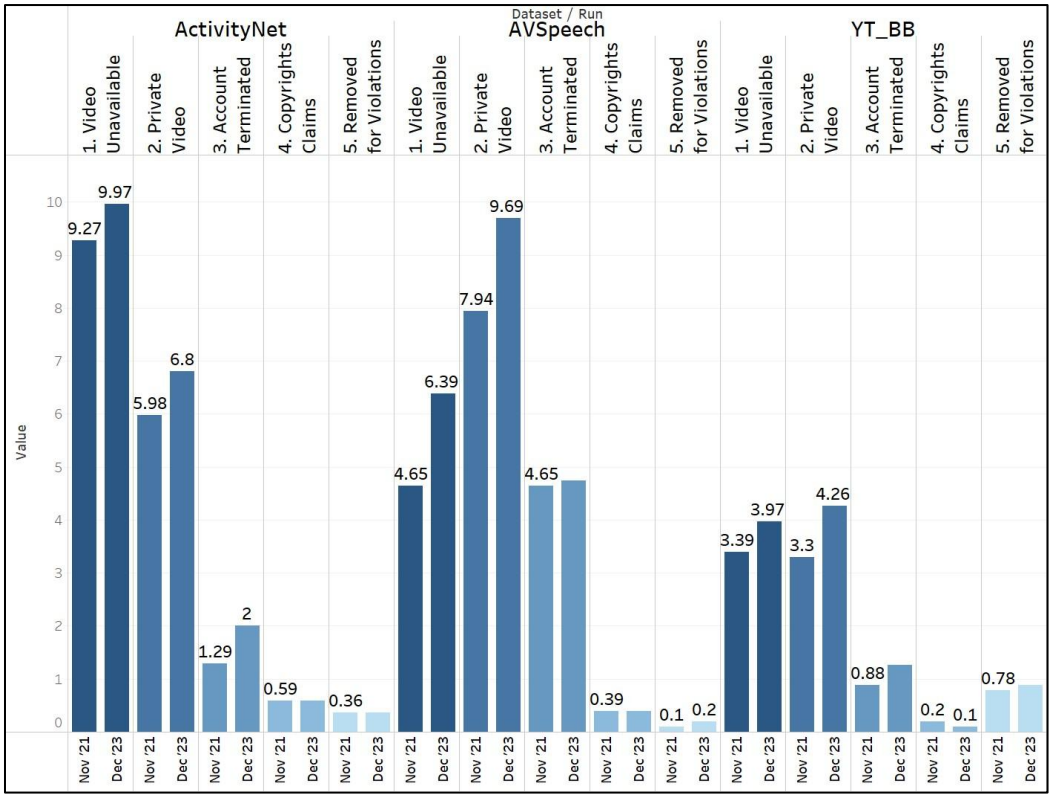

Figure 5. Unavailability Reasons for ActivityNet, AVSpeech, and YT_BB

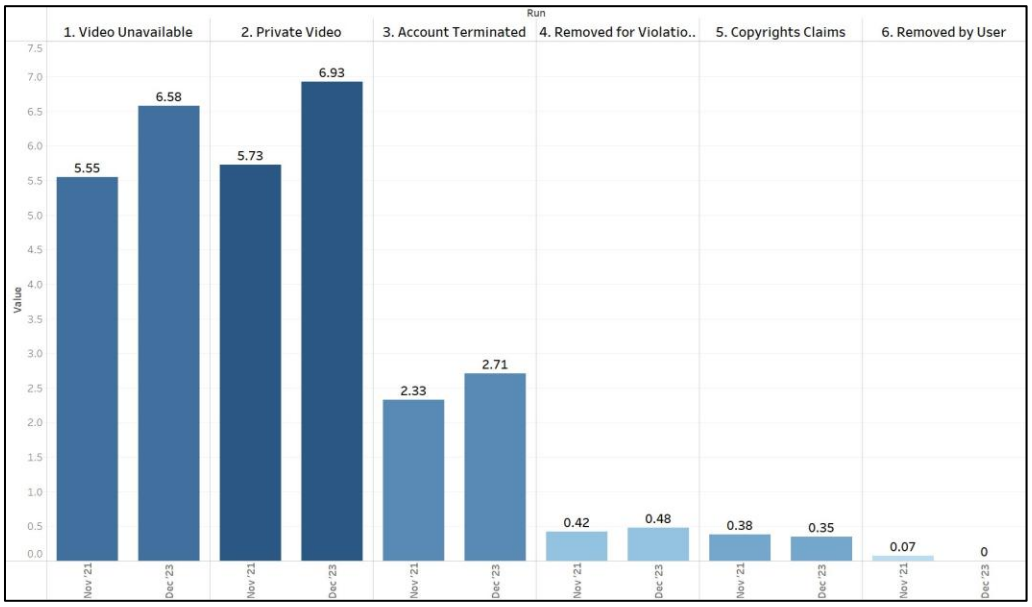

Figure 6. Summary of Unavailability Factors for Videos

Figure 5 illustrates the percentages of videos removed for each of these reasons across the three datasets. It shows, for instance, that in December 2023, 9.69% of the videos in AVSpeech and 6.8% in ActivityNet were unavailable because they had been made private. Two videos in YT_BB were flagged as 'removed by user' in November 2021; these instances are not included in Figure 5. It is important to note that changes implemented by YouTube could influence the reasons given for unavailability, as their classification policies may have evolved over time. However, no significant evidence was observed during the analysis to indicate this. The figure also shows the increases in unavailability for each category between the two times the tests were conducted.

For the Open Images dataset, Flickr provided only two reasons for unavailability: 1) “It appears the photo or video you seek no longer exists,” likely indicating removal by the uploader, although other factors could be at play too, and 2) “It appears you don’t have permission to view this photo or video,” suggesting that the uploader changed the visibility settings to private. The analysis revealed that 9.94% of the images were unavailable for the first reason and 2.89% for the second. It is possible that some images categorized under the first reason might be unavailable due to other issues identified in the video datasets, such as content removal for violations or copyright claims.

Ultimately, users making their content private emerged as one of the most significant reasons for the unavailability of images and videos. For these four datasets, the percentages of content made private were 6.8% for ActivityNet, 9.69% for AVSpeech, 2.89% for Open Images, and 4.26% for YT_BB. Focusing on these factors, Figure 6 illustrates the total percentages for the categories of reasons across the three video datasets at the two time points when the test was conducted. The results highlight that users making their videos private was a significant factor, with 6.93% of videos from the three samples marked as private as of December 2023. As the total number of unavailable videos for the three datasets increased from 422 to 497, this corresponds to a 17.7% rise in unavailability. Among the reasons for this increase, the “private video” category experienced the highest percentage increase at 20.9%, resulting from a rise in unavailable videos due to this reason from 167 to 202. This was followed by “video unavailable” (18.5%), removals due to policy violations (16.6%), and account terminations (16.1%). Overall, these findings offer valuable insights into the causes of video unavailability and their progression over time, providing a foundation for further research.

5 Discussion and Implications

This research investigates the current unavailability of images and videos in five open research datasets. The key findings include: 1) all five datasets contain a large percentage of images and videos that are no longer accessible from their original sources; 2) a longitudinal analysis across four time intervals reveals a statistically significant increase in the unavailability of images and videos; and 3) several factors for

content removal emerged as primary reasons for unavailability, including users making their content private, account terminations, and policy violations. These insights enhance our understanding surrounding user-generated images and videos on the web, particularly in the context of large vision AI models developed using such data.

5.1 Implications for Research

This research has several important implications for research. First, the findings contribute to the discussion about ethical considerations in utilizing user-generated content for open datasets and AI model training, especially when such images and videos are removed by the uploader or taken down by platforms due to policy violations. One implication of this research is that it provides evidence highlighting the extent of this issue. Reflecting on the concerns discussed in the first two sections of this paper, these findings add to existing research across several relevant areas. These include how users may be unaware that their data are being used in research studies (D. Wang et al., 2015), machine unlearning and the ability of researchers to reveal original images from trained AI models (Carlini et al., 2023), and users' right to be forgotten laws (Villaronga et al., 2018).

For each of these areas and others, the findings of this study offer additional support and context. For instance, if users are unaware that their data is included in datasets, they may also be unaware that their removed content has been copied and continues to be used in datasets and AI models. Similarly, if users learn that such content can be reverse-engineered from AI models, even after its removal from external datasets, it raises further ethical concerns. Finally, this research contributes to the right to be forgotten literature by highlighting numerous cases where data, likely intended to be forgotten by their uploaders or hosting platforms, continues to be used in datasets and AI models.

Second, this study raises an important question: If a significant portion of images and videos become unavailable from their original sources, what actions should dataset curators take? This issue is particularly relevant when reproducibility or additional data collection is needed by other scholars interested in the same dataset. While using archived versions of webpages for images or videos might seem like a viable solution, it still introduces concerns about user privacy. The analysis in this study revealed that making content private was one of the reasons for unavailability. Although this does not necessarily indicate privacy concerns as the sole motivation because factors such as regret or embarrassment may also be involved, it does suggest that users may prefer their content to remain inaccessible online. This, in turn, raises further ethical questions about the inclusion of such data in AI models and open datasets.

One possible approach is to generate synthetic data from the original sources and create entirely new datasets consisting solely of synthetic data. The advantage of synthetic data is that it may help protect user privacy by masking identities (Sivizaca Conde et al., 2024). However, synthetic data have faced criticism, particularly in healthcare contexts, where concerns about privacy and potential identity revelation persist even after data anonymization (Arora & Arora, 2022). Another option is to leverage citizen science research, which engages non-experts in data collection. Although often criticized for data quality issues (Lukyanenko et al., 2020), citizen science techniques could be used to build a dataset from users who voluntarily share their data. While four of the five datasets contain hundreds of thousands or even millions of images and videos, ActivityNet has a more manageable number of videos. Thus, by applying methods from citizen science research, it may be feasible to compile a dataset of similar size, which could then be expanded into a larger, synthesized dataset. Ultimately, a suitable approach could involve informing users that their content may be included in scientific datasets and seeking their consent, even if they later remove their content or accounts. However, this raises questions about the representativeness of the datasets, as they may only include data from consenting users.

Third, the study highlights how Information Systems scholars can leverage open datasets commonly used in other disciplines to explore issues relevant to IS research. Since this study focuses on analyzing the unavailability of user-generated images and videos, collecting the data from scratch would have required significantly more time and effort. By utilizing high-quality open datasets, scholars can conduct their work more efficiently, reducing the time and resources dedicated to data collection. For instance, the AVSpeech dataset, which consists of videos of speakers, is predominantly used in computer science research for model development and evaluation. However, IS scholars can repurpose it to investigate topics such as how individuals choose to present themselves in educational videos or explore correlations between content popularity and audio clarity. Since the dataset is already curated, scholars exploring such questions may not need to collect a new dataset, making their research more manageable.

As interest in contributing to AI research grows within IS, scholars are seeking ways to make meaningful contributions. One potential area of impact is the examination and identification of concerns within these open research datasets. Aaltonen et al. (2023) recently identified areas where IS scholars could uniquely contribute to research on data. They suggested "Future research may be directed toward making case study comparisons of data journeys across different contexts or engaging more closely with issues of meaning and knowledge production in AI-powered systems" and "IS scholars could contribute to broader managerial and societal issues by carefully studying not only how and by whom data are governed but also how different domains of socio-economic life become governed by digital data." One potential implication of this paper is demonstrating how IS scholars can provide meaningful contributions in data research by utilizing these open datasets.

Fourth, the sampling strategy used in this research could benefit scholars working on similar projects involving user-generated content. Researchers often face challenges in collecting large datasets from online sources. This sampling method enables efficient data collection, allowing scholars to focus on research inquiries without allocating excessive time to data collection. As long as researchers can identify high-quality, verified open datasets, random sampling from these datasets can be an effective approach, depending on the research study's design.

Given that companies frequently implement measures to deter or prohibit web scraping and restrict API usage, relying solely on these methods for data collection is becoming increasingly challenging. As companies recognize the immense value of user data for training AI models, and perhaps for other strategic or legal reasons, they may be moving toward even stricter restrictions for researchers attempting to access data from these platforms. Consequently, rigorous sampling methods offer a practical and efficient alternative, enabling researchers to conduct a wide range of studies effectively.

Finally, this work has implications for other research areas. For example, previous studies have addressed URL decay, which concerns the persistence of web URLs (Howell & Burtis, 2023; Loan & Shah, 2020). This paper offers a fresh perspective on this topic and presents findings that may benefit researchers studying URL decay and its current status. Additionally, the insights from this study are valuable for scholars examining issues related to open research datasets. These datasets, freely provided by the scientific community, have been analyzed from various angles. This paper's findings contribute to a deeper understanding of how these datasets should be maintained.

An observation made during this study was the unavailability of some open datasets not selected for this study. While searching for suitable open research datasets, in some cases, original websites for datasets were accessible, but download links had been removed or were no longer functional. This issue mirrors previous findings regarding open software links in published papers becoming inaccessible (Alsudais, 2021a). Many URLs for open datasets were either removed or remained accessible but no longer provided the data for download. This trend poses broader concerns for open science, raising questions about the integrity of published papers that promise open data or software but later restrict or remove access.

5.2 Implications for Practice

This research offers several practical implications. First, online platforms that rely on user-generated content typically state in their privacy policies that public users' data may be sold, shared, or harvested by external entities. For instance, Reddit's policy explains: "Reddit also allows other third parties to access public Reddit content using Reddit's developer services," and "content and information may also be available in search results on Internet search engines like Google or in responses provided by an AI chatbot like OpenAI's ChatGPT. You should take the public nature of the Services into consideration before posting" (Reddit, 2024). However, a critical question arises: Should these platforms also consider what happens to users' content when uploaders remove or restrict public access? Put differently, should platforms develop mechanisms to automatically delete all copies of users' data from their services as well as from any entities with which the data was shared?

As the value of user-generated content in the age of AI becomes increasingly evident, platforms may allocate additional resources to strengthen their control over external access to data. They might invest in systems that detect and block web scrapers, ensuring that external entities must pay for data access. Such efforts could also include developing mechanisms for users to remove all instances of their data. For example, if Reddit shares data with OpenAI, they could implement systems to track and delete content when users remove it from Reddit. A future solution might involve blockchain-like infrastructure, where

uploaders retain exclusive rights to their content and control its distribution. This would allow users to license their own data, potentially profiting from it. Such a solution might come from new platforms that gain popularity by implementing these types of mechanisms that empower users.

Second, the scientific community has increasingly focused on machine unlearning, which addresses the potential for users or data owners to request the exclusion of their data from trained machine learning models or future models. This study raises a similar issue: Should developers of AI models reconsider the perpetual inclusion of content that has since been removed or made private? This is a pressing question, as previous studies have shown that user data can be re-identified or extracted from AI models using various techniques.

The analysis in this study found that users making their content private was one of the factors in the unavailability of content. Other reasons, such as account terminations or Flickr's message that "It appears the photo or video you seek no longer exists," suggest that users may not want their data to remain accessible. This may indicate a preference for their content to be forgotten, raising questions about the ethical responsibilities of platforms and AI developers to respect users' wishes and ensure that data can be erased or excluded from training models. Greater awareness of these issues could encourage users to engage more with platforms that support ephemeral content. For example, platforms where content automatically disappears after a set period may see increased participation and user satisfaction. A recent study found that sharing ephemeral content in online dating contexts, led to higher engagement and more image sharing (He et al., 2024). This suggests that ephemeral content strategies can be mutually beneficial for users and platforms.

Finally, several other ethics-related observations relevant to practice were made during the study. For instance, many images and videos viewed did not have faces blurred, raising concerns about the inclusion and dissemination of identifiable public data, including that of children who could not have consented to their data being shared. The analysis also revealed content removal reasons that highlight further complexities. For example, numerous videos were removed by YouTube for violating terms of service, such as policies on violent or graphic content. One video was removed for breaching YouTube's harassment and bullying policy, indicating the presence of such problematic content in the datasets. Other videos were taken down due to copyright complaints. These findings highlight the nuances and ethical considerations that platforms and AI companies should address when using user-generated data for training models.

5.3 Limitations and Future Research

This study has several limitations. First, while the findings suggest a significant number of images and videos become unavailable over time, it does not explore whether uploaders also wish to have their data excluded from open datasets and AI models. It remains possible that a user might remove their public image or video but still approve of its use for AI training or in open datasets. A recent study by Hemphill et al. (2022), which surveyed 1,018 social media users, found that while people were "generally okay with researchers using their data in social research," they preferred that researchers clearly articulate the benefits and seek explicit consent. A similar study focusing specifically on uploaded images and videos might reveal similar findings.

Second, the dataset selection process was based on a general set of criteria rather than starting with a comprehensive review of datasets from a single source to exclude non-compliant ones. Instead, selection criteria were established first, followed by a search that identified the five datasets used as case studies. Third, the analysis of reasons for removal was conducted only twice for three datasets and once for one dataset, meaning that the study does not fully leverage longitudinal insights as it does for unavailability trends. Additionally, this analysis was not completed for ImageNet, presenting an opportunity for future research to focus specifically on this dataset and the status of its URLs. Lastly, much of the content from these datasets has few views. This may suggest that users who deleted or made their content private may have done so upon noticing unusual or unexpected views from certain IP locations or upon discovering their content's inclusion in publicly available datasets. As a result, the actual percentages of unavailability might be lower than reported if a random sample of videos from YouTube were analyzed.

These limitations present opportunities for future research. First, contacting randomly selected users from these datasets to investigate their attitudes toward the inclusion of their content could provide valuable insights. Such an analysis might reveal that many users are unaware their data is copied and used in these open datasets. Second, expanding the selection process to include a broader range of datasets

featuring user-generated content could reveal additional findings. Broadening the search to cover other types of platforms may also uncover further insights into user behavior and content availability.

Additional avenues for future research include monitoring changes over time. As awareness of AI models' use of public user-generated content grows, repeating the unavailability analysis in the future could show even greater trends of content becoming unavailable. Given this study's strength in longitudinal analysis, annual reexaminations could provide deeper insights into the evolving issue of content unavailability and users' attitudes toward their public data. Quantifying the reasons for removal on an annual basis could further enhance our understanding of this complex issue.

6 Conclusion

This paper addressed the issue of the current unavailability of images and videos in five selected open research datasets. The findings of this research point to the need to consider what should happen after these images and videos become private or are removed in the context of large AI models and open datasets. This research is important because of the present state of AI and the future direction of AI development. As competition among large AI models grows, the belief that "more data leads to better models" likely continues to drive advancements. Additionally, the idea that AI models need carefully crafted datasets labeled by knowledgeable human annotators may no longer be optimal when training these advanced models capable of digesting and interpreting large datasets. Thus, it could be advantageous for the creators of these models to collect everything they can get their hands on. Furthermore, these models are becoming more capable and continue to produce intelligent applications. Therefore, what could happen with these public images and videos could be beyond current comprehension. Finally, the movement towards the "right to be forgotten" has suggested policies and frameworks for handling data that users do not wish to continue to be available in systems. However, questions arise when the presence of the data is not disclosed, and when copies of videos, images, or text are immediately copied and prepared to be included in large AI models.

Acknowledgments

This project was funded by the Deanship of Scientific Research at Prince Sattam bin Abdulaziz University award number 2022/01/22216.

Declaration of AI

During the preparation of this work, GPT 4 and GPT 4o from OpenAI were used in order to proofread and fix mistakes in the text. After using this tool, the content was reviewed and edited as needed and full responsibility is taken for the content of the publication.

References

Aaltonen, A., Alaimo, C., Parmiggiani, E., Stelmaszak, M., Jarvenpaa, S., Kallinikos, J., & Monteiro, E. (2023). What is missing from research on data in Information Systems? Insights from the inaugural workshop on data research. *Communications of the Association for Information Systems*, *53*, 475–490.

Abbasi, A., Parsons, J., Pant, G., Sheng, O. R. L., & Sarker, S. (2024). Pathways for design research on artificial intelligence. *Information Systems Research, INFORMS, 35*(2), 441-459.

Agerfalk, P. J., Conboy, K., Crowston, K., Lundstrom, J. S. Z. E., Jarvenpaa, S., Ram, S., & Mikalef, P. (2022). Artificial intelligence in information systems: State of the art and research roadmap. *Communications of the Association for Information Systems*, *50*, 420–438.

AlNoamany, Y., & Borghi, J. A. (2018). Towards computational reproducibility: Researcher perspectives on the use and sharing of software. *PeerJ Computer Science*, *4*, e163.

Alsudais, A. (2019). Image classification in Arabic: Exploring direct English to Arabic translations. *IEEE Access*, *7*, 122730–122739.

Alsudais, A. (2021a). In-code citation practices in open research software libraries. *Journal of Informetrics*, *15*(2), 101139.

Alsudais, A. (2021b). Incorrect data in the widely used inside Airbnb dataset. *Decision Support Systems*, *141*, 113453.

Alsudais, A. (2022). Extending ImageNet to Arabic using Arabic WordNet. *Multimedia Tools and Applications*, *81*, 8835–8852.

Alsudais, A., Alotaibi, W., & Alomary, F. (2022). Similarities between Arabic dialects: Investigating geographical proximity. *Information Processing & Management*, *59*(1), 102770.

Arora, A., & Arora, A. (2022). Generative adversarial networks and synthetic patient data: Current challenges and future perspectives. *Future Healthcare Journal*, *9*(2), 190–193.

Arora, A., Barrett, M., Lee, E., Oborn, E., & Prince, K. (2023). Risk and the future of AI: Algorithmic bias, data colonialism, and marginalization. In *Information and organization* (Vol. 33, Issue 3, p. 100478). Elsevier.

Ashok, M., Madan, R., Joha, A., & Sivarajah, U. (2022). Ethical framework for artificial intelligence and digital technologies. *International Journal of Information Management*, *62*.

Benhamou, Y., & Andrijevic, A. (2022). The protection of AI-generated pictures (photograph and painting) under copyright law. In *Research handbook on intellectual property and artificial intelligence* (pp. 198–217). Edward Elgar Publishing.

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, *45*(3).

Bertram, T., Bursztein, E., Caro, S., Chao, H., Chin Feman, R., Fleischer, P., Gustafsson, A., Hemerly, J., Hibbert, C., Invernizzi, L., Donnelly, L. K., Ketover, J. Laefer, J., Nicholas, P., Niu, Y., Obhi, H., Price, D., Strait, A., Thomas, K., & Verney A. (2019). Five years of the right to be forgotten. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* 9pp. 959–972).

Besançon, L., Peiffer-Smadja, N., Segalas, C., Jiang, H., Masuzzo, P., Smout, C., Billy, E., Deforet, M., & Leyrat, C. (2021). Open science saves lives: Lessons from the COVID-19 pandemic. *BMC Medical Research Methodology*, *21*(1), 1–18.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., & Papernot, N. (2021). Machine unlearning. *2021 IEEE Symposium on Security and Privacy (SP)*, (pp. 141–159).

Brasse, J., Broder, H. R., Förster, M., Klier, M., & Sigler, I. (2023). Explainable artificial intelligence in information systems: A review of the status quo and future research directions. *Electronic Markets*, *33*(1), 26.

Caba Heilbron, F., Escorcia, V., Ghanem, B., & Carlos Niebles, J. (2015). ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 961–970).

Cantador, I., Cortés-Cediel, M. E., & Fernández, M. (2020). Exploiting open data to analyze discussion and controversy in online citizen participation. *Information Processing & Management*, *57*(5), 102301.

Cao, Y., & Yang, J. (2015). Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy* (pp. 463–480).

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., & Wallace, E. (2023). Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)* (pp. 5253–5270).

Cheong, M., Abedin, E., Ferreira, M., Reimann, R., Chalson, S., Robinson, P., Byrne, J., Ruppanner, L., Alfano, M., & Klein, C. (2023). Investigating gender and racial biases in DALL-E mini images. *ACM Journal on Responsible Computing, 1*(2), 1 - 20.

Dedema, M., & Herring, S. C. (2023). How cover images represent video content: A case study of Bilibili. In *Hawaii International Conference on System Sciences 2023 (HICSS-56).*

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255).

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W. T., & Rubinstein, M. (2018). Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics*, *37*.

Fecher, B., Friesike, S., & Hebing, M. (2015). What drives academic data sharing? *PloS One*, *10*(2), e0118053.

Gaffney, D., & Matias, J. N. (2018). Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. *PloS One*, *13*(7), e0200162.

Georgopoulos, M., Oldfield, J., Nicolaou, M. A., Panagakis, Y., & Pantic, M. (2021). Mitigating demographic bias in facial datasets with style-based multi-attribute transfer. *International Journal of Computer Vision*, *129*(7), 2288–2307.

Giermindl, L. M., Strich, F., Christ, O., Leicht-Deobald, U., & Redzepi, A. (2022). The dark sides of people analytics: Reviewing the perils for organisations and employees. *European Journal of Information Systems*, *31*(3), 410–435.

He, Y., Xu, X., Huang, N., Hong, Y., & Liu, D. (2024). *Enhancing user privacy through ephemeral sharing design: Experimental evidence from online dating*. SSRN. https://ssrn.com/abstract=3740782

Hemphill, L., Schöpke-Gonzalez, A., & Panda, A. (2022). Comparative sensitivity of social media data and their acceptable use in research. *Scientific Data*, *9*(1), 643.

Howell, S., & Burtis, A. (2023). The continued problem of URL decay: An updated analysis of health care management journal citations. *Journal of the Medical Library Association*, *110*(4), 463–470.

Jetzek, T., Avital, M., & Bjørn-Andersen, N. (2019). The sustainable value of open government data. *Journal of the Association for Information Systems*, *20*(6), 702–734.

Juliussen, B. A., Rui, J. P., & Johansen, D. (2023). Algorithms that forget: Machine unlearning and the right to erasure. *Computer Law & Security Review*, *51*, 105885.

Kaur, A., Maheshwari, S., Bose, I., & Singh, S. (2024). Watch out, you are live! Toward understanding the impact of AI on privacy of employees. *Communications of the Association for Information Systems*, *55*(1), 24.

Kim, Y., & Nah, S. (2018). Internet researchers' data sharing behaviors. *Online Information Review*, *42*(1), 124–142.

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388–409.

Kotlarsky, J., Oshri, I., & Sakar, S. (2024). The bumpy road to becoming a data-driven enterprise. *Communications of the Association for Information Systems*, *55*(1), 9.

Krasikov, P., & Legner, C. (2023). Introducing a data perspective to sustainability: How companies develop data sourcing practices for sustainability initiatives. *Communications of the Association for Information Systems*, *53*(1), 162–188.

Kurdi, M., Albadi, N., & Mishra, S. (2020). "Video unavailable": Analysis and prediction of deleted and moderated YouTube videos. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 166–173).

Kurdi, M., Albadi, N., & Mishra, S. (2021). "Think before you upload": An in - depth analysis of unavailable videos on YouTube. *Social Network Analysis and Mining*, *11*(1), 1–21.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., & others. (2020). The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, *128*(7), 1956–1981.

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. *Psychological Methods*, *21*(4), 475.

Lee, M. C., Scheepers, H., Lui, A. K., & Ngai, E. W. (2023). The implementation of artificial intelligence in organizations: A systematic literature review. *Information & Management*, *60*(5), 103816.

Leschanowsky, A., Rech, S., Popp, B., & Bäckström, T. (2024). Evaluating privacy, security, and trust perceptions in conversational AI: A systematic review. *Computers in Human Behavior*, *159*, 108344.

Liu, B., & Wei, L. (2023). Unintended effects of open data policy in online behavioral research: An experimental investigation of participants' privacy concerns and research validity. *Computers in Human Behavior*, *139*, 107537.

Lnenicka, M., Luterek, M., & Nikiforova, A. (2022). Benchmarking open data efforts through indices and rankings: Assessing development and contexts of use. *Telematics and Informatics*, *66*, 101745.

Loan, F. A., & Shah, U. Y. (2020). The decay and persistence of web references. *Digital Library Perspectives*, *36*(2), 157–166.

Luccioni, S., Akiki, C., Mitchell, M., & Jernite, Y. (2024). Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, *36*.

Lukyanenko, R., Wiggins, A., & Rosser, H. K. (2020). Citizen science: An information quality research frontier. *Information Systems Frontiers*, *22*, 961–983.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35.

Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and 'the dark side' of AI. *European Journal of Information Systems*, *31*.

Mirbabaie, M., Brendel, A. B., & Hofeditz, L. (2022). Ethics and AI in information systems research. *Communications of the Association for Information Systems*, *50*(1), 726–753.

Park, H., & Wolfram, D. (2017). An examination of research data sharing and re-use: Implications for data citation practice. *Scientometrics*, *111*, 443–461.

Raffaghelli, J. E., & Manca, S. (2023). Exploring the social activity of open research data on ResearchGate: Implications for the data literacy of researchers. *Online Information Review*, *47*(1), 197–217.

Rajabi, A., Yazdani-Jahromi, M., Garibay, O. O., & Sukthankar, G. (2023). Through a fair looking-glass: Mitigating bias in image datasets. In *International Conference on Human-Computer Interaction* (pp. 446–459).

Real, E., Shlens, J., Mazzocchi, S., Pan, X., & Vanhoucke, V. (2017). YouTube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5296–5305).

Reddit. (2024, August 16). *Reddit privacy policy*. https://www.reddit.com/policies/privacy-policy

Sadiq, S., & Indulska, M. (2017). Open data: Quality over quantity. *International Journal of Information Management*, *37*(3), 150–154.

Santos, M. L. B. dos. (2022). The "so-called" UGC: An updated definition of user-generated content in the age of social media. *Online Information Review*, *46*(1), 95–113.

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017). No classification without representation: Assessing geodiversity issues in open data sets for the developing world. In *NIPS 2017 Workshop on Machine Learning for the Developing World*.

Shen, Z. (2023). Learner engagement with YouTube videos in informal online learning: An investigation of the effects of segmenting, signaling, and weeding. *Communications of the Association for Information Systems*, *53*(1), 342–363.

Sivizaca Conde, D. J., Kämpf, N. L., Rößler-von Saß, D., Schurig, T., & Kliewer, N. (2024). Privacy-preserving data sharing: A systematic review and future research areas. In *ECIS 2024 Proceedings*.

Stock, P., & Cisse, M. (2018). Convnets and ImageNet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 498–512).

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1–9).

Varsha, P. (2023). How can we manage biases in artificial intelligence systems–A systematic literature review. *International Journal of Information Management Data Insights*, *3*(1), 100165.

Villaronga, E. F., Kieseberg, P., & Li, T. (2018). Humans forget, machines remember: Artificial intelligence and the right to be forgotten. *Computer Law & Security Review*, *34*(2), 304–313.

Wang, A., Liu, A., Zhang, R., Kleiman, A., Kim, L., Zhao, D., Shirai, I., Narayanan, A., & Russakovsky, O. (2022). REVISE: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision*, *130*(7), 1790–1810.

Wang, D., Abdelzaher, T., & Kaplan, L. (2015). *Social sensing: Building reliable systems on unreliable data*. Morgan Kaufmann.

Wang, X., Duan, Q., & Liang, M. (2021). Understanding the process of data reuse: An extensive review. *Journal of the Association for Information Science and Technology*, *July 2020*, 1161–1182.

Wang, Y., Pan, Y., Yan, M., Su, Z., & Luan, T. H. (2023). A survey on ChatGPT: AI-generated contents, challenges, and solutions. *IEEE Open Journal of the Computer Society*, *4*, 280-302.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., & others. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1–9.

Xu, D., Indulska, M., Someh, I. A., & Shanks, G. (2024). Time to reassess data value: The many faces of data in organizations. *The Journal of Strategic Information Systems*, *33*(4), 101863.

Xu, J., Mei, T., Yao, T., & Rui, Y. (2016). MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5288–5296).

Yang, K., Qinami, K., Fei-Fei, L., Deng, J., & Russakovsky, O. (2020). Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 547–558).

YouTube. (2024). *Robots.txt file for YouTube*. https://www.youtube.com/robots.txt.

About the Authors

Abdulkareem Alsudais is an Associate Professor of Information Systems at Prince Sattam bin Abdulaziz University in Saudi Arabia. He received his Ph.D. in Information Systems and Technology from Claremont Graduate University in 2017. He has published as a first author in journals such as *Decision Support Systems, Information Processing & Management, and Journal of Informetrics*. His research interests include data quality, open data, vision and language, text mining, and AI ethics.

Copyright © 2025 by the Association for Information Systems. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and full citation on the first page. Copyright for components of this work owned by others than the Association for Information Systems must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or fee. Request permission to publish from: AIS Administrative Office, P.O. Box 2712 Atlanta, GA, 30301-2712 Attn: Reprints or via e-mail from publications@aisnet.org.