Public, Identifiable, and Effectively Immutable: Ethical Lessons from Research on arXiv

Jan Pennekamp*, Johannes Lohmöller*, Peter Mayer†, Arjun Arunasalam§
*Communication and Distributed Systems, RWTH Aachen University, Germany · {lastname}@comsys.rwth-aachen.de
†Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark
§Knight Foundation School of Computing and Information Sciences, Florida International University, Miami, FL, US

***Abstract*—Ethical decision-making is central to safe cybersecurity research. Recent studies, one of which is our own, analyzing the source files of arXiv papers bring this challenge to light—arXiv publicly distributes the source files, to the surprise of many authors. Unique aspects reveal ethical challenges and dilemmas that extend beyond established ethical frameworks, institutional processes, and community best practices. In this work, we examine this case and provide insights for similar research on other public live resources, such as Zenodo or the Wayback Machine. While such empirical studies are particularly realistic, they must account for personally identifiable information, leaked secrets, and other sensitive details that could expose, shame, or otherwise negatively affect individuals, possibly leading to both reputational and practical harm.**

Drawing on our collective experience and discussions among researchers, affected authors, and readers, we make three contributions: First, we define the ethical setting and highlight that even intentionally-public data with deliberate real-name attribution can raise ethical concerns when repurposed for empirical analysis. Second, we discuss the ethical issues surrounding the use of arXiv data and derive four higher-level dilemmas that researchers may face when working with similar data. Third, we derive practical guidance for ethically navigating such settings and, in doing so, distill four lessons for research on arXiv and similar public, live datasets: public availability should serve as the starting point of ethical reasoning; research should minimize the amplification it induces on exposed information; disclosure should be calibrated to the harm, confidence, and actionability of a finding; and effective immutability must be accounted for. Together, these lessons contribute to community resources for ethically-sound cybersecurity experimentation.

***Index Terms*—ethical considerations; remediability; sensitivity; privacy; awareness; research ethics; public datasets**

1. Introduction

Cybersecurity research often touches on issues that raise ethical questions [1]. Beyond empirical studies, researching and disclosing vulnerabilities arguably also benefits attackers who can use the information to compromise vulnerable systems. To that end, responsible disclosure practices have been established in the field [2] and are widely used [3]. In contrast, public datasets used in cybersecurity research can come with different trade-offs and warrant careful ethical considerations. Passwords leaked by hackers after data breaches are used in research, e.g., [4], despite profiting from malicious actors [1]. Similarly, data on Reddit [5] or GitHub [6], for example, is widely used in research, even though individuals contributing to these live datasets may not be aware that their data is used for research [7], [8] and that their consent is rarely, if ever, obtained [9], [10].

The arXiv preprint repository [11] represents one such public dataset and is an especially complex ethical case. On arXiv, researchers publish their work under their real names, resulting in trivial author identifiability. Prior work [12]–[18] extensively sourced this resource to understand citation dynamics, for example. Another peculiarity of arXiv is that authors must also upload their LaTeX source files [19] when posting preprints to foster research transparency. Yet authors might not be aware that the paper sources they upload will be fully openly available, including all comments and other artifacts they might have forgotten to delete before uploading. At the same time, arXiv intentionally does not permit the correction of information disclosed by accident in such source files [20]–[22]. Despite this situation, a growing body of literature has used these sources for empirical studies.

The immediate implications are twofold. First, due to its increasing popularity as a dataset [20]–[22], associated ethical challenges are becoming more widespread. Second, given its popularity [23] and market share [24], many researchers are (unknowingly) affected by these studies and their practices. In turn, these observations motivate our assessment.

Paper Story. Based on our experience of working with paper sources that arXiv serves publicly, an ethics-focused analysis of existing work [20]–[22], and discussions with affected authors, we map out the landscape of ethical issues and considerations surrounding the use of this data. Specifically, in Section 2, we describe the background and related work, and, in Section 3, we then outline the unique properties of the arXiv repository as a dataset. Continuing with Section 4, we detail how prior empirical studies approached the emerging ethical aspects. Based on this assessment, in Section 5, we subsequently derive ethical dilemmas that can arise in research using similar data. Finally, we highlight important takeaways and remaining challenges in Section 6 before concluding this experience paper in Section 7.

©2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Contributions. In particular, we make the following primary contributions to ethical cybersecurity experimentation:

- **Mapping the ethical setting:** We describe the unique properties of arXiv and map the ethical issues that can arise in research that uses its data. This setting outlines how even data intentionally made public with real-name attribution can raise ethical concerns when used in research. In doing so, we contribute to a better understanding of how to work ethically with live resources.
- **Identifying dilemmas:** Rooted in our discussion of prior work and their ethical decisions, we abstract the ethical issues surrounding the use of arXiv data into four distinct dilemmas that researchers might face when using similar data sources, supporting future decision-making.
- **Guidance:** We debate the ethical responsibilities of researchers and provide practical guidance on navigating the issues that can arise. This way, we provide broader lessons for researchers working with arXiv and similar datasets.

2. Background and Related Work

In Section 2.1, we first introduce the arXiv repository and research that utilizes source files hosted there. Subsequently, in Section 2.2, we shift our focus to ethical guidance most relevant to analyzing arXiv-like knowledge infrastructures.

2.1. arXiv as a Live Knowledge Infrastructure

Since its inception in 1991, arXiv [11] has become a dominant open-access preprint service, now holding over 3 M submissions and serving more than 60 M connections per day, with computer science accounting for the largest share of preprints added each month [20], [21]. Rooted in a principle of openness, arXiv also distributes the LaTeX sources behind roughly 90 % of submissions that take its recommended source-based route, originally to enable the reuse of notation, and today also framed as supporting transparency and long-term reproducibility. It further offers this corpus for bulk retrieval via Amazon S3 [20]–[22], [25]. Authors, in turn, grant a "perpetual" and "irrevocable" distribution license, while arXiv's privacy policy places the burden of removing anything not intended for publication on the submitter [22].

While preprint adoption is difficult to quantify, studies paint a similar picture: Even back in 2017, specific subfields saw about 23 % of papers on arXiv. Another study reports that 6 % of *all* publications across disciplines were available on arXiv in 2021 [24], with physics, maths, and computer science above that average. Specifically in computer science, 77 % of posted preprints between 2008 and 2017 were eventually published [26]. To put the exposure to the scientific community in numerical terms, already by 2016, 53 % of arXiv users were themselves authors [27]. Borgman [28] characterizes such systems as *knowledge infrastructures* that evolve over time and, in doing so, shape scientific practice. Prior work mostly treats arXiv as a fixed snapshot (cf. Section 3), disregarding indirect consequences: Findings on such knowledge infrastructures can directly shape and impact the science landscape and its actors.

Three recent studies analyze this corpus for security and privacy issues. Apruzzese and Fass [20] report that across the source files of 600 k submissions, 27 % of the uploaded data is unnecessary for LaTeX compilation, including offensive text and details of ongoing research. Similarly, Pennekamp et al. [21] discovered credentials, private keys, and links to third-party resources after scanning the entire corpus, and propose a sanitization tool as a mitigation. Lastly, Dubniczky et al. [22] apply large language models to detect information leakage across 100 k submissions. All three include ethics statements yet reach different conclusions on the same corpus, e.g., on whether to release tooling, and on whether following embedded links is acceptable. These different perspectives on an unambiguously public dataset motivate our work.

2.2. Ethical Guidance for Research on Public Data

Given that the arXiv corpus spans multiple dimensions, covering both contained and referenced data, several ethical considerations are also important. Next up, we will thus feature those directions that are generally of interest.

Ethics Frameworks. The Association of Internet Researchers (AoIR) guidelines are an established reference for Internet research ethics. Version 1.0 [29] is deliberately pluralistic and case-based; Version 2.0 [30] adds three commitments with implications to our setting: (i) greater vulnerability of the studied party implies a greater obligation of the researcher; (ii) "human subject" as a condition fits text-based research poorly and one should instead consider caused harm, vulnerability, and identifiability; and (iii) public/private boundaries are not binary, as people act in public venues while retaining expectations about how their data flows. Version 3.0 [31] extends them to Big Data and machine learning, introduces a taxonomy spanning design through dissemination and project close, and poses dedicated questions for *existing datasets*: whether a dataset is explicitly open for research use, whether any initial consent covers the intended reuse, and whether data subjects could conceivably object to it. Its accompanying impact model separates the immediate treatment of people from side effects and from downstream use of data beyond the initial analysis.

Complementarily, Metcalf and Crawford [32] argue that data science is frequently research *about* humans without qualifying as human-subjects research, so the regulatory exemption for publicly-available data is based on a superseded assumption. Nissenbaum's contextual integrity [33] supplies an alternative, framing privacy as appropriate information depending on context. Zimmer [9] specifies a framework for Big Data that addresses concerns around consent, anonymity, and harm that a binary public/private test leaves unaddressed.

Data statements [34], datasheets [35], and accountability practices adapted from software engineering [36] have also been proposed. All assume a curated artifact with an identifiable creator who can state provenance, composition, and intended use, which, however, is not the case for arXiv.

Empirical Practice. Surveying 263 online-data researchers, Vitak et al. [37] find that disagreements about ethical practices are uncorrelated with discipline, method, or

workplace. Their follow-up with staff at 59 US institutional review boards (IRBs) [38] finds no consensus on which social computing protocols require review, alongside a self-reported lack of technical expertise. Bruckman [39] models Internet users as amateur artists producing "semi-published" work and concludes that the resulting material has to be treated carefully. Regarding ethicality, they argue that being too lax risks violating individual subjects' rights and disrupting the community, while strictness might push researchers to avoid IRB oversight in the first place.

Data subjects, when asked, seem to not generally endorse a public-data exemption: 61.2 % of surveyed Twitter users did not know that researchers use tweets at all, and 64.9 % noted that researchers should not do so without permission, with comfort dropping sharply if entire histories instead of isolated posts are studied [40], contrasting IRB-centered studies that report IRB respondents feel that researchers do not need informed consent for public data [10], [38]. Comparable expectations hold in online fandom [41]. Similarly, users accept population-level depression monitoring far more readily than individual-level assessment [42], suggesting that aggregation may be a potential mitigation. Reported practice, nonetheless, is scarce: Across 727 Reddit studies, only 13.9 % mention any ethics review, mostly to record an exemption. 6.1 % mention consent, 9.4 % name identifiable usernames, and 28.5 % quote users verbatim [5].

Internet Measurement Ethics. Network measurement has developed its own norms, centered on the Menlo Report [43] and on scanning practices such as rate limiting, attributable probes, and opt-out [44]. Yet Fiebig [45] argues that this practice systematically undermines both the ethics and the reliability, and Amin et al. [46], based on 16 interviews, report that researchers treat ethics as craft knowledge passed on by mentors and peers, because review boards lack the technical grounding to assess such work. Studies of exposed secrets on the web [47], [48] come closest to analyzing arXiv submissions in subject, but their data subjects are usually service operators and only rarely named individuals.

Closest to our work, Kohno et al. [1] structure security-specific dilemmas as trolley problems to make them explicit. Their scenarios, like most work on dataset ethics, concern—by design—data deliberately released by a third party with foreseeable negative implications, or data obtained illegitimately (e.g., stolen or leaked). However, arXiv source files are neither: They were deliberately released by the affected individuals themselves under a norm of openness and without any expectation that residual data would be read.

3. Problem Statement: arXiv as a Live Dataset

From an abstract perspective, the arXiv "dataset" is comparable to other live resources that are available on the Internet, such as social media content or snapshots from Internet scans (e.g., Censys [49] or Shodan [50]). Because these resources continue to evolve, they can in principle support longitudinal monitoring of individuals, including observations of newly contributed data as it becomes publicly available. They all have in common that they are not released, curated datasets (no official identifier/reference), but living publicly-available datasets. Further, they neither disclose to affected data subjects that they are part of the dataset nor do they offer effective means to opt out of studies, since the compiled data is publicly available. In health research involving identifiable patient data, for example, access and reuse are commonly governed by formal consent, authorization, or waiver mechanisms [51], [52]. However, as we discuss in Section 3.1, arXiv as a dataset comes with unique aspects; hence, established ethical guidelines are not comprehensive enough. Our assessment of related work in Section 3.2 further underlines that lacking or superficial ethical statements are the norm rather than an exception, calling for a behavior change.

3.1. arXiv's Uniqueness and Dataset Dimensions

Different from the other live resources we have just discussed, the arXiv dataset comes with three unique aspects:

▶ *Intentional Behavior.* Unlike in many measurement settings, where individuals become part of a dataset incidentally or without taking deliberate action, arXiv authors intentionally submit their papers and associated source files for dissemination. However, intentionally submitting a paper for dissemination does not necessarily imply awareness of, or intent regarding, every piece of information they contain, e.g., when considering its associated source files. Authors may intend to make their paper publicly available while remaining unaware that residual comments, obsolete text, or other source artifacts are also publicly retrievable, as suggested by prior work [20], [21].

▶ *Strong Identifiability.* All arXiv submissions and their source files are, by design, directly associated with a fixed set of named authors. This differs from many other datasets in which individuals are anonymous, pseudonymous, or identifiable only through additional inference. Consequently, findings derived from a submission can readily be associated with identifiable researchers or research teams, even when the authorship of a particular source-file artifact cannot be established. Such identifiability may therefore increase the reputational or professional implications of sensitive findings.

▶ *Effective Immutability.* While arXiv, as a service and dataset, is not per se immutable, its operation effectively results in established immutability. Authors have little means to remove or sanitize sensitive information from the dataset, as arXiv opts for amendments and updates rather than replacing data, thereby circumventing permanent deletion.

Moreover, affected individuals are not merely data subjects: they may simultaneously be authors of the work, contributors to the submitted source files, submitters responsible for the arXiv record, and private individuals. As a result, ethical assessments may differ across the roles they affect.

"Similar" Datasets. Researchers frequently use GitHub [53], the Wayback Machine [54], and Zenodo [55] as datasets for studies [6], [56]–[63]. All of them match a subset of arXiv's unique aspects: GitHub typically serves author-provided data that includes a history and provides strong identifiability, often including embedded author email addresses.

In contrast to arXiv, pseudonyms are officially supported, and hosted repositories are not immutable. Besides, different visibility settings enable access control. On a different note, the Wayback Machine also provides an immutable dataset with version history. However, unlike arXiv, the recorded data does not necessarily provide strong identifiability, and information is only available in the exceptions that the authors themselves provide. Lastly, Zenodo most closely matches arXiv's aspects; however, authors can also be organizations, slightly relaxing the strong identifiability.

Dataset Dimensions. Source files served by arXiv—*the dataset*—specifically touch upon three distinct dimensions.

► *Self-Contained Data.* This dimension covers information confined to the dataset itself, primarily conversational LaTeX comments (e.g., author conversations), as well as todo lists and obsolete words, sentences, and paragraphs. Other examples include details on submission histories or (prior) template use, reviews and revisions letters, or use of profanity.

► *Externally-Actionable Data.* Data belonging to the second dimension links information from the dataset with third-party services, i.e., only in combination is the data of value. The most prominent example covers access tokens or passwords that can be used to authenticate users. Other, less straightforward cases are identifiers (e.g., email or cryptocurrency addresses and paper submission IDs), credit card details, or GPS coordinates.

► *External References.* Finally, the third dimension captures data that solely points to external services. Common examples include links to cloud services (including Google Docs or Drive), FTP servers, and IP addresses. This information is not actionable within the dataset itself; it is independent and largely exclusive to the referenced service.

We do not find a strict mapping between these dimensions and the different types of "hidden" information. That is, each dimension touches on the file, metadata, and content/text levels [21]; however, the *Externally-Actionable Data* dimension is particularly prominent as part of dangling files [21]. Lastly, these dimensions (Table 1) mandate individual ethical considerations to account for the diversity in sensitivity, potential harm, and affected stakeholders.

3.2. Ethics Statements in Related Work Using arXiv

In the context of using arXiv as a dataset, the research landscape reveals that most studies do not discuss ethical implications at all. Overall, we come across various research directions, working with LaTeX [64], [65], quantifying the impact of preprints [12]–[17], sourcing content from papers hosted at arXiv [66]–[70], interacting with served papers [71]–[74], and understanding authorship and publication forms [75]–[80]. This situation is particularly surprising since some works [71], [72], [74], [79], [80] study arguably sensitive matters like plagiarism detection and scientific misconduct, or publish arXiv-extracted datasets themselves [76]. Interestingly, all papers investigating arXiv source files for secrets [20]–[22] feature ethics statements of varying length (900, 1098, and 420 words, respectively), distinguishing the decisions to conduct the research and to publish the findings. To the best of our knowledge, current assessments are limited to selected papers [20]–[22], [81] that indeed discuss privacy implications.

We further looked at 35 papers that utilize the *ogb(n)-arxiv* [82], [83] dataset, which contains the citation network of computer science research, originally extracted from papers published at arXiv. Most of them omit an ethics section entirely; three papers disregard negative consequences, albeit two without discussing privacy. The remaining paper argues that no new data collection took place (cf. Section 2.2).

Research Gap. In our view, the nuances of the arXiv ecosystem and their intersection with ethics are not yet sufficiently covered by prior work, warranting a closer look. In fact, research areas differ in their stance toward ethics: While studies working with social media content largely argue that the data is public information and thus their work is exempt from data subject consent and IRB approval [5], Internet measurements usually discuss their ethical implications (cf. Section 2.2) but do not take specific action when solely relying on readily-available datasets from by third-party services, such as Censys or Shodan, i.e., they also conclude that they are not directly interacting with the data subject.

4. Assessing Ethics Decisions in Prior Work

The ethics statements of prior work [20]–[22] are the focus of our assessment. In line with this paper's scope (cf. Section 1), we limit ourselves to studies that work with paper sources, i.e., that empirically analyze the arXiv dataset. In particular, we highlight the reported ethical considerations and decision-making, and further discuss the corresponding outcomes to understand the implications of these decisions.

To this end, in Section 4.1, we first briefly examine how the studies affect arXiv's operation before individually covering the three dataset dimensions in Section 4.2. Finally, we conclude our assessment in Section 4.3 by focusing on how the authors account for arXiv's effective immutability.

4.1. Interacting with a Live Service

Given that arXiv is a service used by millions of users daily [23], ethical research must ensure it does not disrupt the service. Compliant with arXiv's recommendation for bulk access [25], prior work almost exclusively relied on the official Amazon S3 endpoint, rather than directly interacting with the production system's infrastructure. Our interpretation is that researchers likely avoided straining the service's resources, which could otherwise be an ethical concern. Only Pennekamp et al. [21] accessed the official arXiv API to complete their dataset. Nevertheless, they reported that they remained within the documented, appropriate usage rates [84]. Collectively, the decisions made are in line with other analyses, e.g., of citation graphs (cf. Section 3.2). Thus, the research-induced impact on arXiv and its operation appears to have been marginal in all prior studies, not raising deeper-rooted ethical issues. In the following, we shift our attention to more challenging matters.

TABLE 1. EXAMPLES OF THE DATASET DIMENSIONS INTRODUCED IN SECTION 3.1, ALONG WITH THEIR RETROSPECTIVE REMEDIABILITY AND POTENTIAL SENSITIVITY.

Dimension	Remediability	Sensitivity	Representative Examples
Self-Contained Data	Low / ✗	L–M	Author conversations, profanity, in-file todo lists, obsolete text and markup, LLM response patterns
Externally-Actionable Data	Partial / (✓)	M–H	Access tokens, credentials, private keys, (user) identifiers, credit card details, GPS coordinates
External References	High / ✓	L–H	Links to cloud services, websites, file servers, IP addresses

Remediability describes how readily an issue can be addressed retrospectively, given arXiv's effective immutability (cf. Section 3.1). Sensitivity can only be classified subjectively; in line with prior work [21], [85], we assign high (H), medium (M), or low (L) according to potential impact following public exposure of information.

4.2. arXiv's Dataset Dimensions and Implications

The identified dataset dimensions (Section 3.1) also come with unique implications. Table 1 provides an overview of the different dimensions and whether discovered issues are fixable, given arXiv's effective immutability.

As prior work [21], we *coarsely* label findings as high (H), medium (M), or low (L) to indicate their *potential* severity. Although such a labeling is highly subjective [21], a coarse categorization remains useful for distinguishing the relative severity of different findings and structuring the subsequent ethical assessment. The examples detail which findings *may* trigger negative implications. While the exposure of access tokens or links to publicly accessible cloud services (without enforced access control) is simple to grasp, other aspects are less trivial. Aspects such as using profanity, bashing competitors, or "exposing" scientific misconduct can be career-altering. Similarly, GPS coordinates, address details, or the use of GeoIP databases in combination with IP addresses can have real-world implications, potentially leading to in-person stalking. In particular, since authors cannot address all issues retrospectively, researchers must carefully consider the (real-world) implications when conducting their analyses.

Given their differences, we now assess the three dimensions commonly featured in paper sources on arXiv, step by step, and outline the corresponding ethical considerations.

Self-Contained Data. The combination of the scale (millions of unique comments) and context-dependent interpretation (obsolete or "only" uncommented remarks) generally impairs accurate identification of the respective issues with certainty. Notwithstanding, author conversations, obsolete text fragments, or outdated comments have mostly local implications. Likely, given this context dependence and the limited ability to remediate identified instances retrospectively, prior work has reported on the general prevalence rather than highlighting individual cases. As a safety measure, prior work collectively refrained from pointing to individual instances (and their identifiers), retaining the needle in the haystack. Due to these factors, we perceive Self-Contained Data to be low sensitivity (L). Still, scientific misconduct could even be career-altering, contributing to higher sensitivity (M; possibly even H in extreme cases).

Externally-Actionable Data. Here, all studies decided against "testing" any authentication tokens or credentials, which matches with prior work that discovered "hidden" secrets [6]. This approach appears to be in line with the content the Menlo report [43] expresses [22]. The same holds for (user) identifiers, credit card details, or information that potentially reveals the author(s)' whereabouts. Interestingly, unlike the studies we cover, recent work [48], [86] instead decided to verify their findings by "unintrusively" connecting to affected services. Still, we posit that connecting to third-party services greatly increases the risk of harm and may also unintentionally trigger safety guards (which is positive as a responsible disclosure mechanism) or notifications (which may be negative, as they can stress affected/notified users).

External References. This externally-focused dimension is very closely related to Internet measurements and their guidelines (cf. Section 2.2). Given the diversity of findings and referenced services, servers, and documents, processing these pointers touches upon a broad(er) set of external endpoints. Unlike the credentials and tokens reported previously, two studies [20], [21] accessed Google Docs links to determine whether authors had enabled access control and, if not, possibly to understand what kind of data is exposed in this way. Their decisions to restrict themselves to a subset minimize harm and avoid the use of exposed credentials.

Users accessing the source files usually have no way to determine whether accessing a referenced resource is "unauthorized" or not. While "hidden" references are not part of the compiled PDF, it is fair to assume that authors are unlikely to annotate the sensitivity of these references either. The remaining risks are two-fold. First, researchers may access sensitive information that was not intended for public access. Second, researchers may unwillingly be exposed to illegal or disturbing content. Still, the benefits of learning what kind of information may be (unwillingly) exposed in a small sample outweigh the arguably marginal harm to affected authors, particularly because malicious actors—with no intent to inform affected authors—could easily do so as well. The discoveries vary greatly in sensitivity from L to H, which we attribute to the diversity of referenced content. However, this decision of sensitivity is disputed among prior work [20]–[22] and should be reevaluated frequently.

Better understanding which other resources and with which permissions are exposed could further characterize the scope and nature of these external references. However, we expect that such an extended validation would reveal similar findings at greatly amplified risks, given potentially less predictable service behavior. Most likely, such attempts would even again reveal transitively-exposed leaks [21].

4.3. Accounting for arXiv's Effective Immutability

As summarized in Table 1, the three dataset dimensions differ not only in their potential sensitivity but also in their remediability. These differences are closely tied to arXiv's *effective immutability*, a distinguishing property of the dataset. Next, we examine how prior work navigates this constraint.

Immutability Challenge. Given that arXiv is effectively immutable and arXiv argues that authors are responsible [21], mitigation strategies to combat "hidden" information are quite limited. Certainly, for *External References*, authors can likely (re-)configure access control and authorization directly at the exposed third-party service, leading to comparably good remediability (✓). Similarly, in terms of *Externally-Actionable Data*, certain tokens and credentials can be disabled or revoked, though they are not fully removed from the arXiv dataset. Unfortunately, for other aspects, e.g., included (user) identifiers, or location data, no such "convenient" option is available, resulting in only (✓) remediability. Lastly, any data of *Self-Contained Data* is not fixable at all (✗), simply because it is confined to the "immutable" data at arXiv.

While an (undocumented) removal of sources is possible in extreme cases, arXiv is not interested in broader action [21]. As of 10/2026, we are in talks with arXiv to have them address a few severe issues. Regardless, (un)authorized copies of the dataset result in data still floating, forcing affected authors to live with the fact that sensitive data is publicly exposed—in line with the mantra *"The Internet never forgets"*. Fortunately, actions like access-control adjustments or the reissuing of credentials leave some room for containment.

Mitigation Support. Despite this somewhat grim outlook, two of the three papers [20], [21] initiated a responsible disclosure to inform affected authors. Here, a key question is whether this outreach is ethically warranted (a) given that many issues are not fixable and (b) sensitivity assessments of issues are highly subjective, possibly resulting in stressful and undesired situations. Given that the dataset remains publicly accessible, we conclude that the early warning paired with the option to apply some fixes (✓ and (✓)) is still beneficial and probably in the broader interest of affected authors.

5. Touched Ethical Dilemmas

After detailing the ethical considerations that affected the aforementioned arXiv studies, we now take a step back and abstract the reported situation into higher-level dilemmas. That is, we derive dilemmas that may also apply to other research that involves live datasets. Specifically, we want to raise awareness by highlighting four dilemmas that warrant careful evaluation in future studies. While the first two dilemmas touch upon the accuracy of the research, namely quantifying and validating findings, the third dilemma concerns the scale of empirical studies. Lastly, we turn to a fourth dilemma that concerns the responsible disclosure of research findings and the associated benefits, or lack thereof.

Dilemma 1: Sensitivity vs. Need for Quantifying.

The *ACM Code of Ethics* [87] calls for respecting privacy, yet quantifying how much sensitive information a dataset exposes, and how severe it is, requires classifying that information, which in turn requires reading it. This dilemma thus concerns the tension between accessing potentially sensitive private information and the need to quantify its prevalence and sensitivity. The dilemma itself persists across studies, whereas a legitimate research interest is a case-specific factor that may resolve it: For example, establishing that exposed links indeed lead to unprotected documents requires opening a sample of them, while counting such links avoids storing private information, but the resulting estimates may significantly over- or underreport the actual exposure. In the studies we assessed, all three [20]–[22] processed *Self-Contained Data* at scale to report prevalence but published aggregates only; for *External References*, two [20], [21] opened a small sample of referenced documents to assess sensitivity (cf. Section 4.2), while the third [22] did not.

Dilemma 2: Validation Attempts vs. Overreporting.

The second dilemma is closely related and boils down to intentionally "working" with credentials or other data to validate their authenticity, e.g., for authorization purposes or to delimit access boundaries. Different views have emerged (cf. Section 4.2), with some [48], [86] arguing that careful validation attempts are key to avoiding excessive reporting of false positives. We certainly acknowledge the risk of overreporting findings but generally believe that accessing services and resources using data meant to remain private without consent usually contradicts desirable norms [43], [87]. Across the three studies we investigated, none attempted to validate user credentials, access tokens, or credit card information, to our understanding. Instead, Pennekamp et al. [21] confined their validation to non-invasive approaches, such as checking for valid semantic patterns or recalculating embedded checksums, to assess how authentic they appear. The other two studies [20], [22], in contrast, did not report corresponding validation attempts in their papers.

Dilemma 3: Problem Scale vs. Available Resources.

Particularly when dealing with massive datasets, research is challenged by the scale of the problem and may sacrifice the usual precision and extensive coverage for two reasons. First, if (fully) automated processing is not possible, for example, because human reasoning or contextualization is required, the analysis can quickly exceed the available "processing" resources. Likewise, the human attention span is quite limited [88], which may lead to unintentional bias or mental overload for the conducting researchers. Second, from an ethical perspective, sustainability considerations may also play a key role, e.g., the desire to avoid excessively "burning" computing power [89]—an aspect that is increasingly relevant in the age of large language models (LLMs). While this dilemma is certainly not confined to the arXiv, prior work reveals significant differences in how it is addressed. On the one hand, the considered dataset sizes vary significantly, as we introduced in Section 2.1. On the other hand, resource utilization also differs, ranging from no reported use of LLMs [20], through selected use [21], to work that places LLMs at the center of a research question [22]. While a direct comparison of the environmental footprint is not possible, this overview highlights that researchers come to different conclusions concerning what is sustainable.

Dilemma 4: Awareness Benefit vs. Attention.

Moving to a human-centered dilemma, responsible disclosure intends to raise awareness among affected individuals, but the benefits may not always outweigh the harms it can cause. Li et al. [90] argue that when individuals are notified, the potential good outweighs the potential harm. Still, learning that one's data was part of a data breach can evoke a range of negative emotions, from concern [91], [92] to severe emotional distress [93]. Thus, when informing affected stakeholders, researchers should also consider how beneficial the resulting awareness is, particularly when weighed against the stress and time expenditure that a notification may cause. Past research [94] shows that even good intentions can have negative consequences. The wording of the notification can also be essential: legal framings of notification messages can be effective in eliciting a response [95], but may also cause severe reactions [96]. Compounding this issue, poorly crafted notifications may be misclassified as spam by recipients [96].

However, the benefits of raising awareness cannot be considered in isolation, as the same disclosure or publication activities may also direct additional attention toward the exposed information itself. Research outputs, datasets, or analysis tools may make information that was previously difficult to discover substantially easier for third parties to locate, attribute, or act upon. Pennekamp et al. [21], for example, deliberately refrained from releasing their analysis code to prevent potential misuse and avoid enabling third parties to easily identify submissions containing sensitive information. This decision illustrates how efforts to raise awareness of a problem can simultaneously increase the practical visibility of the underlying exposure. Dubniczky et al. [22], in contrast, did not conduct a responsible disclosure. This choice highlights the other side of the dilemma: foregoing notification avoids imposing additional burden on affected individuals and may limit further attention to the exposure, but it also forgoes any remediation benefit that awareness could provide. However, the reasons for their decision are not reported, and we therefore cannot determine whether such considerations influenced it.

Consequently, while notifying affected individuals about identified threats or issues should generally be considered, the decision to do so must be a conscious one that weighs both the potential benefit and the potential harm. For instance, in the case of arXiv, authors are not always able to remediate discovered issues while leftover comments might impact careers (e.g., leftover harsh criticism of related work). This situation might lead to feelings of helplessness, resignation, or a sense of inevitability that cyber incidents will always happen and that one is at their mercy [92], which, in turn, may unnecessarily expose individuals to stress or lead them to spend time attempting fixes that are no longer possible. At the same time, researchers should consider whether the manner in which they communicate or publish their findings increases the practical visibility of exposed information beyond what is necessary to achieve the intended awareness benefit.

6. Discussion, Challenges, and Takeaways

The previous sections illustrate that the ethical challenges surrounding arXiv do not arise simply because researchers encounter sensitive information. Rather, they emerge from the interaction between public availability, strong identifiability, effective immutability and the possibility that research itself changes how exposed information can be discovered or acted upon. In this section, we synthesize these observations into broader lessons for researchers working with arXiv and related "immutable" live datasets. We deliberately address researchers rather than repository operators, as repository-side mitigations (e.g., no longer serving source files publicly) have already been discussed in prior work [20]–[22].

6.1. Public Availability and Ethical Responsibility

Public availability should be the starting point of ethical reasoning rather than its conclusion. Particularly in light of the increasing prevalence of open-source intelligence [97], information leaks pose increasing risks [22]. Prior work [5], [40]–[42] on Twitter, Reddit, and other online platforms shows that technical accessibility does not necessarily imply that individuals expect their information to be collected or analyzed for research. arXiv presents a nuanced case because authors intentionally submit their papers for public dissemination (cf. Section 3), but this intentional action does not necessarily imply equivalent awareness or expectations regarding how broadly the submission will be inspected. For example, authors may post a preprint while it is under double-blind review without anticipating substantial readership at that stage [98]. Moreover, authors may understand that the paper itself is public while remaining unaware that the accompanying source files are likewise publicly retrievable, as evidenced by survey respondents in Pennekamp et al.'s study [21]. Residual artifacts within those sources, such as comments, TODOs, obsolete text, or internal discussion, may therefore not have been intended for broad inspection. The resulting tension is therefore between open science and reproducibility, on the one hand, and privacy and data minimization, on the other.

Informed consent [99] provides limited practical guidance at this scale. Researchers could theoretically request consent from individuals, but doing so across hundreds of thousands of submissions is unlikely to be feasible [100] and would likely result in substantial non-response. This situation also raises the familiar tension between opt-in and opt-out models for public-data research [9], [10]. Similarly, arXiv's scale and heterogeneity challenge identifying vulnerable populations in advance [30]. Researchers should therefore remain attentive to whether particular findings, contexts, or groups introduce vulnerability as analyses progress.

These observations return us to the sensitivity-versus-quantification dilemma (Dilemma 1, Section 5). Rather than asking only whether information is publicly accessible, researchers should consider whether the proposed use is consistent with the context in which it was shared and whether inspecting more sensitive information is necessary to

answer the research question. Researchers should thus justify why the anticipated research benefit warrants the chosen level of inspection. In particular, the potential harm to individuals, including feelings of helplessness or distress in the face of "unfixable" public disclosure, should be considered with care.

6.2. Minimize Research-Induced Amplification

Researchers may not be responsible for creating the original exposure, but their work can amplify it. Information hidden among millions of files differs from information that research makes easier to discover, validate, attribute, or redistribute. The ethical question is therefore not only whether researchers may access information but also how their actions change its practical accessibility, particularly for actors with malicious intent, and impact. This concern directly reflects the attention aspect of Dilemma 4: efforts to raise awareness of an exposure can simultaneously draw additional attention to the underlying information and make it easier for others to discover or act on.

Amplification can occur throughout the research process. Automated discovery can transform "needle in the haystack" information into searchable findings. Validation, such as following external references or interacting with credentials, can shift from passive observation to active intervention (cf. Dilemma 2). Publication can further amplify exposure when examples, datasets, or artifacts make sensitive findings easier to locate. At arXiv scale, automated processing is likely unavoidable given limited human resources (cf. Dilemma 3), but false positives and missing context should constrain how confidently findings are interpreted and attributed. In turn, research might even trigger real-world actions.

Particular caution is warranted for reputationally-sensitive findings. A source-file comment stating that "we fudged certain numbers" may appear to indicate misconduct, yet researchers may not know who wrote it, whether it was serious or obsolete, or whether other authors were aware of it. Since arXiv's strong identifiability links every source file to named authors (cf. Section 3), turning ambiguous evidence into an attributable claim risks creating a digital pillory. Researchers should therefore favor aggregate reporting, avoid unnecessary identifiers, limit redistribution of sensitive examples, and constrain validation to what is scientifically necessary. The greater the potential harm of a false positive, the stronger the evidence should be before attribution.

6.3. Disclosure: Harm, Confidence, Actionability

Responsible disclosure should depend on the nature of the finding: Actionable security issues, such as exposed credentials or private keys, may warrant private notification to enable revocation, access-control changes, or other containment measures. Conversely, ambiguous reputational findings, such as misconduct or sensitive author conversations, may cause substantial harm if misinterpreted or incorrectly attributed.

Disclosure activities can range from private notification or escalation to an appropriate intermediary or institution, to limited public attribution or non-disclosure. The recipient of a notification is likewise not always evident, as affected individuals act simultaneously as authors, submitters, and private individuals (cf. Section 3), and the submitter of a paper need not be the person who wrote a sensitive comment. Researchers must further consider that notification itself may also create stress when affected parties have little ability to remediate the underlying exposure (cf. Dilemma 4). Decisions should therefore consider severity, confidence, who may be affected, false-positive risk, and whether recipients can meaningfully respond. Thus, disclosure should be treated as a context-dependent harm-reduction decision.

6.4. Immutability Influences the Ethical Calculus

Effective immutability, as prevalent for some live datasets, matters throughout the research lifecycle because findings differ in remediability: credentials or permissions may be changed. Then again, comments or obsolete text may remain permanently accessible. Research should therefore account for fixability when collecting, validating, and disclosing findings, favoring minimized collection, limited validation, and aggregate reporting when remediation is difficult.

Repositories may also consider narrowly-scoped remediation or redaction mechanisms for sensitive historical material, balancing privacy and harm reduction against provenance and reproducibility. Similar concerns may arise for GitHub [101], the Wayback Machine [102], or Zenodo, where contributors may not expect technically public material to be systematically examined. However, differences in platform purpose, user expectations, versioning, and residual information require a dataset-specific assessment of remediation, as they generalize poorly across all "public repositories."

7. Conclusion

Public availability alone is not a sufficient basis for ethically unproblematic research use. In this paper, we examined prior work [20]–[22] that utilizes arXiv paper sources as a particularly illustrative case. Even though the data is intentionally submitted, publicly distributed, directly attributable to named individuals, and effectively immutable, it may still contain sensitive information that was unintentionally disclosed. This combination creates ethical tensions that are not fully covered by existing frameworks, review processes, or community norms.

Our work stresses that ethical decision-making in cybersecurity research requires more than checking whether data is publicly accessible. Instead, researchers must treat public availability as the starting point of ethical reasoning, minimize the amplification that their work induces on exposed information, calibrate disclosure to the harm, confidence, and actionability of each finding, and account for the limited remediability that effective immutability entails. Alongside debating these challenges, we further support future studies by deriving four higher-level dilemmas that warrant consideration. Collectively, we contribute a resource that aids the design of ethical cybersecurity experimentation.

Acknowledgment

We appreciate the openness of and the valuable exchange with the computer security community when putting together this work: Fruitful discussions enabled this exhaustive work. We are further grateful for the valuable feedback from Sophie Stephenson and Tadayoshi Kohno. We also thank the anonymous reviewers for their valuable feedback and suggestions.

Research Artifact Statement

This work did not produce any artifacts. Its contributions are of an intellectual nature; it discusses works that released their artifacts. Prior work [21], [22] released partial artifacts.

LLM Usage Statement

No LLMs or other generative AI tools have been used to prepare this work. The text has been edited using suggestions provided by Grammarly. The author(s) manually approved these suggestions and approve this version of the paper.

References

[1] T. Kohno, Y. Acar, and W. Loh, "Ethical Frameworks and Computer Security Trolley Problems: Foundations for Conversations," in *Proceedings of the 32nd USENIX Security Symposium (SEC '23)*. USENIX Association, 2023, pp. 5145–5162.

[2] A. D. Householder, G. Wassermann, A. Manion, and C. King, "The CERT® Guide to Coordinated Vulnerability Disclosure," Carnegie Mellon University, Special Report CMU/SEI-2017-SR-022, 2017.

[3] T. Walshe and A. Simpson, "An Empirical Study of Bug Bounty Programs," in *Proceedings of the 2020 IEEE 2nd International Workshop on Intelligent Bug Fixing (IBF '20)*. IEEE, 2020, pp. 35–44.

[4] M. Golla and M. Dürmuth, "On the Accuracy of Password Strength Meters," *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS '18)*, pp. 1567–1582, 2018.

[5] N. Proferes, N. Jones, S. Gilbert, C. Fiesler, and M. Zimmer, "Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics," *Social Media + Society*, vol. 7, no. 2, 2021.

[6] M. Meli, M. R. McNiece, and B. Reaves, "How Bad Can It Git? Characterizing Secret Leakage in Public GitHub Repositories," in *Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS '19)*. Internet Society, 2019.

[7] K. Shilton, E. Moss, S. A. Gilbert, M. J. Bietz, C. Fiesler, J. Metcalf, J. Vitak, and M. Zimmer, "Excavating awareness and power in data science: A manifesto for trustworthy pervasive data research," *Big Data & Society*, vol. 8, no. 2, 2021.

[8] R. Shahriari and E. D. Ragan, "A Systematic Survey of Empirical User Studies of Unintentional Information Disclosure in Everyday Digital Interaction," *International Journal of Human–Computer Interaction*, 2026.

[9] M. Zimmer, "Addressing Conceptual Gaps in Big Data Research Ethics: An Application of Contextual Integrity," *Social Media + Society*, vol. 4, no. 2, 2018.

[10] C. Fiesler, "Ethical Considerations for Research Involving (Speculative) Public Data," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. GROUP, 2019.

[11] arXiv, "arXiv.org e-Print archive," https://arxiv.org/, 1991.

[12] M. J. Kurtz, G. Eichhorn, A. Accomazzi, C. Grant, M. Demleitner, E. Henneken, and S. S. Murray, "The effect of use and access on citations," *Information Processing & Management*, vol. 41, no. 6, pp. 1395–1402, 2005.

[13] P. M. Davis and M. J. Fromerth, "Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles?" *Scientometrics*, vol. 71, no. 2, pp. 203–215, 2007.

[14] H. F. Moed, "The effect of "open access" on citation impact: An analysis of ArXiv's condensed matter section," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 13, pp. 2047–2054, 2007.

[15] V. Larivière, C. R. Sugimoto, B. Macaluso, S. Milojević, B. Cronin, and M. Thelwall, "arXiv E-prints and the journal of record: An analysis of roles and relationships," *Journal of the Association for Information Science and Technology*, vol. 65, no. 6, pp. 1157–1169, 2014.

[16] A.-u. Haque and P. Ginsparg, "Positional effects on citation and readership in arXiv," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 11, pp. 2203–2218, 2009.

[17] A.-u. Haque and P. Ginsparg, "Last but not least: Additional positional effects on citation and readership in arXiv," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2381–2388, 2010.

[18] K. Okamura, "Scientometric engineering: Exploring citation dynamics via arXiv eprints," *Quantitative Science Studies*, vol. 3, no. 1, pp. 122–146, 2022.

[19] arXiv, "Requiring TeX when possible," https://info.arxiv.org/help/faq/whytex.html, 2005.

[20] G. Apruzzese and A. Fass, "X-raying the arXiv: A Large-Scale Analysis of arXiv Submissions' Source Files," arXiv:2601.11385, 2026.

[21] J. Pennekamp, J. Lohmöller, D. Schütte, J. Loos, and M. Henze, "Hidden Secrets in the arXiv: Discovering, Analyzing, and Preventing Unintentional Information Disclosure in Source Files of Scientific Preprints," in *Proceedings of the 2026 IEEE Symposium on Security and Privacy (SP '26)*. IEEE, 2026, pp. 3358–3377.

[22] R. A. Dubniczky, B. Borsos, and N. Tihanyi, "You Have Been LaTeXpOsEd: A Systematic Analysis of Information Leakage in Preprint Archives Using Large Language Models," in *Proceedings of the 20th USENIX WOOT Conference on Offensive Technologies (WOOT '26)*. USENIX Association, 2026.

[23] arXiv, "Monthly Download Rates," https://arxiv.org/stats/monthly_downloads, 2014.

[24] B. Xie, Z. Shen, and K. Wang, "Is preprint the future of science? A thirty year journey of online preprint services," arXiv:2102.09066, 2021.

[25] arXiv, "Full Text via S3," https://info.arxiv.org/help/bulk_data_s3.html, 2010.

[26] J. Lin, Y. Yu, Y. Zhou, Z. Zhou, and X. Shi, "How many preprints have actually been printed and why: a case study of computer science preprints on arXiv," *Scientometrics*, vol. 124, no. 1, pp. 555–574, 2020.

[27] O. Y. Rieger, G. Steinhart, and D. Cooper, "arXiv@25: Key findings of a user survey," arXiv:1607.08212, 2016.

[28] C. L. Borgman, *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press, 2017.

[29] C. Ess and AoIR Ethics Working Committee, "Ethical decision-making and Internet research: Recommendations from the AoIR ethics working committee," The International Association of Internet Researchers, Tech. Rep., 2002.

[30] A. Markham and E. Buchanan, "Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0)," The International Association of Internet Researchers, Tech. Rep., 2012.

[31] A. S. Franzke, A. Bechmann, C. M. Ess, and M. Zimmer, "Internet Research: Ethical Guidelines 3.0," The International Association of Internet Researchers, Tech. Rep., 2020.

[32] J. Metcalf and K. Crawford, "Where are human subjects in Big Data research? The emerging ethics divide," *Big Data & Society*, vol. 3, no. 1, 2016.

[33] H. Nissenbaum, *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2009.

[34] E. M. Bender and B. Friedman, "Data statements for natural language processing: Toward mitigating system bias and enabling better science," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 587–604, 2018.

[35] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé, III, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.

[36] B. Hutchinson, A. Smart, A. Hanna, R. Denton, C. Greer, O. Kjartansson, P. Barnes, and M. Mitchell, "Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. ACM, 2021, pp. 560–575.

[37] J. Vitak, K. Shilton, and Z. Ashktorab, "Beyond the Belmont Principles: Ethical Challenges, Practices, and Beliefs in the Online Data Research Community," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, 2016, pp. 941–953.

[38] J. Vitak, N. Proferes, K. Shilton, and Z. Ashktorab, "Ethics Regulation in Social Computing Research: Examining the Role of Institutional Review Boards," *Journal of Empirical Research on Human Research Ethics*, vol. 12, no. 5, pp. 372–382, 2017.

[39] A. Bruckman, "Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet," *Ethics and Information Technology*, vol. 4, no. 3, pp. 217–231, 2002.

[40] C. Fiesler and N. Proferes, ""Participant" Perceptions of Twitter Research Ethics," *Social Media + Society*, vol. 4, no. 1, 2018.

[41] B. Dym and C. Fiesler, "Ethical and Privacy Considerations for Research Using Online Fandom Data," *Transformative Works and Cultures*, vol. 33, 2020.

[42] J. Mikal, S. Hurst, and M. Conway, "Ethical issues in using Twitter for population-level depression monitoring: a qualitative study," *BMC Medical Ethics*, vol. 17, no. 1, p. 22, 2016.

[43] D. Dittrich and E. Kenneally, "The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research," U.S. Department of Homeland Security, Tech. Rep., 2012.

[44] Z. Durumeric, E. Wustrow, and J. A. Halderman, "ZMap: Fast Internet-wide Scanning and Its Security Applications," in *Proceedings of the 22nd USENIX Security Symposium (SEC '13)*. USENIX Association, 2013, pp. 605–620.

[45] T. Fiebig, "Crisis, Ethics, Reliability & a measurement.network: Reflections on Active Network Measurements in Academia," in *Proceedings of the 2023 Applied Networking Research Workshop (ANRW '23)*. ACM, 2023, pp. 44–50.

[46] S. F. Amin, S. Athar, A. Feldmann, H. Dao, and M. Kaur, "Navigating the Ethics of Internet Measurement: Researchers' Perspectives from a Case Study in the EU," arXiv:2511.10408, 2025.

[47] M. Dahlmanns, C. Sander, R. Decker, K. Wehrle, and J. Pennekamp, "Secrets Revealed in Container Images: An Internet-wide Study on Occurrence and Impact," in *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security (ASIACCS '23)*. ACM, 2023, pp. 797–811.

[48] N. Demir, Y. Vekaria, G. Smaragdakis, and Z. Durumeric, "Keys on Doormats: Exposed API Credentials on the Web," arXiv:2603.12498, 2026.

[49] Censys, "Censys," https://www.censys.io, 2017.

[50] Shodan, "Shodan," https://www.shodan.io, 2013.

[51] A. Ballantyne and G. O. Schaefer, "Consent and the ethical duty to participate in health data research," *Journal of Medical Ethics*, vol. 44, no. 6, pp. 392–396, 2018.

[52] Department of Health and Human Services, "U.S. HHS Common Rule, 45 CFR §46.116," https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46/subpart-A/section-46.116, 2018 (accessed September 10, 2026).

[53] GitHub, Inc., "GitHub," https://github.com/, 2008.

[54] Internet Archive, "Wayback Machine," https://web.archive.org/, 2001.

[55] European Organization For Nuclear Research and OpenAIRE, "Zenodo," https://www.zenodo.org/, 2013.

[56] S. K. Arora, Y. Li, J. Youtie, and P. Shapira, "Using the Wayback Machine to Mine Websites in the Social Sciences: A Methodological Resource," *Journal of the Association for Information Science and Technology*, vol. 67, no. 8, pp. 1904–1915, 2016.

[57] V. Cosentino, J. Luis, and J. Cabot, "Findings from GitHub: Methods, Datasets and Limitations," in *Proceedings of the IEEE/ACM 13th International Conference on Mining Software Repositories (MSR '16)*. ACM, 2016, pp. 137–141.

[58] I. Peters, P. Kraker, E. Lex, C. Gumpenberger, and J. I. Gorraiz, "Zenodo in the Spotlight of Traditional and New Metrics," *Frontiers in Research Metrics and Analytics*, vol. 2, p. 13, 2017.

[59] M.-A. Sicilia, E. García-Barriocanal, and S. Sánchez-Alonso, "Community Curation in Open Dataset Repositories: Insights from Zenodo," *Procedia Computer Science*, vol. 106, pp. 54–60, 2017, proceedings of the 13th International Conference on Current Research Information Systems (CRIS '16).

[60] S. van de Sandt, L. H. Nielsen, A. Ioannidis, A. Muench, E. Henneken, A. Accomazzi, C. Bigarella, J. B. G. Lopez, and S. Dallmeier-Tiessen, "Practice meets Principle: Tracking Software and Data Citations to Zenodo DOIs," arXiv:1911.00295, 2019.

[61] K. Garg, S. Alam, D. Ayala, M. C. Weigle, and M. L. Nelson, "A Longitudinal Dataset of URLs Sampled From the Wayback Machine," in *Proceedings of the 2025 ACM/IEEE Joint Conference on Digital Libraries (JCDL '25)*. IEEE, 2025, pp. 277–280.

[62] M. Schreiber and P. Tippe, "Security Vulnerabilities in AI-Generated Code: A Large-Scale Analysis of Public GitHub Repositories," in *Proceedings of the 27th International Conference on Information and Communications Security (ICICS '25)*, vol. 16219. Springer, 2025, pp. 153–172.

[63] D. Zarrillo, M. Kelly, E. Yan, and C. Ni, "Web archives for data collection: An ethics case study," *Accountability in Research*, vol. 32, no. 8, pp. 1408–1425, 2025.

[64] G. Cormode, S. Muthukrishnan, and J. Yan, "Scienceography: The Study of How Science Is Written," in *Proceedings of the 6th International Conference on Fun with Algorithms (FUN '12)*, vol. 7288. Springer, 2012, pp. 379–391.

[65] G. Cormode, S. Muthukrishnan, and J. Yan, "Studying the source code of scientific research," *ACM SIGKDD Explorations Newsletter*, vol. 14, no. 2, pp. 59–62, 2013.

[66] W. Dawei, R. Alfred, J. H. Obit, and C. K. On, "A Literature Review on Text Classification and Sentiment Analysis Approaches," in *Proceedings of the 7th International Conference on Computational Science and Technology (ICCST '20)*, vol. 724. Springer, 2020, pp. 305–323.

[67] D. Ginev and B. R. Miller, "Scientific Statement Classification over arXiv.org," in *Proceedings of the 12th Language Resources and Evaluation Conference (LREC '20)*. ELRA, 2020, pp. 1219–1226.

[68] M. Li, Y. Xu, L. Cui, S. Huang, F. Wei, Z. Li, and M. Zhou, "DocBank: A Benchmark Dataset for Document Layout Analysis," in *Proceedings of the 28th International Conference on Computational Linguistics (COLING '20)*. ICCL, 2020, pp. 949–960.

[69] P. Scharpf, M. Schubotz, A. Youssef, F. Hamborg, N. Meuschke, and B. Gipp, "Classification and Clustering of arXiv Documents, Sections, and Abstracts, Comparing Encodings of Natural and Mathematical Language," in *Proceedings of the ACM/IEEE 20th Joint Conference on Digital Libraries (JCDL '20)*. ACM, 2020, pp. 137–146.

[70] N. Meuschke, A. Jagdale, T. Spinde, J. Mitrović, and B. Gipp, "A Benchmark of PDF Information Extraction Tools Using a Multi-task and Multi-domain Evaluation Framework for Academic Documents," in *Proceedings of the 18th International Conference on Information for a Better World: Normality, Virtuality, Physicality, Inclusivity (iConference '23)*, vol. 13972. Springer, 2023, pp. 383–405.

[71] D. Sorokina, J. Gehrke, S. Warner, and P. Ginsparg, "Plagiarism Detection in arXiv," in *Proceedings of the 6th International Conference on Data Mining (ICDM '06)*. IEEE, 2006, pp. 1070–1075.

[72] D. T. Citron and P. Ginsparg, "Patterns of text reuse in a scientific corpus," *Proceedings of the National Academy of Sciences*, vol. 112, no. 1, pp. 25–30, 2015.

[73] B. Safnuk and G. Hu, "Reconstructing LaTeX Source Files from Generated PDFs — a Neural Network Approach," in *Proceedings of the 2018 IEEE 16th International Conference on Industrial Informatics (INDIN '18)*. IEEE, 2018, pp. 890–895.

[74] A. Akram, "Quantitative Analysis of AI-Generated Texts in Academic Research: A Study of AI Presence in Arxiv Submissions using AI Detection Tool," arXiv:2403.13812, 2024.

[75] C. B. Clement, M. Bierbaum, K. P. O'Keeffe, and A. A. Alemi, "On the Use of ArXiv as a Dataset," arXiv:1905.00075, 2019, representation Learning on Graphs and Manifolds (RLGM) Workshop 2019.

[76] T. Saier and M. Färber, "unarXive: a large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata," *Scientometrics*, vol. 125, no. 3, pp. 3085–3108, 2020.

[77] M. L. Bianchi, "Text Mining arXiv: A Look Through Quantitative Finance Papers," *Mathematics*, vol. 13, no. 9, 2025.

[78] B. Hepler, "Modular versus Hierarchical: A Structural Signature of Topic Popularity in Mathematical Research," arXiv:2506.22946, 2025.

[79] J. Pei, L. Yang, and L. Wu, "Hidden Division of Labor in Scientific Teams Revealed Through 1.6 Million LaTeX Files," arXiv:2502.07263, 2025.

[80] L. Yang, J. Pei, and L. Wu, "Writing Patterns Reveal a Hidden Division of Labor in Scientific Teams," arXiv:2504.14093, 2025.

[81] D. Rao, J. Young, T. Dietterich, and C. Callison-Burch, "WithdrarXiv: A Large-Scale Dataset for Retraction Study," arXiv:2412.03775, 2024.

[82] Open Graph Benchmark, "Dataset ogbn-arxiv," https://ogb.stanford.edu/docs/nodeprop/#ogbn-arxiv, 2020.

[83] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec, "Open Graph Benchmark: Datasets for Machine Learning on Graphs," in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*, vol. 33. NeurIPS, 2020, pp. 22 118–22 133.

[84] arXiv, "Terms of Use for arXiv APIs," https://info.arxiv.org/help/api/tou.html#rate-limits, 2019.

[85] P. G. Leon, B. Ur, Y. Wang, M. Sleeper, R. Balebako, R. Shay, L. Bauer, M. Christodorescu, and L. F. Cranor, "What matters to users?: factors that affect users' willingness to share information with online advertisers," in *Proceedings of the 9th Symposium on Usable Privacy and Security (SOUPS '13)*. ACM, 2013.

[86] S. El Yadmani, O. Gadyatskaya, and Y. Zhauniarovich, "The file that contained the keys has been removed: An empirical analysis of secret leaks in cloud buckets and responsible disclosure outcomes," in *Proceedings of the 2025 IEEE Symposium on Security and Privacy (SP '25)*. IEEE, 2025, pp. 3180–3198.

[87] ACM, "ACM Code of Ethics and Professional Conduct," ACM, Tech. Rep., 2018.

[88] R. Langner and S. B. Eickhoff, "Sustaining attention to simple tasks: A meta-analytic review of the neural mechanisms of vigilant attention," *Psychological Bulletin*, vol. 139, no. 4, pp. 870–900, 2013.

[89] L. Lannelongue, J. Grealey, A. Bateman, and M. Inouye, "Ten simple rules to make your computing more environmentally sustainable," *PLOS Computational Biology*, vol. 17, no. 9, 2021.

[90] F. Li, Z. Durumeric, J. Czyz, M. Karami, M. Bailey, D. McCoy, S. Savage, and V. Paxson, "You've Got Vulnerability: Exploring Effective Vulnerability Notifications," in *Proceedings of the 25th USENIX Security Symposium (SEC '16)*. USENIX Association, 2016, pp. 1033–1050.

[91] P. Mayer, Y. Zou, F. Schaub, and A. J. Aviv, "'Now I'm a bit angry:' Individuals' Awareness, Perception, and Responses to Data Breaches that Affected Them," in *Proceedings of the 30th USENIX Security Symposium (SEC '21)*. USENIX Association, 2021, pp. 393–410.

[92] P. Mayer, Y. Zou, B. M. Lowens, H. A. Dyer, K. Le, F. Schaub, and A. J. Aviv, "Awareness, Intention, (In)Action: Individuals' Reactions to Data Breaches," *ACM Transactions on Computer-Human Interaction*, vol. 30, no. 5, 2023.

[93] D. J. Solove and D. K. Citron, "Risk and Anxiety: A Theory of Data Breach Harms," *Texas Law Review*, vol. 96, pp. 737–786, 2017.

[94] Princeton University, "Princeton-Radboud Study on Privacy Law Implementation," https://privacystudy.cs.princeton.edu/, 2021 (accessed December 20, 2025).

[95] M. Maass, A. Stöver, H. Pridöhl, S. Bretthauer, D. Herrmann, M. Hollick, and I. Spiecker, "Effective Notification Campaigns on the Web: A Matter of Trust, Framing, and Support," in *Proceedings of the 30th USENIX Security Symposium (SEC '21)*. USENIX Association, 2021, pp. 2489–2506.

[96] A. Hennig, M. Veit, L. Schmidt-Enke, F. Neusser, D. Herrmann, and P. Mayer, ""I believe it's incredibly difficult to fight against this flood of spam": Towards enhancing strategies for creating effective vulnerability notifications," *Computers & Security*, vol. 160, p. 104682, 2025.

[97] J. Pastor-Galindo, P. Nespoli, F. G. Mármol, and G. M. Pérez, "The Not Yet Exploited Goldmine of OSINT: Opportunities, Open Challenges and Future Trends," *IEEE Access*, vol. 8, pp. 10 282–10 304, 2020.

[98] A. Pranav, S. Storks, and A. Lauscher, "The Double Bind: Revisiting Preprinting and Peer Review Two Years After the Removal of the ACL Anonymity Period," in *Findings of the Association for Computational Linguistics (ACL '26)*. ACL, 2026, pp. 4566–4584.

[99] L. P. Nijhawan, M. D. Janodia, and P. B. Musmade, "Informed consent: Issues and challenges," *Journal of Advanced Pharmaceutical Technology & Research*, vol. 4, no. 3, pp. 134–140, 2013.

[100] S. Golder, S. Ahmed, G. Norman, and A. Booth, "Attitudes Toward the Ethics of Research Using Social Media: A Systematic Review," *Journal of Medical Internet Research*, vol. 19, no. 6, 2017.

[101] N. E. Gold and J. Krinke, "Ethics in the mining of software repositories," *Empirical Software Engineering*, vol. 27, no. 1, 2022.

[102] J. A. Hodges, "How to Collect a Corpus of Websites With a Web Crawler," in *Sage Research Methods: Doing Research Online*. SAGE, 2022.