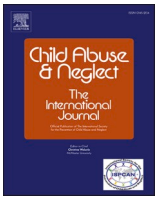
A content analysis of metrics on online child sexual exploitation and abuse used by online content-sharing services

Child Abuse & Neglect 157 (2024) 107046

Contents lists available at ScienceDirect

Child Abuse & Neglect

journal homepage: www.elsevier.com/locate/chiabuneg



Research article

A content analysis of metrics on online child sexual exploitation and abuse used by online content-sharing services

Mengyao Lu [a,*], Maria Lamond [a,b], Deborah Fry [a]

[a] Childlight – Global Child Safety Institute, Moray House School of Education and Sport, University of Edinburgh, St John's Land, Holyrood Road, Edinburgh EH8 8AQ, UK
[b] Division of Psychology, Abertay University, Bell Street, Dundee DD1 1HG, UK

ARTICLE INFO

Keywords:
Online child sexual exploitation and abuse
Online content-sharing services
Child sexual abuse material
Metrics

ABSTRACT

Background: A critical step in improving the response to and monitoring of online child sexual exploitation and abuse (OCSEA) is the need to standardize the data that are collected, stored, and analyzed that effectively measure change in the frequency, nature and risk of OCSEA over time.
Objective: The objective of the content analysis was to investigate the metrics used by online content-sharing platforms in their efforts to combat OCSEA.
Methods: A content analysis was undertaken on 19 online content-sharing services' transparency reports on their metrics related to OCSEA.
Results: From the 19 transparency reports, 132 data points in relation to OCSEA were identified with 22 distinct metrics on OCSEA. Findings revealed a disparity of appropriate metrics and reporting mechanisms employed, particularly, there is a lack of standardized approaches to metrics reporting and an absence of time related measures. Furthermore, very few online content-sharing services disclosed metadata on the data reported and its capture methodology.
Conclusion: This study highlights the critical need for standardized metrics reporting to enable comparability across services. Without such an evidence base, there are no objective measures to assess the progress and effectiveness in addressing OCSEA.

1. Introduction

1.1. Background

Online Child Sexual Exploitation and Abuse (OCSEA) has been seen both as a public health problem and criminal justice problem (Ali et al., 2021). OCSEA is defined as crimes that include production, dissemination and possession of child sexual abuse material (CSAM), online grooming of children for sexual purposes, sexting, sexual extortion of children, revenge pornography, commercial sexual exploitation of children, exploitation of children through online prostitution, and live streaming of sexual abuse (Quayle, 2016). Furthermore, different aspects and definitions that may fall under OCSEA including sexual harassment and online solicitation, exposure to sexual content, sexual bullying, pressure to share sexual images of themselves, wider sharing of sexual images, grooming, sexual abuse and exploitation (Livingstone et al., 2017). Child sexual exploitation is distinguishable from child sexual abuse with the underlying notion involving some form of exchange and often encompasses a broader range of activities. This can include activities

* Corresponding author.
E-mail address: Mengyao.Lu@ed.ac.uk (M. Lu).

https://doi.org/10.1016/j.chiabu.2024.107046
Received 6 July 2024; Received in revised form 2 September 2024; Accepted 11 September 2024
Available online 23 September 2024

such as enticing, manipulating, or threatening a child into performing sexual acts, grooming potential victims, and the production, distribution, buying, selling, possession, or transmission of CSAM (ECPAT International, 2016; Wager et al., 2018).

The prevalence of OCSEA is deeply concerning. A comprehensive meta-analysis of 88 publications, including findings from 125 studies across 57 countries and published in English, Chinese, and Spanish between January 1st, 2010, and September 30th, 2023, revealed that approximately 1 in 8 children globally (12.6 %) experienced non-consensual taking, sharing, or exposure to sexual images and videos within the past year. Similarly, nearly the same proportion (12.5 %) reported experiencing online solicitation during the same period. Moreover, an overall prevalence rate of OCSEA at 8.1 % was observed across 15 included studies that measured exposure to at least one type of OCSEA when multiple types of OCSEA were assessed within the same sample (Krzeczkowska et al., 2024). Regarding the extent and characteristics of CSAM, over 36 million online cases of CSAM were identified and reported from 2018 to 2022 by five organizations that publish public reports on the accessibility and dissemination of CSAM. The presence of CSAM perpetuates ongoing violations of the rights of each victim-survivor, with its distribution, production, and commercialization further exacerbating harm to these individuals (Stevenson et al., 2024).

1.2. Legislation and regulations on mandatory reporting OCSEA

Legislation addressing OCSEA is growing globally. For example, the recent Online Safety Bill in the UK requires online content-sharing services to report OCSEA to the National Crime Agency within specified timeframes based on when OCSEA content is detected (Online Safety Bill, 2021). Services are required to publish annual transparency reports to inform users about the measures providers are implementing to enhance safety and empower the office of communications (OFCOM) to hold them accountable (Online Safety Act, 2023). In the European Union, the proposed Digital Services Act (DSA) aims to create a safer digital space by requiring online platforms to remove illegal content, including OCSEA, and to cooperate with law enforcement authorities. The DSA also mandates transparency reports and sets out clear obligations for online platforms to protect children from illegal and harmful content (European Commission, 2022).

The eSafety commissioner in Australia provides guidance to online content-sharing services on expectations related to OCSEA and reporting. The guidance provided pertains to the Basic Online Safety Expectations established under the Online Safety Act 2021 in Australia. These Expectations outline the Australian Government's requirements for online service providers. Online content-sharing services are expected to prioritize responding to the most harmful risks on their service, particularly where these involve unlawful material or activity including OCSEA (eSafety Commissioner, 2023).

The African Union (AU) (African Union, (n.d.)) detailed a five-year strategic plan to tackle OCSEA following a poor response from member states in ratifying the AU cyber convention 2014. Currently there is no legislation that mandates online content-sharing services to remove, block, and report instances of OCSEA (African Union, n.d.). The plan details that industry should collaboratively develop and share innovative tools to identify and remove OCSEA, address OCSEA in polices and process to identify, measure, prevent, and mitigate against OCSEA. Several indicators have been established including legal obligations with internet industries to report, remove and block CSAM.

In North America, the US Kids Online Safety Act (2023) requires large platforms to provide transparency reports including foreseeable risks to children. Platforms are required to implement reasonable measures in the design and operation of their products or services used by minors to prevent and address certain potential harms, such as sexual exploitation. Canada's most recent proposed amendments to their Online Harms Bill mandates that social media platforms are required to make non-consensually distributed intimate images and CSAM inaccessible within 24 h. Additionally, compliance with the Act mandates services to develop a plan that includes statistics related to moderating harmful content and managing electronic data (Parliament of Canada, 2023).

Complementing these legislative efforts, the National Center for Missing & Exploited Children (NCMEC; National Center for Missing & Exploited Children, n.d) is a private, non-profit organization established in 1984 by the United States Congress to serve as the nation's resource on issues related to the abduction, sexual exploitation, and victimization of children. One of its key roles is managing the CyberTipline, a central mechanism for electronic service providers to report instances of online child sexual exploitation (Grossman et al., 2024). The organization has played a pivotal role in increasing public awareness, advocating for legislative changes, and enhancing technological tools to combat OCSEA.

1.2.1. Current evidence gap and review question

A crucial step in enhancing the response to and monitoring of OCSEA involves standardizing the data collection, storage, and analysis processes. This also involves creating uniform metrics that can effectively measure changes in OCSEA over time, enabling us to track trends, identify emerging threats, and assess the impact of interventions of OCSEA prevention. As such, this study was conducted with the objective of bridging the existing evidence gap, thorough examination of the metrics currently employed by online content-sharing services in relation to OCSEA. By delving into these metrics, the study sought to shed light on their effectiveness, limitations, and potential areas for improvement. Ultimately, the study aimed to provide valuable insights and recommendations that would contribute to a better understanding and more robust strategies for combating OCSEA. The research question is: what are the current practices in the reporting metrics of OCSEA by online content-sharing services according to their most recent published transparency reports?

2. Method

2.1. Content analysis and study selection

Content analysis is a "research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use" (Krippendorff, 2018, p.18). The goal is to achieve a concise yet comprehensive description of the phenomenon. The analysis results in concepts or categories that describe the phenomenon. Typically, these concepts or categories are used to develop a model, conceptual system, conceptual map, or categories (Elo & Kyngäs, 2008).

Given the rapid advancement of technology in the field of CSEA metrics, relying solely on academic literature may not capture the most relevant or current information, as these metrics are typically reported in services' transparency reports. A transparency report is "a public communication document released by an internet company that discloses key metrics and information about digital governance and enforcement measures on its platform(s)" (Trust and Safety Professional Association, 2024). Our approach was to focus specifically on mapping the key metrics outlined in the online content-sharing services included in a recent report by the Organization for Economic Co-operation Development (OECD) on transparency reporting on OCSEA (OECD, 2023). This benchmarking study provided a robust evidence base on how leading content-sharing service providers publicly report on their efforts to combat OCSEA. Online content-sharing services are defined as "any online service that enables the transfer, transmission and dissemination of content, in whatever form, whether one-to-one, one-to-few or one-to-many and irrespective of whether the content is public-facing, semi-private or private. All of the services profiled in this report are Online Content-Sharing Services" (OECD, 2023, p.156). The report analyzed how the top 50 global content-sharing services address CSAM and offers a detailed analysis of the policies and procedures for addressing CSAM on their platforms and services (OECD, 2023). To identify the most popular social media platforms, video-sharing sites, and communication services, the metric of monthly active users (MAU) was selected as the most appropriate measure. MAU is commonly used in the industry to gauge online engagement and platform reach, making it a suitable basis for ranking the most frequently used services. However, MAU is less relevant for other types of services. Therefore, market share was used to identify the leading cloud-based file-sharing services. Additionally, two services - the WordPress content management system and the Wikipedia reference site - were included in the original top-50 ranking, even though their relative popularity compared to other services could not be directly determined (OECD, 2023).

By leveraging the OECD study and transparency reports from various online content-sharing services, we aimed to provide a comprehensive overview of the current landscape of OCSEA metrics. Two authors conducted comprehensive searches on the websites of all 50 online content-sharing services identified by the OECD report, as well as Verizon, a member of Tech Coalition,[1] and downloaded the latest transparency reports where available.

2.2. Data extraction and analysis

Following the retrieval of transparency reports, two authors reviewed each report to extract information including: 1) the name of the online content-sharing service, 2) the specific metrics pertaining to OCSEA, 3) publication period of the transparency report. To ensure accuracy and reliability, each author was assigned half of the total number of reports for initial extraction, while cross-checking the other half extracted by the other reviewer. Each metric was recorded as a single data point; subsequently, the authors collaboratively grouped similar metrics into broader categories and synthesized them based on thematic areas. This process facilitated a comprehensive analysis of the identified metrics and allowed for the identification of overarching trends and patterns in the strategies employed by online content-sharing services to address OCSEA.

The study was registered with Open Science Foundation: https://osf.io/fsp5n

3. Results

3.1. Characteristics of included transparency reports

The majority of online content-sharing services do not publish transparency metrics on OCSEA data. Even among the companies that provide transparency metrics, differences in the metrics used often make it challenging to compare data across all services. Among the 50 online content-sharing services examined, transparency reports were successfully located and downloaded from 19 services. Table 1 presents the name, country, and publication period of these services and Table A1 details the services that do and do not publish transparency reports.

As outlined in the table, the publication periods of transparency reports vary among services. Specifically, four services publish their transparency reports annually, nine services release transparency reports biannually, six services publish transparency reports quarterly.

3.1.1. Definitions of OCSEA by online content-sharing services

Most services do not provide details about what aspects of OCSEA are being reported. Typically, these services do not go beyond the

[1] The Tech Coalition (https://www.technologycoalition.org/) is an alliance of technology companies that collaborate to combat OCSEA.

Table 1
Characteristics of online content-sharing services.

Name of Service	Country	Publication period
Amazon		
Reddit		
Microsoft	United States	Annually
Microsoft (Bing)		
Dropbox		
LINE	Japan	
LinkedIn	United States	
Snap		
Twitch		
X		Half year
Verizon		
Yubo	France	
Zoom		
Discord		
Google	United States	
Meta		
Pinterest		Quarterly
Tiktok (ByteDance Co.)	China	
YouTube	United States	

traditional focus on the CSAM, namely depictions of nudity, abuse, and exploitation such as images and videos. As well as lack of definitions of OCSEA, online content-sharing services do not define "child" or "age" in their transparency reports. Furthermore, there is wide variation and categorisation between services definitions related to their OSCEA metrics. See Table 2 for a summary of definitions that online content-sharing services use to define their OCSEA related metrics.

TikTok provides additional information about the definitions of CSEA underpinning its metrics with reported violations indexed to the respective policies and sub-policies. The metrics reported by TikTok under the minor safety policy can be traced to the different sub-policies. The most relevant for CSEA are sexual exploitation of minors, grooming behavior and nudity and sexual activity involving minors. TikTok's report states:

The "sexual activity involving minors" sub-policy prohibits a broad range of content, including "minors in minimal clothing", and "sexually explicit dancing". These two categories represent the majority of content removed under that sub-policy. We report CSAM and supporting evidence to the NCMEC and to any additional relevant legal authorities.

Such detailed descriptions allow a much more fine-grained understanding of CSEA dynamics in the platform.

3.2. Overview of OCSEA metrics

The initial mapping exercise of OCSEA metrics resulted in 132 data points, indicating the heterogeneity across different online content-sharing services regarding the reporting of OCSEA (data points are available upon request). After grouping them into broader themes, a total number of 22 metrics on OCSEA were identified. Based on the thematic metrics, we further synthesized them based on the target objectives of CSEA; as a result, six categories were identified, including 1) content-based metrics, 2) account-based metrics, 3) detection-based CSEA, 4) report, 5) time related metrics, 6) escalations or legal requests. Table 3 presents the categories of the metrics identified from the transparency reports. Additionally, a table of which services report each of these categories can be found in

Table 2
Summary of services definitions of OCSEA metrics in their transparency reports.

Definition of OCSEA	Services
Child sexual Abuse Material (CSAM)	Amazon, Discord, Google, and YouTube, Verizon, Yubo, Pinterest
CSAM including content in images, videos, or text	Amazon, Pinterest
CSAM as "visual depictions, including but not limited to photos, videos, and computer-generated imagery" of sexually explicit conduct involving minors.	Google and YouTube
Child sexual exploitation and abuse imagery (CSEAI).	Snap, Microsoft, Bing
Content depicting sexual exploitation of a minor or grooming behavior.	Verizon
Child exploitation (content depicting or promoting sexual activity or abuse of minors, solicitation of such materials, or the solicitation of minor)	LinkedIn
Content depicting sexual exploitation of a minor or grooming behavior	Twitch
Child nudity, physical abuse and child sexual exploitation	Meta
Child sexual exploitation, which includes media, text, illustrated, or computer-generated images, and URLs	X
Grooming	Yubo
Minor sexualization	Reddit
Abuse of Children	Line
Sexual exploitation of minors, grooming behavior and nudity and sexual activity involving minors	TikTok

Table 3
Overview of metrics on CSEA reported by online content-sharing services.

Category	Thematic metrics	Number of services
Detection-related metrics ($n = 9$)	Proactive rate	5
	Content detected (manually, with Photo DNA/hash matching or with hybrid tools	4
Content-related ($n = 27$)	Appealed content	1
	Restored content	1
	Content removed/actions	14
	Content detected	6
	Impressions	1
	URLs deindexed	2
	Prevalence of child endangerment	1
Report ($n = 19$)	Reports to NCMEC	13
	Reports per month	1
	Disclosed data on child abuse	1
	User reports	4
	Removal before any views	1
	Removal within 24 h	1
Time-related metrics (n = 3)	Reach of content deactivated for child sexual exploitation	1
Escalations or legal requests (n = 13)	Grooming or endangerment escalations	1
	Legal requests, including info disclosed	12
Account-related metrics (n = 18)	Account reinstated	2
	Account appealed	4
	Accounts Actioned	12

Table A3.

3.2.1. *Content-related metrics (n = 27)*

The analysis identified a total of 27 data points regarding content-related metrics. The predominant focus lies on actions taken regarding content actions or removal ($n = 14$), followed by metrics related to content detection ($n = 6$). Other commonly seen metrics include appealed content, restored content, percentage of content removed manually versus. With hybrid tools, and URLs deindexed.

Content actioned refers to the total number of pieces of content that have been acted upon by the platform. This includes not only the content that has been removed but also content that may have been flagged, labelled, reduced in visibility, or otherwise modified to comply with the platform's policies. It may also include issuing warnings, suspending accounts, and permanently banning users. Meta (both Facebook and Instagram), Snap, Microsoft (consumer services and Bing) report this metric. Another important metric is content removed. Reporting on removed content sheds light on the amount of content that is deemed unacceptable by the platform. This metric is reported by Microsoft, LinkedIn, Amazon, Tik Tok, Twitch, YouTube, Reddit, and LINE. This metric involves removing chat messages (Twitch), removal of pieces of content and private message (Reddit), videos and comments (YouTube) and removing and disabling content (Amazon). Similarly, Google also reports on URLs deindexed, which appears to be an equivalent to removed user generated content for search engine indexing of the web.

3.2.1.1. Appealed content. Content appealed refers to the number of times users have challenged removed content. YouTube provides metrics related to videos that have been appealed and videos that have been reinstated, however this metric is not disaggregated by content type. Instead, they represent all removals for any content violations. Meta provides metrics regarding actioned content appeals, for both Facebook and Instagram. These metrics may provide some insights regarding fairness and accuracy of content moderation practices. However, Meta warns of limitations:

This metric should not be interpreted as an indicator of the accuracy of our decisions on content, as people may choose to appeal for many different reasons. We report the total number of pieces of content that had an appeal submitted in each quarter – for example, 1 January to 31 March. Bear in mind that this means that the numbers can't be compared directly to content actioned or to content restored for the same quarter. Some restored content may have been appealed in the previous quarter, and some appealed content may be restored in the next quarter.

3.2.1.2. Other content related metrics. Several additional metrics have been reported by various online content-sharing services, providing further insights into their approaches to reporting OCSEA. For example, Meta and TikTok have metrics related to "restored content", representing the number of pieces of content reinstated after initial removal or warning. Meta has also incorporated a metric concerning the prevalence of child endangerment, although no data has been collected thus far. Meta states that "we are working on estimating prevalence for child endangerment violations. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data" (Meta Platforms, Inc., 2023).

Pinterest used a multifaceted approach involving automated tools, manual review, and hybrid methods, combining elements of both, to deactivate policy-violating content including OCSEA. Pinterest also reports the percentage of content deactivated manually versus hybrid tools. Meanwhile, X introduces the concept of "impressions," defined as any time at least half of the area of a given Tweet

is visible to a user for at least half a second (including while scrolling) Additionally, X provides metrics on content deleted by country and removed globally for OCSEA. Lastly, Google provides data on URLs deindexed for CSAM from its search results, reflecting its efforts to remove harmful content reported on third-party web pages, although it lacks control over the content itself. These diverse metrics contribute to a more comprehensive understanding of content moderation strategies employed by online content-sharing services and their efforts to combat OCSEA.

3.2.2. Account-related metrics (n = 18)

The transparency reports offer insights into metrics related to user accounts on their platforms. One of the most commonly reported metrics in this category is "accounts actioned". This includes various actions taken by the platform in response to policy violations, such as disabling accounts for CSAM violations (e.g., Google), suspending or reporting accounts (e.g., X), and imposing sanctions for minor sexualization (e.g., Reddit). Other online content-sharing services such as Amazon, Snap, Discord, and Yubo also report on disabled, banned, or suspended accounts, with Yubo providing specific CSEA-related metrics, such as the number of permanent bans related to grooming and CSAM. Similarly, YouTube report the total accounts disabled for CSAM by the top 10 countries offending. While some services report on actions taken on accounts, only a few provide information on appeals and reinstated accounts. For example, Microsoft report on accounts reinstated and X provide a metric related to suspensions overturned. Discord provides metrics on appeals, including the percentage of accounts who submitted an appeal and percentage of accounts that appeals granted. YouTube and Snap have metrics related to reinstated accounts; however, the metric is not disaggregated by content type.

3.2.3. Report (n = 19)

Four metrics were identified under the Report category, including reports to NCMEC, user reports, reports per month, disclosed data on child abuse.

3.2.3.1. Reports to NCMEC (n = 13). NCMEC reporting is mandatory for online content-sharing services operating in the U.S. This is the case of Google (for both Google and YouTube), Snap, Yubo, Microsoft, Verizon and Amazon (for both their consumer services and Twitch). Some of the services provide the number of NCMEC reports in their transparency reports. However, simply reporting on the total number of reports made to NCMEC does not provide sufficient insight into the nature and extent of CSEA in these online environments. In addition, the number of reports by the services are already made public by NCMEC's annual reports. In addition to the absolute number of NCMEC reports, some services provide further information regarding this reporting. Twitch, for instance, publishes the metric relative to the hours watched. The NCMEC Cyber Tipline signals how many reports were made for every 1000 h of live-streamed content watched. This metric is particularly useful because it measures effectiveness in real-time. The ratio of reports to viewership seeks to reveal how the magnitude of CSEA relates to the total amount of streamed content in the platform. In addition, Google reports not only the number of reports made but also the total pieces of content reported to NCMEC, as well as the CSAM hashes contributed by the service to the NCMEC database. This type of contribution strengthens the ability of quickly locating remerging material using automated hash matching techniques. Google's strategy provides a more complete picture of the scope of the efforts to tackle OCSEA and efforts to fight OCSEA.

3.2.3.2. User reports (n = 4). Overall, there appears to be a disparity on user report on OCSEA metrics with only a few online-content sharing services reporting on this. Challenges to self-reporting OSCEA have been identified on services with lack of intuitive methods of reporting, or specificity in reporting OCSEA.

3.2.4. Time-related metrics (n = 3)

Among all the metrics, only three were time-related, coming from TikTok and Pinterest, including removal before any views, removal within 24 h, and reach of content deactivated for child sexual exploitation (i.e., seen by 0 users, seen by 1 to 9, seen by 10 to 100 and seen by more than 100). Previous research has shown how quickly an abuser can initiate contact with a child online, sometimes in just a few minutes (Álvarez-Guerrero et al., 2024). This highlights the urgent need for online content-sharing services to include more time-related metrics in their reports on OCSEA. By capturing this temporal data, online content-sharing services can better understand and respond to the rapid dynamics of these harmful interactions, ultimately enhancing their efforts to combat OCSEA effectively.

3.2.5. Detection-related metrics (n = 7)

Detection-related metrics are closely intertwined with other metrics discussed earlier and should not be viewed in isolation, as detection inherently involves action and various other aspects of CSAM. This section particularly focuses on aspects of detection-related metrics that were not addressed previously.

Some services report on proactively detected CSEA content. Proactive detection involves using different detection strategies, from human moderation efforts to advanced algorithms and machine learning techniques that identify potentially problematic content. Meta reports the percentage of all content or accounts acted upon that they identified and flagged before receiving user reports. This metric serves as an indicator of the service's effectiveness in detecting violations. During October to December, 2023, Facebook detected and took action on 96.70 % of child sexual exploitation content before user reports, with the remaining 3.30 % detected and acted upon based on user reports (Meta Platforms, Inc., 2023). Microsoft Bing further reported the percentage of CSEAI content detected through various approaches, such as PhotoDNA, proactive measures, and manual searching. Similarly, Yubo (2023) reports

that 85 % of the reports processed by the company during the second half of 2023 resulted from proactive detection strategies (Yubo, 2023). TikTok also provides information on the proactive removal rate of harmful content, with 96.70 % of harmful content being removed via proactive detection (TikTok, 2023). However, Zoom reports that PhotoDNA technology has been inefficient with 96.85 % of all reports found were false positives after being reviewed by the Trust and Safety team and are therefore currently seeking better CSAM hash-matching and other meta driven technologies (Zoom, 2023).

In addition, Meta (for both Facebook and Instagram), Snap, and Microsoft (for both its consumer services and Bing). Amazon also reports on the number of images detected using Safer found on Amazon Photos. Safer is a CSAM classifier, a machine learning tool developed by Thorn which can detect known and unknown CSAM in images and videos. Reporting on proactively detected content offers of insight into a platform's commitment to keeping its community safe and secure. In addition, Amazon provides two metrics which can be seen as complementary to the proactive detection. These are the number of reports of other content such as chat interactions and URLs from third parties, and the reports by trusted reporters for content quickly removed. These metrics illustrate how Amazon has a system in place whereby external agents (e.g., NCMEC, Internet Watch Foundation (n.d.), Canadian CyberTipline, and INHOPE hotlines) can flag content directly to the company, resulting in swift removal.

3.2.6. Escalations and legal or government requests (n = 13)

This metric primarily revolves around two categories: legal requests for the removal of content and legal requests for account information. These requests are typically submitted by government and law enforcement agencies to online content-sharing services. Typically, this metric is not disaggregated by content type, while likely including OCSEA activity it does not shed light on escalations and legal requests specific to OSCEA. Identifying what legal and government requests are related to OCSEA would be beneficial to identify further actions taken in response to OCSEA. Thirteen services, including Zoom, Discord, Reddit, Pinterest, and Verizon, report metrics on the legal requests they receive and respond to. For instance, Pinterest receives legal requests from law enforcement and government entities for Pinterest account information. Additionally, Pinterest also receives content removal requests from government entities. Reddit, on the other hand, may disclose specific account information in response to valid legal requests from government and law enforcement agencies or private parties, such as civil litigants and criminal defendants, when required by law and in certain emergency situations. During the reporting period of July–December 2022, there was a notable 29 % decrease in the volume of global law enforcement and government removal requests compared to the previous reporting period. However, there was a 3.4 % increase in the rate of removal, which includes both Content Policy violations and geo-blocking actions, compared to the previous reporting period (Reddit, 2023). Furthermore, Discord provided information regarding grooming or endangerment escalations, indicating a focus on addressing potentially harmful content or interactions on their platform.

4. Discussion

This content analysis offers a comprehensive evaluation of the metrics related to OCSEA as reported in transparency reports by online content-sharing services. Building on insights from the OECD report, this study further refined these metrics to present a nuanced analysis of the data provided by these platforms. Of the 19 included transparency reports, 132 data points and 22 distinct metrics were identified. However, despite the number of metrics, it became clear by looking at the data that any given metric may only have a couple of services reporting it. Similar metrics are also not measured in the same way or even with enough clarity on how they are measured to determine if they are comparable. In essence, there is a data landscape, but it is not comparable and not usable for enacting change to keep children safe across the sector.

4.1. Key findings

4.1.1. Various definitions of OCSEA used across transparency reports

The variety of definitions of OCSEA adopted by online content-sharing services has made comparing findings and estimates of prevalence across reports and data sources challenging and complicated. The scope of definitions varies between online content-sharing services. For example, some online content-sharing services define OCSEA narrowly, focusing solely on explicit sexual content involving minors, while others may adopt broader definitions that include a wider range of behaviors, such as grooming, coercion, or non-explicit forms of exploitation. Additionally, some online content-sharing services have clear guidelines on OCSEA criteria, while other online content-sharing services may have more ambiguous criteria including prohibited content in broad or general terms without specific reference to OCSEA or child endangerment. Few services include AI-generated OCSEA content within their definitions. Overall, the lack of consistency in definitions of OCSEA among online content-sharing services complicates efforts to compare findings and estimates of prevalence across reports and data sources. As such, harmonizing definitions and standardizing criteria for identifying and reporting CSEA could facilitate more accurate measurement and assessment of the problem. There are several global initiatives underway for this including an update on the by ECPAT (2021) on terminolgy guiedlines to include online and technology-facilitated CSEA. Additionally, there has been a proposal for a global classification system aimed at improving the comparability of data across organizations (INHOPE, 2023).

4.1.2. Lack of consistency of metrics across online content-sharing services

The majority of transparency reports discuss the prevention, detection and removal measures they utilize to respond to OCSEA. However, analysis on the 19 transparency reports revealed a wide variation in metrics. Metrics related to content, including actions taken; ranging from overall content actioned that encompassed the total content enforced upon from suspension to removal to more

discrete content related measures such as content removed only. Findings from the content analysis highlight the importance of effective and proactive detection metrics in the fight against CSAM. It emphasizes the proactive nature of detection measures, which are essential for intercepting and combating the dissemination of CSAM before it can inflict further harm. Through a comprehensive review of detection-related metrics extracted from transparency reports, findings highlight the important role of proactive detection in safeguarding children. Notably, online content-sharing services such as Meta, Yubo, TikTok, and Microsoft Bing have made progress in identifying and addressing CSAM through proactive detection measures. A proactive approach not only enables online content-sharing services to detect and remove CSAM promptly but also plays a crucial role in disrupting the cycle of abuse associated with CSAM. In essence, the detection-related metrics highlight the proactive detection strategies employed by leading online content-sharing services, recognizing their instrumental role in combating CSAM and fostering a safer online environment for children.

Similarly, metrics related to accounts took various forms; including overall actioned account metrics to individual metrics related to accounts being suspended or warned. Few transparency reports gave specific measures related to detection, with the mechanism of detection incorporating multiple strategies including technology approaches, content moderators and user reports. There is a paucity of comparable data in the field to allow for accurate estimation of the prevalence of OCSEA, with variation in how online content-sharing services produce metrics with limited information on how the metrics are calculated. In addition, some of the metrics are not mutually exclusive, which might lead to an over-estimation of the prevalence rate.

4.1.3. Challenges in locating OCSEA data, reports and user reports

The inconsistency in report publication is another key barrier to estimating the prevalence identified in the rapid assessment. Transparency reports relating to OCSEA are not produced consistently across online content-sharing services and data are difficult to locate as they are published in multiple websites, with different titles and sometimes lack specific categories for OCSEA and CSAM reporting. This aligns with a study by the Canadian Centre for Child Protection (2020), which found that although most platforms have general reporting mechanisms, they rarely offer a CSAM-specific process or menu options for users to report content that is (or is suspected to be) CSAM. However, it is encouraging that during 2020–2021, 86 % of Tech Coalition Members regularly publish transparency reports that include CSEA data or have stated an intention to do so starting 2022 (Tech Coalition, 2022). However, there may be a bigger role here that regulators such as OFCOM and e-Safety Commission, among other initiatives (e.g., Lantern Project at Tech Coalition), can play in setting minimum standards for transparency reporting metrics to ensure comparability across sources.

4.1.4. Lack of disaggregated data

Existing metrics are not disaggregated by sex, age, location etc., which prevents researchers and public from developing a deeper understanding of the scope of online CSEA. Lack of such disaggregated data also hampers efforts at effective prevention and evidence-based programming, especially for vulnerable groups such as children with disabilities, who face unique challenges in OCSEA victimization due to their vulnerabilities (Álvarez-Guerrero et al., 2024). As such, access to such data could strengthen efforts to create targeted educational programs on consent, helping children recognize inappropriate behaviors and understand the consequences of sharing personal information online (Álvarez-Guerrero et al., 2024). In addition, disaggregated data can also provide insights into the characteristics and behaviors of perpetrators of OCSEA. Understanding the demographics and patterns associated with perpetrators can inform law enforcement strategies, educational initiatives, and technological solutions aimed at combating OCSEA. Moreover, the absence of a clear definition of “child” or “age” when reporting OCSEA content poses major including inconsistent enforcement, legal ambiguities, ineffective detection, ethical concerns, and difficulties with international compliance. Explicitly defining these terms is essential to enhance content moderation effectiveness and ensure comprehensive protection for all minors.

4.1.5. Lack of time-focused metrics

There is a lack of metrics reporting time-focused indicators such as content actioned before views and uploading. Time-focused metrics are essential for evaluating the efficiency and effectiveness of online content-sharing services' responses to reports of CSEA. Metrics that track the time it takes for content to be actioned such as removed, flagged, or reported, provide insights into the speed at which online content-sharing services are addressing harmful material. This is crucial for minimizing the exposure of users, particularly children, to abusive content. Time-related metrics contribute to establishing online content-sharing services accountability by reporting on the timeliness of their responses as well minimizing harm to victims. A survivor survey, which included 150 participants, sought to understand the experiences of child sexual abuse survivors. The findings revealed insights into the unique needs of survivors, particularly those whose abuse was recorded or distributed. Many survivors reported that the recording and potential distribution of their abuse added another layer of trauma, with over 70 % expressing constant fear of being recognized by someone who has seen the images. These findings emphasized the urgent need for rapid detection and removal of CSAM/OCSEA material to limit the time harmful content accessible online, thereby reducing further harm to victims (Canadian Centre for Child Protection, 2017).

4.1.6. Limitations

In addition to the limitations identified from the metrics, the content analysis has several other limitations. Firstly, it only included available transparency reports from the top 50 online content-sharing services. Future research should aim to incorporate transparency reports from a broader range of services. Secondly, similar metrics are often not measured in a consistent manner. This inconsistency hampers the ability to determine if these metrics are truly comparable. Therefore, any comparability of these metrics should be interpreted with caution.

5. Conclusion

Online content-sharing services seek to demonstrate a commitment to transparency and accountability by publishing transparency reports. It is essential that they continue to refine their methods of reporting exploitative content, so that this reporting can effectively reflect emerging trends in OCSEA. We should note that simply reporting numbers and ratios over time do not necessarily indicate how much offending behavior takes place, nor how the platform's algorithms and moderation efforts are successfully identifying and acting on offending material. Insights into the patterns and trends associated with OCSEA as reported by online content-sharing services lacks the contextual information that would allow to infer how much of observed change over time is due to changes in policies, procedures, and technological innovation, and how much relates to offending behavior and victimization. Reporting could provide more specific details, namely the types of content being actioned, the number of unique images or videos, the age of the victims depicted, and the frequency with which certain types of content appear. By providing this kind of granular data, online content-sharing services could better equip researchers, policymakers, law enforcement agencies, and civil society organizations to understand and address the problem. Furthermore, this approach would allow consumers to make informed decisions about which products and services to use based on the services record of accomplishment when it comes to protecting vulnerable populations such as children.

Funding source

This work is supported by a research grant by the Human Dignity Foundation.

CRediT authorship contribution statement

Mengyao Lu: Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Maria Lamond:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Deborah Fry:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization.

Data availability

Data will be made available on request.

Acknowledgement

The authors wish to express their gratitude to Dr. Pedro Jacobetty for his valuable feedback and contributions to the original draft. They also extend their thanks to the authors of the original OECD study that served as the foundation for this research.

Author contributions

Mengyao Lu: conceptualization, methodology, data extraction and analysis, original draft preparation. Maria Lamond: data extraction and analysis, editing of the manuscript. Deborah Fry: conceptualization, reviewing, and editing of the manuscript.

Appendix A

Table A1
. Online content-sharing services that do and do not have transparency reports.

Online content sharing service with transparency reports	Online content sharing service without transparency reports
Ask.fm (IAC [InterActiveCorp])	Baidu Tieba (Baidu, Inc.)
Discord (Discord, Inc.)	DeviantArt (DeviantArt, Inc.)
Dropbox (Dropbox, Inc.)	Douban (Information Technology Company, Inc.)
Facebook (Meta Platforms, Inc.)	Flickr (SmugMug, Inc.)
Facebook Messenger (Meta Platforms, Inc.)	Huoshan (ByteDance Technology Co.)
Google Drive (Alphabet, Inc.)	iMessage/FaceTime (Apple, Inc)
Instagram (Meta Platforms, Inc.)	IMO (PageBites, Inc.)
LINE (Line Corporation)	iQIYI (Baidu, Inc.)
LinkedIn (Microsoft, Inc.)	KaKao Talk (Daum Kakao Corporation)
Microsoft OneDrive (Microsoft, Inc.)	Kuaishou (Beijing Kuaishou Technology Co., Ltd)
Microsoft Teams (Microsoft, Inc.)	Likee (BIGO Technology PTE. LTD.)
Pinterest (Pinterest, Inc.)	Medium (A Medium Corporation.)
Reddit (Reddit, Inc.)	Odnoklassniki (Mail.Ru Group)
Skype (Microsoft, Inc.)	Picsart (Picsart, Inc.)

(*continued on next page*)

Table A1 (*continued*)

Online content sharing service with transparency reports	Online content sharing service without transparency reports
Snapchat (Snap, Inc.)	QQ (Tencent Holdings Ltd.)
Tik Tok (ByteDance Technology Co.)	Quora (Quora, Inc.)
Twitch (Amazon.com, Inc.)	QZone (Tencent Holdings Ltd.)
Twitter (Twitter, Inc.)	Smule (Smule, Inc.)
YouTube (Alphabet, Inc.)	Telegram (Telegram Messenger LLP)
Zoom (Zoom Video Communications, In	Tumblr (Automattic, Inc.)
Verizon	Viber (Rakuten, Inc.)
	Vimeo (Vimeo, Inc.)
	VK (Mail.Ru Group)
	Weibo (Sina Corp.)
	Weixin/WeChat (Tencent Holdings Ltd.)

Note. Adapted from OECD, 2023 Report. Transparency reporting on child sexual exploitation and abuse online (oecd.org).

Appendix B

Table A2
Links to included transparency reports.

Name of service	Link to report
Discord	https://discord.com/safety-transparency-reports/2023-
Microsoft Bing	2023 October Microsoft Bing EU DSA Report
Pinterest	https://policy.pinterest.com/en/transparency-report
Reddit	2022 Transparency Report - Reddit (redditinc.com)
Dropbox	https://help.dropbox.com/transparency/reports
LinkedIn	https://about.linkedin.com/transparency/community-report
Microsoft	https://www.microsoft.com/en-us/corporate-responsibility/digital-safety-content-report?activetab=pivot_1:primaryr3
Google	https://transparencyreport.google.com/youtube-policy/featured-policies/child-safety?
Meta	https://transparency.fb.com/reports/community-standards-enforcement/
Youtube	https://transparencyreport.google.com/youtube-policy/removals?hl=en_GB
Twitch	https://safety.twitch.tv/s/article/H1-2023-NetzDG-Transparency-Report?language=en_US
Snap	https://values.snap.com/privacy/transparency
Yubo	https://yubo.cdn.prismic.io/yubo/4bb10550-0506-4f82-a455-20ba06fd9ecd_Yubo_Transparency+Report_Second_Half_2023.pdf
X (formally Twitter)	https://transparency.twitter.com/en/resources.html
Tiktok	https://www.tiktok.com/transparency/en/community-guidelines-enforcement-2023-2/
Verizon	https://www.verizon.com/about/sites/default/files/International-Transparency-Report-1H-2023.pdf
LINE	https://linecorp.com/en/security/transparency/2022h2
Zoom	https://explore.zoom.us/en/trust/transparency/
Amazon	https://brandservices.amazon.com/transparency

Appendix C

Table A3
Online content-sharing services metrics reported by category.

		Metrics		Category			
		Detection-related	Content-related	Reports	Time-related	Escalations or legal requests	Account-related
	Amazon		x	x		x	x
	Reddit	x	x	x		x	x
	Microsoft	x	x	x			x
	Microsoft (Bing)		x	x			
	Dropbox			x			x
	LINE		x				
	LinkedIn		x	x		x	
	Snap		x	x		x	x
Online	Twitch	x	x	x		x	x
Content	X		x			x	x
Sharing	Verizon			x			
Service	Yubo		x	x			x
	Zoom	x	x	x		x	x
	Discord			x		x	x
	Google	x	x	x		x	x
	Meta	x	x	x			

(*continued on next page*)

Table A3 (*continued*)

			Metrics	Category		
Pinterest	x	x	x	x	x	x
Tiktok	x	x		x	x	
YouTube	x	x		x		

References

African Union, (n.d.) African Union Initiative on: Strengthening Regional and National Capacity and Action against Online Child Sexual Exploitation and Abuse in Africa Strategy and Plan of Action 2020–2025 Online-Child-Sexual-Exploitation-and-Abuse-OCSEA-2020-2025-Strategy-1.pdf (aucecma.org).

Ali, S., Haykal, H. A., & Youssef, E. Y. M. (2021). Child sexual abuse and the internet—A systematic review. *Human Arenas*, 1–18. Child Sexual Abuse and the Internet—A Systematic Review | Human Arenas springer.com.

Álvarez-Guerrero, G., Fry, D., Lu, M., & Gaitis, K. K. (2024). Online child sexual exploitation and abuse of children and adolescents with disabilities: A systematic review. *Disabilities, 4*(2), 264–276.

Canadian Centre for Child Protection (2017). Survivors' Survey Full Report. C3P_SurvivorsSurveyFullReport2017.pdf.

Canadian Centre for Child Protection (2020). Reviewing Child Sexual Abuse Material Reporting Functions on Popular Platforms. C3P_ReviewingCSAMMaterialReporting_en.pdf.

ECPAT International. (2016). *Terminology Guidelines for the Protection of Children from Sexual Exploitation and Sexual Abuse*. UK: Terminology guidelines | ECPAT.

ECPAT. (2021). *Terminology guidelines: Terminology related to sexual exploitation of children*. ECPAT International. Retrieved from https://ecpat.org/wp-content/uploads/2021/05/Terminology-guidelines-396922-EN-1.pdf.

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing, 62*(1), 107–115.

European Commission. (2022). Digital Services Act. Retrieved from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065.

Grossman, S., Pfefferkorn, R., Thiel, D., Shah, S., Stamos, A., DiResta, R., Perrino, J., Cryst, E., & Hancock, J. (2024). The strengths and weaknesses of the online child safety ecosystem: Perspectives from platforms, NCMEC, and law enforcement on the CyberTipline and how to improve it. *Stanford Internet Observatory*. https://purl.stanford.edu/pr592kc5483.

Internet Watch Foundation. (n.d.). IWF. https://www.iwf.org.uk/).

Kids Online Safety Act (2023) S.1409 - 118th Congress (2023–2024): Kids Online Safety Act | Congress.gov | Library of Congress.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). Sage Publications.

Krzeczkowska, A., Fry, D., Anderson, N., Ren, J., Lu, M., Lu, Y., … Fang, X. (2024). *Indicator 1: The prevalence of online victimisation, technical note in into the light: Childlight's global child sexual exploitation and abuse index* (p. 2024). Edinburgh: Childlight. https://childlight.org/sites/default/files/2024-05/technical-note-1.pdf.

Livingstone, S., Davidson, J., Bryce, J., & with Batool, S., Haughton, C., & Nandi, A.. (2017). *Children's online activities, risks and safety: A literature review by the UKCCIS evidence group*. London: London School of Economics (LES) consulting.

Meta Platforms, Inc.. (2023). Integrity and transparency reports: First quarter 2023. In *Meta*. https://about.meta.com/newsroom/.

National Center for Missing & Exploited Children. (n.d.). NCMEC. https://www.missingkids.org/.

OECD. (2023). *Transparency reporting on child sexual exploitation and abuse online", OECD Digital Economy Papers, 357*. Paris: OECD Publishing. https://doi.org/10.1787/554ad91f-en

Online Safety Act 2023- Parliamentary Bills - UK Parliament.

Online Safety Bill. (2021). *Parliament of Australia*. aph.gov.au.

Parliament of Canada. (2023). *Bill C-63: An act to enact the Online Harms Act, to amend the Broadcasting Act and to make related amendments to other acts*. Parliament of Canada. https://www.parl.ca/DocumentViewer/en/44-1/bill/C-63/first-reading.

Quayle, E. (2016). Researching online child sexual exploitation and abuse: Are there links between online and offline vulnerabilities?. http://globalkidsonline.net/wp-content/uploads/2016/05/Guide-7-Child-sexual-exploitation-and-abuse-Quayle.pdf.

Reddit. (2023). *2023 H1 transparency report*. Reddit Inc.. Retrieved from https://www.redditinc.com/policies/2023-h1-transparency-report-.

Stevenson, J., Vermeulen, I., & Fry, D. (2024). *Indicator 3: The global scale and nature of child sexual abuse material (CSAM) online, technical note for into the light 2024: Childlight's global child sexual exploitation and abuse index*. Edinburgh: Childlight. https://intothelight.childlight.org/indicator-3.html.

Tech Coalition (2022) Tech Coalition | Annual Report (technologycoalition.org).

TikTok. (2023). Transparency reports. *TikTok Transparency Center*. https://www.tiktok.com/transparency/en/reports.

Trust and Safety Professional Association (2024). Transparency report categories. Retrieved August 5, 2024, from https://www.tspa.org/curriculum/ts-fundamentals/transparency-report/transparency-report-categories/.

Wager, N., Armitage, R., Christmann, K., Gallagher, B., Ioannou, M., Parkinson, S., et al. (2018). Rapid evidence assessment: Quantifying the extent of online-facilitated child sexual abuse: Report for the independent inquiry into child sexual abuse. Available at http://cdn.basw.co.uk/upload/basw_103534-9.pdf.

Yubo. (2023). Transparency report. *Yubo*. https://www.yubo.live/safety/transparency-report.

Zoom. (2023). Transparency report. *Zoom Trust Center*. https://explore.zoom.us/en/trust/transparency/.