



An OSINT-Based Mobile Number Intelligence Framework for Ethical Investigations

Mrs. Gandhi R.S.¹, Mrs. Dalavi M.T.², Mr. Karan Bhavar³, Mr. Akshay Bhawar⁴

Assistant Professor, Dattakala Group of Institutions Faculty of Engineering, Tal-Daund, Pune¹

Author, Dattakala Group of Institutions Faculty of Engineering, Tal-Daund Dist-Pune²⁻⁴

Abstract: This paper addresses the critical challenge of conducting digital investigations in an era of ubiquitous data and stringent privacy regulations. Open-Source Intelligence (OSINT) provides powerful capabilities for gathering information, with the mobile number serving as a unique digital identifier and a crucial pivot point. However, existing tools are often fragmented, lack integrated ethical guardrails, and risk violating legal frameworks like the GDPR. We propose a novel, integrated software architecture for a mobile number intelligence tool designed on the principle of "Ethical-by-Design." The framework features a modular architecture comprising data collection, processing, AI-augmented analysis using an on-premise Large Language Model (LLM) to ensure operational security, and a visualization dashboard. We present a structured investigative workflow that embeds legal compliance checks, such as documenting the lawful basis for processing, directly into the process. Through conceptual case studies in law enforcement and cybersecurity, we demonstrate the framework's efficacy in generating actionable intelligence while adhering to principles of data minimization and proportionality. This research contributes a blueprint for the next generation of OSINT tools that harmonize advanced investigative capabilities with the fundamental right to privacy.

Keywords: Data privacy, digital forensics, ethical hacking, mobile number intelligence, open-source intelligence (OSINT).

I. INTRODUCTION

THE DIGITAL age has fundamentally transformed the landscape of intelligence gathering and investigation. The exponential growth of publicly available information (PAI)—disseminated across social media networks, public government records, news outlets, and the deep web—has elevated Open-Source Intelligence (OSINT) from a supplementary practice to an indispensable discipline.¹ OSINT is no longer confined to the domain of national intelligence agencies; it now serves as a foundational methodology for a diverse range of professionals, including cybersecurity analysts, law enforcement officers, corporate investigators, and investigative journalists.² The discipline's origins, which can be traced to military intelligence operations during the Cold War, have evolved dramatically with the ubiquity of the internet, rendering it a more potent and universally accessible tool than ever before.² This rapid evolution presents both unprecedented opportunities for discovery and significant challenges related to the sheer volume of data, its veracity, and the profound implications for individual privacy.⁴

A. The Evolving Landscape of Digital Investigations

The modern investigative environment is characterized by a data deluge. Every online interaction, from a social media post to a domain registration, creates a digital trace that can be collected and analyzed.⁵ This vast repository of PAI offers a powerful resource for understanding and responding to a wide array of threats and events, including cybercrime, disinformation campaigns, corporate fraud, and human rights violations.² The structured collection and analysis of this information are what define the modern practice of OSINT. This process allows security teams to proactively identify vulnerabilities in their own organizations, such as exposed metadata or unpatched software, before they can be exploited by malicious actors.¹ Similarly, law enforcement agencies leverage OSINT to gather evidence, identify suspects, and monitor criminal networks operating in plain sight on public platforms.

B. Defining Open-Source Intelligence (OSINT) and its Methodologies

Formally, OSINT is defined as intelligence derived exclusively from publicly or commercially available information that is collected, exploited, and disseminated to address a specific intelligence requirement.² It is a structured and ethical approach that transforms raw, open-source data into actionable insights through a systematic process.² This process is often modeled on the traditional intelligence cycle, a systematic and iterative framework comprising several distinct phases: planning and direction, data collection, processing, analysis and production, dissemination, and feedback. Methodologies for data collection in OSINT are typically categorized based on the level of interaction with the target.



Passive collection, the most common and lowest-risk method, involves gathering information from public sources without directly engaging with the target's systems. This includes activities like scraping publicly accessible websites, utilizing public APIs (e.g., for social media platforms), or searching through archived data, all of which ensure the investigation remains covert and leaves no digital footprint.¹ In contrast, *semi-passive* and *active* collection methods involve more direct interaction, such as sending carefully crafted network traffic to a target's server to elicit a response. While these methods can yield more specific technical information, they carry a significantly higher risk of detection and must be employed with caution.

C. The Mobile Number as a Critical Pivot Point in Digital Footprinting

Within the expansive universe of PAI, the mobile phone number has emerged as a uniquely powerful digital identifier. It functions as a critical bridge, connecting an individual's identity in the physical world of telecommunications with their multifaceted persona in the digital realm.¹¹ Mobile devices have become central repositories for vast quantities of sensitive personal information, ranging from contact lists and private communications to financial credentials and location history.⁵ As a result, a single mobile number can serve as the initial thread to unravel an individual's entire digital footprint. An OSINT investigation commencing with a phone number can uncover a wealth of associated data points, including the owner's name, physical address, location history, carrier details, and, most critically, linked social media profiles and other online accounts where the number was used for registration or verification.¹² This makes "mobile number intelligence" a pivotal and highly effective sub-discipline of OSINT, enabling investigators to map an individual's digital presence with remarkable precision.

D. Problem Statement: The Dichotomy of Powerful OSINT Tools and Privacy Imperatives

The immense power of OSINT creates an inherent and challenging dichotomy. The very same techniques and data sources that empower cyber defenders to identify and mitigate organizational weaknesses—such as open ports, leaked employee credentials, or exposed infrastructure details—are the exact same ones exploited by malicious threat actors for reconnaissance.¹ Attackers use OSINT to gather personal and professional information about employees from social media to craft highly targeted and effective spear-phishing campaigns.¹ This dual-use nature of open-source information gives rise to a fundamental tension: the legitimate and necessary pursuit of information for security and justice versus the individual's fundamental and inalienable right to privacy.¹⁴

This conflict is being amplified by two parallel and accelerating trends. On one hand, OSINT tools are becoming more powerful and automated, with the integration of artificial intelligence (AI) and machine learning (ML) promising deeper and more efficient analysis of vast datasets.¹⁰ On the other hand, a global movement towards stronger data protection has resulted in the enactment of stringent legal frameworks, most notably the European Union's General Data Protection Regulation (GDPR) and India's new Digital Personal Data Protection (DPDP) Act.¹⁵ These regulations impose strict rules on the processing of personal data, demanding lawfulness, transparency, and accountability. The central problem this paper addresses is the conspicuous absence of an integrated framework that equips investigators with advanced mobile number intelligence capabilities while systematically embedding ethical principles and legal compliance checks into the investigative process itself.

E. Contribution and Paper Organization

This paper introduces a comprehensive framework for an OSINT-based mobile number intelligence tool meticulously designed for ethical investigations. The primary contributions of this research are threefold: 1) A systematic review of the current state-of-the-art in OSINT tools, which identifies a critical gap in process-oriented, ethically-grounded frameworks that move beyond fragmented, single-purpose utilities. 2) A novel, modular system architecture founded on the principles of "Privacy by Design" and operational security, featuring the use of an on-premise Large Language Model (LLM) for secure, deep semantic analysis. 3) An integrated investigative workflow that aligns with the requirements of major international data protection regulations, transforming legal compliance from a post-investigation consideration into an integral part of the process.

The remainder of this paper is organized as follows: Section II provides a review of related work in the field of mobile number OSINT. Section III details the proposed system architecture and its associated investigative workflow. Section IV conducts an in-depth analysis of the ethical and legal frameworks governing such investigations. Section V presents conceptual case studies to demonstrate the practical application and evaluation of the proposed framework. Finally, Section VI concludes the paper and discusses avenues for future work.

II. STATE OF THE ART IN MOBILE NUMBER OSINT

A thorough understanding of the current landscape of OSINT tools and techniques is essential to identify existing capabilities and pinpoint critical gaps. This section presents a systematic review of existing frameworks, details the



primary data sources for mobile intelligence, and defines the research gap that motivates the development of the proposed system.

A. Systematic Review of Existing OSINT Frameworks and Tools

The current ecosystem of OSINT tools is both vast and highly fragmented, comprising comprehensive frameworks, specialized utilities, and curated resource lists. General-purpose platforms like Maltego and SpiderFoot represent the more sophisticated end of the spectrum. They offer modular designs that allow investigators to construct custom data collection and analysis workflows.¹⁸ Maltego, in particular, excels at link analysis, providing powerful visualizations that map the relationships between disparate data points such as phone numbers, email addresses, social media profiles, and corporate entities.¹ However, the power of these frameworks is often contingent on significant user configuration and the purchase of commercial "transforms" or modules, which are required to access premium or aggregated data sets.

Alongside these overarching frameworks, a plethora of highly specialized tools exist for the express purpose of mobile number intelligence. Academic literature and practitioner guides frequently cite utilities such as PhoneInfoga, Zlookup, Truecaller, and Sync.ME.¹⁶ These tools primarily perform reverse lookups, querying various databases to identify a number's registered owner, associated carrier, and line type (e.g., mobile, landline, VoIP). Other tools, like Email2phonenumber, are designed for pivoting, attempting to discover a phone number from a known email address by scraping various online platforms.²⁰ While these tools are often effective at their specific, narrow tasks, they typically operate in silos. This forces investigators to manually pivot between different browser tabs and applications, painstakingly collating data and attempting to correlate findings on their own. Curated resources like the OSINT Framework website and "Awesome-OSINT" lists on platforms like GitHub are invaluable repositories for discovering these tools, but their very structure—a categorized list of individual utilities—underscores the fragmented nature of the current toolchain.

This fragmentation is not merely an inconvenience; it represents a systemic risk. The manual process of transferring data between disparate, often third-party, online services creates multiple potential points of data leakage. Each query to an external service can expose the investigator's IP address, the target's information, and the nature of the investigation itself, thereby compromising operational security (OPSEC). Furthermore, maintaining a clean and verifiable chain of evidence—a critical requirement for legal admissibility—becomes exceedingly difficult when data is being copied and pasted across numerous interfaces without a unified, automated logging mechanism. The need for dedicated evidence-capture tools like Hunchly or meticulous manual note-keeping applications like KeepNote is a direct symptom of this systemic fragmentation; they are functional patches for a problem that should be solved at an architectural level.²³ The most significant contribution a new tool can therefore make is not simply to add another data scraper to the list, but to unify these disparate functions into a single, secure, and auditable workflow, transforming a collection of ad-hoc tactics into a coherent and defensible investigative strategy.

Tool Name	Primary Function	Key Data Sources	Integration Capability	Cost Model	Noted Limitations
PhoneInfoga	Technical Number Analysis	Google Search, Public APIs, Numverify	Command-line, Web UI	Free (Open Source)	Requires API keys; siloed functionality; no integrated case management.
Maltego	Link Analysis & Visualization	Social Media, Public Records, Breach DBs (via Transforms)	API (via Transforms)	Freemium/Commercial	Core functionality relies on paid, third-party transforms; steep learning curve.



Truecaller	Reverse Phone Lookup (Caller ID)	Crowdsourced User Data	API (Limited)	Freemium	Data accuracy depends on crowdsourcing; significant privacy concerns with data collection.
Sync.ME	Reverse Lookup & Social Media Correlation	Social Media Platforms, Public Data	API	Freemium	Primarily focused on contact enrichment; limited deep investigative features.
SpiderFoot	Automated OSINT Aggregation	200+ Public Sources (Modules)	API (via Modules)	Free/Commercial	Can generate significant noise; requires careful configuration to be effective.
Moriarty Project	Phone Number Search (Python Tool)	Social Media, Spam DBs, Search Engines	N/A (Standalone Tool)	Free (Open Source)	Requires technical proficiency; lacks a graphical interface and evidence management.
Table I. Comparative Analysis of Existing Mobile Number OSINT Tools					

B. Data Sources and Collection Techniques for Mobile Intelligence

Effective mobile number intelligence is predicated on the ability to aggregate and correlate data from a wide and diverse array of public sources. The proposed tool's data collection module must be capable of systematically querying these sources to build a comprehensive profile. Key source categories include:

1. *Public and Government Records*: Many government agencies at national and local levels publish reports, datasets, and public records such as court documents, property records, and business registration filings. These documents can often contain phone numbers associated with individuals or corporate entities, providing a highly reliable data point.² People search engines such as WhitePages, Spokeo, and TruePeopleSearch specialize in aggregating this publicly available data, making it searchable through a single interface.¹⁸
2. *Social Media and Online Platforms*: A vast number of individuals link their mobile numbers to social media accounts on platforms like Facebook, X (formerly Twitter), and LinkedIn for purposes such as account recovery, friend-finding features, or two-factor authentication.¹² A direct search for a phone number within these platforms'



search bars can sometimes reveal an associated user profile. Even if direct search is disabled, the number can be used in password reset workflows to confirm the existence of an associated account.

3. *Data Breach Corpora*: Leaked databases from historical data breaches represent a significant, albeit ethically complex, source of intelligence. Services like HaveIBeenPwned (HIBP) and Dehashed maintain massive, searchable repositories of data from thousands of breaches.¹ An investigator can query these services with a phone number to determine if it has been exposed in a known breach. The results often include associated data such as usernames, email addresses, plaintext or hashed passwords, and the source of the breach. This information is invaluable for mapping a target's digital footprint across different services and assessing their overall security posture.
4. *Technical Data Sources*: Beyond identifying the owner, technical details about the number itself can provide crucial context. Free online tools and commercial APIs, such as those provided by PhoneValidator.com or Twilio, can be used to validate a number's existence and determine its technical characteristics.¹² This includes the issuing carrier (e.g., AT&T, Verizon), the line type (mobile, landline, or Voice over IP - VoIP), and the geographic region associated with its registration. Identifying a number as a disposable or VoIP service is a critical indicator, as these are commonly used by scammers or individuals seeking to obscure their identity.¹²
5. *Search Engine Dorking*: Advanced search operators, commonly known as "Google Dorks," are a powerful technique for uncovering information that is not easily found through simple keyword searches. By using operators like filetype:pdf, inurl:contact, and site:, an investigator can craft highly specific queries to find documents, web pages, forum posts, or pastebin entries where a target phone number has been inadvertently exposed.

C. Identifying the Research Gap: The Need for an Ethically-Grounded, Integrated Tool

Our systematic review of the existing OSINT landscape reveals a significant and critical research and development gap. The current ecosystem of tools is overwhelmingly **capability-centric**, offering a fragmented collection of powerful but disconnected utilities. There is a profound lack of **process-oriented** frameworks that are designed to guide an investigator through an entire ethical investigation lifecycle from initiation to reporting. This systemic fragmentation leads to several critical, interlocking issues that undermine the effectiveness and legitimacy of digital investigations:

1. *Operational Security (OPSEC) Risks*: The manual process of pivoting between numerous online services and websites dramatically increases the investigator's digital footprint. Each query leaves a trace, and without a centralized and anonymized access point, the investigator risks exposing their identity, location, and the focus of their investigation to the target or to third-party service providers.
2. *Evidence Integrity Challenges*: Collating data from a dozen different tools into a single coherent case file is a recipe for error and omission. Without a unified logging and data-stamping mechanism, it becomes exceptionally difficult to maintain a clear, chronological, and admissible chain of evidence.¹² This can jeopardize the use of the collected intelligence in legal proceedings.

III. PROPOSED SYSTEM ARCHITECTURE AND WORKFLOW

To address the identified gaps of fragmentation, operational insecurity, and lack of integrated ethical governance, we propose a novel system architecture and an accompanying investigative workflow. The framework is founded on the principle of "Ethical-by-Design," a proactive approach that treats legal and ethical compliance not as external constraints but as core functional requirements of the system itself.

A. Conceptual Framework: The 'Ethical-by-Design' Principle

The proposed system is architected from the ground up based on the principle of "Ethical-by-Design." This is a direct extension of the well-established concept of "Privacy by Design," which advocates for the incorporation of privacy safeguards from the very beginning of the development process.¹⁵ In our framework, this means that every module, data flow, and user interaction is intentionally designed to promote transparency, enforce accountability, and facilitate compliance with key data protection principles. The system is not merely a passive tool for data retrieval; it is an active governance framework. It is designed to compel the investigator to operate within predefined ethical and legal boundaries. For example, the system's workflow makes it impossible to initiate data collection without first formally documenting a lawful basis for the investigation. This creates an auditable, defensible, and contemporaneous record from the outset, ensuring that the justification for the privacy intrusion is considered and recorded before it occurs.

B. System Architecture

The proposed architecture is modular, scalable, and engineered to minimize the investigator's digital footprint while maximizing analytical power. It is designed to leverage distributed processing for performance and on-premise data analysis to ensure the highest level of operational security and data confidentiality.¹⁰ The architecture consists of five core modules, as depicted in Fig. 1.



!(placeholder_fig1.png)

Fig. 1. High-Level System Architecture Flowchart

1. *Data Collection Module*: This module serves as the system's sole interface to the external internet. It integrates a multitude of APIs and robust web scrapers to query a wide range of pre-defined, high-value OSINT sources in a coordinated and automated fashion.¹⁸
 - *Source Integration*: The module contains a library of connectors for key data source categories, including social media platforms (Facebook, X, LinkedIn), people search engines, data breach services (Dehashed, HIBP), domain and IP intelligence tools (WHOIS, DNSDumpster), and public cryptocurrency blockchain explorers.¹ This centralized approach obviates the need for the investigator to visit these sites manually.
 - *Anonymization Layer*: All outgoing requests from the collection module are mandatorily routed through a managed proxy network or the Tor network. This obfuscates the true origin IP address of the investigation, providing a critical layer of operational security (OPSEC) and protecting the investigator's identity and location.²³
2. *Data Processing and Normalization Engine*: Raw data collected from heterogeneous sources is often unstructured, inconsistent, and noisy. This engine acts as a data refinery. It employs a suite of parsers and data cleaning algorithms to parse, structure, and transform the incoming raw data into a standardized, machine-readable format, such as JSON objects, which is essential for effective analysis.² During this stage, it automatically extracts and tags key entities like usernames, email addresses, physical locations, phone numbers, and IP addresses.
3. *Analysis and Correlation Core*: This is the analytical heart of the system. It utilizes a graph database (e.g., Neo4j) to store the normalized data. This data structure is ideal for modeling and exploring the complex relationships between different entities.¹
 - *Identity Stitching*: A key function of this core is to perform automated identity stitching. It systematically searches for common data points—such as a recurring username, email address, or profile picture hash—across different data sources. When a match is found, it creates a relationship (an "edge") in the graph database, effectively linking disparate online profiles to a single, unified digital identity.²⁹
 - *Pattern Recognition*: The core also analyzes temporal data, such as the timestamps of social media posts, to identify patterns-of-life. This can help determine a subject's likely time zone, periods of activity, and potential anomalies that deviate from their established baseline behavior.²⁹
4. *AI/LLM Augmentation Layer (On-Premise)*: To achieve deep semantic analysis without compromising operational security, this layer integrates a locally-hosted Large Language Model (LLM), such as a fine-tuned version of an open-source model like Llama 3 or Mistral. This is a critical architectural decision. By keeping the LLM on-premise, the system ensures that sensitive investigative queries, prompts, and data are never transmitted to third-party cloud-based AI services, thus preventing data leakage and maintaining the confidentiality of the investigation.¹⁰
 - *Semantic Analysis*: The LLM can process large volumes of unstructured text from sources like social media posts or forum comments to infer sentiment, identify key topics of discussion, and generate concise summaries, saving the analyst significant time.
 - *Advanced Entity Extraction*: The LLM can identify and extract non-obvious entities and relationships from text that simple rule-based parsers would likely miss, such as implied relationships or affiliations.
5. *Visualization and Reporting Dashboard*: The final module is the user-facing interface that presents the correlated and analyzed data to the investigator in an intuitive and interactive manner.
 - *Interactive Graph Visualization*: The primary view is a dynamic, interactive graph that displays entities (people, phone numbers, social media accounts) as nodes and their relationships as edges. This allows the investigator to visually explore the network of connections, identify clusters of activity, and uncover non-obvious links.¹
 - *Automated Report Generation*: At any point in the investigation, the system can generate a comprehensive, timestamped report. This report includes all collected data points, their original sources, the generated visualizations, and a complete audit log of all queries and analytical actions taken by the user. This ensures the entire process is transparent, reproducible, and documented in a format suitable for use as evidence.



Target Data Point	OSINT Source Category	Specific Tools/Techniques	Ethical Consideration/Flag
Owner's Name	Reverse Lookup Services, Public Records	Truecaller API, WhitePages Scraping	High Privacy Impact
Current/Past Addresses	Public Records, People Search Engines	Spokeo API, Voter Registration DBs	High Privacy Impact
Associated Usernames	Social Media Platforms, Breach Databases	Facebook Graph Search, HIBP API, NameChk	Medium Privacy Impact
Associated Email Addresses	Breach Databases, Domain Registries	Dehashed API, WHOIS Lookup	Medium Privacy Impact
Carrier & Line Type	Technical Lookup Services	Twilio API, PhoneValidator.com	Low Privacy Impact
Data Breach Exposure	Breach Databases	HaveIBeenPwned (HIBP) API	Medium Privacy Impact
Linked Social Media Profiles	Social Media Platforms, Search Engines	Direct Platform Search, Google Dorking	High Privacy Impact
Cryptocurrency Addresses	Pastebins, Dark Web Forums, Blockchain Explorers	Pastebin Scrapers, Blockchain.com API	Varies; Requires Strong Justification
Table II. Mapping Data Points to OSINT Sources and Techniques within the Data Collection Module			

C. Investigative Workflow Model

The tool enforces a structured, four-phase investigative workflow. This model is a practical implementation of the standard intelligence cycle, but with critical ethical and legal checkpoints embedded directly into the process.² This workflow ensures that investigations are conducted in a systematic, accountable, and defensible manner.

1. *Phase 1: Case Initiation and Scope Limitation:* The process begins with the investigator creating a new, encrypted case file within the system. The system prompts for the initial target identifier—the mobile number. Crucially, the workflow cannot proceed until the investigator completes a mandatory form to declare and document the **lawful basis** for the investigation. This involves selecting from a predefined list of justifications (e.g., "Prevention or Detection of Crime," "Journalistic Purposes," "Consent from Data Subject") that are derived from relevant data protection legislation like the GDPR.³⁰ The investigator must also provide a brief, written rationale. This initial step enforces accountability and purpose limitation from the very start of the investigation.
2. *Phase 2: Automated Data Aggregation:* Once the lawful basis is documented and saved to the audit log, the investigator can initiate the automated collection process with a single click. The Data Collection Module begins querying all relevant, integrated sources in parallel. As data is returned, it is automatically processed, normalized, and populated into the Analysis Core's graph database. This phase is designed to be both rapid and comprehensive, automating the most time-consuming part of OSINT and freeing the analyst to focus on higher-level cognitive tasks.
3. *Phase 3: Human-in-the-Loop Analysis and Verification:* With the initial data aggregation complete, the investigator interacts with the results via the Visualization Dashboard. They can explore the automatically generated entity graph, clicking on nodes to view detailed information and source data. The role of the human analyst in this phase



is paramount. They use their domain expertise and critical thinking to validate the connections suggested by the system, dismiss false positives, and interpret the significance of the findings.³² The analyst can manually add new entities or relationships based on their insights, ensuring that human intelligence remains central to the process. This "human-in-the-loop" approach is essential for mitigating the risks of purely automated, and potentially biased, conclusions.¹⁵

4. *Phase 4: Evidence Documentation and Reporting:* Upon concluding the analysis, the investigator uses the system to generate a final, comprehensive report. The system compiles all collected data, the final state of the entity graph visualization, a complete, unalterable audit log of every query made and every action taken by the user, and the initial lawful basis declaration into a single, secure, and portable document (e.g., an encrypted PDF). This ensures that the entire investigation is transparent, reproducible, and packaged in a format that is ready for legal scrutiny or dissemination to relevant stakeholders.

IV. ETHICAL AND LEGAL FRAMEWORK FOR MOBILE INTELLIGENCE

The development and use of any OSINT tool that processes personal data must be firmly grounded in a comprehensive understanding of the prevailing ethical standards and legal obligations. An investigation that is technically successful but legally non-compliant is a failure that can lead to inadmissible evidence, civil liability, and significant reputational damage. This section analyzes the key data protection regulations and ethical principles that must govern mobile number intelligence.

A. Navigating Global Data Protection Regulations

Any OSINT tool with the potential to be used in cross-border investigations must be designed with a global perspective on data protection. While laws vary by jurisdiction, several key frameworks have set international benchmarks.

1. *The General Data Protection Regulation (GDPR):* As the most comprehensive and influential data protection law globally, the GDPR's principles are of paramount importance. The regulation applies whenever the personal data of an individual located within the European Union is processed, regardless of where the investigator or the tool is based.¹⁵
 - *Lawful Basis for Processing:* One of the GDPR's core tenets is that all processing of personal data must have a lawful basis (Article 6). For most OSINT investigations, obtaining the data subject's consent is not feasible. Therefore, investigators must typically rely on the "legitimate interest" basis.³⁰ This is not a carte blanche; it requires the investigator to conduct and document a balancing test, demonstrating that their legitimate interest (e.g., in preventing crime or ensuring cybersecurity) is not overridden by the data subject's fundamental rights and freedoms. The entire legitimacy of a non-consensual OSINT investigation under GDPR rests on this subjective, context-dependent balancing act. This creates a significant legal "gray zone" and a point of risk for investigators. A tool designed for ethical investigations must directly address this risk. By building a mandatory, structured "Lawful Basis Assessment" module into the initial workflow, the tool provides a mechanism for the investigator to create a contemporaneous record of their justification, significantly strengthening their legal position and promoting a culture of deliberate, accountable decision-making.
 - *Data Subject Rights:* Investigators must be cognizant of the rights granted to data subjects under GDPR, which include the right to access their data, the right to rectification of inaccurate data, and the right to erasure (the "right to be forgotten").³¹ While Article 23 of the GDPR provides for exemptions to these rights for purposes such as the prevention, investigation, detection, or prosecution of criminal offenses, these exemptions are not absolute and must be applied on a case-by-case basis.³⁰
2. *The Indian Data Protection Framework:* India's legal landscape for data privacy is undergoing a significant transition. The long-standing framework under the **Information Technology (IT) Act, 2000**, is being replaced by the new **Digital Personal Data Protection (DPDP) Act, 2023**.¹⁷
 - *IT Act, 2000:* This foundational act provided the initial rules for data security in India. A particularly relevant provision for OSINT investigators is Section 69A, which grants the Indian government broad powers to block public access to any online information and conduct surveillance, creating a precarious and potentially monitored environment for investigative activities.³⁴
 - *DPDP Act, 2023:* This new, comprehensive law introduces a consent-based framework with principles similar to the GDPR, such as purpose limitation and data minimization.³³ However, it also contains broad exemptions that are highly relevant to investigative work. Section 17(1)(c) of the Act exempts the processing of personal data that is "necessary for the prevention, detection, investigation or prosecution of any offence or contravention of any law".³⁶ The precise scope and interpretation of this exemption have yet to be fully defined by the Data Protection Board of India and the courts, creating a zone of legal ambiguity that necessitates a cautious and meticulously documented approach from any investigator processing the data of Indian residents.³⁶
3. *Other Legal Frameworks (e.g., CCPA):* In the United States, the legal landscape is a patchwork of federal and state



laws. The **California Consumer Privacy Act (CCPA)**, as amended by the CPRA, grants consumers rights regarding their personal information, including the right to know what data is being collected about them and the right to have it deleted.¹⁵ Investigators must consider these and other state-level privacy laws when their activities involve the data of residents of those jurisdictions.

Legal Principle	GDPR Requirement	Indian DPDP Act Requirement	Implication for OSINT Tool Design
Lawful Basis for Processing	Must be established (e.g., legitimate interest) and documented via a balancing test.	Consent is the default, but broad exemptions exist for crime investigation (Sec. 17(1)(c)).	Mandatory "Lawful Basis" declaration and justification module at case initiation.
Data Minimization	Only process data that is adequate, relevant, and limited to what is necessary.	Collect only personal data necessary for the specified purpose.	Features for selective data collection and export; analyst must confirm relevance.
Purpose Limitation	Data collected for one purpose cannot be used for an incompatible purpose.	Data must be processed only for the purpose specified at the time of consent/collection.	Case files must have a clearly defined and documented scope; system logs purpose.
Data Security	Implement appropriate technical and organizational measures to ensure data security.	Obligation to protect personal data with reasonable security safeguards.	End-to-end encryption for data at rest and in transit; role-based access controls.
Data Subject Rights	Right to access, rectification, erasure, etc. (with exemptions for law enforcement).	Right to access, correction, and erasure (with exemptions).	Mechanism to flag data subject requests and facilitate compliance where required.
Breach Notification	Mandatory notification to supervisory authority (within 72 hours) and data subjects.	Mandatory notification to the Data Protection Board and affected individuals.	Automated system monitoring and alerting for potential internal or external breaches.
Table III. Key Compliance Requirements of Major Data Protection Laws and Their Architectural Implications			

B. Defining Ethical Boundaries for Investigators

Legal compliance provides the minimum standard of conduct, but ethical investigation requires adherence to a higher set of principles that prioritize professional integrity and respect for individual dignity.¹⁴

1. *Distinguishing Ethical Intelligence Gathering from Illegal Surveillance:* There exists a bright, unequivocal line between ethical OSINT and illegal surveillance. Ethical OSINT is confined strictly to the collection and analysis of PAI—information that is legally and publicly accessible.³⁷ It does not involve illegal acts such as hacking into private accounts, deploying malware, trespassing on digital systems, or using deception and social engineering to trick individuals into divulging private information.³⁷ The use of "sock puppets" or anonymous online personas is a common and accepted OSINT practice for protecting the investigator's identity, but it must be employed carefully



to gather public information without engaging in fraudulent misrepresentation or harassment.²⁴

2. *Principles of Data Minimization, Proportionality, and Respect for Privacy:*
 - *Data Minimization:* Investigators have an ethical obligation to collect and retain only the data that is strictly necessary to achieve their specified investigative objective.³⁰ The proposed tool is designed to support this principle by focusing the investigation on a specific initial identifier and providing functionality that allows the analyst to select and export only the relevant data points for the final report, leaving irrelevant or incidental data behind.
 - *Proportionality:* The intrusiveness of an investigation must be proportionate to the gravity of the matter being investigated.²⁸ A routine pre-employment background check should employ far less intrusive methods and access fewer data sources than a high-stakes counter-terrorism or child exploitation investigation. The tool's modular design allows for the creation of different investigation templates with varying levels of data collection enabled, facilitating a proportionate response.
 - *Respect for Privacy:* Even when data is technically "public," its aggregation and analysis can reveal deeply sensitive patterns and insights about an individual's life, beliefs, and vulnerabilities that were never intended for public disclosure.¹⁵ An ethical investigator has a duty to handle this inferred intelligence with the utmost responsibility, protecting it from unauthorized disclosure and respecting the privacy and dignity of all individuals involved, including victims, witnesses, and even the subjects of the investigation.¹⁴
3. *Maintaining an Auditable Trail for Legal Admissibility:* For the findings of an OSINT investigation to be considered credible and admissible in a court of law, the entire process of collection and analysis must be meticulously and contemporaneously documented.³⁹ This includes recording every source that was accessed, the precise timestamps of data collection, the data that was retrieved, and the analytical steps taken to reach a conclusion. The proposed tool's automated, unalterable logging and one-click reporting features are designed specifically to meet this critical requirement, ensuring the integrity, transparency, and defensibility of the entire investigative process.¹²

V. CONCEPTUAL APPLICATION AND EVALUATION

To demonstrate the practical utility and ethical advantages of the proposed framework, this section presents two conceptual case studies. These scenarios, drawn from the domains of law enforcement and cybersecurity, illustrate how the integrated tool and its structured workflow can accelerate investigations while adhering to legal and ethical principles.

A. Case Study 1: Law enforcement - Missing Persons Investigation

- **Scenario:** A high-risk missing person case is reported. The individual has been out of contact for 48 hours, and the only immediate lead available to investigators is their last known mobile number. In such situations, time is a critical factor, and the ability to rapidly develop new leads is paramount.
- **Workflow Demonstration:**
 1. *Phase 1 (Initiation):* An investigator in the missing persons unit creates a new case file in the tool. They input the mobile number of the missing individual. The system then requires them to document the lawful basis for processing this personal data. The investigator selects "Protection of Vital Interests" from the GDPR-aligned dropdown menu, citing the immediate risk to the individual's life and safety in the mandatory rationale field. This action is automatically timestamped and logged.
 2. *Phase 2 (Aggregation):* The investigator initiates the automated data aggregation. The tool's collection module begins its work. A reverse lookup identifies the registered owner's name and the mobile carrier. Simultaneously, the module queries integrated data breach services. A hit is found in a large credential leak from a popular forum, linking the phone number to a specific email address and a username: "Wanderer88".
 3. *Phase 3 (Analysis):* The tool's Analysis and Correlation Core automatically pivots on the newly discovered username. It searches for "Wanderer88" across integrated social media platforms and other public forums.²⁴ It discovers a public Twitter account and an account on a specialized hiking forum under the same handle. The on-premise AI/LLM layer is triggered to analyze the text of recent posts on the hiking forum. The LLM extracts mentions of a specific national park and a planned multi-day hiking route, information that would be difficult to find with simple keyword searches. The Visualization Dashboard presents a clear, interactive graph linking the initial Phone Number -> Owner's Name -> Username "Wanderer88" -> Twitter Account -> Hiking Forum Account -> Mentioned National Park and Route.
 4. *Phase 4 (Reporting):* The investigator, having validated the connections, generates a comprehensive report. The report includes the correlated data, screenshots of the relevant social media and forum posts, a map of the likely search area derived from the hiking route, and the complete audit trail. This report is immediately and securely disseminated to field search and rescue units. The entire process, from possessing only a single phone



number to generating actionable intelligence for a physical search, is conducted and fully documented within minutes, showcasing the tool's potential for speed and efficacy in a time-critical, ethical investigation.

B. Case Study 2: Cybersecurity - Threat Actor Profiling

- **Scenario:** A Security Operations Center (SOC) analyst is investigating a sophisticated phishing campaign targeting company executives. Forensic analysis of the phishing email reveals the command-and-control (C2) server's domain. A WHOIS lookup of the domain provides a contact phone number, which is likely a disposable or "burner" number used to obscure the threat actor's identity.
- **Workflow Demonstration:**
 1. *Phase 1 (Initiation):* The SOC analyst inputs the burner phone number into the tool, citing "Legitimate Interest - Prevention and Investigation of Cybercrime" as the lawful basis, and documents the details of the phishing campaign as justification.
 2. *Phase 2 (Aggregation):* The automated collection begins. The tool's technical lookup service immediately identifies the number as belonging to a VoIP provider, confirming the analyst's suspicion of its disposable nature.¹² As expected, queries to public records and social media platforms yield no results. However, a query against a database of pastebin leaks reveals a text file where the number was used to register an account on a specific cryptocurrency exchange.¹⁹
 3. *Phase 3 (Analysis):* The analyst now has a new pivot point: the cryptocurrency exchange account. While the account itself is pseudonymous, the pastebin leak also included a deposit wallet address associated with it. The analyst instructs the tool to pivot on this cryptocurrency wallet address. The tool queries public blockchain explorers and other OSINT sources that track illicit crypto activity.¹ It discovers that the same wallet address was used to make a payment on a prominent dark web forum specializing in the sale of malware and phishing kits.² The forum username associated with the transaction is "CyberGhost_X". The tool's entity graph now displays a clear chain of connections: Phishing Domain -> Burner Phone Number -> VoIP Service -> Crypto Exchange Account -> Cryptocurrency Wallet Address -> Dark Web Forum -> Threat Actor Username "CyberGhost_X".
 4. *Phase 4 (Reporting):* The analyst has successfully unmasked a key operational detail of the threat actor, linking the phishing infrastructure to a specific dark web persona. The generated report, containing all data points, sources, and the entity graph, is exported and added to the organization's threat intelligence platform (TIP). This enriches the organization's understanding of the adversary, allowing for better-informed defensive strategies and potential future attribution.¹⁸ This case demonstrates the tool's ability to navigate the more technical and anonymous realms of cybersecurity investigations, connecting disparate clues across the clear, deep, and dark web.

C. Limitations and Future Work

While the proposed framework offers a significant advancement over existing tools, it is not without limitations. The evolving tactics of privacy-conscious individuals and malicious actors present continuous challenges to OSINT methodologies. The increasing use of privacy-enhancing technologies, such as disposable VoIP numbers, SIM swapping attacks to take over legitimate numbers, and the widespread adoption of end-to-end encrypted messaging platforms, can create significant roadblocks for mobile number intelligence.²⁵ While the tool can effectively identify a number as a VoIP service, unmasking the true user behind it often remains an insurmountable challenge using OSINT alone.

Future work on this framework should be directed toward several key areas of development:

1. *Advanced Anonymity Detection:* Research and development should focus on creating more sophisticated machine learning models to detect and flag the use of anonymization services. This could involve analyzing patterns in number allocation, carrier data, and usage history to assign a "privacy score" to a given number, helping investigators prioritize leads.
2. *Cross-Lingual and Multimodal Analysis:* To be effective in global investigations, the on-premise LLM's capabilities must be expanded. Future versions should incorporate state-of-the-art models for performing effective analysis on data from multiple languages and modalities, including text, images, and video, which is a necessity for complex international cases.²⁹
3. *Secure Integration with Closed Sources:* A significant enhancement would be to develop a secure, privacy-preserving mechanism to allow organizations to fuse the tool's OSINT findings with their own internal, proprietary data. For example, a corporation could correlate OSINT on a threat actor with its internal network security logs, or a law enforcement agency could cross-reference findings with its confidential informant database. This fusion of open and closed sources would create a more holistic and powerful intelligence picture.⁴¹
4. *Proactive AI-Driven Threat Alerting:* The framework could evolve from a reactive, query-based tool into a proactive monitoring platform. Users could define a set of numbers or digital personas of interest, and the system would continuously monitor public sources for new information, generating real-time alerts based on predefined risk indicators or changes in online activity.



VI. CONCLUSION

This paper has advocated for a necessary paradigm shift in the design and deployment of Open-Source Intelligence tools, urging a move from a purely capability-driven approach to one that is fundamentally grounded in the principles of "Ethical-by-Design." We have presented a comprehensive framework for a mobile number intelligence tool that successfully integrates advanced data aggregation and AI-driven analysis with a structured workflow that enforces legal and ethical compliance as a core function. The proposed modular architecture, distinguished by its use of an on-premise LLM for secure analysis and a graph-based correlation core, is designed to provide investigators with powerful identity-stitching and pattern-recognition capabilities. Simultaneously, it is engineered to minimize operational security risks and ensure adherence to the complex tapestry of global data protection laws.

The conceptual case studies presented herein illustrate how such an integrated tool can dramatically accelerate investigations in critical domains like law enforcement and cybersecurity, effectively transforming a single, isolated data point—a mobile number—into a rich and actionable intelligence product. The true contribution of this work, however, is not a specific piece of technology but a methodological blueprint for the future of digital investigation. The central imperative for the next generation of investigative tools is the co-design of systems that treat ethical and legal guardrails not as optional constraints or afterthoughts, but as non-negotiable, core functional requirements. By embedding accountability, transparency, and a profound respect for privacy into the very code of our investigative tools, we can better equip ourselves to navigate the complex and often perilous terrain of the digital age, ensuring that the vital pursuit of security and justice does not come at the unacceptable cost of fundamental human rights.

ACKNOWLEDGMENT

The authors wish to thank the anonymous reviewers for their insightful feedback, which significantly improved the quality of this paper.

REFERENCES

- [1]. Open-Source Intelligence (OSINT) | Techniques & Tools | Imperva, accessed on October 27, 2025, <https://www.imperva.com/learn/application-security/open-source-intelligence-osint/>
- [2]. What is OSINT [Open-Source Intelligence]? Complete Guide ..., accessed on October 27, 2025, <https://shadowdragon.io/blog/what-is-osint/>
- [3]. (PDF) Open Source Intelligence Opportunities and Challenges: a Review - ResearchGate, accessed on October 27, 2025, https://www.researchgate.net/publication/381074245_Open_Source_Intelligence_Opportunities_and_Challenges_a_Review
- [4]. (PDF) The Art of Open Source Intelligence (OSINT): Addressing ..., accessed on October 27, 2025, https://www.researchgate.net/publication/392404120_The_Art_of_Open_Source_Intelligence_OSINT_Addressing_Cybercrime_Opportunities_and_Challenges
- [5]. Mobile Phone Data: A Survey of Techniques, Features, and ... - MDPI, accessed on October 27, 2025, <https://www.mdpi.com/1424-8220/23/2/908>
- [6]. OSINT Case Studies & Investigations | OSINT Industries, accessed on October 27, 2025, <https://www.osint.industries/case-studies>
- [7]. OSINT for Investigations - LexisNexis Risk Solutions, accessed on October 27, 2025, <https://risk.lexisnexis.com/insights-resources/article/osint-for-investigations>
- [8]. Open Source Intelligence Strategy - United States Department of State, accessed on October 27, 2025, <https://2021-2025.state.gov/open-source-intelligence-strategy/>
- [9]. Full article: Balancing National Security and Privacy: Examining the Use of Commercially Available Information in OSINT Practices, accessed on October 27, 2025, <https://www.tandfonline.com/doi/full/10.1080/08850607.2024.2387850>
- [10]. (PDF) Redefining OSINT Software Architecture with ... - ResearchGate, accessed on October 27, 2025, https://www.researchgate.net/publication/390998249_Redefining_OSINT_Software_Architecture_with_System-Centric_Architecture_Design_A_Framework_Shaped_by_QAW_ADD_and_ATAM
- [11]. lampyre.io, accessed on October 27, 2025, <https://lampyre.io/blog/osint-phone-number-investigations-a-comprehensive-guide/#:~:text=In%20short%2C%20phone%20numbers%20serve,about%20a%20person%20or%20entity.>
- [12]. OSINT Phone Number Investigations: A Comprehensive Guide ..., accessed on October 27, 2025, <https://lampyre.io/blog/osint-phone-number-investigations-a-comprehensive-guide/>



- [13]. Mobile Device Security: A Systematic Literature Review on Research Trends, Methods and Datasets - ResearchGate, accessed on October 27, 2025, https://www.researchgate.net/publication/361275356_Mobile_Device_Security_A_Systematic_Literature_Review_on_Research_Trends_Methods_and_Datasets
- [14]. The Ethical Considerations of OSINT: Privacy vs. Information ..., accessed on October 27, 2025, <https://medium.com/@scottbolen/the-ethical-considerations-of-osint-privacy-vs-information-gathering-63b5b2f76c55>
- [15]. Preserving Privacy: An Impact Framework for Open-Source ..., accessed on October 27, 2025, <https://www.newamerica.org/future-security/reports/preserving-privacy-an-impact-framework/exploring-the-intersection-of-osint-and-data-privacy-in-the-digital-world/>
- [16]. Open-source intelligence: a comprehensive review of the current ..., accessed on October 27, 2025, <https://pubmed.ncbi.nlm.nih.gov/37362900/>
- [17]. Data protection laws in India - Data Protection Laws of the World, accessed on October 27, 2025, <https://www.dlapiperdataprotection.com/?t=law&c=IN>
- [18]. Best OSINT Tools for Intelligence Gathering (2025) Free and Paid - ShadowDragon.io, accessed on October 27, 2025, <https://shadowdragon.io/blog/best-osint-tools/>
- [19]. jivoi/awesome-osint: :scream: A curated list of amazingly ... - GitHub, accessed on October 27, 2025, <https://github.com/jivoi/awesome-osint>
- [20]. Review OSINT tool for social engineering - PMC, accessed on October 27, 2025, <https://pubmed.ncbi.nlm.nih.gov/articles/PMC10504660/>
- [21]. 🔍 The Insider's Guide to Mastering OSINT Techniques for Phone Number Tracking 📞 | by Efim Lerner | Medium, accessed on October 27, 2025, <https://medium.com/@efim.lerner/the-insiders-guide-to-mastering-osint-techniques-for-phone-number-tracking-da61dd004c7c>
- [22]. OSINT Framework, accessed on October 27, 2025, <https://osintframework.com/>
- [23]. Comprehensive OSINT Resources Guide | by James Henning - Medium, accessed on October 27, 2025, <https://voodootomato.medium.com/comprehensive-osint-resources-guide-ee2b639969f7>
- [24]. A Practical OSINT Methodology — Tools, Notes, and Workflow | by ..., accessed on October 27, 2025, <https://medium.com/@pizzasteve/a-practical-osint-methodology-tools-notes-and-workflow-fbf027fde0bc>
- [25]. 4 Ways a Phone Number Exposes Your Private Data - Lifelines Neuro, accessed on October 27, 2025, <https://lifelinesneuro.com/phone-number-exposes-your-private-data/>
- [26]. Advancements in Open Source Intelligence (OSINT) Techniques and the role of Artificial Intelligence in Cyber Threat Intelligence (CTI), accessed on October 27, 2025, https://dione.lib.unipi.gr/xmlui/bitstream/handle/unipi/16306/Gioti_mte2205.pdf?sequence=3&isAllowed=y
- [27]. A systematic review on research utilising artificial intelligence for open source intelligence (OSINT) applications - ResearchGate, accessed on October 27, 2025, https://www.researchgate.net/publication/381194926_A_systematic_review_on_research_utilising_artificial_intelligence_for_open_source_intelligence_OSINT_applications
- [28]. Compliance in OSINT | Cyber | Proelium Law LLP, accessed on October 27, 2025, <https://proeliumlaw.com/open-source-intelligence-and-privacy/>
- [29]. OSINT Techniques: Complete List for Investigators (2025) ..., accessed on October 27, 2025, <https://shadowdragon.io/blog/osint-techniques/>
- [30]. Staying GDPR Compliant When Using OSINT for Fraud Prevention, accessed on October 27, 2025, <https://trustfull.com/articles/staying-gdpr-compliant-when-using-osint-for-fraud-prevention>
- [31]. OSINT and GDPR - OSINT Central, accessed on October 27, 2025, <https://osint-central.com/osint-gdpr/>
- [32]. Open Source Intelligence 2023-24 - Semi Automated ... - JETIR.org, accessed on October 27, 2025, <https://www.jetir.org/papers/JETIR2408276.pdf>
- [33]. Data Protection Laws and Regulations Report 2025 India - ICLG.com, accessed on October 27, 2025, <https://iclg.com/practice-areas/data-protection-laws-and-regulations/india>
- [34]. OSINT in India: Conducting OSINT Across India and the Subcontinent, accessed on October 27, 2025, <https://www.osint.industries/post/osint-in-india-conducting-osint-across-india-and-the-subcontinent>
- [35]. India Passes Long Awaited Privacy Law - WilmerHale, accessed on October 27, 2025, <https://www.wilmerhale.com/en/insights/blogs/wilmerhale-privacy-and-cybersecurity-law/20230818-india-passes-long-awaited-privacy-law>
- [36]. Internal Investigations Under the Digital Personal Data Protection Act, 2023, accessed on October 27, 2025, <https://disputeresolution.cyrilamarchandblogs.com/2025/01/internal-investigations-under-the-digital-personal-data-protection-act-2023/>
- [37]. Understanding The Nuances Of Pi Surveillance - Stillinger Investigations, Inc., accessed on October 27, 2025, <https://www.investigatesc.com/understanding-the-nuances-of-pi-surveillance/>



- [38]. How Private Investigators Conduct Surveillance Legally and Ethically, accessed on October 27, 2025, <https://investigations-nbi.com/surveillance-legally-ethically/>
- [39]. The Legal and Ethical Challenges of Mobile Data in Investigations - BlueForce Learning, accessed on October 27, 2025, <https://www.blueforcelearning.com/blog/the-legal-and-ethical-challenges-of-mobile-data-in-investigations>
- [40]. OSINT Tools And Techniques | OSINT Technical Sources - Neotas, accessed on October 27, 2025, <https://www.neotas.com/osint-tools-and-techniques/>
- [41]. Open-source intelligence: a comprehensive review of the current ..., accessed on October 27, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC10014398/>
- [42]. Create your own OSINT flowcharts - Aware Online Academy, accessed on October 27, 2025, <https://www.aware-online.com/en/create-your-own-osint-flowcharts/>
- [43]. Examples of Data Points Used In Profiling - Privacy International, accessed on