POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Argus: A Platform for Multi-source Cyber Threat Intelligence Collection and Analysis

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Attilio Polito, Biagio Marra

Advisor: Prof. Michele Carminati

Co-advisor: Lorenzo Bossi, Francesco Panebianco, Tommaso Paladini

Academic year: 2025-2026

1. Introduction

Cyber Threat Intelligence (CTI) supports security decision-making by turning sparse evidence (indicators, vulnerabilities, attacker discussions, and campaign reports) into actionable knowledge. In practice, CTI is fragmented across a heterogeneous ecosystem that includes high-frequency structured feeds, official vendor bulletins, long-form reports, high-velocity communities, and underground forums and dark leak sites.

Each channel is useful but incomplete. Feeds provide scale and precision on atomic indicators, while narrative sources provide semantics (actors, techniques, motivations). High-velocity and underground channels can surface early signals (proof-of-concept exploits, victim leaks, newly leaked credentials), but they are hard to monitor systematically and are affected by noise, duplication, and volatility.

This thesis addresses two practical gaps. First, an operational gap: most CTI tooling does not provide continuous, comparable, multi-source collection across both structured and unstructured ecosystems while preserving traceability to original evidence. Second, a measurement gap: organizations lack quantitative evidence about which source families provide unique intelligence, which are redundant, and which anticipate official disclosure cycles.

2. Related Works

Recent CTI research has advanced in three complementary directions, starting from engineering techniques that enable collection in adversarial settings (unstable services, anti-bot defenses, JavaScript-rendered websites), proceeding with NLP pipelines that turn unstructured text into machine-actionable intelligence via relevance filtering, Transformer-based NER, and increasingly CTI-oriented summarization; concluding with studies that assess the "value" of specific source families. However, most evaluations remain single-source and do not provide an integrated, repeatable framework for longitudinal, cross-source comparison.

The closest work to our goal is Schröer et al. [5], which comparatively analyzes first-hand cybercrime sources (forums, chats, darknet websites) and shows that underground CTI requires automated filtering and semantic processing due to extreme noise. Unlike our approach, their study relies on pre-collected data and mainly compares platforms in terms of how much of their content

is CTI-relevant versus noise, and whether the extracted intelligence is predominantly technical or strategic. Argus instead continuously collects from live sources and produces structured, traceable entities and normalized signals to support source-to-source measurements over time.

3. Goals and Contributions

The thesis proposes *Argus*, a platform designed to (i) continuously collect CTI from multiple source families, (ii) transform raw text into structured intelligence, and (iii) enable a rigorous comparative evaluation of the CTI ecosystem over time.

Argus provides a multi-source CTI platform with a separation between acquisition and intelligence transformation, enabling robust ingestion from heterogeneous sources and end-to-end traceability. It introduces a validated intelligence pipeline combining a *Relevance Gate* (binary classifier that discards non-CTI noise) with *hybrid extraction* (Named Entity Recognition plus deterministic parsers) and entity canonicalization. It produces a longitudinal dataset spanning from January 2023 to January 2026, collecting approximately 500,000 artifacts, enabling analysis of source evolution and the empirical role of each source family. Finally, it proposes a normalized evaluation framework that compares sources (Chapter 5) beyond volume using Information Density, Exclusivity, Overlap (raw and strategic-semantic), and Timeliness (First-to-Report and lead time), aligned with prior work in comparative CTI [2, 4].

4. Approach

4.1. Sources Selection

Argus adopts a portfolio strategy that covers different CTI layers (from indicators to higher-level behavior), consistent with the intuition captured by the "Pyramid of Pain" [1]. It ingests data from categories that are deliberately complementary, summarized in Table 1.

4.2. Argus Architecture

At a conceptual level, Argus follows an asynchronous ETL paradigm. It *extracts* content continuously and stores raw evidence with timestamps and provenance; it *transforms* data through intelligence services that decide relevance, extract entities, normalize them, and build higher-level relationships; and it *loads* structured intelligence into query-friendly storage for analytics and cross-source comparison. The key design decision is traceability: every extracted entity or relationship maintains a pointer to the original evidence, which is essential for analyst verification and contextual interpretation. The overall architecture is summarized in Figure 1.

Table 1: Argus source categories and typical intelligence produced.

Category	Typical output and role
Intelligence (Structured Feeds)	High-volume atomic indicators (IPs, URLs, hashes), strong tactical signal, low narrative context.
Bulletins (Official Advisories)	Validated CVEs, remediation guidance, attribution clues, high trust anchor for the disclosure cycle.
Reports (Long-form PDFs)	Rich semantics and relationships; contextualize known events; low novelty but high explanatory value.
Telegram and Discord	High-velocity dissemination; can behave as real-time bulletins and early leak amplifiers; variable noise.
Forums (Underground)	Discussions, PoCs, exploit trading; noisy but can anticipate key events.
DLS (Dark Leak Sites)	Structured victim leak announcements; episodic and campaign-driven signals.

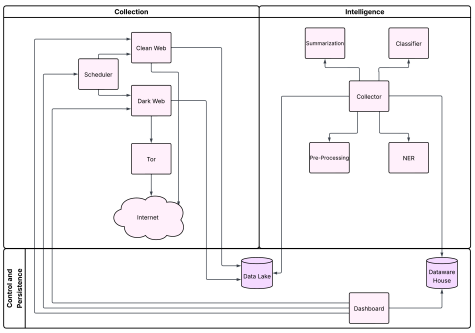

Figure 1: System Architecture Overview

4.3. Extraction

The *Extraction* layer is responsible for continuously acquiring heterogeneous CTI content from both structured and unstructured ecosystems. A scheduler orchestrates the crawling jobs and dispatches them to dedicated acquisition backends for *Clean Web* sources (surface web feeds, bulletins, reports, forums) and *Dark Web* sources, where a Tor-enabled crawler is used to

access onion services. Each collected item is persisted as raw evidence in a staging store (Data Lake) together with timestamps, source identifiers, and provenance metadata, enabling replayability and full traceability to the original evidence.

4.4. Transform

The *Transform* layer is orchestrated by a collector that retrieves new evidence from the Data Lake and acts as a router: items are dispatched to different processing tracks depending on their source family (e.g., bulletins, reports, structured feeds, forums, chats, dark-web), since each category has different noise levels, formats, and enrichment requirements.

Across these tracks, Argus applies a set of AI and NLP components. A *Pre-processing* to convert heterogeneous inputs into a model-ready text representation (e.g., HTML stripping, boilerplate removal, and normalization), ensuring consistent downstream behavior. The *Relevance Gate* then performs binary CTI detection to filter noisy unstructured content by discarding non-CTI items, using a Stacking Ensemble that combines a fine-tuned `deberta-v3-small` classifier with a TF-IDF + LinearSVC baseline through a Logistic Regression meta-learner; training and performance are reported in Section 5.4. For items that require semantic enrichment, a *Summarizer* condenses verbose narratives while preserving operational facts (IoCs, CVEs, actor/tool names), and a hybrid *NER* pipeline extracts cybersecurity entities from natural language by combining a neural Transformer tagger (`xlm-roberta-large`, fine-tuned with BIOES labels and a sliding-window strategy) with a deterministic Regex extractor for fixed-format IoCs; NER performance is reported in Section 5.5. Once an item completes its predefined track, the resulting structured records are persisted into a relational database.

4.5. Load

The *Load* layer provides *persistence and observability*: it consolidates the outputs of the pipeline into two complementary stores and exposes system status and results for analysis. Raw evidence is stored in the *Data Lake* as a pure backup of newly collected items, keeping the original content unchanged while enriching it with provenance metadata (source, timestamps, collection context, and processing identifiers) to ensure traceability and replayability. In parallel, the *Data Warehouse* stores the curated view of the same evidence, processed items, extracted entities, and relationships, so that analysts can run efficient longitudinal and cross-source queries without reprocessing raw data.

On top of these stores, a control plane offers an operator-facing interface that keeps track of the global crawler lifecycle, showing which jobs are scheduled, running, or completed and summarizing collection activity through basic counters such as throughput and number of collected items. The same interface also exposes the curated outputs of the pipeline together with the analysis views used in this work (e.g., density, exclusivity, overlap, and timeliness), allowing analysts to understand what is being collected and how the measured signals evolve over time.

4.6. Canonicalization and Semantic Relationships

To compare sources fairly, Argus normalizes entities to canonical forms. This is especially critical for Threat Actor aliases (e.g., different names used by vendors or communities). Canonicalization supports accurate Exclusivity and Overlap computations, aggregation of actor-related statistics, and semantic comparisons via higher-level links (e.g., APT–TTP relationships). This is achieved in two stages, starting from the preprocessing phase (through dictionary mapping), proceeding with querying the final database, considering MITRE ATT&CK aliases and similarity between single entities (as a whole and at the token level – considering two entities as the same, if the similarity score is above a certain threshold, depending on the entity type considered).

5. Experimental Evaluation

5.1. Dataset

The Argus platform was operated continuously to build a longitudinal corpus covering January 2023 to January 2026. The experimental campaign ingested and processed approximately 500,000 artifacts, with the ingestion pipeline running continuously and no manual filtering applied before persistence, so that the dataset

reflects the real noise level of the monitored environments.

The dataset is stratified into seven primary source categories: Structured Feeds, Cybersecurity Bulletins, Vendor/Institutional Reports, Forums, Telegram, Discord, and Ransomware Data Leak Sites. In terms of monitored endpoints, Argus integrates 7 structured intelligence feeds, monitors 9 Telegram channels and 29 Discord channels across selected CTI-oriented servers, and ingests underground discussions from Dread (i.e., a dark web forum) plus multiple Surface-Web forums. For ransomware leak sites, Argus relies on Ransomware.live as a dynamic directory and ingests only those flagged as active at collection time (typically around 120 entries).

5.2. Research Questions

The evaluation is structured around four research questions: RQ1 quantifies volume versus concentrated signal using normalized richness and density metrics; RQ2 measures exclusivity and redundancy through exclusivity and cross-source overlap; RQ3 assesses timeliness and anticipation via first-seen timestamps and First-to-Report analyzes; and RQ4 studies how the role of source families evolves over time across the longitudinal window.

5.3. Evaluation Metrics

Raw volume is misleading because sources differ in verbosity and structure. Argus therefore uses normalized metrics: entities per item (richness proxy), information density (unique entities per 1,000 words), exclusivity (entities found only in one category), overlap (Jaccard similarity), semantic overlap (shared APT–TTP links), and timeliness (First-to-Report ranking based on global first-seen timestamps).

5.4. Relevance Gate Validation

The Relevance Gate decides whether a collected item is CTI-relevant or just noise. We tested two ways of building the training set: (i) *weak labels*, automatically obtained at a large scale but noisy, and (ii) a small *gold* set, manually labeled. Because most collected content is non-CTI, the task is strongly imbalanced, so we select models by prioritizing PR-AUC and especially recall, to avoid missing relevant threats (false negatives). The best setup uses 600 gold labels and a stacking ensemble, achieving PR-AUC 0.78, Recall 0.87, and F1 0.77. In contrast, models trained only on weak supervision perform worse (PR-AUC 0.66; Recall 0.57).

5.5. NER Validation

Named Entity Recognition (NER) performance is measured on the held-out test set of the adopted CTI dataset [3]. The neural NER component used in the pipeline achieves 79.60% F1-score (Precision 77.11%, Recall 79.74%) with overall Accuracy 91.99%.

A per-entity analysis motivates the hybrid design: semantic entities (e.g., threat actors and techniques) are handled effectively by the neural model, while strict indicators may degrade due to subword tokenization and formatting variability. Deterministic parsers act as a safety net for rigid IoCs (e.g., hashes, IPs, URLs, CVEs), improving reliability.

6. Main Results

The following section summarizes the key findings from the 2023–2026 campaign, reporting the most relevant results for each research question, followed by the MOVEit case study.

6.1. R1: Volume, Richness, and Density

The 2025 window is analyzed as the most mature phase of the campaign. The ecosystem is strongly dominated by structured feeds: the Intelligence category contributes 881,593 entities, consistent with high-frequency OSINT streams. However, entities per item highlight that Telegram and Bulletins converge as high-density dissemination vectors (26.96 vs. 24.61 entities/item), while Forums are lower (11.60) due to conversational dilution. A deterministic leak-site constraint is observed for DLS (exactly 3.00 entities/item) because the ingestion logic extracts a fixed set of structured fields per victim announcement. Finally, information density strongly separates "pure signal" feeds from narrative sources (Intelligence 502.65 vs. Reports 27.56 and Forums 17.92 entities/1K words). Table 2 summarizes the key normalized metrics.

These measurements demonstrate a core finding: volume does not equal value. Feeds dominate volume and density, but provide limited strategic context; human-generated and high-velocity

Table 2: Key normalized metrics in 2025.

Metric (2025)	Value
Total entities (Intelligence)	881,593
Avg. entities/item (Telegram)	26.96
Avg. entities/item (Bulletins)	24.61
Avg. entities/item (Forums)	11.60
Avg. entities/item (DLS)	3.00
Information density (Intelligence)	502.65
Information density (Reports)	27.56
Information density (Forums)	17.92

Table 3: Exclusivity evolution.

Category	2024 exclusive	2025 exclusive
Bulletins	25,978	26,175
Telegram	11,851	29,825
Reports	195	126

sources provide fewer entities overall but often richer, more structured artifacts per item.

6.2. R2: Exclusivity and Overlap

Exclusivity reveals which sources contribute intelligence that is not available elsewhere. In 2025, Telegram provides 29,825 exclusive entities, surpassing Bulletins (26,175), indicating that Telegram has evolved into a primary generator of unique intelligence rather than a mere repost channel. Reports show negligible exclusivity (126 in 2025), confirming their role as contextualizers. In 2024, the hierarchy is different (Bulletins 25,978 vs. Telegram 11,851), showing that Telegram's role as a primary source is an evolving phenomenon. Table 3 summarizes the main results.

Overlap analysis quantifies redundancy and corroboration. In 2025, the Bulletin–Telegram pair is the central axis of the ecosystem with 16.87% Jaccard similarity and 11,064 shared entities, indicating strong synchronization between official advisories and high-velocity channels; the same axis is already visible in 2023 (14.16%). Table 4 reports the comparison.

Beyond raw indicators, Argus evaluates semantic overlap via shared APT–TTP links. Shared links between Bulletins and Telegram increase from 347/338 (2023/2024) to 1,038 in 2025, showing that Telegram increasingly mirrors (and sometimes contributes to) strategic attribution narratives.

Table 4: Raw overlap of Bulletins and Telegram.

Year	Jaccard similarity	Shared entities
2023	14.16%	6,454
2025	16.87%	11,064

Table 5: Shared APT–TTP links between Bulletins and Telegram.

Year	Shared APT–TTP links
2023	347
2024	338
2025	1,038

6.3. R3: Timeliness

Timeliness is measured by computing, for each normalized entity, the global first-seen timestamp across all sources, then counting how often each category is the earliest observer (First-to-Report). Bulletins dominate the standard disclosure cycle, growing from 7,189 (2023) to 8,286 (2024) and 9,663 (2025). Telegram shows a strong upward trend in anticipation capability, from 3,011 (2023) to 2,735 (2024) and 6,342 (2025), confirming its evolution into a high-impact early-warning and dissemination channel (Table 6).

6.4. R4: Longitudinal Evolution

The CTI ecosystem is not static. Telegram shifts from a secondary repost channel (2024) to a major generator of unique intelligence (2025). The strategic coupling between official and high-velocity channels strengthens over time, as measured by shared semantic links. Some channels exhibit episodic value (e.g., leak sites), reinforcing the need for continuous monitoring rather than fixed assumptions. The implication is that CTI effectiveness depends on continuous measurement and adaptive portfolios: new platforms emerge, old ones decay, and a robust CTI program must be able to onboard and evaluate sources over time.

7. Case Study: MOVEit Transfer campaign

To validate the effectiveness of the Platform in a concrete scenario, the thesis analyzes the MOVEit Transfer campaign. This widely exploited SQL injection vulnerability in Progress MOVEit Transfer enabled unauthorized data

Table 6: First-to-Report evolution.

Category	2023	2024	2025
Bulletins	7,189	8,286	9,663
Telegram	3,011	2,735	6,342

Table 7: MOVEit timeline reconstructed from multi-source timestamps

Phase	Date	Source	Event
Pre-Exploitation	24/04/2023	Forum	Release of SQLi PoC
Day of Attack	27/05/2023	Intelligence	Start of mass exploitation
Disclosure	31/05/2023	Bulletin	Vendor advisory
Official Alert	02/06/2023	Bulletin	CISA KEV update
Peak Leak	June 2023	Telegram	Mass posting of victims

access and triggered large-scale victim disclosures. We reconstruct the timeline by correlating first-seen timestamps across source categories. The dataset reveals a critical pre-disclosure window: a forum PoC is detected on 24/04/2023; the first official bulletins appear around 31/05/2023; CISA adds the vulnerability to KEV on 02/06/2023, resulting in a 37-day gap between underground PoC circulation and institutional alerting. Table 7 summarizes the reconstructed timeline.

This case study illustrates the operational value of a multi-source approach: official bulletins remain essential for validated remediation, but underground and high-velocity streams can reduce blind spots and shorten time-to-insight.

8. Conclusions

This thesis shows that modern CTI should not be approached as "collect as much as possible": raw volume can be misleading and often reflects low-context tactical streams. Argus demonstrates a rigorous alternative by combining continuous multi-source collection with traceable, structured intelligence extraction and by providing a longitudinal measurement of the ecosystem with normalized metrics. The evidence suggests that official bulletins anchor trust and remediation, while high-velocity and underground ecosystems can materially reduce time-to-insight for specific high-impact cases.

In summary, Argus provides both a practical CTI monitoring instrument and an empirical framework to answer a critical question for security operations: *which sources are worth monitoring, and why, now?*

References

[1] D. Bianco. The pyramid of pain, 2014. URL `http://detect-respond.blogspot.com/2013/03/the-pyramid-of-pain.html`. Conceptual model for CTI indicators.

[2] X. Liao, K. Yuan, X. Wang, Z. Li, L. Xing, and R. Beyah. Acing the ioc game: Toward automatic discovery and analysis of open-source cyber threat intelligence. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 755–766, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341394. doi: 10.1145/2976749.2978315. URL `https://doi.org/10.1145/2976749.2978315`.

[3] I. Mouiche and S. Saad. Ti-nermergerv2: Automating the integration of threat intelligence ner datasets via stix standard. *Computers & Security*, 163:104838, 2026. ISSN 0167-4048. doi: https://doi.org/10.1016/j.cose.2026.104838. URL `https://www.sciencedirect.com/science/article/pii/S0167404826000143`.

[4] T. Paladini, L. Ferro, M. Polino, S. Zanero, and M. Carminati. You might have known it earlier: Analyzing the role of underground forums in threat intelligence. In *Proceedings of the 27th International Symposium on Research in Attacks, Intrusions and Defenses (RAID)*, Padua, Italy, 2024. ACM. doi: 10.1145/3678890.3678930.

[5] S. L. Schröer, N. Canevascini, I. Pekaric, P. Widmer, and P. Laskov. The dark side of the web: Towards understanding various data sources in cyber threat intelligence. In *2025 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 79–89, 2025. doi: 10.1109/EuroSPW67616.2025.00015.