CIRCL
Computer Incident
Response Center
Luxembourg

GCVE.eu

# What's Happening Around Vulnerability-Lookup

## **EPSS SIG** - Collaboration between Vulnerability-Lookup, VLAI, and GCVE

https://vulnerability-lookup.org

team@circl.lu

January 16, 2026

CIRCL https://www.circl.lu

# Contents

1. Why we build these tools
2. Vulnerability-Lookup
    1. Who is behind Vulnerability-Lookup?
    2. Origin and Challenges we aim to address
    3. Where the Data Comes From
3. AI strategy
    1. Mission
    2. Open Data Approach
    3. AI Datasets
    4. Generation workflow
4. Global CVE Allocation System

# Why we build these tools

# Why We Build These Tools

- Enabling better vulnerability metrics & research
- A shared challenge: improving **vulnerability risk metrics & prioritization** requires **real, reproducible, openly accessible data**.
- Metrics like **CVSS, EPSS, vendor scores, exploitability signals** only improve when results can be **validated across organizations**.

# Our Approach: An Open, End-to-End Research Stack

- **Vulnerability-Lookup:** normalization, enrichment, deduplication; research-friendly **APIs & dumps**.
- **GCVE & DB.GCVE.EU:** decentralized identifiers & interoperable publishing at scale.[1]
- **VLAI:** extraction, structuring, and comparison of textual & contextual signals.
- **Sightings:** real-world exposure & observation data as **time-series signals**.
- **Crowdsourced + open data pipelines:** broader coverage, reduced single-source bias.

[1] https://gcve.eu

TLP:CLEAR

# What This Enables for the Community

- **Transparent datasets** for benchmarking scoring models & prioritization strategies.
- **Faster iteration** on new metrics, and honest evaluation of existing ones.
- **Reproducible research:** same inputs → comparable outputs → verifiable conclusions.

TLP:CLEAR

# Vulnerability-Lookup

# Who is behind Vulnerability-Lookup?

Vulnerability-Lookup[2] is an **Open Source** project led by **CIRCL**.
It is co-funded by **CIRCL** and the **European Union**[3].
Used by many organisations including CSIRTs and ENISA (EUVD).

[2] https://www.vulnerability-lookup.org

[3] https://github.com/ngsoti

# What is Vulnerability-Lookup?

- A unified place to **search**, **triage**, and **track** software and product vulnerabilities from different sources.
- Brings together vulnerability information, correlation of identifiers (e.g., CVE, GHSA, OSV), references, timelines, and risk signals in one view.
- Designed for **CSIRTs**, **SOCs**, **vulnerability managers**, and **developers** in mind.
- **Web UI first** with strong **API** and automation options.
- Support a complete **CVD process management**[4] along with the ability to fork vulnerability information[5]. Distributed GNA directory.

[4] CNA Program, GCVE GNA Publication

[5] Implemented before the GNA initiative!

TLP:CLEAR

# Origin of Vulnerability-Lookup

- `cve-search`[6] is an open-source tool initially developed in late 2012, focusing on maintaining a **local** CVE database.
- `cve-search` is widely used as an **internal** tool.
- The design and scalability of cve-search are limited. Our operational public instance has reached a hard limit of 20,000 queries per second.
- Vulnerability sources have **diversified**, and the **NVD CVE is no longer the sole source** of vulnerability information.

[6] `https://github.com/cve-search/cve-search`

TLP:CLEAR

# Initial Challenges

- **Volume of data:** Handling a substantial dataset and heavy network traffic, currently over **1,802,699** security advisories and more than **183,000** sightings collected in approximately a year [7].
- **Flexibility:** Balancing ongoing development with legacy issues while designing a future-proof architecture. It's complex and yes, sometimes chaotic[8].
- **Robustness:** Validating data even when external entities don't comply with their own JSON schemas. It's not always pretty.
- **Fast lookup:** Rapidly correlating identifiers across **diverse sources**, including unpublished advisories.

[7] The first sighting on Exploit-DB dates back 26 years.
[8] We enjoy challenges, especially when they lead to practical solutions.

TLP:CLEAR

# Ongoing Challenges and Development

- **CVD process:** Developing an open-source tool to support Coordinated Vulnerability Disclosure. [9]
- **Vulnerability numbering:** Enabling a distributed approach via the Global CVE Allocation System. [10]
- **Scoring vulnerabilities:** Aggregating large volumes of observations from diverse advisory types to improve scoring. Automatic **weighting** of sightings by source or type (e.g., exploited, PoC available) and **detection of type** independent of the source.
- **Data quality:** Transforming messy data into clean datasets suitable for model training and analysis.

[9] Aligned with NIS 2 and the Cyber Resilience Act.

[10] `https://gcve.eu`

TLP:CLEAR

# Current Sources

- **CVE Program** Git
- **NIST NVD CVE** API 2.0
- **Fraunhofer FKIE CVE** Git
- **GCVE** API

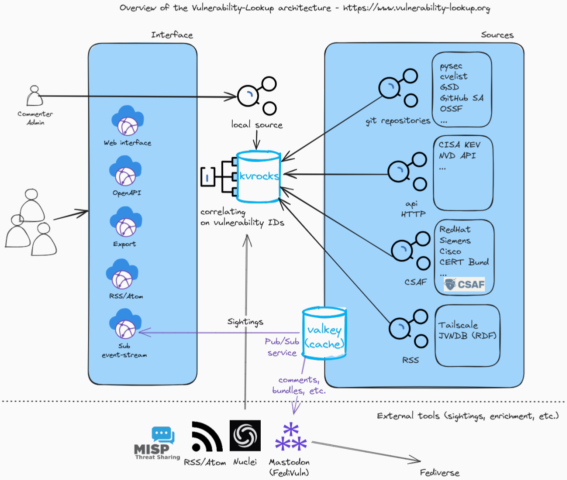
  GNA already publishing
- **GitHub Advisory Database** Git
- **Python Advisory Database** Git
- **Cloud Security Alliance - GSD** Git
- **VARIoT** API

- **CSAF 2.0** (HTTP CSAF)

  CERT-Bund, Cisco, Siemens, Red Hat, Suse, OpenSuse, Microsoft, NCSC-NL, CISA, ABB, etc.
- **Japan - JVN DB** HTTP
- **China - CNVD** HTTP
- **CERT-FR** HTTP
- **Tailscale** RSS
- **CISA KEV** HTTP
- **EU KEV** HTTP
- **Growing...**

# Vulnerability-Lookup High-Level Architecture

- Collects and aggregates vulnerability data.
- Preserves data integrity (never alters the original content).
- Harmonizes data using kvrocks.
- Correlates information across multiple sources (CVE, PySec, etc.).
- Provides both APIs and a Web interface.
- Supports advanced queries and filtering.



TLP:CLEAR

# AI strategy

# AI strategy

- CIRCL AI approach[11]: we enhance existing solutions rather than replacing functional systems with NLP/ML/LLM solutions.
- We actively participate in collaborative research and development efforts, such as the EU-funded AIPITCH (AI-Powered Innovative Toolkit for Cybersecurity Hubs) project[12]
- Improve operational outcomes in threat intelligence and incident response
- AI-powered enrichment of vulnerability descriptions (e.g., severity, CWE, CPE information)
- VLAI: A RoBERTa-Based Model for Automated Vulnerability Severity Classification[13]

[11] https://circl.lu/pub/ai-strategy
[12] https://www.science.nask.pl/en/research-areas/projects/12456
[13] https://arxiv.org/abs/2507.03607

TLP:CLEAR

# Why We Share Datasets

- **Open Data Initiative**: CIRCL's commitment to making data openly available[14].
- Consistent open approach applied across all our projects.
- Regularly updated JSON dumps [15] and "AI" datasets [16].
- Public, unauthenticated API access for Vulnerability-Lookup.

[14] https://data.public.lu/en/organizations/computer-incident-response-center-luxembourg

[15] https://vulnerability.circl.lu/dumps/

[16] https://huggingface.co/CIRCL/datasets

# Building AI Datasets

- Our experience with large datasets is not recent (Passive DNS[17], BGP ranking[18], MISP[19], AIL[20], Lookyloo[21], etc.). And we learned from our past mistakes.
- Turn messy data into structured, actionable insights.
- Link related vulnerabilities via enrichment, correlation, and crawling.
- Support the process with **VulnTrain**[22]: a tool to build AI Datasets and Models for vulnerability management.

[17] https://www.circl.lu/services/passive-dns/
[18] https://github.com/D4-project/BGP-Ranking
[19] https://github.com/MISP
[20] https://github.com/ail-project
[21] https://github.com/Lookyloo
[22] https://github.com/vulnerability-lookup/VulnTrain

# Current datasets

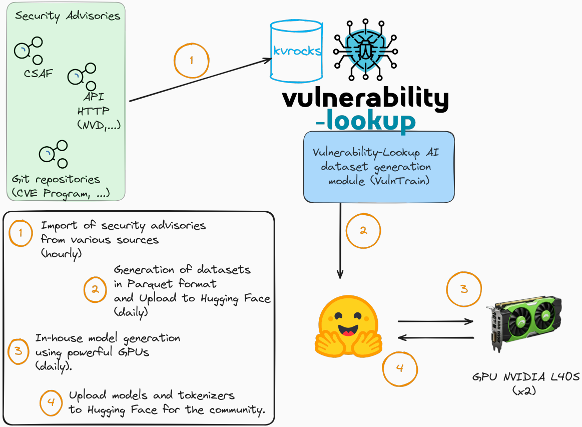
| Dataset | Size (rows) | Generation Time | Features |
|---|---|---|---|
| vulnerability-scores[23] | 655,258 | 10m45s | Descriptions (en), CVSS, CPE |
| vulnerability-CNVD[24] | 126,127 | 1m32s | Descriptions (cn), CVSS |
| vulnerability-cwe-patch[25] | 39,260 | 350m | Descriptions (en), CWE, patches (commit id + url + full diff) |

[23] https://huggingface.co/datasets/CIRCL/vulnerability-scores

[24] https://huggingface.co/datasets/CIRCL/Vulnerability-CNVD

[25] https://huggingface.co/datasets/CIRCL/vulnerability-cwe-patch

# Model generation workflow with VulnTrain



Security Advisories

CSAF

API HTTP (NVD,...)

Git repositories (CVE Program, ...)

1

kvrocks

vulnerability -lookup

Vulnerability-Lookup AI dataset generation module (VulnTrain)

2

3

4

GPU NVIDIA L40S (x2)

1 Import of security advisories from various sources (hourly)

2 Generation of datasets in Parquet format and Upload to Hugging Face (daily)

3 In-house model generation using powerful GPUs (daily).

4 Upload models and tokenizers to Hugging Face for the community.

- local training
- models are publicly shared
- regular update

TLP:CLEAR

# Current Models

| Model | Size | Epochs | Accuracy | Training Time |
|---|---|---|---|---|
| Severity classification[26] | 125M params | 5 | 0.8289 | 3.41h |
| Severity classification (CNVD)[27] | 102M params | 5 | 0.7817 | 34.741m |
| CWE guessing[28] | 125M params | 36-40 | 0.7127 | 60m |

With 4 × NVIDIA L40S.

GPU Efficiency in VLAI Model Training:
`https://www.vulnerability-lookup.org/2025/12/12/gpu-efficiency-in-vlai-model-training`

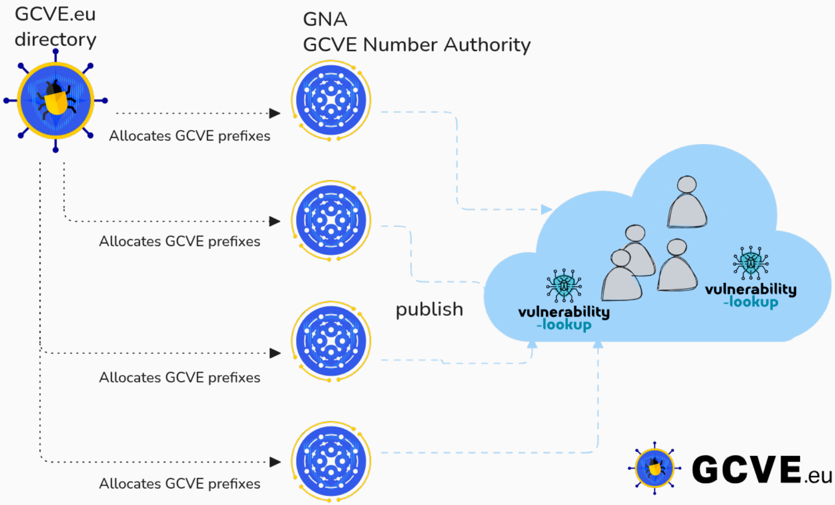
[26] https://huggingface.co/CIRCL/vulnerability-severity-classification-roberta-base
[27] https://huggingface.co/CIRCL/vulnerability-severity-classification-chinese-macbert-base
[28] https://huggingface.co/CIRCL/cwe-parent-vulnerability-classification-roberta-base

# Global CVE Allocation System

# Role

- The primary role of GCVE is to provide **globally unique identifiers** to GCVE Numbering Authorities (GNAs).
- **GNAs operate autonomously**, with full control over how they assign and manage identifiers.
- **GCVE publishes Best Current Practices (BCPs)** on directory management, Coordinated Vulnerability Disclosure (CVD), and publication protocols.
- GCVE maintains and publishes the **official directory of all GNAs**, including their publication endpoints.

# Overview



GCVE.eu directory

GNA
GCVE Number Authority

Allocates GCVE prefixes

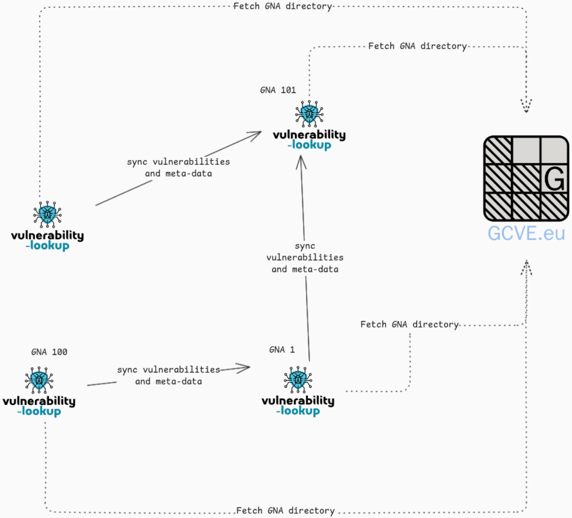
Allocates GCVE prefixes

Allocates GCVE prefixes

Allocates GCVE prefixes

publish

# A Distributed Network



TLP:CLEAR

# Who Can Be a GNA?

- You are an existing CNA recognized by the CVE Program.
- You are not a CNA, but **meet at least one of the following conditions**[29]:
  - You are a registered CSIRT or CERT listed on FIRST.org, part of the EU CSIRTs Network, or a member of TF-CSIRT.
  - You are a software, hardware, or service provider that regularly discloses vulnerabilities affecting your own products or services, and you have an official CPE vendor name assigned.
  - You have a public vulnerability disclosure policy and maintain a publicly accessible source for newly disclosed vulnerabilities.

[29] `https://gcve.eu/about/#eligibility-and-process-to-obtain-a-gna-id`

# Benefits of Becoming a GNA

- **Fast, straightforward onboarding**: as soon as the eligibility criteria are met, registration is quick and simple.
- **Flexible identifier usage**: you may publish new CVE entries immediately and apply your assigned prefix to both current and historical identifiers.
- **Autonomy over publication**: each GNA determines for itself what constitutes a vulnerability and what information is made public.
- **Incremental adoption of BCPs**: additional GCVE Best Current Practices can be adopted over time; implementing every BCP is encouraged but not mandatory.

# Best Current Practices (BCPs)

| ID | Title | Status | Ver. | Published |
|---|---|---|---|---|
| GCVE-BCP-01 | Signature Verification of the Directory File | **PUB** | 1.1 | 25 Apr 2025 |
| GCVE-BCP-02 | Practical Guide to Vulnerability Handling and Disclosure | **PUB** | 1.3 | 09 Dec 2025 |
| GCVE-BCP-03 | Decentralized Publication Standard | *Draft* | 1.0 | 10 Jun 2025 |
| GCVE-BCP-04 | Recommendations and Best Practices for ID Allocation | **PUB** | 1.3 | 02 Oct 2025 |
| GCVE-BCP-05 | GCVE Vulnerability Format (Updated CVE Record Format) | *Draft* | 1.6 | 02 Jan 2026 |
| GCVE-BCP-07 | Known Exploited Vulnerability - KEV Assertion Format | *Draft* | 1.0 | 03 Jan 2026 |

*Draft and public documents are always open for public review.*

https://gcve.eu/bcp

TLP:CLEAR

# Closing

# Ongoing Development

- Deeper analysis of the content and context surrounding sightings and exploited vulnerabilities[30].
- Insights generation based on patches dataset.
- CPE Guessing across sources.
- Full-text search capabilities across all sources.
- Synchronization between multiple Vulnerability-Lookup instances allow **real-time sightings accross instances**.

> The project is evolving rapidly — we always welcome feedback[31] and feature suggestions!

[30] https://github.com/vulnerability-lookup/TARDISsight

[31] https://discourse.ossbase.org/c/gcve/14 - https://discourse.ossbase.org/c/vulnerability-lookup/6

TLP:CLEAR

# References

- 🏠 https://www.vulnerability-lookup.org
- 📕 https://vulnerability.circl.lu https://db.gcve.eu/
- https://github.com/vulnerability-lookup/vulnerability-lookup
- https://social.circl.lu/@circl

TLP:CLEAR

# Thank you for your attention

- Issues, new sources of advisories or ideas: `https://github.com/vulnerability-lookup/vulnerability-lookup`
- For support and questions, contact: info@circl.lu

TLP:CLEAR