# Ask the Right Queries: Improving Search Engine Retrieval of Vulnerable Internet-Connected Devices Through Interactive Query Reformulation

Andrea Bernardini[1,*,†], Claudio Carpineto[1,†], Simone Angelini[1,†], Giovanna Dondossola[2,†] and Roberta Terruggia[2,†]

[1] *Fondazione Ugo Bordoni, Viale del Policlinico, 147, 00161, Rome, Italy*
[2] *RSE S.p.A., Via Raffaele Rubattino, 54, 20134, Milan, Italy*

**Abstract**

An IoT search engine collects and indexes a plethora of information associated with individual devices exposed on the internet, which theoretically can be combined with analogous information present in vulnerability databases to attempt to discover the presence of certain types of devices exhibiting known vulnerabilities. However, in practice, this is a challenging task. Indeed, the difficulty of handling and cross-referencing often incomplete or erroneous textual descriptions typically results in many false positives and false negatives in the obtained results, undermining the usefulness of such systems. This paper focuses on refining the query formulation to maximize retrieval effectiveness. The proposed interactive methodology relies on leveraging various security-related OSINT tools and data to refine queries based on insights gained from initial results, thus yielding new relevant findings. In a case study concerning photovoltaic generation monitoring systems, it is demonstrated that employing the proposed methodology allows for the non-intrusive identification of numerous internet-connected devices hosting such services, which can plausibly be exploited to carry out cyber-attacks against energy communities or renewable generation plants.

**Keywords**

IoT Search Engine (IoTSE), internet-connected devices, vulnerabilities, OSINT tools, query reformulation

## 1. Introduction

In our increasingly interconnected era, internet-connected devices (ICDs) have fundamentally transformed how we interact with the surrounding world, enabling remote control and automation of a wide array of both household and industrial devices.

Despite the numerous benefits offered by ICDs, significant challenges emerge, particularly concerning security and privacy. Among these, a primary concern revolves around vulnerabilities present in ICDs which can be exploited by malicious individuals for harmful purposes [1].

* Corresponding author.
† These authors contributed equally.
abernardini@fub.it (A. Bernardini); ccarpineto@fub.it (C. Carpineto); sangelini@fub.it (S. Angelini); giovanna.dondossola@rse-web.it (G. Dondossola); roberta.terruggia@rse-web.it (R. Terruggia)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Firstly, the increased connectivity often leads to inadvertent exposure due e.g. to keeping default password and network configuration [2].

Additionally, the widespread reuse of hardware/software components, employed to keep production costs low for ICDs, can facilitate vulnerability propagation [3]. Once a vulnerability is identified in one service, there exists a tangible risk of finding it in other services exposed on internet by devices produced by the same vendor. Lastly, the diverse range of ICDs from various manufacturers entails a diversity of security standards and communication protocols, heightening the risk of exposure to cyber threats.

It is therefore crucial to develop methodologies for swiftly identifying installations of exposed devices as soon as new vulnerabilities are discovered. In particular, the integration of IoT search engines [4] [5] [6], etc. with vulnerability databases [7] [8] appears to hold potential for identifying internet-exposed services and devices of interest to users that exhibit known vulnerabilities, starting from simple keyword searches.

However, a series of conceptual and practical issues renders such an approach highly unsatisfactory, as the identified hosts often pertain to different services, while many vulnerable hosts of the sought-after type remain unidentified. In information retrieval terms, the results yield a high number of false positives (low precision) and false negatives (low recall).

To address this situation, we propose focusing on the quality of input queries rather than the intelligence of indexing, matching, and ranking algorithms. The proposed solution draws inspiration from methodologies for query reformulation or expansion used in web search engines, wherein queries are refined using highly specific information extracted from web pages obtained in response to generic queries as in [9] and [10].

In the case of IoT search engines, the challenge lies in identifying fingerprints of hosts with specific vulnerable services and subsequently utilizing these fingerprints as search keys to retrieve other hosts of the same type.

The proposed methodology integrates several publicly available search tools and databases. The concept is to employ company specific information from vulnerability databases as an additional input for IoT search engines. Subsequently, once the presence of a vulnerable service of the sought-after type is detected, its fingerprints are extracted, and queries are reformulated using advanced search features.

Using this methodology in an energy-related case study, we demonstrate that it is possible to identify many photovoltaic (PV) production monitoring systems accessible on the internet that indeed appear vulnerable to known Common Vulnerabilities and Exposure (CVE), well beyond those retrievable through more elementary queries.

Solar photovoltaics (PV) have achieved widespread adoption across numerous countries and regions globally [11], primarily due to their already established status as the most cost-effective power source. Furthermore, within the European Union [12], PV technology stands as a cornerstone of the transition towards achieving a zero-carbon energy supply by 2050, primarily attributable to its remarkably low carbon dioxide ($CO_2$) footprint. However, the observed rise in reported cyber-attacks targeting PV systems highlights a growing and alarming issue [13].

The main contributions of this article are as follows:

- The proposal to focus on constructing better queries in the process of searching for vulnerable hosts, rather than treating input queries as some sort of independent variable and concentrating solely on their post-processing.
- A methodology that assists the user in the query reformulation process, primarily based on the extraction of fingerprints of vulnerable hosts and their utilization as advanced search features.
- An experimental study of the effectiveness of the methodology in a critical sector that has not been explored in conjunction with IoT search engines thus far, namely the PV production monitoring systems domain, resulting in the retrieval of a multitude of hosts hosting presumably vulnerable systems.

The rest of the article is structured as follows: Section 2 discusses related work on the topic of vulnerability search using IoT search engines. Section 3 analyzes the limitations associated with the use of IoT search engines for vulnerability discovery and identifies the reasoned use of advanced search language as a possible remedy. Section 4 describes in detail the methodology supporting the user in the interactive formulation of effective queries, which particularly utilizes vulnerable hosts' fingerprints as advanced search features. Section 5 covers the experimental analysis conducted on photovoltaic production monitoring systems: the usage scenario, the process of retrieving the hosts of interest, and the verification of their vulnerability are described. Finally, Section 6 provides some conclusions.

## 2. Related work

There are a few recent works dealing with the utilization of IoT search engines for the identification of vulnerable exposed devices. These works are characterized by a variety of approaches concerning data sources, search tools employed, types of devices and vulnerabilities addressed, and potential validation thereof.

One approach involves integrating IoT search engines with more traditional vulnerability analysis tools, as seen in [14] and [15], where the combined use of Shodan and Nessus is proposed for medical devices, or in [16] and [17], where Shodan is employed alongside the CVE database, or even in [18], where Shodan is utilized in conjunction with Octave Allegro. In [19], the authors introduced an approach termed Banner-CPE-CVE, where information from banners obtained via Shodan or Google dorking are associated with data contained in corresponding Common Platform Enumeration (CPEs) and CVEs using regular expressions applied to banner content. The authors also sought to validate their approach by directly querying the identified IPs and services and calculating precision and recall measures. In [20], the focus is on N-days vulnerabilities utilizing various IoT search engines to obtain data through a semi-automatic multi-phase process, where manual refinement of initial queries is accompanied by result validation based on manual checks and the use of a customized Nmap scanner with device fingerprinting.

Device fingerprinting is a well-known technique, although generally applied to enhance the retrieval process of exposed devices regardless of their vulnerabilities. Classic fingerprints include the protocol [21] [22], TCP/IP header [23], or elements of network traffic [24], as well

as groups of fingerprints like-the triplets <device type, vendor, product> [25], and <device type, version, port> [26], often inferred from communication packets using deep neural network (DNN) techniques.

Keyword search refinement is also pursued in [27], specifically referring to "firmware version" and mapping it with data derived from vendor homepages, and in [28], where the authors conduct an exploratory research via a full-text search of common terms related to smart grid, industrial control system (ICS), and ICD devices on Shodan, subsequently identifying a field search on the HTML title field.

In contrast to this variety of approaches, our study is essentially characterized by the attempt to utilize query results as the primary source of information guiding the reformulation process and to make more systematic use of the Advanced Search Language provided by IoT search engines.

Additionally, the application domain of PV production monitoring systems had not yet been explored in connection with IoT search engines (to the best of our knowledge), and non-intrusive verification of actual exposure to attacks is an aspect generally lacking in previous studies.

Another line of research connected to this work is query reformulation or expansion in web information retrieval. Various additional sources of information are used to choose more effective queries when conducting web searches, such as thesauri or query logs, or directly the results of an initial query.

This latter approach, which inspired our work, is well described in [9] and in [10], typically resulting in a significant increase in recall [29]. Thus far, to our knowledge, it has never been explored for IoT search. In this case, the problem is more complicated than web page retrieval because the quality of initial results is lower, and therefore manual intervention seems necessary; on the other hand, even though methods for web query reformulation/expansion are fully automatic, it should be noted that some searches may be penalized by their utilization.

## 3. Limitations and potentials of IoT Search Engines for vulnerability detection

The task under analysis involves identifying internet-accessible hosts hosting certain types of services or products of user interest with known vulnerabilities. To execute this task, IoT search engines can be utilized in conjunction with public vulnerability databases.

The most common strategy entails locating the desired hosts through queries to IoT search engines, followed by extracting a series of information from the associated data, and finally cross-referencing this information (typically via CPE) with vulnerability databases to identify CVEs.

The processing flow described also forms the basis of vulnerability discovery services provided (at a premium price) by the same search engines, but which may result unsatisfactory. The difficulty lies both in locating hosts hosting the desired products and in subsequently cross-referencing the information extracted from these hosts with vulnerability databases.

The effectiveness of the initial retrieval essentially depends on the quality of the queries submitted to the system. Users may be tempted to rely on the product or vendor name. However, doing so may lead to retrieving irrelevant results (with false positives and low precision, in Information Retrieval terms) while simultaneously failing to retrieve relevant

results (with false negatives and low recall). On the other hand, increasing the number of keywords in an AND operation may result in an empty set of results.

These issues are well-known when conducting keyword searches in large textual databases and are primarily due to the polysemy and synonymy of natural language, incomplete or erroneous information (such as in banners or web pages) due to banner obfuscation techniques, as well as the difficulty of screening and sorting through a plethora of potentially relevant results. For example, conducting searches on the source code of a web page without considering its context and without using restrictive filters may yield false positives because, for instance, a random number could be mistakenly identified as a firmware version.

The presence of false positives and false negatives is well documented in the literature [30] [31] [32] [33]. Matching with vulnerability databases also presents many challenges. In addition to being conditioned by the accuracy with which the information of interest (product name, vendor, version, etc.) is extracted from textual banners and other collected data, this operation is hindered by the difficulty of matching this information with the textual descriptions of CVEs [34].

The two difficulties just described compound each other, heavily affecting the vulnerability identification process. In fact, vulnerabilities provided by IoT search engines have relative reliability, so much so that, for example, Shodan has taken precautions by offering unverified (the vast majority) and manually verified CVEs (only for a very limited number of queries [6].

The automatically found CVEs, besides having modest coverage, suffer from high inaccuracy because they often relate to communication protocols or the use of specific servers or certain operating systems, rather than the specified devices of interest in the query [17]. Even when the initial search is performed using the CVE ID (as allowed by Shodan and other tools), things do not improve because the matching issues between the textual descriptions of CVEs and those associated with the devices persist, and the results continue to be unsatisfactory.

Another factor to consider is backporting, which involves applying a patch for a vulnerability at the operating system (Linux) distribution level rather than updating the software where the vulnerability exists. If this circumstance is not considered, software with backported patches may be mistakenly flagged as vulnerable based on the version number, even if the vulnerability has been effectively mitigated.

It should also be considered that search engines focus, for efficiency reasons, on identifying vulnerabilities associated with the most common services (OpenSSH, IIS, Apache, etc.), effectively penalizing those related to more specific devices/services, as is the case of specific components of PV production monitoring systems discussed in Section 5.

However, alongside these challenges, IoT search engines have two features that, if better exploited, could significantly improve performance. The first is that they collect a wide variety of content during network scanning. The Censys search engine, for example, examines over 3500 ports of the entire IPv4 address space and can detect over 100 Layer 7 protocols. Depending on the type of protocol, Censys collects various types of data: HTTP(S) root pages, banners of lightweight protocols, MQTT messages, etc.

Data collection is accompanied by intelligent indexing, which facilitates subsequent retrieval through an advanced search language. In addition to traditional search tools (full text, boolean operators, wildcards, regular expressions), it provides a range of filters that utilize structured information retrieved by the IoT search engine during network scanning.

Censys filters, for instance, cover hosts, services, DNS, location, operating systems, protocols, certificates, specific services, vendors, and products. These filters can be further refined through nested searches. For example, using Censys, regarding the "service" filter, it is possible to specify, among others, banners, banner hashes, port, and transport protocol. However, these capabilities are not adequately exploited.

In the next section, a methodology will be presented that leverages the engines' ability to retrieve very specific and accurate information at scanning time, aiming to use this information to formulate more targeted queries.

## 4. Proposed methodology for retrieving vulnerable hosts

In Figure 1, a general outline of the proposed vulnerable host retrieval process is depicted. The input consists of very general information, such as the name of a product or a vendor. The first step involves using these keywords to find in vulnerability databases, particularly the National Vulnerability Database (NVD), any CVE associated with that device, utilizing the information retrieval system linked to the database.

The rationale behind this initial step is twofold. On one hand, it allows us to direct the overall search towards known vulnerabilities, while on the other hand, it enables us to easily retrieve some more detailed information, contained both in structured data (such as CVE ID and CPE) and in the textual CVE description (e.g., product details and version range affected by the vulnerability).

The gathered information is utilized in the second step of the procedure, wherein progressively more specific queries are formulated to an IoT search engine, stopping before the set of results becomes empty. This step serves to reduce false positives, to retrieve a reduced set of potentially relevant hosts that are easy to manually inspect to identify a relevant result, i.e., a host related to the sought-after device and vulnerable according to the initially selected CVE.

Retrieving a relevant result constitutes the premise of the subsequent phase. Indeed, a series of information associated with the relevant result can be considered as "fingerprints" of that vulnerable device. These fingerprints can be mapped to specific constructs (filters) of the IoT search engine's advanced search language, facilitating the retrieval of other relevant results by the engine. Censys was chosen as IoTSE (Internet of Things Search Engine) of reference for our experiments due to its flexibility in usage through APIs with a free search account, a rich set of structured data characterizing results, along with a strong expressive power of its query language.

This choice inherently introduces a bias concerning the capabilities of each individual IoT engine regarding sampling frequency, analyzed ports, and identified devices. However, it is substantiated by preliminary comparative tests conducted with alternative search engines.

Below are listed some useful fingerprints and their associated constructs, with reference to Censys:

- Banner metadata, which are information sent to the web server and related to operating system, IP address, ports, serial number, hardware specifications, geographic location, organization, etc. (*software.uniform_resource_identifier, services.banner, services.software.version, location, operating_system, services.transport_fingerprint name, IP, services.software.vendor, services.software.product, services.software.version*)

- Components of the HTTP response header including the Etag field, an identifier of the specific version of a resource (*services.http.response.headers*)
- Web page title (*services.http.response.html_title*)
- Portions of the web page URL as in the case where HTML templates have a specific encoding of page names: login, dashboard, install, etc. (*services.http.request.uri*)
- Components of the web page such as favicon, copyright, product names (*services.http.response.favicons.md5_hash*, *services.http.response.body*)
- Labels, which are distinctive service tags inferred by the IoT search engine (*labels*).

The utilization of footprints in conjunction with associated filters can prove highly effective in expanding the set of relevant results, as we will demonstrate in Section 5. However, it should be noted that the precise determination of the most suitable queries may require a combination of filters and other constructs of the query language, suggesting the difficulty of automating the task. Section 5 presents examples where it can be observed that constructing the right query poses varying degrees of difficulty.

A final consideration pertains to the vulnerability assessment of ICD retrieved. Although typically performed manually, it can be partially automated by employing systems that provide CVEs associated with a specific IP address (those associated with individual results in our case) and verifying that the set of CVEs returned as output includes the one of interest. Examples of such systems include Netlas, Vulners, and Shodan InternetDB. In the subsequent section, concerning the case study, this aspect will be revisited.

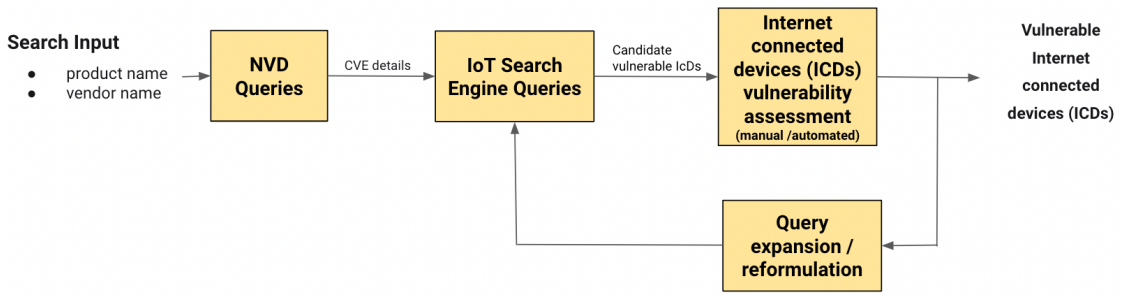


**Figure 1:** Main blocks of the proposed interactive process for retrieving vulnerable internet connected devices

# 5. Case study: finding vulnerable photovoltaic production monitoring systems

## 5.1. PV production monitoring systems are vulnerable

Energy insecurity, besides involving the lack of reliable and affordable access to energy sources, is also threatened by vulnerabilities of the components and devices responsible for its production and consumption. The increasing interconnection of renewable energy resources to electrical grids and the widespread adoption of smart technologies today make energy infrastructures potentially susceptible to cyber-attacks.

In the race for energy generation, there has been a proliferation of devices and monitoring tools accessible from the internet and controlled by end users, third parties, and utility companies, creating a vast attack surface vulnerable to threats at the level of individual devices for gaining access to the electrical grid [35]. This is a trend indicated by ENISA experiencing strong growth [36], demonstrating how from the early coordinated attacks on thousands of devices to block consumption accounting, we now face scenarios of compromise of the energy distribution network at various levels, from critical infrastructures to energy communities and up to distributed energy resources (DER), which are power production units operating locally and connected to the distribution grid.

Among the sensitive devices targeted by cyber-attacks are PV production monitoring systems, which are digital platforms using sensors, logs, and other components to conduct monitoring, maximize energy production, and ensure operational efficiency of photovoltaic generation production plants, including household use.

Additionally, among other available functions, they allow for anomaly reporting, control and variation of key parameters, generation of performance reports, and coordination of integration with the energy grid. In general, among the most frequent vulnerability categories for ICD devices there are the use of default credentials, unprotected communications, lack of software update plans, poor or absent access control, which if exploited allow access to sensitive information, system configurations, and control functions including device power on and off. From these data, further information can be derived, including, in the case of domestic users, space occupancy through energy consumption analysis.

Focusing the attention on PV production monitoring systems, a recent study [37] estimated around 130,000 exposed PV production monitoring systems. This study, although providing a broad overview of the issue, does not delve into detailing the methodologies by which hosts are identified and does not provide a more precise evaluation of the identified vulnerabilities.

## 5.2. Searching for vulnerable PV production monitoring systems

In Table 1, the phases and results of applying our methodology to the PV production monitoring systems sector are summarized. The choices and results reported in the table, obtained at the end of 2023, are now subject to detailed analysis. Concerning the first column (*NVD query*) the three manufacturers Contec, Solar-Log, and Enphase were selected due to their market significance and the presence of vulnerabilities associated with their products. By querying the NVD with these search keys, several CVEs emerged for each manufacturer. For each manufacturer, a specific CVE was selected, listed in the second column (*CVE ID*) of Table 1, which could be verified non-intrusively and did not fall under those patched at the operating system level.

The CVEs shown in Table 1 refers to the following vulnerabilities:

- For Contec, reference was made to the possibility of accessing a file upload webpage without credentials, related to CVE-2022-44354, corresponding to the CWE-434 weakness: Unrestricted Upload of File with Dangerous Type.
- For Enphase, reference was made to an outdated version of the device identifiable from the service homepage examination, related to CVE-2020-25755, corresponding to the CWE-119 weakness: Improper Restriction of Operations within the Bounds of a Memory Buffer.

- For Solar-Log, reference was made to credential-less access to the control panel, related to CVE-2021-34543, corresponding to the CWE-306 weakness: Missing Authentication for Critical Function.

In columns 3 (*Censys query*) and 4 (*Censys results*), the queries conducted on the Censys IoT search engine for each manufacturer and the corresponding number of results obtained are reported. Censys was chosen because it is a state-of-the-art system and for its flexible usage policies for scientific research purposes.

The queries were progressively specialized using information contained in the description of each CVE, and the data in column 4 clearly demonstrate the benefit in terms of result set reduction. Subsequent manual analysis of these sets allowed for the identification of a relevant result for each product. By "relevant," we denote a host hosting the respective product and potentially susceptible to attacks as indicated in the corresponding CVE in Table 1.

This vulnerability assessment operation is described in detail in Section 5.3 for each product. In column 5 (*Fingerprint-based Censys query*), the reformulated queries using the fingerprints extracted from each relevant result are shown.

In column 6 (Vulnerable ICD), the number of relevant and potentially vulnerable results obtained through the reformulated queries with the fingerprints is shown.

We now describe more in detail the construction of such queries. For Enphase, this step was straightforward, as it merely required utilizing the unique information contained in the service banner and conducting a search based on its hashed version. For Solar-Log, reference was made to the distinctive characteristics of the product homepage.

Therefore, the search was based on the page title, a portion of the URL, the associated page icon (favicon), and verification that the page was indexed categorized by Censys as a login page. Finally, in the case of Contec Solarview, the query construction process was more laborious. Starting from unique characteristics such as the page title and product name, the search was then narrowed down through boolean clauses to identify product versions characterized by the copyright year in the footer preceding 2022.

The reported data clearly demonstrate the overall effectiveness of the research methodology and the key role played by fingerprint-based queries, which enhance recall without sacrificing precision.

It is noteworthy that the validation process was conducted manually, while attempts to support it with automatic vulnerabilities identification methodologies were unsuccessful. Specifically, an ad hoc python program was written to query Censys and validate the resulting hosts with three automatic vulnerability identification tools [38] [39] [40], leveraging the interfacing APIs provided by IoT engines and vulnerability analysis tools.

In no case was the analyzed CVE found, presumably due to the high specificity of CVEs and the difficulty of the tools in identifying the fingerprints of the sought services. Instead, the tools flagged other CVEs, which, after accounting for false positives, appeared to be associated with the lack of host updates and various types of services hosted on them.

In contrast, fingerprint-based queries were able to retrieve a high number of vulnerable hosts compared to the service of interest, while simultaneously distinguishing them from other types.

**Table 1**
Retrieval of vulnerable PV production monitoring systems through interactive query reformulation

| NVD query | CVE ID | Censys query | Censys results | Fingerprint-based Censys query | Vulnerable ICD |
|---|---|---|---|---|---|
| Contec | 2022-44354 | solarview | 3373 | | |
| | | solarview compact | 926 | | |
| | | solarview compact 4.0 | 18 | (SolarView Compact) and ((2020) or (2021) or ((201*))) and services.http.response.html_tags="<title>Top</title>" | 573 |
| Enphase | 2020-25755 | enphase | 781 | | |
| | | enphase envoy | 431 | | |
| | | enphase envoy R3 | 13 | services.banner_hashes="sha256:42334785e3e0b1437a78bc9e032a19daff0e26e4235fde365f7dcfb2ad503e9d" | 331 |
| Solar-Log | 2021-34543 | solar-log | 6136 | | |
| | | solar-log 2.8 | 321 | | |
| | | solare solar-log 2.8 | 10 | services.http.response.favicons.md5_hash="c6e83fd6894b1de92c19e25fb668919b" and services.http.response.html_title="Solar-Log™" and labels=`login-page` and (p_live_cockpit) | 1285 |

## 5.3. Vulnerability analysis of search results

In this section, we focus on a crucial aspect often not explicitly addressed in other studies, namely, ensuring that the host retrieved through the IoT search engine is indeed potentially susceptible to attack as indicated by the associated CVE, taking also into account the backporting issue. For each vendor in Table 1 the actual exposed service and the methods used to verify the vulnerability are shown.

These verification methods were deliberately non-intrusive, thus preserving the integrity and availability of the service itself, also making usage of cached webpage version from search engines and Internet Archive [41].

**Contec**. Starting from the results obtained from Censys, we proceeded to verify whether the identified hosts contained a file upload page, as the CVE-2022-44354 under analysis indicated that various versions of the Contec Compact system are affected by a vulnerability related to the possibility for an unauthenticated user to upload files into the system through an HTML prompt. In Figure 2 an example of the results showing a file upload page exposed by a vulnerable service accessible without credentials can be observed. Additionally, it is noted how (Figure 3) other pages exposed by the service display sensitive information regarding the status of energy production.

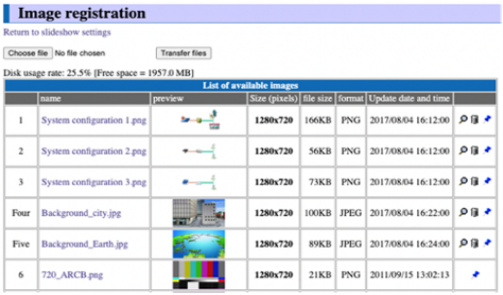


**Figure 2:** File upload

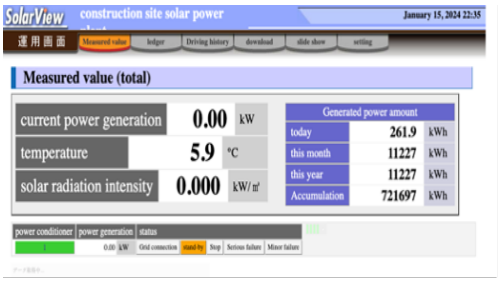


**Figure 3:** Control panel

**Enphase**. The analysis focused on identifying administration pages that reported one of the software versions indicated as vulnerable by CVE-2020-25755, namely version R3.17.3. In Figure 4 the homepage of an Enphase product (with a zoomed detail of the system statistics) is shown, which displays numerous sensitive pieces of information including the serial number, version, number of inverters, total energy produced, and the current production status.

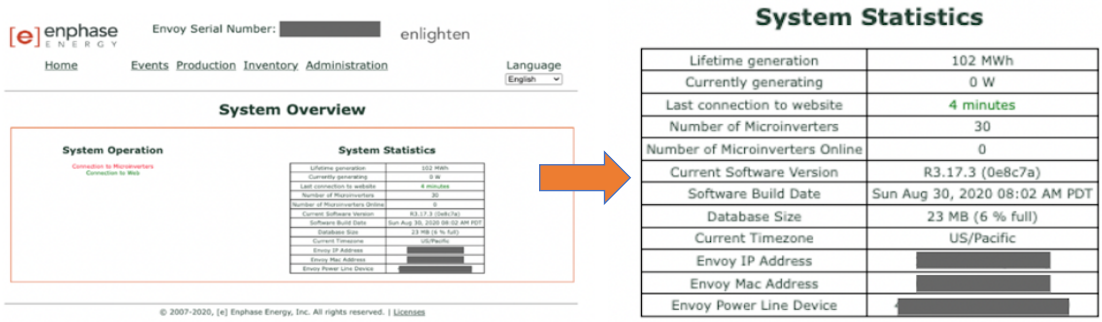


**Figure 4:** System overview

**Solar-Log**. In this case, the vulnerability to ascertain was CVE-2021-34543. To maintain a non-intrusive approach, the use of default credentials was not verified, and the focus was placed on checking for the presence of services lacking credentials (Figure 5) and a service page providing the option for even an unauthenticated user to configure a new password (Figure 6).

Among other pages accessible without credentials is the monitoring interface, which allows access to sensitive information related to the power flow of the system (power produced,

purchased, and consumed) as shown in Figure 7, including detailed economic data year by year (Figure 8).


**Figure 5**: Reporting of password absence

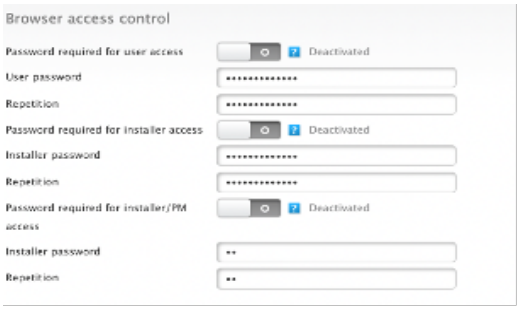


**Figure 6:** New password setting

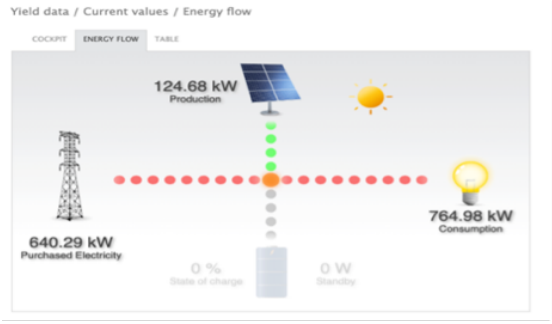


**Figure 7**: Energy production summary

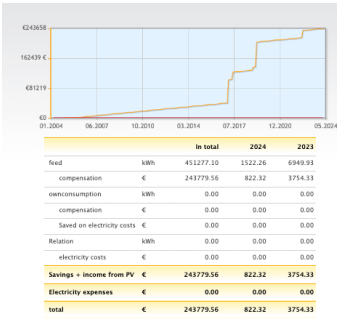


**Figure 8:** Economic summary

## 6. Conclusions

In this study, we explored the feasibility of identifying, through an IoT search engine, internet-exposed devices that are of interest to the user and susceptible to attacks using known vulnerabilities.

The approach undertaken emphasizes the formulation of effective queries, aiming to better exploit the capabilities of the advanced search language, which is likely underutilized by users. Within the proposed interactive methodology, a key role is played by the analysis of results obtained by simple queries and the subsequent extraction of a series of fingerprints from them to be used as advanced search features in reformulated queries.

Applying this methodology to PV production monitoring systems, a type of device experiencing rapid proliferation and documented vulnerability issues, we succeeded in identifying a significant number of exposed systems exhibiting characteristics indicative of known vulnerabilities.

A potential extension of this research, presently under investigation, involves attempting to automate the primary phases of the interactive reformulation process, particularly the selection of terms from CVE descriptions (using information retrieval techniques) and the extraction of fingerprints from IoT search engine results (employing machine learning techniques).

## Acknowledgments

The authors would like to express gratitude to Censys, Vulners and Netlas for their collaboration and generosity in providing access to their search engines and vulnerability scanners. This work is original and has been supported by a collaboration between RSE S.p.A. and Fondazione Ugo Bordoni, financed by the Research Fund for the Italian Electrical System under the Three-Year Research Plan 2022-2024 (DM MITE n. 337, 15.09.2022), in compliance with the Decree of April 16th, 2018.

## References

[1] S. Baho and J. Abawajy, "Analysis of Consumer IoT Device Vulnerability Quantification Frameworks," *Electronics,* 2023.

[2] F. Gordy, "The State of BAS Cybersecurity," 2019. [Online]. Available: https://automatedbuildings.com/news/apr19/articles/ib/190318022808ib.html.

[3] X. Wang, Y. Wang, X. Feng, H. Zhu, L. Sun and Y. Zou, "IoTTracker: An enhanced engine for discovering internet-of-thing devices.," in *In 2019 IEEE 20th International Symposium on A World of Wireless, Mobile and Multimedia Networks*, 2019.

[4] Censys, 2024. [Online]. Available: https://censys.com/. [Accessed 19 April 2024].

[5] Zoomeye, 2024. [Online]. Available: https://www.zoomeye.org/. [Accessed 19 April 2024].

[6] Shodan, "Shodan Vulnerability assessment," 2024. [Online]. Available: https://help.shodan.io/mastery/vulnerability-assessment. [Accessed 2024 april 2024].

[7] NIST, "NVD," 2024. [Online]. Available: https://nvd.nist.gov/. [Accessed 19 April 2024].

[8] Mitre, "CVE," 2024. [Online]. Available: https://cve.mitre.org/. [Accessed 19 April 2024].

[9] C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval," *Acm Computing Surveys (CSUR), 44(1),* pp. 1-50, 2012.

[10] V. Gupta and A. Dixit, "Recent Query Reformulation Approaches for Information Retrieval System-A Survey. Recent Advances in Computer Science and Communications," *Recent Patents on Computer Science 16(1),* pp. 94-107, 2023.

[11] A. Jäger-Waldau, "Snapshot of photovoltaics– May 2023," *EPJ Photovoltaics ,* vol. 14, no. 23, 2023.

[12] A. &. J.-W. A. Chatzipanagi, "The European Solar Communication—Will It Pave the Road to Achieve 1 TW of Photovoltaic System Capacity in the European Union by 2030?," *Sustainability,* Vols. 15(8), 6531.

[13] F. Harrou, B. Taghezouit, B. Bouyeddou and Y. Sun, "Cybersecurity of photovoltaic systems: challenges, threats, and mitigation strategies: a short survey," *Frontiers Media SA,* 2023.

[14] R. Williams, E. McMahon, S. Samtani, M. Patton and H. Chen, "Identifying vulnerabilities of consumer Internet of Things (IoT) devices: A scalable approach," in *Proceedings of the 2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*, Beijing, China, 2017.

[15] E. McMahon, R. Williams, M. El, S. Samtani, M. Patton and H. Chen, "Assessing medical device vulnerabilities on the Internet of Things," in *IEEE international conference on intelligence and security informatics (ISI)*, 2017.

[16] J. Bugeja, D. Jönsson and A. Jacobsson, "An investigation of vulnerabilities in smart connected cameras," in *IEEE international conference on pervasive computing and communications workshops (PerCom workshops*, 2018.

[17] S. Mulero-Palencia and V. Monzon Baeza, "Detection of Vulnerabilities in Smart Buildings Using the Shodan Tool," in *Electronics 2023 12*, 2023.

[18] V. Rajasekar and S. Rajkumar, "A Study on Internet of Things Devices Vulnerabilities using Shodan," *International Journal of Computing, 22(2),* pp. 149-158, 2023.

[19] K. Simon, C. Moucha and J. Keller, "Contactless Vulnerability Analysis using Google and Shodan.," *J. Univers. Comput. Sci., 23(4),,* pp. 404-430, 2017.

[20] B. Zhao, S. Ji, W. H. Lee, C. Lin, H. Weng, J. WU and R. Beyah, "A large-scale empirical study on the vulnerability of deployed IoT devices," *IEEE Transactions on Dependable and Secure Computing, 19(3),* pp. 1826-1840, 2020.

[21] X. Feng, Q. Li, H. Wang and L. Sun, "Characterizing industrial control system devices on the internet," in *24th International Conference on Network Protocols (ICNP)*, 2016.

[22] A. Keliris and M. Maniatakos, "Remote field device fingerprinting using device-specific modbus information," in *59th international Midwest symposium on circuits and systems (MWSCAS)*, 2016.

[23] A. Tanaka, C. Han, T. Takahashi and K. Fujisawa, "Internet-wide scanner fingerprint identifier based on TCP/IP header.," in *In 2021 Sixth International Conference on Fog and Mobile Edge Computing (FMEC)*, 2021.

[24] K. Yang, Q. Li, H. Wang, L. Sun and J. Liu, "Fingerprinting Industrial IoT devices based on multi-branch neural network," *Expert Systems with Applications, 238,* 2024.

[25] M. Bures, M. Klima, V. Rechtberger, B. . S. Ahmed, H. Hindy and X. Bellekens , "Review of specific features and challenges in the current internet of things systems impacting their security and reliability,"," *Trends and Applications in Information Systems and Technologies,* vol. 3 9, pp. 546-556, 2021.

[26] J. Song, S. Wan, M. Huang, J. Liu, L. Sun and Q. Li, "Toward Automatically Connecting IoT Devices with Vulnerabilities in the Wild," *ACM Transactions on Sensor Networks, 20(1),* pp. 1-26, 2023.

[27] F. Ebbers, "A large-scale analysis of IoT firmware version distribution in the wild.," *IEEE Transactions on Software Engineering, 49(2),* pp. 816-830, 2022.

[28] D. Ackley and H. Yang, "Exploration of smart grid device cybersecurity vulnerability using Shodan," in *IEEE Power & Energy Society General Meeting (PESGM)*, 2020.

[29] R. Nogueira and K. Cho, "Task-oriented query reformulation with reinforcement learning," *ArXiv preprint arXiv:1704.04572,* 2017.

[30] G. Barbieri, M. Conti, N. Tippenhauer and F. Turri, "Assessing the use of insecure ics protocols via ixp network traffic analysis," in *International Conference on Computer Communications and Networks (ICCCN)*, 2021.

[31] C. Mathas, C. Vassilakis, N. Kolokotronis, C. Zarakovitis and M. Kourtis, "On the design of IoT security: Analysis of software vulnerabilities for smart grids," *Energies, 14(10), 2818,* 2021.

[32] S. A. Alsaeed and I. Siddiq.Patent 11,381,590., 2022.

[33] J. Luo and J. Wang, "Vulnerability assessment of iot devices through multi-layer keyword matching.," in *International Conference on Computer, Internet of Things and Control Engineering (CITCE)*, 2021.

[34] B. Genge and C. Enăchescu, "ShoVAT: Shodan-based vulnerability assessment tool for Internet-facing services," *Security and communication networks, 9(15),* pp. 2696-2714, 2016.

[35] I. Zografopoulos, N. D. Hatziargyriou and C. Konstantinou, "Distributed energy resources cybersecurity outlook: Vulnerabilities, attacks, impacts, and mitigations," *EEE Systems Journal,* 2023.

[36] ENISA, "Identifying emerging cyber security threats and challenges for 2030.," European Union Agency for Cybersecurity (ENISA), Athens-Heraklion, Greece, 64., 2023.

[37] Cyble, "Security Gaps in Green Energy Sector: Unveiling the Hidden Dangers of Public-Facing PV Measuring and Diagnostics Solutions,," July 2023. [Online]. Available: https://cyble.com/blog/security-gaps-in-green-energy-sector/.

[38] "Shodan InternetDB," 2024. [Online]. Available: https://internetdb.shodan.io/. [Accessed 19 april 2024].

[39] Netlas, 2024. [Online]. Available: https://netlas.io/. [Accessed 19 April 2024].

[40] Vulners, 2024. [Online]. Available: https://vulners.com/. [Accessed 19 April 2024].

[41] "Wayback Machine," Internet Archive, [Online]. Available: https://web.archive.org/. [Accessed 17 April 2024].