‘I’m the mess that you wanted’: Evaluating the accuracy of WHOIS asset discovery against self-reported data

Yana Angelova
Delft University of Technology,
Delft, the Netherlands
y.y.angelova@tudelft.nl

Aksel Ethembabaoglu
Delft University of Technology,
Delft, the Netherlands
a.m.ethembabaoglu@tudelft.nl

Rolf van Wegberg
Delft University of Technology,
Delft, the Netherlands
r.s.vanwegberg@tudelft.nl

Carlos Gañán
Delft University of Technology,
Delft, the Netherlands
c.hernandezganan@tudelft.nl

Michel van Eeten
Delft University of Technology,
Delft, the Netherlands
m.j.g.vaneeten@tudelft.nl

***Abstract*—Security practices, such as internet-wide vulnerability scanning, need to identify the entity responsible for an asset so that the entity can be informed. In many cases, practitioners and researchers use IP WHOIS to map assets to owners. However, little is known about the accuracy of these methods due to the lack of ground truth data. In this paper, we conduct a case study in which we compare the WHOIS results against self-reported asset data from a set of public organizations. We retrieved IP ranges of Dutch municipalities from public IP WHOIS databases using keyword matching and then compared these results to the IP ranges that the municipalities reported to their sectoral CERT organization. Our findings show a modest degree of overlap between the two sets of IP addresses, with a considerable amount of false positive results present in the public data. To understand the implications of these attribution issues for security research, we retrieve Shodan scan results of both the public and the self-reported ranges in order to observe the differences in public-facing assets and their vulnerabilities. The found inaccuracies in asset mapping result in a large overestimation of the vulnerability exposure of Dutch municipalities, which, ironically, aligns with ongoing narratives about local governments not properly securing their infrastructure. We conclude that WHOIS-based asset discovery is viable but faces serious limitations that are rarely recognized and warrant further investigation by the security community.**

I. Introduction

A fundamental building block of network security practices and research is the ability to correctly identify an organization’s internet-facing assets. A wide range of security activities, including external network risk assessment, regulatory oversight, and incident attribution, rely on accurate asset attribution. Use cases of organization-to-IP asset mapping appear in academic work [17], [10], [14], as well as in cybersecurity solutions such as Bitsight [2] and SecurityScorecard [26], which provide security risk ratings of organizations. Other use cases are in regulatory oversight or government support services, like vulnerability notification programs. While authorities know which organizations fall under their jurisdiction, they typically lack direct insight into what assets belong to those organizations and instead rely on WHOIS or other mechanisms to identify the relevant assets to monitor.

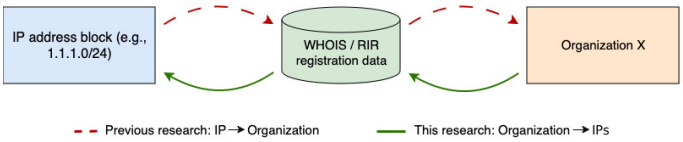


Fig. 1: IP-to-org vs. org-to-IP asset mapping

Working from organizations to IP addresses and other assets has been an area of limited academic work. This stands in stark contrast to advances in techniques for working in the other direction: from IP addresses to organizations (‘IP-to-org’).Recent papers presented new tools to systematically attribute ownership of IP addresses [13], [16], [29] and Autonomous System [32], [27]. While impressive, these tools are solving a different problem – as visualized in Figure 1. The use cases we mentioned above need to function in the opposite direction: moving from a set of organizations to their IP addresses. This requires a comprehensive view of all network assets belonging to a given entity, which is a fundamentally different task from attributing a single IP to its most likely owner. IP-to-org tools cannot simply be inverted, as organizations often leave incomplete or inconsistent footprints across registries, routing data, and certificates, and their identifiers are rarely homogeneous or unique across data sources, making reverse aggregation incomplete and error-prone.

Org-to-IP attribution process is notoriously time-consuming and error-prone. Some researchers have explicitly opted to work around this problem and extract risk signals from the web domains of organizations, which are much easier to

identify [25]. This alternative approach has been shown to yield effective risk prediction, but it does not require the need to undertake org-to-IP mapping. The properties of the broader set of digital assets impact risk rating, attack surface monitoring, incident response, and regulatory oversight.

Although a range of asset discovery techniques are deployed in practice [30], IP WHOIS plays a key role in both org-to-IP and IP-to-org mapping. The WHOIS protocol is widely used and provides the same data as its successor, the RDAP (Registration Data Access Protocol). The WHOIS system is a distributed registration database containing the contact details of the companies to which an IP range was assigned by a Regional Internet Registry (RIR), the entities allocating IP addresses. Unlike the case of domain WHOIS data, where the domain registrar is responsible for the accuracy of the data, for IP WHOIS data, this responsibility falls on the RIR and their specific regulations for LIRs (local internet registries), i.e., the network operators to which IP addresses were assigned. So a natural question is: how accurate is that registration data for the purposes of asset discovery?

Prior work on IP WHOIS has focused on mapping IP addresses to autonomous systems [1], conducting historical WHOIS lookups [29], automated parsing of WHOIS records [5], and comparison of the consistency of WHOIS records with the alternative RDAP protocol [7]. Remarkably, as far as we can tell, no prior work has evaluated the accuracy of IP WHOIS registration data – in contrast to the accuracy of domain WHOIS, where a study found that it is riddled with erroneous and false information [4]. In short, there is reason to worry about WHOIS accuracy. These concerns have been discussed in previous works for years now [6], but not systematically researched and evaluated.

While there is value in evaluating the accuracy of IP WHOIS to identify an organization's IP addresses, it is also understandable that this research gap exists. Evaluating IP WHOIS accuracy faces a critical hurdle: the lack of ground truth against which to compare the WHOIS data. In this paper, we study the accuracy of WHOIS records for discovering the IP addresses of specific organizations by leveraging the best available approximation of ground truth: the IP ranges that organizations themselves have registered as belonging to them. We received access to self-reported IP ranges for one specific set of organizations, namely Dutch municipalities. Nearly 300 Dutch municipalities have registered their IP assets with their sectoral Computer Emergency Response Team (CERT) called IBD. We compare these assets against the IP ranges that we were able to identify using IP WHOIS through keyword matching as belonging to a Dutch municipality in the time period between 2009 and 2022. We find that only 38% of all publicly discovered IP addresses (383,831 in total) match the self-reported addresses. Furthermore, 42% of the self-reported municipality assets are never reported by public sources.

To quantify the security implications of these discrepancies, we examine a representative use case: identifying externally exposed and potentially vulnerable hosts for a given organization. We use Shodan [28] to discover the present vulnerabilities for both sets of IP assets. We find that although querying WHOIS records correctly maps public-facing assets in 58% of the self-reported IP cases, it also produces a considerable amount of false positive results, leading to a serious overestimation of the vulnerability landscape of the municipalities. This incorrectly reinforces an existing narrative that claims that local governments are especially vulnerable when it comes to their internet-facing systems [11], [21]. These findings show the implications of WHOIS inaccuracies for security processes relying on asset attribution, such as vulnerability and network traffic analysis.

In sum, we make the following contributions:

- We present the first empirical study to evaluate the accuracy of IP WHOIS data for asset discovery by comparing the results of a keyword matching approach with leveraging self-registered data for 292 Dutch municipalities.
- We find that using WHOIS records to identify organization assets correctly identifies 58% of the IP addresses in the self-registered dataset (true positive rate) and misses 42% of them (false negative rate). Of the total number of IPs attributed to the municipalities in WHOIS data, 63% were false positives. Most of the false negatives are related to network names not reflecting the municipalities but other entities, like internet service providers and IT shared service centers working for local governments.
- We quantify the impact of these errors for security research, and vulnerability discovery in particular, and find that public data sources tend to overestimate how up-to-date software products are, as well as the number of vulnerable hosts discovered.

Flaws in attribution can result in organizations failing to receive critical security warnings for their assets, being blamed incorrectly for security issues, or delaying the containment of ongoing attacks due to confusion over ownership. While the problem of asset attribution is often seen as a technical problem, it has real-world economic consequences. Errors in asset mapping can lead to distorted risk signals and misalignment of cybersecurity resources.

II. Related Work

Although we have been unable to identify any other work directly assessing the effectiveness of keyword matching in WHOIS records as compared to the self-reported data, there has been earlier work that examines the different aspects of using WHOIS data in the research community.

In their work, Bianzino et al. [1] have compared consistency between WHOIS records from different databases when mapping IP addresses to their respective autonomous system. They show that different sources have varying success in terms of WHOIS query responses, underlining the importance of source selection and variation. Corneo and Di Francesco [5], on the other hand, have looked into the efficiency of parsing WHOIS records and compared this to using RDAP. Although the latter is viewed as a future replacement of the WHOIS protocol, when it comes to automatic processing of registration data,

their findings suggest that the new proposed solution still lacks consistency when looking into result retrieval. The work of Fernandez et al. [7] in the same area investigates the consistency between the two protocols and highlights that although the majority of the records are the same between WHOIS and RDAP, in 7.6% of the cases, there are inconsistencies in record fields like IANA ID, creation date, and nameserver. In their work, Lu et al. [18] have looked at domain registration data as provided by the WHOIS protocol and what the effect has been of the implementation of the General Data Protection Regulation (GDPR) on the available data.

Finally, Vermeer et al. [30] have shown the importance of WHOIS records for asset discovery techniques, further strengthening the need for accuracy evaluation. Moreover, work like that of Clayton and Mansfield [4] and Streibelt et al. [29] shows both the need and use cases of WHOIS data in the research community. In all these cases, the lack of more accurate (ground truth) data limits our ability to quantify how reliable the retrieved information is, a limitation we try to explore in this work.

III. Methodology

We now detail the data collection and research methodology of our study. In this section, we will describe the approach we used for the data analysis and elaborate on the different aspects of data retrieval and dataset creation.

A. Approach

The research approach can be summarized as follows:

1) We searched two different public WHOIS-based datasets for municipality-related IP ranges. We then verified those ranges against known municipality names and country locations to retain the relevant Dutch municipalities.
2) Next, we combined the retrieved IP addresses in a single dataset, and we assessed its accuracy by comparing it against a list of municipality IP addresses as registered with the sectoral CERT.
3) We retrieved scan data from Shodan for the public and self-reported datasets.
4) Finally, to observe the effects of the inaccuracies in the WHOIS-based asset discovery, we compare the findings on the detected software and vulnerabilities of the municipalities across the public and self-reported datasets, using attack surface monitoring as a typical use case.

B. Data

For this study, we leverage two different types of data: IPv4 ranges associated with Dutch municipalities and network scan data for these IP ranges. We limit our IP range discovery to IPv4 addresses because of the composition of our self-reported data, which only contains IPv4 assets. We believe this only has a limited impact on the generalizability of our approach, as WHOIS for IPv6 addresses can be queried in the same way.

1) Municipal Network Ranges: The municipal IPv4 addresses are identified via two datasets: (a) a publicly compiled dataset based on IP registration data retrieved via WHOIS and keyword matching, and (b) an approximation of the ground truth data in the form of the assets that municipalities registered with their sectoral CERT.

a) Public Data: For the collection of our public data, we have used two well-known sources: MaxMind GeoIP ISP Database [20] and RIPE NCC's Database [24]. We have chosen these two data sources so that we can best collect all available IP addresses related to Dutch municipalities. They both differ in subtle ways. RIPE is the Regional Internet Registry (RIR), the authoritative source for WHOIS records, and its position is supposed to provide accurate information regarding IP address allocation and ownership. MaxMind is a commercial entity, and we have no insight into their data collection methods or the sources for their databases. While it is primarily based on public WHOIS, the database contains more information than is visible in the authoritative WHOIS records at RIPE. It also contains historical WHOIS information, which we need for our purposes. For these reasons, we include it to complement the RIPE data. Previous work [15] suggests that MaxMind has limitations in terms of precision for fine-grained asset mapping, which is why we also include the RIPE data. We acknowledge that our choice to use both sources for our data mapping implies that we accept a certain degree of overestimation of assets in order to prevent missing relevant data.

MaxMind provided network names, CIDR ranges, and observation dates, while RIPE NCC contributed names, IP range records, and last modified dates. In the case of MaxMind, we have made use of their GeoIP database by querying it for all IPv4 addresses and saving the associated organization. After that, we have aggregated the IP addresses per organization such that we create a dataset that can be searched using the organization name to retrieve all associated IPs. An important aspect of the data retrieved from MaxMind is that it sometimes differs from the data present in RIPE. This suggests that MaxMind makes use of additional resources and techniques to construct its database, a process that is not known to us due to the proprietary nature of the product.

When it comes to RIPE, we used their "RIPE Database" and, in particular, we searched the table containing registration information for INETNUM objects. These objects are IP ranges containing starting and ending IP addresses and the associated allocation data for those addresses, such as network name, allocation date, last modified date, and network description. In its role as a Regional Internet Registry, the data in the RIPE database is updated whenever new IP address allocations or changes take place. However, the responsibility of keeping the data up-to-date falls on the organizations owning the IP resources, making it a best-effort source of data.

Each of these sources was queried for IP ranges that included the term "gemeente" (the Dutch word for 'municipality') either in their network name or in the description of the

associated network range allocation records. This approach does not capture cloud-hosted assets or other assets managed by third parties. Neither does it find networks that are not explicitly labeled as municipality-owned by using the keyword "gemeente" in either their network name or description.

To verify the correctness of the returned networks, we conducted manual filtering and verification on all found ranges to confirm that they belong to a Dutch municipality, excluding results that might mention our chosen keyword, but that are not related to these governmental organizations. Using the official lists of active municipalities per year, taking into account the structural changes in terms of mergers or separation of organizations [8], we checked that the retrieved results correspond to these municipality names. We have removed false positive results for municipalities located in Belgium as well as IT networks owned by church organizations related to the Reformed Congregations, which in Dutch use the name "Gereformeerde Gemeenten" [9]. In short, because the list of Dutch municipalities is finite and known, we could confidently exclude false positives based on the network name, but there is no way to reduce false negatives caused by WHOIS records that have no explicit relation to the municipalities. This approach means that the dataset is a lower-bound estimate of municipal IP ranges. As mentioned before, our multi-source approach is susceptible to producing false positive results. To mitigate this issue as much as possible, we have implemented the manual checks and data sanitization outlined above. Although imperfect, we believe this has significantly improved the quality of our publicly retrieved dataset.

The mapping process focuses on assets in the period between 2009 and 2022 for MaxMind records and December 2022 for RIPE. Due to the nature of the RIPE database, the retrieved results are for IP assets assigned to a municipality at one point in time in the past, and that, on paper, are still owned by the organization. It is possible that some records were never updated and still contain ranges for municipalities from years back, potentially including organizations that no longer exist due to structural changes in the local government. Using the last modified date of the RIPE records, we have observed that some ranges have not been changed since as early as 2001, whereas the most recent update records were from 2022. The long period chosen for the asset mapping also relates to the slow turnover of ICT infrastructure in our self-reported data, as we will explain below. Some municipalities registered their ranges with the sectoral CERT more than 10 years ago, and these registrations are still active.

At the end of the data retrieval and mapping, we combine both sources and create a single dataset to act as our publicly collected ranges, from now on referred to as "public data".

b) Self-reported CERT Data: The second dataset of IP ranges data comes directly from the sectoral CERT for Dutch municipalities – called the 'Informatiebeveiligingsdienst' (IBD)[12]. This data consists of ranges that the municipalities themselves have registered with the CERT as their assets. When the IBD was formed, they had an active outreach campaign to stimulate the municipalities to register their assets. Since then, the CERT has been putting in efforts to regularly seek updates from the already registered organizations and stimulate any potentially missing ones to sign up. Since these ranges are self-reported by the municipalities, and it's in their interest to keep them up-to-date, we regard this data source as the closest approximation of ground truth.

This self-reporting process comes with its limitations regarding the completeness and how up-to-date the asset information is, heavily relying on the motivation of the municipal organizations to collaborate with the CERT. We believe that the additional benefits in terms of security information sharing and vulnerability notifications that municipalities receive from the CERT are enough motivation for them to keep their registered assets as true as possible. This would ensure that the quality of these self-reported records is higher compared to the data we could retrieve from public sources. The data we were provided with consists of the latest ranges reported by municipalities up to 2022. Our data also includes the last update time for the IP

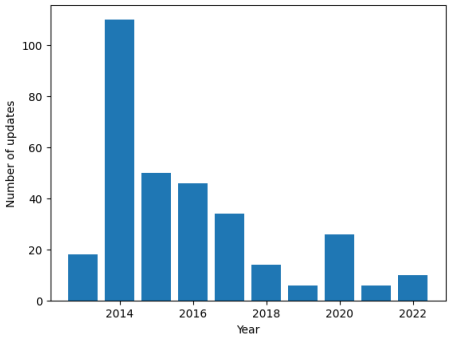


Fig. 2: Yearly counts of latest asset updates (CERT)

ranges, which shows that some municipalities have not updated their reported assets since as early as 2013, as seen in Figure 2. Because of this finding, we have chosen to use the full time period of our public asset mapping since we expect that some IPs have remained unchanged since that period.

2) Software and Vulnerability Data: Shodan [28] was used to collect network scan data for the municipal IP ranges, retrieving details such as open ports, timestamps, and software product and version information [19]. For each IP address in both the public and CERT datasets, we queried Shodan for the most recent record from 2022 for available information about the host. We have opted for the passive measurement approach of using Shodan to ensure that our study does not cause any disruptions or impacts on municipal IT staff. It is important to acknowledge that organizations can opt out of being scanned by network scanners such as Shodan or manually hide the underlying network information. Both of these factors contribute to the limitations of this study's approach as discussed later in Section VII.

To enhance our analysis, additional data was manually collected on software release dates and vulnerabilities. For the

most popular open source products, we have also manually collected release dates of present versions, as well as, if available, the next version of that product and its release information. The set of resources we have used for this data collection can be found in the Appendix A.

Next, each software product was identified using the Common Platform Enumeration (CPE) system as present in the host's Shodan record. Using the retrieved CPEs, we query the National Vulnerability Database (NVD) for all associated vulnerabilities as defined through the Common Vulnerabilities and Exposures (CVE) entries [23]. This data allows us to analyze the perceived security vulnerability on an organizational level as seen from a network scanning point of view.

Together, these datasets form the foundation for the investigation of the cybersecurity posture of Dutch municipalities, offering insights into the management of network infrastructure and providing the opportunity to evaluate the asset discovery technique using keyword mapping against the CERT data.

C. Ethical Considerations

One ethical consideration in this research is the decision not to perform active network scans. Active scans could put an unnecessary load on the IT systems and staff of the municipalities in our study. Instead, we make use of the data collected by the established network scanning service Shodan. This service provides guidance on how organizations can request their networks to be excluded from the scanning procedures, thus providing an opt-out mechanism from further research.

A further risk related to this type of research is the possibility that the findings on the presence of vulnerable hosts could put the underlying organizations at greater risk of attacks by malicious actors. To mitigate this issue, all results are anonymized, and no organization names or identifying information will be disclosed. All presented results are in the form of aggregate data in order to prevent consequent de-anonymization.

IV. WHOIS-Based Asset Discovery Evaluation

Next, we present the evaluation of our asset discovery technique by comparing it to the data received from the sectoral CERT on Dutch municipalities. We present in greater detail the composition of the retrieved data and observe how well it matches with what the municipalities report as their IP ranges.

A. Identified Network Ranges

As explained in Section III, the two sources of public IP range information have been RIPE NCC (RIPE) and MaxMind. Each data source has provided a different number of network ranges associated with Dutch municipality networks when using keyword matching in WHOIS records. It is important to note that there are also differences in their time period coverage. The MaxMind ranges span over the period between 2009 and 2022, whereas the RIPE data was obtained for 2022, the period during which the CERT data was collected. However, these WHOIS records in practice cover a much larger time period, with some being last modified in 2001 and the most recent ones in 2022. Implying that through the data from RIPE, we are receiving ranges that were allocated before 2009, discovering organizations that no longer exist, even though the WHOIS record is still treated as active by RIPE.

In the time period between 2001 and 2022, in the Netherlands, there have been a total of 567 municipalities, when accounting for all structural reorganizations such as municipalities splitting or merging with others [8]. In total, we have discovered 383,831 IP addresses allocated to 430 current and former Dutch municipalities. This allows us to cover 76% of the organizations in the extended time period. From the undiscovered organization, the majority are ones that no longer exist due to the above-mentioned structural changes.

If we focus only on the municipalities that existed in 2022, using the keyword matching approach, we have discovered 366,271 IP assets corresponding to 327 out of 345 municipalities in total, or covering 95%. The remaining municipalities that our mapping process fails to discover are a mix of medium and small organizations. We observe that the asset discovery technique of keyword matching fails in these cases because these organizations do not explicitly label their assets and potentially rely on collaboration with bigger organizations, thus not having internet-facing IPs registered in their own name.

If we look at the two data sources separately, we observe that through MaxMind we have discovered 369,675 IP addresses, while from RIPE we have retrieved 242,881 IPs. Both sources have a large overlap in IP addresses of 228,725 addresses. Roughly speaking, the identified RIPE IP ranges are a subset of the identified MaxMind ranges.

From a set of a few manual WHOIS lookups, we have established that the main reason for differences between the two sets lies in the absence of the keyword "gemeente" in the network range names for the latest updates of the RIPE records. In our experience, MaxMind has been able to label these IP ranges using the full word "gemeente", while the official RIPE records mention only "gem" or the name of the municipality, without the keyword "gemeente". Unfortunately, we have not been able to identify how MaxMind is able to make the inference that these ranges do actually belong to a "gemeente", since it is a commercial product and its sources and inferences are not disclosed. In our limited checks, the majority of the assets that were only found by MaxMind contained the keyword "gemeente" in a previous version of the WHOIS record, which could suggest that MaxMind combines information from different WHOIS records over time. For the exploratory nature of this study, we proceed by merging these two data sources together to create an overall dataset of publicly-identifiable municipal IP addresses. It consists of 383,831 IP addresses. From hereon, we shall refer to this as the "public" dataset.

B. Self-reported Data Comparison

After the creation of the public dataset, it is important to evaluate how well the keyword matching discovery method performs by comparing it to the self-reported asset dataset we were provided with. As explained earlier in Section III, the

comparison data consists of IP addresses that the municipalities self-reported to the sectoral CERT (IBD). The data we had access to consists of the most up-to-date assets as of January 2022, allowing us to compare it with the public dataset that also spans the same year.

In total, our CERT data consists of 250,936 IP addresses corresponding to 292 municipalities out of 345 active organizations in 2022. Although this is not a perfect coverage of all municipal organizations, it is worth noting that some municipalities may not have assets to their name due to collaboration efforts between (smaller) organizations. To better understand the overlap between the public data and the self-reported, we looked into the intersection between the two sources as presented in the confusion matrix in Table I. What we see is that 58% of the self-reported IPs have been correctly identified using the public data retrieval. This means that 42% of the actual municipality IP addresses are never discovered using our public sources. In terms of municipalities, 270 of them are identified in both the public and the CERT sets.

	In public WHOIS		Not in public WHOIS	
	IPs	% IPs WHOIS	IPs	% IPs CERT
In CERT asset data	144,290	38%	106,646	42%
Not in CERT asset data	239,541	62%	N/A	N/A

TABLE I: The overlap between the public and self-reported to the CERT data sources in terms of IP addresses and percentage

When looking at the false negative results, we have identified assets that have no explicit mention of our keyword in their network name or description, hence escaping our search. One big example, which accounts for 31% of the false negative IPs, is a network block assigned to a large internet service provider (ISP) in the Netherlands. In this case, all the available information in both the network name and description points to the ISP, preventing us from discovering these IP addresses as municipality-owned. Other unmapped assets are hiding behind the names of shared service centers, again escaping our keyword search due to the lack of an explicit mention of their relation to the municipalities. We also encountered one very strange case, where one municipality registered an IP range that is part of the private IP address space, as specified by IANA [22]. This is clearly a mistake.

We have also seen that only 38% of the publicly mapped IP addresses are actually part of the self-reported data. This leaves 62% as potential false positive results, even though the ranges are registered explicitly to a municipality. The IP addresses discovered by the public data, which are not present in the CERT data, could contain assets that are no longer used by the municipalities, but have not been deregistered with the RIR. There could also be shadow IT present in this subset, assets that the municipalities actively own and use, but have not reported to the CERT, hence their absence in the CERT dataset. In both cases, the underlying organization has failed to keep the information about its network ranges updated.

The distribution of results between the two data sources does follow the generalized findings as well. We observe that for both sources, a similar percentage of the self-reported IPs have been mapped correctly, 57% and 54% for MaxMind and RIPE, respectively. Important to note is that the true positive results discovered by both sources share a considerable overlap, where almost 6% of the MaxMind IP addresses are not present in RIPE. Only 0.4% are found only in the RIPE's data and not in MaxMind. The main difference between the two sources is that they have different rates of false positives. In this regard, RIPE performs better by having only 44% IP addresses that are not in the self-registered set, compared to 61% in MaxMind. These findings show that MaxMind contributes only a marginal increase in discovering true municipality assets, at the expense of adding more false positives and noise.

We also compared the found assets in the self-reported data to those received from public sources on an organizational level. So we analyzed for each individual municipality, how many IP addresses the public datasets attribute to it, and how many IP addresses it has registered with the CERT. We observed an overlap of 270 municipalities present in both datasets, which means that in 84% of the cases, the public sources contain some data about the required organizations from the self-reported data. For the consecutive analysis, we focus on these municipalities present in both datasets.

On this level, the public mapping correctly finds all self-reported IP addresses in 14% of the cases, while producing on average 75% false positive results for the same organizations. Surprisingly, for more than half of the municipalities that are present in both datasets (60%, or 161 out of 270), all of the discovered public assets are false positives. So, while the comparison of the overall sets of IP addresses suggests modest success for WHOIS-based discovery of IP addresses, at the level of the organizations, for more than half of the organizations, the attribution is wrong. This can be understood by the fact that this mostly affects smaller municipalities. They contribute relatively few IP addresses to the total set. The total set is more dominated by the larger municipalities, so at that level, the performance of our method is better.

If we compare the successful with the unsuccessful identification of municipal IP assets, we see that the public mapping performs better for medium (between 50,000 and 100,000 inhabitants) and large municipalities (100,000+ inhabitants): between 42% and 33% of the IPs have been correctly identified, respectively. Our method performs particularly poorly for small municipalities with a population of less than 50,000, where the average proportion of false positive results is 91%.

Overall, these results suggest that the technique of keyword matching in public WHOIS records has modest success in terms of finding true positives, but does come with a substantial amount of noise and inaccuracy at an organizational level. The level of noise is especially high for smaller municipalities, to the point of completely drowning out the signal.

V. Perceived Vulnerability of Municipality Networks

To assess the impact of errors associated with WHOIS-based asset attribution, we explore a typical use case: monitoring the attack surface of an organization to identify the version of software running on hosts and the presence of vulnerabilities in that software. Here, we use scan data from Shodan to identify these features for both the public and self-registered datasets.

We employ Shodan to search for available network information, and the relevant banner data was retrieved. These searches reflect the networks' state in 2022. Next, we measure how the asset mapping inaccuracies discovered earlier influence the perception of the vulnerability landscape of Dutch municipalities.

A. Active Hosts

Using publicly collected IPs tied to municipality networks, we retrieved 16,792 Shodan records (hosts) linked to 8,872 unique IPs active in 2022. A host is defined by its IP address and port. Only 2% of all discovered IP assets are publicly visible to Shodan.

To identify which software services and products are present, we rely on Shodan records containing valid product and version tags. Out of 16,792 records, 89% lack enough information (either no product info or no version info), leaving 1,815 records with both product and version.

We repeated these steps for self-reported IP addresses, yielding 3,393 Shodan records for 1,897 unique IPs, which is 1% of all considered IPs. Here, 87% of records lack complete info, leaving 455 records with both product and version. Shodan derives its data from network banners, which can be manually modified by administrators. This can reduce publicly available product and version data.

We then examined product distributions as shown in Table II. The most common products include both open source and commercial tools. Overall, the public dataset suggests a larger attack surface than municipalities think they are operating.

Product	# Records in public	# Records in CERT data
MS IIS httpd	490	154
ntpd	446	42
MS HTTPAPI httpd	174	101
OpenSSH	137	31
Apache	107	46
DrayTek	90	11
nginx	77	14
lighttpd	27	3
Dropbear sshd	25	0
BGP	20	8
Total	1,593	410

TABLE II: Count for top 10 most popular products

B. Software Version

Next, we investigate product versions. This allows us to understand to what extent software is kept up-to-date for internet-facing hosts in municipality networks.

From the set of active products with version information, we have selected the most popular open-source products. We have excluded commercial products such as those from Microsoft since we cannot easily track granular-level changes of the version updates, like most open source software allows us to.

We manually retrieved information about the update release dates and vulnerabilities associated with each version seen in the Shodan records. For the public and CERT datasets, we analyze the number of hosts running up-to-date products versus outdated ones. For the latter, we calculate the software age as the time (in days) since the next version of the given software was released, as at that point the software became outdated. In the rest of the section, we will use the terms software age and outdatedness interchangeably.

When analyzing software age in the public dataset, our data shows that 15% of all hosts are already running the newest version of the underlying software. Surprisingly, we observe that the maximum software age in days is 6,796 days or about 18 years. The same steps of analysis were carried out for the self-reported dataset as well. Compared to the public dataset, only 12% of hosts in the CERT set have their systems up-to-date – a difference of 3%. In this set, the maximum software age amounts to 4,617 days or approximately 12 years.

This long tail of extremely outdated software might be surprising, if not shocking, in light of the widely shared best practice of patching vulnerabilities quickly. The Cybersecurity and Infrastructure Security Agency (CISA) has recommended that organizations apply patches within 24 to 48 hours – not years – when vulnerabilities are disclosed, especially if they are actively exploited or are present in internet-facing systems [3]. Surprising as it may be, earlier research also reported very long, sometimes multi-year, delays in deploying patched versions of software [31] [14].

To better illustrate the findings concerning outdatedness, we constructed a cumulative distribution function (CDF) representing the percentage of hosts with a certain number of days since a newer version has been available, as seen in Figure 3. The dotted line shows the distribution of outdatedness over hosts in the public dataset. Notably, about 29% of hosts have updated their systems in the last 365 days. In contrast, more than 50% of systems are outdated by more than 1,161 days. The hosts in the self-reported data are plotted with the solid line. We can see that 22% of hosts have software with an age of less than 365 days, while more than 50% of products are 1,446 days outdated. In Figure 3, we can also observe the differences between the two datasets. We can see that the self-reported hosts are, on average, more outdated than those in the public dataset. In turn, this suggests that relying on public data carries the risk of underestimating the outdatedness of the hosts.

C. Vulnerabilities

Software age alone does not reveal the vulnerability landscape. We use the number of CVEs (via product/version CPE data in the Shodan data) to measure vulnerable hosts.

In the public and self-reported datasets, we have found 221 and 67 unique vulnerable hosts, respectively. In Figure 4a, we

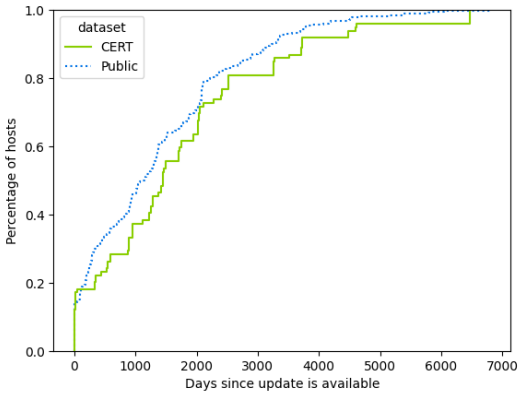
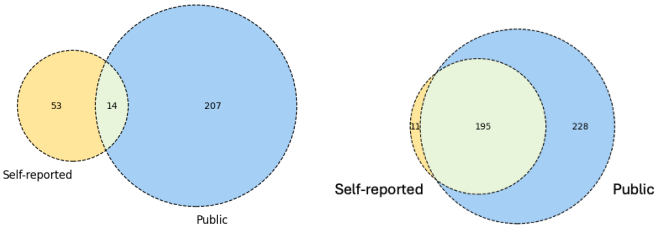


Fig. 3: A cumulative distribution function (CDF) of the percentage of hosts over the software age of the products in days, comparing public and self-reported data

observe the overlap between these hosts for the two sources. It is visible that the public data can correctly identify 14 out of 67 true vulnerable municipal hosts, which accounts for 21%. This shows that 94% of the vulnerabilities in the public data come from IP addresses that the municipalities do not see as theirs. In other words, any outside observer relying on WHOIS data is likely to heavily overestimate the number of vulnerable hosts in municipal networks.



(a) Vulnerable hosts (b) Unique CVE IDs

Fig. 4: Venn diagrams presenting overlap between the two datasets in terms of vulnerable hosts and unique CVE IDs

We also performed a simple linear regression analysis to see if the size of the municipality, as measured by the number of inhabitants, correlates with the number of vulnerable hosts in its network. For both groups, the r^2 value retrieved from the test is below 0.1, indicating no correlation.

If we look into which vulnerabilities were identified (CVE IDs), irrespective of which hosts they resided on, we see a bit more agreement. As is visible in Figure 4b, 95% of the CVEs discovered by the self-reported municipality IP addresses were also discovered using the public data. However, the latter also attributes an additional 228 other CVE IDs to the municipalities.

When looking into the publicly retrieved dataset, we observe that hosts have, on average, 16 CVEs. The average for the self-registered set is considerably higher at 26 CVEs. This high number can be explained by the relatively high outdatedness of the observed products.

These results show that not only do the compiled IP assets have a significant difference with the CERT data, but also when it comes to active hosts and their vulnerability information, the two sets exhibit different patterns in terms of the average number of CVEs per host. Moreover, the perceived vulnerability posture from the public data is not based on the actual vulnerable hosts, according to our self-reported data.

VI. Discussion

In this study, we present the first evaluation of an asset-discovery method based on keyword matching in WHOIS records. We compared the identified assets against self-reported data by Dutch municipalities. This approach aims to evaluate the effectiveness of utilizing public data sources for correctly identifying and attributing IP address network ranges that municipalities own and use. We found that the method produces a set of results with serious limitations. While 58% of the self-registered data is discovered this way (true positives), the remainder (42%) is missed (false negatives). Also, this result comes at the price of a considerable amount of false positives: of the assets in the public dataset, 62% are false positives.

We also observed that, in particular, MaxMind performs more poorly in terms of adding noise to the discovered assets compared to RIPE, while adding a few additional true positives. This suggests that researchers might be better off relying on RIR WHOIS databases to minimize the amount of false positives they receive. It is also a benefit that WHOIS data is freely available. These findings highlight the limitations of the asset discovery method as well as quality problems with the WHOIS data in the public datasets employed in this study.

Our findings suggest that public resources, such as RIPE's WHOIS databases and the commercial MaxMind database, are often not kept up-to-date by the asset owners. These organizations vary in their degree of diligence when it comes to keeping the information up-to-date, creating a fundamental issue regarding these public databases. This raises the question of the suitability of these databases for asset discovery, in light of their use by researchers, but also by a much broader community. This lack of reliability presents a clear issue for a typical use case in cybersecurity research, namely, to identify the owners of vulnerable assets, which is the foundation of processes such as external vulnerability scanning and security notifications. As observed, the false positives present in the public datasets can significantly influence the vulnerability landscape under analysis by under-representing the outdatedness of the used software and overestimating the number and attribution of vulnerable municipal networks. These discoveries highlight the importance of caution when it comes to drawing conclusions about vulnerabilities using public data sources, especially when researchers and other observers, like the national government or international CERTs, use WHOIS data to identify to whom a vulnerable host belongs or to evaluate the security posture of municipalities.

Given the wide use of these sources and the apparent misalignment of incentives, more effort should be put in place to keep registration records as updated as possible. Asset owners rarely see the direct benefits of keeping these records up-to-date, while the costs of the inaccuracies are often paid by third parties in the face of security researchers, CERTs, and regulators. This misalignment helps to understand the prevalence of outdated WHOIS records, despite their central role in cybersecurity operations. To address this issue, incentives should be provided by RIRs to asset owners or pressure from CERTs on their constituents to keep the records updated.

VII. Limitations

A limitations of the keyword asset matching method we used is the inability to find resources that have no explicit mention of the chosen keyword. In our specific case study, this limits our network ranges where the organization is registered as the owner in WHOIS, i.e., that self-hosts its infrastructure.

Another limitation is our approximation of the ground truth. In our case, the sectoral CERT for Dutch municipalities leaves it to the individual organizations to report and update their assets. The CERT performs regular and active outreach campaigns to the municipalities to update their IP data; however, in the end, it is up to the individual organizations to take action. This means that some of the IP addresses we have access to might already be out of use if the municipality forgot to notify the CERT.

We have to acknowledge the phaseout of WHOIS for data retrieval and its subsequent replacement with RDAP. Although for the time period under research, both systems were widely used, in the future, only RDAP will be employed. As shown in previous work [7], both systems have exhibited compatible results. This indicates that our findings can be applied to the newer system as well, but further verification of that is needed.

Further evaluations are needed, encompassing different organizational contexts and geographical locations as well. These could reveal how generalizable our findings are and shed light on the blind spots that different public asset databases exhibit or, more positively, in which contexts they do perform better. Comparing results from different contexts could present insights into the quality of public data on an international and cross-sector scale, potentially pointing to concrete factors contributing to better practices for information updating.

VIII. Conclusions

In this work, we evaluated the accuracy of WHOIS-based asset discovery techniques. We compared a keyword-based WHOIS method for finding Dutch municipality IPs with self-reported IPs from the sectoral CERT (IBD). Our analysis shows that only 38% of the publicly identified addresses matched the CERT data, and 43% of self-reported addresses were missing from the public dataset. The results also highlight the considerable number of false positives retrieved from the public sources. What analyzing active hosts revealed is that the public data overestimates both up-to-date hosts (15% vs. 12%) and vulnerable hosts (94% overestimation). Neither dataset showed any link between municipality size and vulnerability counts. With this case study analysis, we highlight the inaccuracy of public WHOIS-based asset discovery and its implications for typical security processes such as vulnerability scanning. These results show how asset attribution inaccuracies go beyond technical issues and can propagate into the economic and policy domain, by distorting the risk signals, weakening accountability, and undermining effective allocation of cybersecurity resources.

References

[1] Bianzino, A.P., Pezzuolo, D., Mazzini, G.: Who is whois? An analysis of results consistence. In: 2014 22nd International Conference on Software, Telecommunications and Computer Networks (SoftCOM). pp. 289–292 (2014). https://doi.org/10.1109/SOFTCOM.2014.7039137

[2] Bitsight, https://www.bitsight.com/

[3] CISA: #StopRansomware: LockBit 3.0 Ransomware Affiliates Exploit CVE 2023-4966 Citrix Bleed Vulnerability (Sep 2024), https://www.cisa.gov/news-events/cybersecurity-advisories/aa23-325a

[4] Clayton, R., Mansfield, T.: A Study of Whois Privacy and Proxy Service Abuse. In: 13th Workshop on the Economics of Information Security (WEIS 14) (2014)

[5] Corneo, L., Di Francesco, M.: From WHOIS to RDAP: Are IP Lookup Services Getting any Better? In: NOMS 2024-2024 IEEE Network Operations and Management Symposium. pp. 1–10 (2024). https://doi.org/10.1109/NOMS59830.2024.10575906

[6] Elliott, K.: The who, what, where, when, and why of whois: Privacy and accuracy concerns of the WHOIS database, https://scholar.smu.edu/scitech/vol12/iss2/4/

[7] Fernandez, S., Hureau, O., Duda, A., Korczynski, M.: Whois right? an analysis of whois and rdap consistency. In: Richter, P., Bajpai, V., Carisimo, E. (eds.) Passive and Active Measurement. pp. 206–231. Springer Nature Switzerland, Cham (2024)

[8] Gemeentelijke Indelingen per jaar (Feb 2023), https://www.cbs.nl/nl-nl/onze-diensten/methoden/classificaties/overig/gemeentelijke-indelingen-per-jaar

[9] Gergeminfo, https://www.gergeminfo.nl/

[10] Harry, C., Sivan-Sevilla, I., McDermott, M.: Measuring the size and severity of the integrated cyber attack surface across US county governments. Journal of Cybersecurity **11**(1) (01 2025). https://doi.org/10.1093/cybsec/tyae032

[11] Ib&p: Kwetsbaar door software – lessen naar Aanleiding van... (Feb 2022), https://ib-p.nl/download/kwetsbaar-door-software-lessen-naar-aanleiding-van-beveiligingslekken-door-software-van-citrix/

[12] Informatiebeveiligingsdienst, https://www.informatiebeveiligingsdienst.nl/

[13] Izhikevich, K., Chaparala, S., Patel, M., Srinath, S., Zhou, E., Du, B., Izhikevich, L.: Identifying Operators of IP Services At Scale with OperatorSage. In: Proceedings of the 2025 ACM Internet Measurement Conference (IMC '25), October 28–31, 2025, Madison, WI, USA (2025)

[14] Kotzias, P., Bilge, L., Vervier, P.A., Caballero, J.: Mind your own business: A longitudinal study of threats and vulnerabilities in enterprises (Aug 2023), https://www.ndss-symposium.org/ndss-paper/mind-your-own-business-a-longitudinal-study-of-threats-and-vulnerabilities-in-enterprises/

[15] Lee, Y., Park, H., Lee, Y.: IP Geolocation with a Crowd-sourcing Broadband Performance Tool. SIGCOMM Comput. Commun. Rev. **46**(1), 12–20 (Jan 2016). https://doi.org/10.1145/2875951.2875954

[16] Levett, J., Vassilakis, V., Yadav, P.: WHOactuallyIS? Finding the Companies Behind the Networks. In: 2025 9th Network Traffic Measurement and Analysis Conference (TMA). pp. 1–4 (2025). https://doi.org/10.23919/TMA66427.2025.11096964

[17] Liu, Y., Sarabi, A., Zhang, J., Naghizadeh, P., Karir, M., Bailey, M., Liu, M.: Cloudy with a Chance of Breach: Forecasting Cyber Security Incidents. In: 24th USENIX Security Symposium (USENIX Security 15). pp. 1009–1024. USENIX Association, Washington, D.C. (Aug 2015), https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/liu

[18] Lu, C., Liu, B., Zhang, Y., Li, Z., Zhang, F., Duan, H., Liu, Y., Qiongna Chen, J., Liang, J., Zhang, Z., et al.: From whois to WHOWAS: A large-scale measurement study of Domain Registration Privacy under the GDPR (Jun 2023), https://www.ndss-symposium.org/ndss-paper/from-whois-to-whowas-a-large-scale-measurement-study-of-domain-registration-privacy-under-the-gdpr/
[19] Matherly, J.: Complete guide to shodan (Jul 2015), https://leanpub.com/shodan/c/mv15rvJhm05J
[20] MaxMind, https://www.maxmind.com/en/home
[21] Meyer, J.: Local governments are more vulnerable to cyberattacks than ever before. DHS wants mayors to step up. (Feb 2022), https://eu.usatoday.com/story/news/politics/2022/02/08/local-government-cybersecurity-digital-threats/9208951002/
[22] Moskowitz, R., Karrenberg, D., Rekhter, Y., Lear, E., de Groot, G.J.: Address Allocation for Private Internets. RFC 1918 (Feb 1996). https://doi.org/10.17487/RFC1918
[23] NIST. National Vulnerability Database - Home, https://nvd.nist.gov/
[24] RIPE Network Coordination Centre, https://www.ripe.net/
[25] Sarabi, A., Karir, M., Liu, M.: Scoring the Unscorables: Cyber Risk Assessment Beyond Internet Scans. In: Weis 2025 – the 24th workshop on the Economics of Information Security (Tokyo, Japan) (2025), http://kmlabcw.iis.u-tokyo.ac.jp/weis/2025/doc/proceedings/WEIS2025_paper_22.pdf
[26] Securityscorecard, https://securityscorecard.com/
[27] Selmo, C., Carisimo, E., Bustamante, Fabian e. amd Alvarez-Hamelin, J.I.: Learning AS-to-Organization Mappings with Borges. In: Proceedings of the 2025 ACM Internet Measurement Conference (IMC '25), October 28–31, 2025, Madison, WI, USA (2025)
[28] Shodan, https://www.shodan.io/
[29] Streibelt, F., Lindorfer, M., Gürses, S., Gañán, C.H., Fiebig, T.: Back-to-the-future whois: An ip address attribution service for working with historic datasets. In: Brunstrom, A., Flores, M., Fiore, M. (eds.) Passive and Active Measurement. pp. 209–226. Springer Nature Switzerland, Cham (2023)
[30] Vermeer, M., West, J., Cuevas, A., Niu, S., Christin, N., Van Eeten, M., Fiebig, T., Gañán, C., Moore, T.: SoK: A Framework for Asset Discovery: Systematizing Advances in Network Measurements for Protecting Organizations. In: 2021 IEEE European Symposium on Security and Privacy (EuroS&P). pp. 440–456 (2021). https://doi.org/10.1109/EuroSP51992.2021.00037
[31] West, J.C., Moore, T.: Longitudinal Study of internet-facing openssh update patterns. Passive and Active Measurement p. 675–689 (2022). https://doi.org/10.1007/978-3-030-98785-5_30
[32] Ziv, M., Izhikevich, L., Ruth, K., Izhikevich, K., Durumeric, Z.: ASdb: a system for classifying owners of autonomous systems. In: Proceedings of the 21st ACM Internet Measurement Conference. p. 703–719. IMC '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3487552.3487853

APPENDIX

Software	Resource
Apache httpd	https://github.com/apache/httpd/tags https://archive.apache.org/dist/httpd
Jetty	https://github.com/eclipse/jetty.project/releases https://www.eclipse.org//lists/jetty-announce/maillist.html
OpenSSH	https://launchpad.net
Nginx	http://hg.nginx.org/nginx/log
Webmin (MiniServ)	https://github.com/webmin/webmin/releases
ProFTP	https://sourceforge.net/p/proftp/mailman/proftp-announce
Plex	https://forums.plex.tv/t/plex-media-server/30447
Grafana	https://github.com/grafana/grafana/releases
Dropbear sshd	https://matt.ucc.asn.au/dropbear/dropbear.html
Ivanti Endpoint	https://help.ivanti.com/mi/help/en_us/core/11.x/rn/CoreConnectorReleaseNotes/Revision_history.htm
Manager Mobile	https://help.ivanti.com/mi/help/en_us/core/12.x/rn/CoreConnectorReleaseNotes/Revision_history.htm
MariaDB	https://mariadb.org/mariadb/all-releases/
Netatalk	https://netatalk.io/
Samba	https://www.samba.org/samba/history/
lighttpd	https://www.lighttpd.net/releases/
thttpd	https://www.acme.com/updates/

TABLE III: Resources used for software release date retrieval