The use of Whois databases in practice and research

This blog explains, without completeness, how you can make use of Whois data, especially in a form of a Whois database in various areas of marketing, IT security, etc.

In the present blog we give a holistic overview, of at least a part of applications of WHOIS databases. As a practitioner or a researcher, you can broaden your insight get some inspiration and possibly find new ideas and starting points for further study.

While WHOIS data can be obtained by using a dedicated Internet protocol and practice the collection of up-to-date data is far from being straightforward in practice. This data is stored on the Internet in a distributed manner and many of the providers pose cumbersome limitations on their accessibility. WhoisXML API, INC. is specialized in collecting, cleaning and organizing WHOIS data in a central database, including, data back to several years. In the last section of this blog we briefly describe our services in support of setting up a database for applications like the ones described.

Introduction

The fundamental importance of WHOIS data in the operation of the Internet and notably in that of the World Wide Web is broadly overlooked by non-specialist, even though, in various communities it is well known of course. Web hosting providers are, for instance, fully aware of it: if a client intends to register a new domain, of course, they need to know if it’s still available. Domain names bear a crucial importance in marketing, too: as a domain name is essentially a synonym of a brand, and it’s a basis of Search Engine Optimization (SEO). IT security experts consider these data as fundamental, these are the only data connecting entities active on the Internet with physically existing organizations or persons. There are several areas in which domain WHOIS data are required for practical everyday purposes. All of these have their scientific implications, too. Several research groups, especially in IT security and marketing, require a WHOIS database to achieve certain research goals of significant impact.

WHOIS and The domain name ecosystem

The root of the Domain Name System is maintained by the Internet Corporation for Assigned Names and Numbers (ICANN (1)). They run a specialized website (2) where you can find very detailed information on WHOIS, including its history and its provisioned future, a glossary of related terms, the description of the related policies and technology. Here we just give a brief introduction necessary to understand the discussed applications.

On the Internet the devices have a unique identification number, the IPv4 address, currently. In order to find the way to the devices, these numbers have assigned a name. Names are organized hierarchically to domains. The assignment between names and IP addresses is the Domain Name System (DNS), implemented by the distributed network of Domain Name Servers. These servers operate on zones: subsets, often a single domain of the hierarchical domain name structure. The information about a zone is contained in so-called zone files.

Returning to the domain name hierarchy, the top-level domains are maintained by ICANN. They are of two kinds: country-code top-level domains (ccTLDs) such as .us, .uk, etc., and generic top level domains (GTLDs) such as .com, org. The gTLDs themselves can fall into two possible categories. Originally there were a few of these defined such as .com or .org. From the year 2012 on, the set of GTLDs started to significantly extend thanks to the new GTLD program announced by ICANN (3). These gTLDs are termed as new GTLDs (nGTLDs).

The registration of subdomains in ccTLDs as well as lower level domains are done by registrars: these are typically competing companies complying with ICANN’s rules. So if someone wants to register a domain, just contacts are registered. But how do we know to whom an actual domain belongs? And this is where WHOIS comes in: it is a distributed database with a dedicated protocol. WHOIS servers are typically maintained by the registrars and contain detailed information on each domain registration. These data cover: the name and contact data of the registrar as well as the registrant (who has registered the domain), dates of the creation, update and expiry, the address of the primary domain name servers and the date of the last update of the given record. So WHOIS is the only link between the virtual entities in the DNS and the physical actor behind.

So let us now turn our attention to the applications of these data.

Domain branding and marketing research

The lion’s part of business activities is based on the World Wide Web, even for traditional companies. Therefore, the possibility of finding a company in the

cyberspace is maybe the most important requirement of successful business. Of course, it's important that the domain name should be easy to remember, as the company name itself. But how does a potential client find a company? Most typically by using a search engine such as Google. And with this, the domain name is maybe even more important than the company name. In all search engines, domain names have a crucial role in the estimation of how relevant is a website for the particular query. And the goal for the company is, of course, to maximize this possibility. This is called Search Engine Optimization (SEO). The amplification of SEO needs a well-chosen domain name. There are many guides describing how to do it well, see e.g. (4). There can be a lot of psychological, economical, sociological considerations put behind a good choice.

When a new domain name is registered, obviously it should be available. The verification of this is indeed a very prevalent application of WHOIS data. Web service providers do this also for simple private web-pages. For company-owned domains there are more elaborate aspects, however. Many of these belong to the question of brand protection.

One aspect is the reputation of the domain names as well as those which are similar to them. It might be the case that the given domain name or something similar was used in the past and gained a bad reputation. It's important to check for this to make a good decision and requires a complete dataset to perform advanced queries. And it's not only the past but also the future which is relevant. If you have a brand, say, "Amazon", and you own the domain "Amazon.com", you probably want to keep track of other domain registrations on similar domains that potentially violate your trademark or damage your reputation, such as e.g. "amazonsucks.com", "amaz0n.com" or "myamazon101.com". Time by time you need to determine a list of domain names to register in advance to prevent this.

It is not only the name of the domain name which matters. To keep track of registration expirations as well. Especially in case of a corporation which may have thousands or hundreds of thousands of domains to manage spread between many affiliate companies. The expiration times and status of these domains should definitely kept track of. We all heard about how an important domain names gets expired and dropped due to negligence (google.com was dropped and owned by a guy for a minute). Owing to the large number of involved domains, an up-to-date WHOIS database is required.

And what if your company will be discontinued for some reason? Even if you don't care about that business anymore, the domain name can be associated with you. It

might be embarrassing if, for instance, a porn website would be set up under your previous domain. And there are such cases known. The message here is: domain registrations should be taken care of even when a company which used to be behind them ceases to exist.

The access to a domain database can be useful for marketing research purposes too, with a similar approach. From domain name registrations you can identify potential clients, collaborators or competitors. You may follow trends of your business area in order to stay up-to-date with your marketing.

Domain name registrations can also be considered as a business. Having registered potentially popular domains, domain investors can build a portfolio. The valuation of this and the determination of strategies also requires to follow the trends and time by time, to get information on the ownership of certain domains.

In conclusion, if you are in domain branding, domain protection or marketing research, either as an expert or just want to look into these issues yourself, you are indeed in the need of bulk and even historical WHOIS data.

Economics research

So far we have focused on practical activities in marketing. However, the analysis of WHOIS data can also lead to relevant scientific results. In the field of economic research there is a vast amount of scientific publications in which WHOIS data has been used. For instance, as competing companies are doing SEO and manage their web pages accordingly, WHOIS data hold information on the structure of dynamics of companies. Having a WHOIS database of a given scope at hand, one can develop techniques to extract the data with various techniques of data mining. The applicable techniques can include graph-theoretic approaches, heuristics, machine learning, and various types of dynamical analysis. Obviously such approaches need a structured, WHOIS database as a background.

From among the many research papers which use WHOIS data, let us briefly illustrate this with one in which the data used were purchased from WhoisXML API. This is a very timely approach illustrating well the potential and the typical techniques of this kind of research. NESTA, a global innovation foundation based in the UK, recently reported a research on a new kind of measurement of entrepreneurship in a working paper (5). They point out that, as we would expect according to the last section, most growth-oriented businesses begin not with company registration, but with domain name registration. This is often carried out by entrepreneurs before starting the actual business, this domain name

registration data can harbor useful information predicting actual entrepreneurship activity. WHOIS data are reach and sometimes provide a direct indicator of the entrepreneurial intent of the company, including information on the sector of the business and the orientation of growth.

To illustrate their approach, the researchers studied in detail the entrepreneurship in the regions of Oxford and Cambridge, UK. They have supplemented the WHOIS data with information from other sources. For instance, to identify entrepreneurial activity, the contents of actual web pages have to be analyzed with text-mining techniques. The WHOIS data contain addresses, so they augmented the data with geolocation information as well. (This is an opportunity worth emphasizing in general: as WHOIS data contain postal addresses, they can be augmented with GIS data to analyze the entities actual geo locations). Finally, by analyzing the data with advanced machine learning techniques, they have opened a new avenue for the measurement of entrepreneurship.

Of course there are many other questions that can be answered and many other timely techniques that can be used on the basis of a WHOIS database.

IT security: practice

The practitioners of cybersecurity are continually facing new challenges. In the last few years, for instance, several new types of attacks came up which cannot be fought against by “traditional” approaches such as firewalls. Malicious agents frequently collect data by phishing or other methods based on a distributed attack from many sources, targeted at the infrastructure of a company or organizations. The only way to reveal or prevent such attacks requires the use of fraud intelligence.

In a phishing mail campaign, for instance, the adversaries send emails to addresses from various domains to the accounts of the targeted entity with the purpose of obtaining sensitive data. Frequently these emails contain addresses of malignant websites which implement the collection of data or contain viruses to be distributed on the client machine. The detection of such attacks requires a special approach. And WHOIS data are of crucial importance here. From a proper database one can find malicious domains which have a short lifetime and are registered by a given subset of entities with the purpose of hosting the malicious websites and sending phishing mails. Clearly the analysis of bulk WHOIS data with specialized techniques can reveal the activity of malicious agents.

In order to capture potential attack in a sophisticated way, even some industry-

leading solutions recommend the use of the data of WhoisXMLAPI. The basis of IBM's key security solutions is the QRadar Security Intelligence Platform, a security information and event management system (SIEM). This is a unified platform covering many security-related tasks and incorporating a broad spectrum of solutions including the use of X-Force® Threat Intelligence. In an online tutorial (6) it is demonstrated how the big data extension of QRadar can be used to do DNS forensics in order to identify risky domains, risky users ,risky IP addresses and feed this information back to QRadar in order to define new protection rules. The solution requires an API subscription at WhoisXMLAPI.

Based on a WHOIS database one can introduce various quantifications of the reputation of a domain or a website. Reputation scoring is readily made available through APIs. When having a full WHOIS database, one can develop custom approaches to this problem.

But not only this, as we shall point out in the next section, based on bulk WHOIS data, cybersecurity research directs the attention to various potential vulnerabilities having a significant potential impact in some cases.

A very extensive and timely source of information for security experts is IBM X-Force (7). Browsing the documents available it is probable that the relevance of WHOIS data in this field cannot be overestimated. The MITRE corporation, for instance, has developed a pivot table tool for analysts and researchers of cybersecurity to work with WHOIS data obtained from WhoisXMLAPI. Their tool named WhoDat (8) is freely available under the General Public License and can be used for a variety of tasks. In Ref. (9), for instance, it is demonstrated how to use it for searching on a spear-phishing link domain.

IT security: research

IT security has its implications for fundamental research sometimes with important practical implications. There are several scientific papers in the IT security literature in which the results are based on Whois data. We mention a few recent ones of these as an illustration. In all these projects, data from WhoisXMLAPI were actually used. Similar applications can be found also amongst our Customer Success Stories.

The protection against malicious websites is an important task in cybersecurity. A common way of identifying such sites is the use of blacklists which contain a large set of URLs considered as dangerous. There are various techniques for compiling such lists, and there is obviously a need for methods to verify if a suspicious site is

really dangerous. In Ref. (10) researchers of University of Calabria, Italy have recently proposed and demonstrated an efficient machine-learning approach based on WHOIS data to the generation of blacklist.

There is a large variety of emerging challenges of cybersecurity for instance, the prevalent use of Secure Socket Layer (SSL) relies on the private key of the entity to be trusted. When communicating e.g. a web-shop via https our security is guaranteed by the private key of the shop which is frequently stored by a third party, the hosting provider for the shop in our case. This obviously raises cumbersome security questions. The first Internet-wide survey of the situation has just been published recently (11), based on WHOIS data and several other data sources. The research reveals a complex picture on the HTTPS ecosystem and its potential issues due to private key sharing and elucidates a number of challenges to cope with in the future to maintain the possibility of secure communication over the web.

As another example of a security issue which has been identified by researchers recently partly using WHOIS data, in Ref. (12). This is about so-called WPAD (Web Proxy Autodiscovery) protocol which is prevalently used to configure the web proxy settings of end systems such as desktops and other devices belonging to an administrative domain, e.g. a corporate network. This otherwise useful protocol can introduce the possibility of a man-in-the-middle attack when the corporate's internally used domain name coincides with a top level domain. To identify this vulnerability or identify malicious agents preparing for such an attack one needs to query a WHOIS database.

Other areas

In addition to the aforementioned fields there are several other areas in which WHOIS data bear fundamental importance. Here we briefly mention some other cases, without the need of completeness.

It can be considered as a part of cybersecurity, but a very delicate area: bank transaction fraud detection. Here, banks and payment processors definitely need to identify physical entities associated with IP addresses. They frequently augment these data with those from geolocation information systems to find suspicious transactions.

Another important group of experts who use our WHOIS data consists of criminal investigators and lawyers. When untangling the details of some crime or fraudulent activity, the details of digital communication are often encountered. It is important

to probably relate these details of the virtual word of Internet to actual people or organizations. Check out the relevance of e-mail registrations in the Hillary Clinton e-mail controversy (13) to see a really high-impact case...

How to obtain BULK WHOIS data?

There is a dedicated Internet protocol to access WHOIS data. You may of course use this directly with a client. As long as you need the data for a few domains, this works well. But if you are in the need more WHOIS data, you can easily run into a trouble. Many operators pose limits on the frequency of WHOIS queries. You may ask for a few domain names within a time period. If you are a web service provider, you cannot make your client wait just because you ran into this limit. Also, the WHOIS protocol is filtered out by many firewalls on the Internet, so the data maybe just unavailable this way. There is a solution though, subscribe to an API serving WHOIS data. Our APIs provide up-to-date data through RESTFUL interfaces in popular formats such as JSON or XML., however, you do not want to make external API calls, it's a better idea to set up a WHOIS database and keep it up-to-date.

Also, if you are doing some research or analysis requiring WHOIS data for a longer period, you are in a trouble just using WHOIS protocol. In addition to the obstacles in the previous paragraph, the queries will give you actual data only. Historic data are not available this way.

In many of the described applications bulk WHOIS data are queried and analyzed in sophisticated ways. This typically requires to set up a database, either a relational one or some noSQL approach such as MongoDB, ElasticSearch or solr. This requires data readily available in a format which is compatible with the chosen too.

So if you are in the need of a WHOIS database, you can subscribe to a suitable download plan at WhoisXML API. We provide separate technical blogs describing how to set up your database. You can also learn more about business use cases in our guide "WHOIS Database Download: 13 Business, Cybersecurity, and other Applications Explored."

References

1. **ICANN.** Internet Corporation for Assigned Names and Numbers. [Online] https://www.icann.org.

2. —. ICANN WHOIS website. [Online] https://whois.icann.org/en.

4. **Wallace, Tracey.** How To Choose a Good Domain Name that Embodies Your Brand (and Amplifies SEO). [Online]

https://www.bigcommerce.com/blog/how-to-pick-domain-name/.

5. Quantifying Entrepreneurship Using Domain Name Registration Data: Methods And Applications for Oxford and Cambridge, UK. **Abhishek, Nagaraj and Sibo, Wang.** United Kingdom : NESTA, 2017, NESTA Working Paper, Vol. 16/02.

8. **MITRE Corporation.** WhoDat Project. [Online] 2013. [Cited: december 11, 2017.] https://github.com/MITRECND/WhoDat.

9. **Shields, Wesley.** Using WHOIS and Passive DNS for Intelligence. [Online] february 5, 2015. [Cited: december 11, 2017.] https://www.mitre.org/capabilities/cybersecurity/overview/cybersecurity-blog/using-whois-and-passive-dns-for-intelligence.

10. Malicious URL Detection Via Spherical Classification. **Astorino, A., et al.** Suppl. 1., s.l. : Springer, June 2016, Neural Computing and Applications, Vol. 28, pp. 699-705.

11. Measurement and Analysis of Private Key Sharing in the HTTPS Ecosystem. **Cangialosi, Frank, et al.** New York, NY, USA : ACM, 2016. CCS '16 Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria — October 24 – 28, 2016 . pp. 628-640. ISBN: 978-1-4503-4139-4.

12. MitM Attack by Name Collision: Cause Analysis and Vulnerability Assessment in the New gTLD Era. **Qi, Chen Alfred, Eric, Osterweil and Thomas, Matthew.** s.l. : IEEE, 2016. 2016 IEEE Symposium on Security and Privacy (SP), San Jose, CA. pp. 675-690.

13. **Wikipedia.** Hillary Clinton email controversy. [Online] [Cited: december 11, 2017.] https://en.wikipedia.org/wiki/Hillary_Clinton_email_controversy.