WHOIS running the Internet from May 25, 2018 onwards?

Posted on January 15, 2021

The virtual space of the Internet is a relevant scene of our everyday life. And the elements of reality and their virtual counterparts – friends with social media contacts, shops and web shops, companies and websites, etc. – are becoming more and more confusable. Albeit this must have been in principle already expected by the founding fathers and mothers of the Internet, in many respects the Internet has been developing not quite as they had envisaged.

For instance, it had been clear from the very beginning that there should be a link between Internet domains and the real-life people and entities responsible for them. In the beginning, the motivation was mainly technical, of course: if something went wrong on the network, the operators needed to know whom to contact. This demand gave birth to the WHOIS protocol, a standard way to learn who is responsible for a high-level Internet domain.

Table of Contents

1. Introduction

The WHOIS protocol and the policies around it have gone through a number of changes during their history of several decades. Yet it is still a unique link that ties the virtual space of Internet domains to reality. Ironically, in spite of its fundamental relevance, the literature on it is very scarce: there are no textbooks or classroom introductions. One exception is the seminal book of Garth O. Bruhen, whose title is also very expressive: "WHOIS running the Internet? Protocol, Policy, and Privacy". The book offers an excellent account of the topic roughly till about 2015; just till the dawning of the new data regulations like the EU GDPR, which is also mentioned there.

WHOIS has always been amongst the most debated issues in Internet policy, especially because of its privacy and security implications: the data contents can partly be personal. In addition, like certain other protocols of the Internet (including e-mail, plain http, etc.), it is in a way archaic and thus it has technical issues, including the lack of access control mechanisms. Still, it is a fundamental tool for law enforcement, cybersecurity professionals, researchers, and trademark and intellectual property rights holders, to mention a few.

The EU GDPR is a complex legislation whose main goal is to ensure a high level of privacy in a digital era, which is certainly appreciable. However, its interference with the WHOIS system has proved to be destructive by many experts in the last few years. It is broadly accepted that a key unintended consequence of the GDPR is through its effect on WHOIS data "it undermines the transparency of the international systems and architectures that organize the Internet" (as pointed out also in this recent study).

Here we present probably the first research results that quantify the extent of this undermining. We illustrate it with particular examples. We comment on the initiatives that aim to solve these

problems; currently this is a developing field with a number of uncertainties. We also point out why WHOIS data are still fundamental in spite of their partial degradation.

2. Information in WHOIS data

Besides answering the fundamental question of "WHOIS responsible for this domain", the WHOIS service provides additional fundamental data that luckily, nobody seems to have considered as confidential so far. Very important are the registration and expiry dates: they can be used to find out the age of a domain which is an important feature for its appraisal or security assessment, and it can be found out when the domain will possibly be up for sale. The domain's status codes are also relevant, especially if the data source (the registry or registrar) uses ICANN's standard codes. They show if the domain is in a critical phase of its life cycle, and it can also reveal if there are some legal debates about it. The registrar's identity is also typically revealed; this remains as a stable starting point if someone has to investigate who is behind a given domain: these are companies and they have no interest in hiding their name.

However, the question of "who" is related to the registrant in most of the cases. Traditionally we are supposed to have the name of the registrants, the organization, and various contacts including e-mail, phone, and conventional mail, for the registrant, for its administration and the domain's technical personnel. And this is where privacy comes in: to what extent a person's or a company's name should be publicly revealed in the WHOIS system?

A few decades ago it wasn't a problem if an organization had nominated e.g., an administrative contact with a real name, e-mail and other data. Nor was it a problem that WHOIS had revealed the organization the domain belongs to. Of course, our relation to data has changed a lot; and there are also valid concerns that such an open attitude would rise. What if a given person does not want certain individuals to contact them or does not want to reveal who they are working for. Also, what if a company is running various brands and does not want competitors to identify the link between the different brands' pages? Surely, many other reasons can be found.

On the other hand, in many cases, there is no reason to hide just one part of your identity. A company, for instance, unless it has a special reason as the one just mentioned, has typically no interest in hiding the fact that it is in possession of a domain, nor it should be interested in not providing administrative or technical contacts for the domain, without including persons' names

and private data of course. Indeed, the WHOIS data of the official pages of big players like Google, Microsoft, Apple, etc. will all have this information. Also, if an individual owns a domain with benign purposes, in most cases there is no valid reason to hide the identity fully from the public.

So in our research described in what follows, we will check WHOIS data primarily from the point of view of whether the registrant's identity is at least minimally revealed.

3. Measuring the effect of GDPR

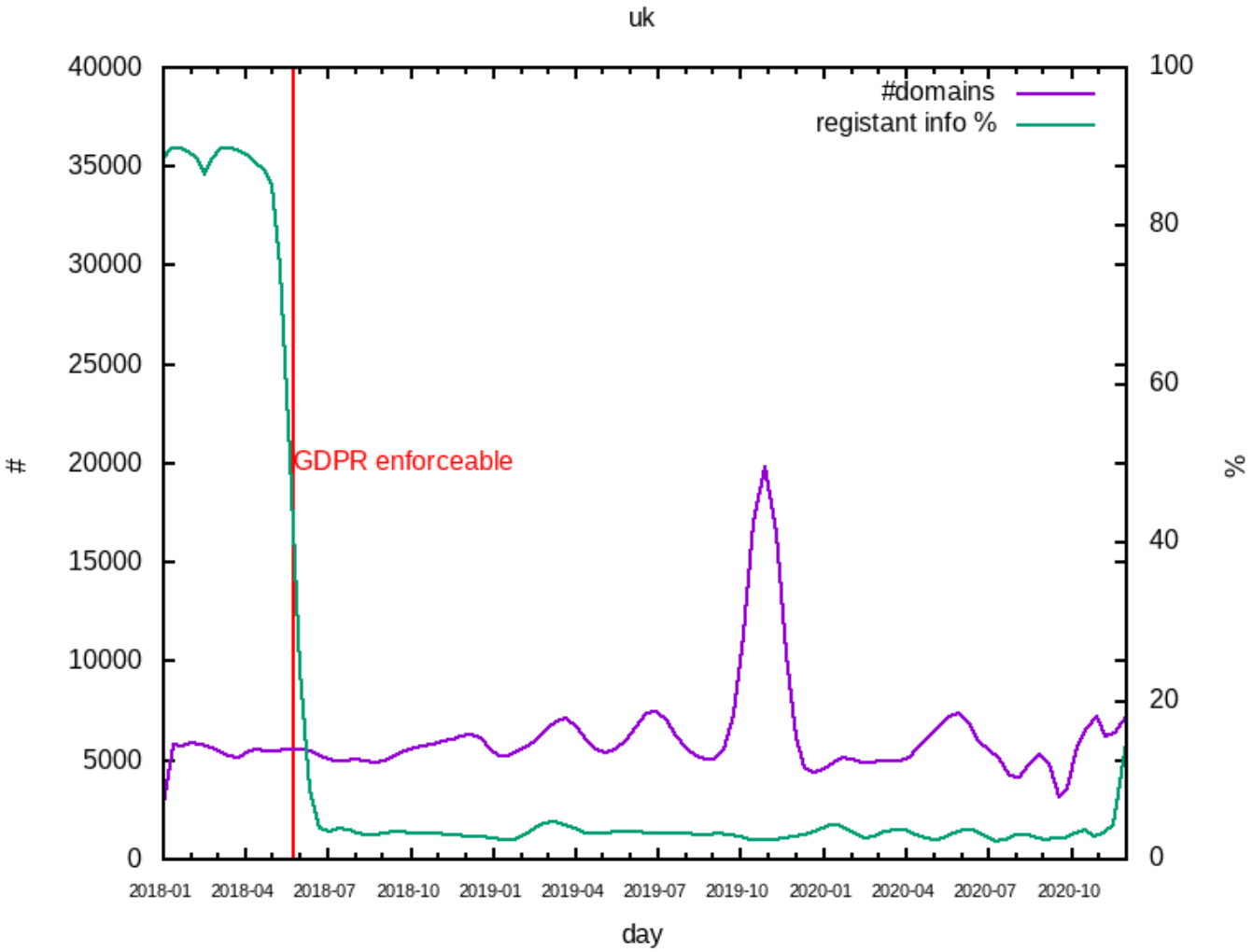
Our data sets come from WhoisXML API's daily data feed archives. While the WHOIS service provides information in a free textual format, WhoisXML API data sets are structured and parsed, so the presence/absence or quality of a given piece of information such as the registrant's name or the registration date can be easily accessed.

These data consist of daily collections of WHOIS records regarding domains that started to resolve in the Domain Name System on a given day. So our dates will not be official registration dates but dates when a given domain that was not technically accessible the day before started to be accessible. This day does not necessarily coincide with the registration date but it is usually very close to it. In some cases, a domain just disappears and reappears (e.g., if it is in its Grace Redemption Period); the reappearance will also be detected. In spite of not dealing with the official registration dates from the WHOIS records, we know from experience that the domain counts are very similar to those calculated from the WHOIS data, and the trends say the same.

Our primary criterion will be the ratio of those WHOIS records on a given day in which either the registrant_name or the registrant_organization field is not empty or not redacted. We verify the fact of being redacted by checking the presence of certain substrings found on an experimental basis. For instance, ICANN recommends using "REDACTED FOR PRIVACY" in redacted fields. However, "GDPR masked" is also prevalent; it is not standardized. Hence our approach has a certain chance of failure: we'll not detect redaction if it is e.g., stated in a different language, and we'll falsely detect a Mr. Redacted's domain as redacted. Still, our research well reflects the trends.

3.1 Results: .uk

We will start with a European country-code domain. We opt for that of the United Kingdom for several reasons. The WHOIS records' text there is typically in English, decreasing our probability of failure. The UK had a tradition in producing informative records (unlike some other ccTLDs in Europe like .de or .hu, which have had a long-standing tradition in hiding much of the information). The country has been a member of the EU in the given period but this might change. Finally, WhoisXML API is in hold of good quality data sets from the beginning of the year 2018 onward. Our results can be seen in the following figure:



The left vertical axis is for the violet curve of the daily number of domains, while the right one is for the green curve of the ratio of WHOIS records meeting our criterion of having minimal registrant information. All the curves are plotted after applying Bezier smoothing to eliminate rapid changes and outliers.

The red vertical line shows the emblematic date of 2020-05-25 when the EU GDPR became enforceable. One can see that the disappearance of registrant information was sudden and dramatic. Of course, in the view of the sanctions that the EU GDPR promises upon its violation,this is partially understandable. However, let’s bear in mind that neither the registrant's name norits organization is personal data. Hence the following questions naturally arise:

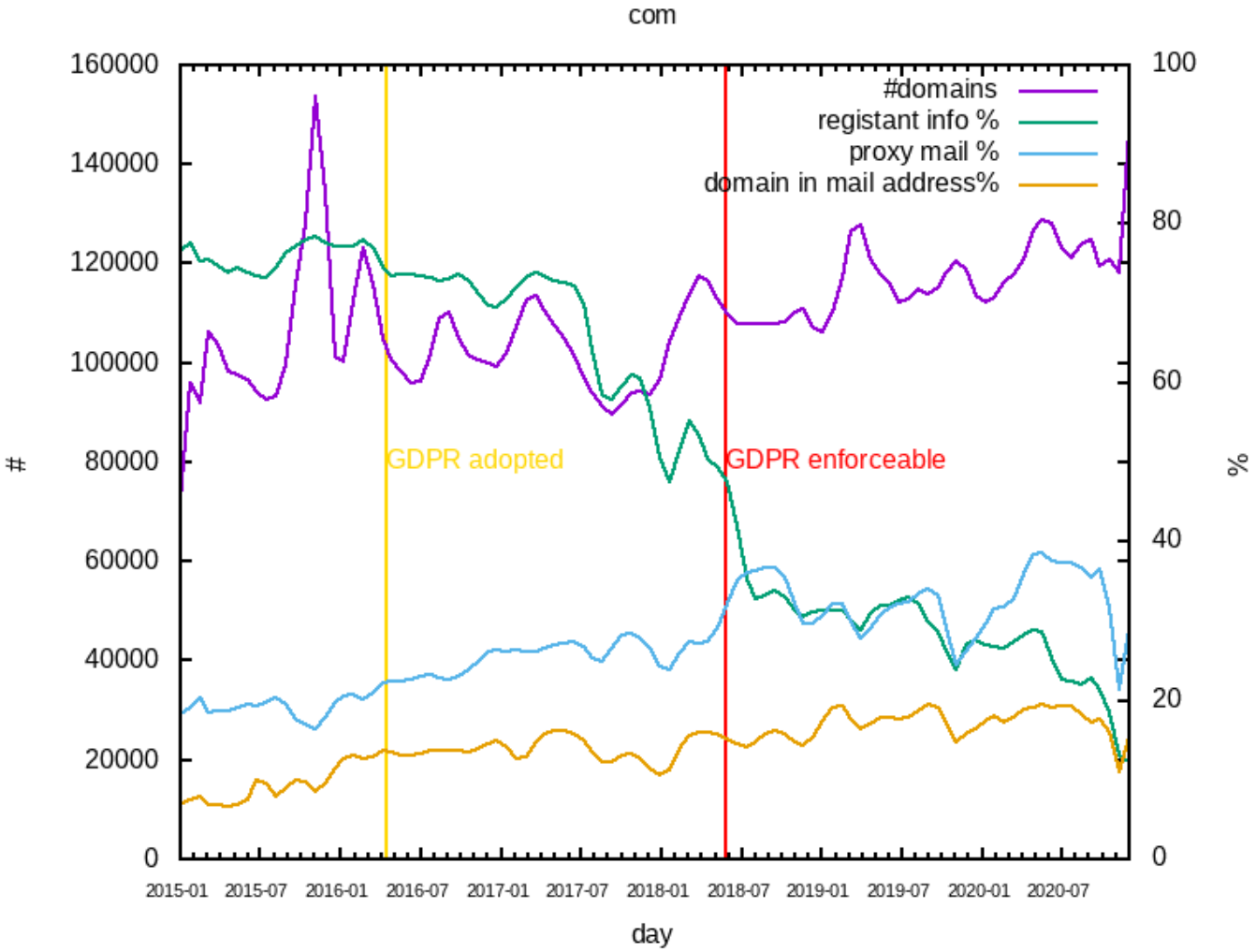
- Isn't the GDPR just a good alibi for hiding information that would not be necessary to hide?
- Has it become part of the registrar's policy to hide everything by default, "just to make sure"?

We shall return to these conjectures later.

One might think that the slight recent increase of the ratio might be due to Brexit, but it is not the case: it is just the nomination of a company interested in the domain business as a registrant in many domains. As a final side remark, in this data set the labels such as "REDACTED FOR PRIVACY" are not used; the records simply became empty.

3.2 Results: .com

Having evaluated a prestigious ccTLD, let us now turn our attention to the biggest and maybe most relevant generic top-level domain: .com. In this case, the WhoisXML API database has data from as early as 2012, but we will look into the last 5 years. Our results are presented in the following figure:



The format is similar to the previous one except for the gold and cyan curves that will be discussed later. Let's focus on the green curve first. Now we have two critical dates that are covered: April 14, 2016, when the EU GDPR was adopted, and the aforementioned May 25, 2018, when it became enforceable. It appears that the typical original ratio of records with basic registrant information, which was around 80% before, had started to decrease gradually after the adoption, and went down to as low as about 20% quickly around the date of enforcement. This is our maybe most important observation. While it cannot be fully considered as unexpected, it does actually quantify the undesired effect of the GDPR.

There are some unexpected features though. As .com is a very common internationally used

domain, the effect could have been smaller. After all, the EU GDPR applies to the personal data of EU citizens, whereas the examined WHOIS data could well be company and organization names, not necessarily belonging to the EU. To have a clearer picture, we have randomly chosen a recent particular day from the GDPR era: November 20, 2020. (Other choices of days show a similar behavior.) On that particular day there were 126,183 .com domains appearing as new in the DNS. Out of these only 19,043 (15%) had either a registrant name or a registrant organization nominated. Meanwhile the top values of the "registrant_country" fields have the following counts:

country	count	percent
Unknown	56221	45
UNITED STATES	27534	22
CANADA	9582	8
CHINA	8439	7
PANAMA	4267	3
INDIA	1598	1
UNITED KINGDOM	1505	1
JAPAN	1439	1
FRANCE	1127	1
NETHERLANDS	1073	1
GERMANY	834	1
TURKEY	819	1
SPAIN	819	1
ITALY	753	1
BRAZIL	653	1
HONG KONG	579	0
UK	560	0
AUSTRALIA	502	0
MEXICO	493	0
KOREA, REPUBLIC OF	430	0
INDONESIA	397	0

While the 45% of unknown registrant countries is also striking, observe also that at least 44% of

the registrants are from non-EU countries. It is known of course that similar legislation is planned in several other countries. Nevertheless, it is in stark contrast with the 15% of the records revealing even very basic identity information.

Other days show similar figures, which seem to underline the conjectures put forward in the previous section. Meanwhile, we can conclude that the EU GDPR destructs the transparency of Internet operation and organization both directly and indirectly. The direct effect is that certain data of European entities has been deleted, as we have seen in the case of the .uk domain. However, perhaps an even more mischievous effect is that it made unnecessary and unreasonable masking of certain WHOIS records accepted and prevalent.

Let us now move on to the gold and cyan curves in our figure. It has been broadly accepted for quite a long time that providing unprotected mail addresses in WHOIS data, like in any public database, will immediately result in a tremendous amount of spam received by that e-mail account. Hence, there has been a significant trend long before the GDPR to hide the mail addresses somehow. Options include:

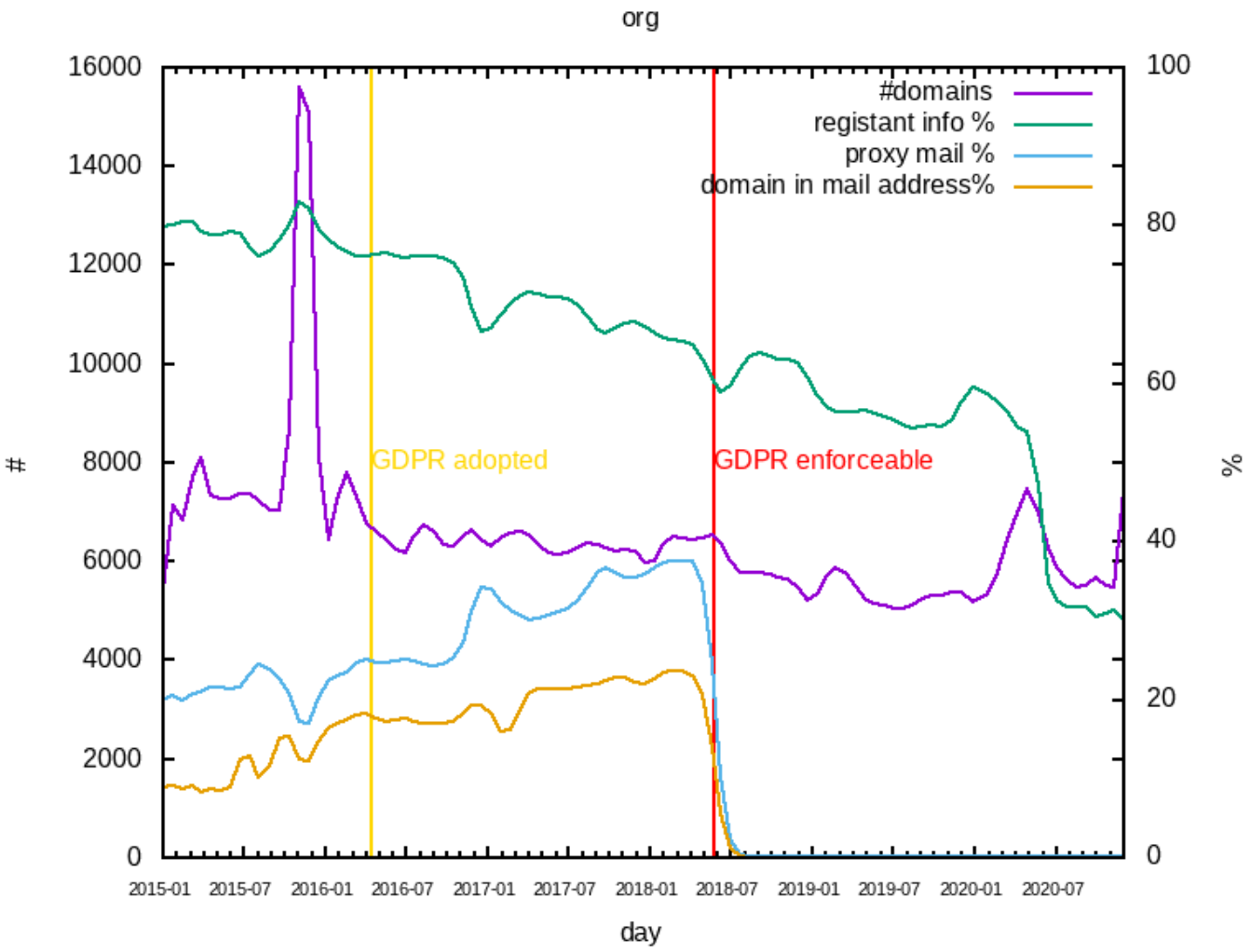
- not to provide any address
- use a generic address, e.g., the abuse address of the registrar, or
- use a specialized service providing a proxy e-mail

From amongst these, maybe the last one is the most cultivated approach: while not revealing an e-mail address used by the registrant for other purposes, it provides a unique e-mail address that can be used as a contact. Our cyan curve shows the approximate ratio of records where such a proxy address is provided. It is good news that this ratio is not very low, and shows a steady increase which has even been accelerated by the GDPR. The gold curve shows the ratio of records with proxy e-mails where the e-mail text contains the domain name as a substring; these are surely specific to the given domain. (Another typical solution is to have a client code in the e-mail address, so many of those represented by the cyan curve are also likely to be specific.) So while the basic information about the registrants in .com is dramatically suppressed, the situation with contact e-mails is not that bad. And the relevant dates, status codes, name servers, as well as the data of the registrant are still mostly available, so even though WHOIS seems to have lost a part of its glory, it is still a crucial source of information. While this latter statement applies to most

of the TLDs, the fine structure of GDPR's effect can be different. Let us turn our attention to another important domain for an illustration.

3.3 Results: .org

We now opt for the .org domain which is also a big, relevant, and traditional gTLD, founded as early as 1985, and is used mostly by non-profit (and some for-profit) organizations. Our curves for this domain look like this:



The increase in the ratio of records without basic registrant information shows a decreasing trend here: rather plausibly, organizations tend to have less interest in hiding themselves. What is somewhat surprising though is that the specific e-mail addresses suddenly disappeared with the enforcement of the EU GDPR. In the records, the contact e-mail is either empty or is directed to the abuse address of the registrar. It seems to be a somewhat obscure new policy.

What we can learn from .org's data is that different domains are differently affected by the EU GDPR and its indirect consequences. In this case the possibility to identify the registrant is less marked. On the other hand, it requires more effort to get a working specific e-mail address. Alternatively, one can look for a contact in WhoisXML API's website contacts and categorization database.

4. Will ICANN (and RDAP) save us?

Aware of the potential problems the EU GDPR will cause, ICANN came up with a Temporary Specification for gTLD Registration Data, which is essentially a working document to cope with these issues and pave the way to a new and proper approach to the regulation and maintenance of the transparency of operation on the Internet. While the problems, as we have just pointed out quantitatively, emerged immediately after the enforcement and their indirect effect is likely to be stronger than it was anticipated, ICANN's specification is still temporary. This means that nobody knows the royal way yet, and the new approach has not settled completely.

It is beyond the scope to analyze in detail the contents and status of the Temporary Specification. Nevertheless, the intention to resolve certain issues of the archaic WHOIS system is apparent. Namely, it defines roles and access rights, and it addresses the specification of the data which must be public and which have to be kept private. Let's hope that it will boil down to a reasonable regulation acceptable to all players (including those who had been in fundamental need of the data that have disappeared from WHOIS for perfectly benign and important purposes), and that they shall succeed in maintaining the aforementioned transparency.

Apart from the operational and legal considerations, there is a highly technical part of this progress: ICANN required gTLD registries to implement an RDAP service by August 26, 2019. RDAP stands for "Registration Data Access Protocol" and it is intended to replace the old WHOIS

as a state-of-the-art solution that can provide data in a parsed form (like JSON directly) and has the appropriate access control solutions to maintain WHOIS data's security and privacy. Obviously, the Temporary Specification envisages the future with RDAP as the primary source of registration data.

While RDAP had been standardized in 2015, it never took off. One would expect that, more than a year after ICANN's requirement, RDAP should now be a valid alternative to WHOIS. So let us go for a test drive.

While a WHOIS client is naturally present or easily installable on any contemporary operating system, with RDAP one has to rely on software downloaded from various sources. For instance, no RDAP client is available from a standard package in the current Ubuntu Linux distribution. ICANN itself has a web page for RDAP users from which a command-line client written in the "go" language can be downloaded. Python users may rely on the "rdap" package coming with a stand-alone command-line utility which is at least easily installable with Python's package manager on any platform.

After having installed the client, one has to face that while it will work for major gTLDs like .com or .org out of the box, for most new GTLDs like .work, for instance, it will not provide any result, not to mention ccTLDs. Tampering with the configuration to include more RDAP servers may improve the situation. However, in the case of WHOIS, it just works trivially.

Another surprising issue is, e.g., when querying google.com with both WHOIS and RDAP, one has to face that while the WHOIS record clearly reveals that this domain is predictably owned by "Google LLC" as a registrant organization, for the same domain, at least at the time of writing of this blog on December 4, 2020, the RDAP query will not reveal this secret. To get this, one still has to rely on a WHOIS query resulting in unstructured textual data, or get it in a processable form e.g., through WhoisXML API's services. Thus, we have to conclude that even though RDAP's coverage has increased significantly, it is still far from being a mature tool to readily replace the good old WHOIS.

We also remark that while RDAP provides structured data, thereby eliminating a significant shortcoming of WHOIS, the RDAP servers still do not facilitate a "reverse WHOIS search", e.g., for the search for all domains expiring on a given date. And it does not facilitate the building of a local database from public data to answer such research questions: frequent queries run into

throttling limits quickly on RDAP servers, just like in the case of WHOIS.

5. Conclusions

We have analyzed quantitatively the structure of WHOIS redaction, with an emphasis on the negative effects of the EU GDPR, through which it undermines the transparency of the international systems and architectures that organize the Internet. To our knowledge, no such quantitative assessment was made so far. We found that the impact is indeed significant, and it has a different structure in different TLDs. One of our significant conclusions is that the effect is partly indirect, and it consists of making deliberate data reduction generally accepted and preferred. The future regulations that are supposed to cope with these effects are not yet settled. These regulations also point towards the replacement of the outdated WHOIS protocol with a new one, RDAP. We have found that RDAP is not yet mature and it does not better support the building of a database from certain public data for research purposes. In spite of the undeniable damages the WHOIS ecosystem has suffered in the last few years, at least for the time being it remains the unique and thus crucial source of fundamental data that is required for the transparent and secure operation of the Internet. And most of its technical shortcomings can be overcome by using various services of WhoisXML API. The development of these services and products follows and continuously incorporates the progress in the implementation of the aforementioned regulations and techniques. WhoisXML API conducts a significant R&D activity in the search for alternative information sources to support legitimate and benign use cases for e.g., scientific research, cybersecurity, or legal investigations, which require data that used to be available in the WHOIS system.