BARCELONA –GAC: Technical Seminar on WHOIS and Data Protection/Privacy Issues
Sunday, October 21, 2018 –17:45 to 18:30 CEST
ICANN63 | Barcelona, Spain

GÖRAN MARBY: Hello. It's the guy sitting up on the podium. Okay. My name is Göran Marby, for the one who makes the transcripts. You identify me on my Swenglish. I'm waiting for that transcript. I have a special translation from Swenglish to English. Hello, down there.

So, I'm happy to see so many people here, and hope that you are here because we invited you and not because you don't know how to leave the room. Because, the reason behind this seminar, is over the years -- I don't know about this, but over the last year or so there's been a lot of discussions about a system called WHOIS, and we have received many, many questions about WHOIS system, how it works technically. So this is not about talking about policy, as you know, probably know, do we have an expedited PDP right now who works with that.

We're not talking to you about the different interests about WHOIS, but for interested parties, with the help of David Conrad, our CTO and Head of IANA, to actually go through, "What is WHOIS; where does it come from; what kind of data is in WHOIS system; what information is today shown in the WHOIS system; where is the information stored?" Many times, there are many of those questions, it's quite fundamental. And I'm also here to answer questions, if there are questions about how we interpret the law today, I can answer those questions well. But it's really to give a basic about one of the oldest systems in the world of Internet, the WHOIS system. Doesn't that sound fun? I can see the joy in your faces. Okay. With that, I'm handing over to David Conrad.

DAVID CONRAD: Thank you, Göran. This hopefully will be entertaining in the sense that it will keep you awake. It is talking about stuff that you may already have a pretty good handle on. Hope to provide sort of a different perspective. I've been doing this Internet stuff since around 1983 and WHOIS actually predates me on this planet. So, there is a lot of history here and I hope, if you have any questions, that I will be able to answer them, at least on the technical side, I will gleefully hand over any policy related questions to my boss. Next slide, please. There we go.

So to start, just an example of some email that people receive occasionally, and this is instructive because of the domain names that you see within them. As you can see on these samples there are names that appear to be somehow related. GoDaddy names that appear to be related to Amazon, names that are related to Netflix. And one of the challenges when you are an end user or an ISP, is trying to figure out whether these are legitimate or not. Next slide, please. There we go.

WHOIS is a mechanism, a tool that will help, at least was designed to help people, in identifying the origin of the resources that are used on the Internet. Those resources are typically identifiers and in this context those identifiers are domain names which are obviously used by people, and the Internet addresses which are generally used by computers. These identifiers are obtained from registries in the domain name system, which most of you are probably very familiar with, there are about 2,500 registries and there are a few more registrars that are selling those names to end users. The registrars are quite familiar, GoDaddy, NameCheap, those sort of folks. The registries are the folks like Verisign, and Afilias, and [inaudible], JPRS and those sorts of guys.

In the addressing world, there are five regional Internet registries, AFRINIC, APNIC, ARIN, LACNIC and RIPE NCC. And these registries' primary job is to remain the registration data for these Internet identifiers. I should say, one of the their primary jobs is to maintain registration data for these identifiers. That box on the right shows what the registration data generally looks like. You can see, within that box, it will have the name, this is a registration data firm, .US. You have the registrar, where the domain name was purchased from, various technical information, and then contact information which, I'm sure, many of you are now much more familiar with because ofJDPR. Next slide, please.

GÖRAN MARBY: Can I make a comment?

DAVID CONRAD: Oh, please do.

GÖRAN MARBY: So, just to point out, something is going to be repeating. There is not one WHOIS database. We will come back to that site later, as well. There are in our domain name system about 2,500 different databases that, through a technical protocol called WHOIS you get access to it. And that's so ICANN doesn't have the WHOIS database. The other thing is, it's not end users, we don't have a system of 4.2 billion users, it's only the ones who bought a domain name who is in the WHOIS database. So, to make it in simple terms, today there are 4.2-4.3 billion users, and as I said, I've actually no idea where that numbers came from, it was given to me. Out of that, there are about 385 million domain names, I think. Whereas about 185 million of them are connected to the ICANN WHOIS system and about 60 million of those domain names are connected to ICANN in Europe.

DAVID CONRAD: Thank you. Next slide, please. So, talking about GDPR, there was, obviously, an impact from GDPR on the availability of this information located within this multitude of databases. Prior to GDPR, is on the left, and this is for EFF.org, and you can see some common information like the domain name itself, when was created, the expiration date, but the contact information has radically changed. It's gone from having a full address and telephone number which were required, optional fax numbers, email address that was required, to merely having the registration organization and the state or province and country from which has registered that domain name.

Now, if you go back to the idea of using the WHOIS database to try to identify resources, you can get an idea of the registrant, who the organization of the registrant is, but you no longer have the contact information and you no longer have a way of finding out from the WHOIS data how to contact the individual who is responsible for that domain name. Next slide, please.

So, why does this matter? Going back into ancient, ancient history, the Internet was very small and comprised primarily of network researchers who worked at universities or research institutions and they dedicated lines and generally everybody knew everyone else and they didn't really need a database to maintain contact information, because they were like ten people, and you knew them by name, and you knew their telephone numbers by heart. Back in those days, connectivity problems were quite common, that was actually surprising, that things were actually kept up for any length of time, and any time there was a problem, you know, one of the researchers would simply call up the another researcher and say, “Well, it looks like your computer is down again, so you go fix it.” Next slide, please.

Moving on in time, I believe this was in the early 80s, no, 1977, the network gets much larger, there are more researchers, it’s still primarily an academic and research network, but connectivity problems still exist, systems go up and down all the time, and we’re starting to go migrate into host names. Not yet into domain names phase, the domain name system wasn’t invented until 1983, but we have host names that are associated with specific machines, but there are way too many researchers for anyone to keep their information in their brain, so  people start writing them down into telephone books and Rolodexes, and that sort of thing, but that doesn't scale very well and people keep moving around and information gets out of date.  So, around this time, people decided to create what they called a reverse telephone book, that was maintained by  SRI-NIC, which was a Stanford research international network information  center, and it would translate the IP addresses that people receive, when they look at a packet they see the source and destination addresses, into the contact information associated with that address in order to help people identify who is responsible for the IP address, if it happened to be not responding or if it happen to be  flooding them with traffic or causing some other problem.

The network access to this reverse telephone book was via a new protocol called WHOIS. Now, the WHOIS protocol is a -- calling it a protocol is actually a bit generous, because it's perhaps the dumbest protocol that you could possibly  have.  It doesn't actually specify anything other than, you send a string and you get back result, it doesn't actually describe what the result is, it doesn’t describe what the query is that you send into it. Next slide.

Around this time the Internet starts becoming a bit more annoying because more and more people are connecting to the Internet, more and more people start doing fun and interesting things to it.  In 1988 a guy named Robert Tappan Morris created the first really large scale worm that went scurrying around the Internet, it was actually a proof of concept to demonstrate vulnerabilities in software at the time, and it happened to escape.  That worm actually ended up taking down quite a number of machines but it did cause network administrators at the time to desperately look up contact information on the people that slamming them with, you know, traffic trying to -- of this worm trying to break into their systems, to try to get that stuff to stop.

In addition, in 1994 the first large scale unsolicited commercial email also known as spam was sent. That was from a legal firm, Canter and Siegel, advertising for the green card lottery in the US and that generated a tremendous amount of traffic, both in relation to the green card spam going out, but also, people complaining about the green card spam, resulting in quite a bit of contact attempts for the email providers that were actually forwarding this stuff along.

WHOIS at the time was used to try to track down the contacts associated with the source of the spam or the source of the attack in the Morris worm to try to get those folks to stop. Next slide, please.

In 1991 the Internet sort of explodes, the worldwide web allowed almost anyone to become a publisher of information and to become a source of content and the Internet grows very rapidly. The centralized allocation of names and addresses couldn't keep up. Prior to this, in general, there was SRI-NIC, moved over a couple of places, eventually ending up in something called InterNIC, and it was more of a centralized database with most of the contact information being kept in one place.

Around 1993-1994, the regional Internet registries sprung up for a distribution of the IP addresses out of the InterNIC. ARIN actually started up after RIPE NCC and APNIC. And these were independently operated databases. The domain name system similarly began to be partitioned out around this time. There was, prior to the '90s there was distribution of the creation of ccTLDs which maintained their own databases and that was just expanded as more top-level domains were added.

There was and is a single point of connection across all of these databases, and that's IANA, but IANA does not contain any information other than referral information. So, for example, if you're looking for a domain name that happens to be in .com, you can go to IANA and have information that will tell you where to go to look up information in .com, but IANA itself does not contain anything other than that referral information. Next slide, please.

Today the Internet has over 700,000 individual networks connecting 4.1 or 2 or 3, depending on how you count, billion end users, over 350,000 domain names, over twenty -- sorry, 350 million, what did I say? Thousand. That's slightly, it's close. Over 2,500 top -- having problem with units today -- over 2,500 top-level domains, that you are all familiar with. The Internet itself remains peer-to-peer. Anyone can create content, and every end point on the Internet can provide services. And if you remember the rationale, the reason that WHOIS was created, it was to be able to track down the producers of content, primarily to help and identify sources of abuse or network interruptions to try to help people remedy that, make sure the Internet continued to operate. But, you know, today, with over 350,000 names connecting 4.1 billion people on 700,000 networks, there is no single database that could support all of that.

And unfortunately, figuring out whether a source of traffic is a friend or foe has become increasingly difficult because of the bad guys being able to mimic anything the good guys do to take advantage of the way people make use of the Internet. But that's become increasingly important, particularly when you look at, you know, the promulgation of fake news, the various denials of service attacks, the various forms of phishing that people see on the Internet today. Next slide, please.

So, what is WHOIS? It's actually an umbrella term, it defines a protocol. As I mentioned, one of the dumbest protocols that has ever been created. It's actually specified post facto, long after WHOIS was actually developed in RFC 3912 by the Internet engineering task force, but WHOIS is also a sort of a generic term for the database of registration records that's defined by ICANN via contractual requirements and by the RIRs within their databases, and by the ccTLDs within each of their databases. This database actually contains information about who owns, and I don't want to get into the question of ownership here, but, you know, conceptually has responsibility for the names and addresses being used on the Internet.

Prior to 25th of May, that information was publicly available and it was used primarily -- well, some people argue, it was used primarily to facilitate network administration and tracking down various abuses and uses of the Internet. It's also true that the database was used as a source of email addresses for spam and other unfortunate purposes.

The uses of the WHOIS includes providing contact information for identifiers and resources on the Internet. You could sort of think of it somewhere to a land ownership registry or registry of companies. And it's primarily used, as I mentioned, for tracking down connectivity problems and abuse. WHOIS has increasingly been used to help verify sources of information on the Internet, to try to identify whether or not the information being sent came from a reputable source, mapping sources of network information into their real world sources and, of course, the WHOIS has been used to identify trademark and copyright violations.

One key thing to note, however, is that for the people who were required to offer WHOIS, it's not generally a profit center, it's a cost center, it's something that they don't actually generally make money for, because it was offered as a public service. It was something that they were required to deploy in order to facilitate network administration. Next slide, please. Might have been the last slide. No, one more.

So, the future of WHOIS. The demand for the functionality that's provided by WHOIS, the ability to identify the sources of traffic and facilitate the contacting of that source in order to make it stop, for example, is increasing. The -- more bad guys are hitting the Internet, more misconfigured devices. It's not always bad guys, you know, sometimes it's just mistakes. Is increasing, just with the growth of the Internet.

The protocol itself is actually being replaced. There is a new protocol called 'The Registration Data Access Protocol', RDAP. Is being deployed actually, eventually replace and hopefully burry in the backyard the protocol known as WHOIS. It has a whole bunch of very useful features, it allows for differentiated access. You can specify credentials to allow people to get in different views of the database depending on their credentials. It allows for automatic referrals.

One of the big problems with WHOIS as it exists, is you have to sort of manually figure out, "Well, okay. I go to IANA and it tells me where com is, I go to com, it tells me where a sub-domain of com exists, I sometimes have to go to the sub-domain to find out additional information." RDAP has an automatic referral mechanism which will make it a lot easier to sort of traverse that tree of information to get to the actual leaf that you're looking for.

There are, obviously, policy changes. The database is no longer fully public as of May 25th. Access to the full data by authorized entities is something that is continuing to be discussed. You probably might have heard that if you come to the ICANN meetings. What is in the database? Will evolve and it will continue to evolve. New fields are going to be added based on demands, naming within the ICANN world and within the addressing communities within the RIR world, and obviously that will have to be in conformance with legislation from various jurisdictions. Next slide, please. And with that, I will hand it back to Göran.

GÖRAN MARBY: So, when we -- there is a couple of things that it's sort of essentially in the discussion also in the framing of this, is that, it's not a single database. It's not even a single WHOIS, there are several WHOIS. In

Europe, for instance, there is also a WHOIS system for RIPE. It distributes the IP addresses, the IP numbers. So, I think that many times in this -- that's why we started to do this, sitting down comedian thinking; it has to go through so people get a sense of what we're actually talking about.

And there was a couple of things David mentioned I want to reiterate. ICANN as an organization [indiscernible] of the contracted parties [inaudible] for commercial usage. As you can see, the uses out of it database or information is used outside ICANN and it's not bylaws because it predates ICANN that we should have this information. So it's used by a third party and that's actually what creates one of those problems with GDPR because the GDPR is a symmetric regulation about you collect data and you are giving it out with a purpose. We tell the contracted party to have this data, they are according to GDPR the data controller responsible for that data which means that we can't tell them what to do with it.

That's why the first piece of information we got from the DPA, so the right of collection of data was important, but also what makes it so impossible to have a unified access model if you can't change some of the underlying assumptions of this.

But if I get questions about this, I actually refer to myself, because I'm a certified nerd. And my background is information security. So I often, when I get an email -- and I should think you can do this as well -- before you go and click on a link on someone you don't know, you can go in and still see some valid information in the WHOIS system.

it's actually behind that because that's what I have done. You should also of course every week clean your underwear and your cache on your web browser. You do that, don't you? I see everybody's nodding.

So in any scenario we're now talking about access to hidden information, about WHOIS information [inaudible] have anyone like me. And another thing is that ICANN the organization, the org or the institution doesn't have access to more information than what you see here. After the temp spec was enforced and part of the information got behind a wall, ICANN org doesn't get access to that information outside the use for compliance and data retention. Data retention we use if some contracted party's goes down and we can replace them so the end user doesn't get hurt.

Was it interesting? Most of you stayed anyway, so what were you doing, emails? Any questions to us? You have one of the world's leading DNS engineers here. And by the way, is the rumor right that you were actually part of doing the WHOIS protocol and that's why you know how bad it is?

DAVID CONRAD: I can neither confirm nor deny.

GÖRAN MARBY: Good. Anyone?

LUXEMBOURG: It’s Claudine from Luxembourg. I have a question regarding RDAP, where does it stand at the moment?

DAVID CONRAD: The protocol has been standardized by the ATF. There have been implementations deployed within the RIR community, there is a pilot program within the gTLD registries, and I don't recall -- is Cyrus in the audience anywhere? I don't recall the status of the pilot at this stage. They’re working on a specific profile for the gTLD registries, profile basically being which fields within the WHOIS data will be represented whether you’re coming in as a uncredentialed request or whether you have credentials. But the protocol itself works, it's in production in a bunch of places. I believe right now they're just finishing off defining the profiles that will be used within the context of the gTLD registries.

GÖRAN MARBY: The question you are asking is also very smart [inaudible], and that is that many existing policies and many existing work that we do right now is related to also how the expedited PDP will come up with its conclusion about the handing of the WHOIS, because that’s going to be sort of an overarching policy when it comes to handing that kind of data in the ICANN remit.

So some of those things we actually put on hold just to wait out this important work. An example for that is the implementation of thick WHOIS, which is a policy set by the community but it’s very hard to implement that right now before we fully understand what the community thinks about the implementation of WHOIS going forward in the future. [indiscernible] temp spec and the expedited PDP. Hope that was an answer to your question. Anyone else?

UNIDENTIFIED SPEAKER: A banal question to understand the volume of what we're talking. Do you have any statistics of what was the number of access to WHOIS before?

DAVID CONRAD: Unfortunately, given the databases distributed, there is no central point of collection for WHOIS as a whole. Each individual registry will have different statistics based on the queries that the individual registries receive. I don't have off the top of my head an aggregation of those statistics. But that's something that we can try to dig up and provide to you.

GÖRAN MARBY: I can give you, it's a lot. Think about it 185 million domain names in the system; plusThe CCs, plus the RIRs. There is a lot of request every second, every minute for this information, so there are gigantic volumes we're talking about.

DAVID CONRAD: And one point to point out; a lot of the registries in order to combat abuse and in particular the combat people scraping their WHOIS database for email addresses, for use and spam or customer collection, [inaudible] limit. They only allow one IP address query every five seconds or one IP that can send a hundred queries a day or that sort of thing.

So the numbers that you get get sort of distorted because people play games, right? Instead of having just one server query WHOIS, they'll have a whole farm of servers querying WHOIS to try to scrape out data from the WHOIS database, so the statistics that you see for the number of queries are sort of complicated that way but like I said, we'll try to dig something up and provide it to the GAC.

GÖRAN MARBY: Thank you. Any other questions? Oh, hi.

UNITED STATES: Hi, this is Ashley with the United States. First of all, thank you, that was really interesting. I mean, that was interesting on a very kind of dull subject.

GÖRAN MARBY: This is not dull. No one in here agrees with you. This is fun, this is what we do.

UNITED STATES: I agree, this was interesting. But I wanted to just first also thank you for making it very clear that there is not a single WHOIS database that's available, and that's something that we often forget but not those who actually rely on this information because that makes it so much more difficult in this new paradigm where if you need information that is now redacted, you have to find who is responsible for that domain and then contact them, and now, at least as things currently stand, you have to kind of figure out what process and procedure that particular party is also using in terms of providing access. So this really I think kind of bolsters the reason why we need a unified access model and kind of just highlights a lot of the points that we're facing as users of WHOIS and why it's really critical for purposes such as dealing with DNS Abuse, so thank you.

GÖRAN MARBY: Thank you. And another comment I would like to make is that one of the things -- we're only talking about the WHOIS database here, one specific database. When the inventors of Internet, the men and women, the fathers or the mothers of the Internet set this up, the accountability part of that was very important. So we have thousands of databases containing names. I mean, all the way from IETF when someone writes an RFC, you can actually see who wrote it, because you want to see if someone wrote a standard who was part of that, and I think that's very, very important, and some of them are old because they should be there over time.

And I think that we in the ecosystem where also the government representatives are a part of, there's something we have to figure out because accountability and transparency goes hand in hand in this, and as ICANN, who is a transparent organization, we can't give out the information we want to give out about how we do things when it comes to certain things because of GDPR. There is a lot of discussions internally on how do we preserve some of the most important databases that exist on the Internet to make sure that we can have that information publicly available.

With that said, some of those databases contains within ICANN, the ICANN org, and we are very happy to be challenged about those databases, but the Internet is built around that you can make people accountable for something that is wrong. So it's built through the system, it's not a bug, it's a feature of how the Internet is built. Thank you. By the way, I'm the worst moderator in the world, I haveno memory, so you have to shout anything to me.

VENEZUELA: Normally when we have a project, we take a look at the domain name and if it’s not in use, we try to take a look in the WHOIS to keep in touch and to request to sell the domain. It is the main purpose for me of the WHOIS, but if I'm a bad guy and I want to create a scam platform to steal money, I can steal also a credential like a passport and register a domain, so it's not useful of the WHOIS because the credentials are also with spoofing.

So now with this new law of the European Commission, we cannot see the owner, the country of each of the owner, we have only the name, so it's more difficult to the user that we need to take a look at the name to go to the public contact information and to keep in touch to request information to sell, to buy the domain. So many many things to do, information to share, it's my honest view when a developer [indiscernible] the project owner -- the main issue.

GÖRAN MARBY: Yes, that's right. What happened since the -- and we are a technical organization. We have no opinion about the legislation itself and especially ICANN org doesn't take stands in the policy discussion. So with that aside, that's not a bug, it's a feature of the legislation, some information is retracted. As many of you know, we're trying to figure out ways after policies set by the community, if it's possible to give access for researchers [indiscernible] in a unified way. So the law is made to prevent some things which are bad, and I think we have to abide by the law. Thank you. Who's next?

INDONESIA: You mentioned about security and [inaudible] should go hand in hand, and you mentioned about the law. Under the situation where so many countries are using the internet and so many laws and regulations vary from one country to another, how you can your WHOIS accountability and so on can look after all these laws? Also, the same question for the unified access model; how can this unified access model accommodate all those different regulations? Thank you.

GÖRAN MARBY: Thank you for asking the toughest question I always don't want to have. But it's a fair question. For good reasons, we see many countries around the world who are now realizing that the Internet has an impact on society, for good and for bad. I believe it's mostly good. But of course they look at some of the things that don't work. Remember that the Internet and what we do in the domain name system has nothing to do with the content or who connects to who or anything, but many of those legislations are about that. They call them fake news, they talk about people who sell bad stuff, or child pornography, or all those bad things that happen on top of the internet, and as you know, we are more the plumbers when it comes to this. I call this the road to hell is paved with good intentions, because we see legislative proposals around the world and it's not for bad and it's not for bad intent that you look upon things to take away some of those bad effects.

The problem is that, in our perspective, sometimes the Internet is a very small box technically wise and that's why it's been so successful to take it from 0 users to 4.2 or whatever it is billion user, because it's a very very contained box and it works according to some very fundamental principles. And sometimes we see legislative proposals that can actually break the box. And what I mean there is that that would mean that Internet users would not be able to connect.

We've seen proposals that disconnect routing, that makes routing impossible, we also think that all the information has to be contained within a country or in a region, and we don't judge on that, it's up to elected politicians who represent their governments to make those decisions. We have no say in that and we shouldn't say. But what we try now to do is to engage, if we are asked, to tell the technical consequences of potential decisions. And then of course politicians have the right to say, ”We agree with you, we disagree with you; we think that what we do now is more important than what you do,” and we will be fine with that, we’re not advocating in that sense.

But you are right. We see a lot of potential different legislations around the world that can have an affect on the domain name systems, and it's your responsibility, really, representing your governments to take that into account or not. It's very hard for us to have a say in that because we shouldn't have a say in that. Thank you. Thank you for the question by the way.

UNIDENTIFIED SPEAKER: In your presentation there was one statement that accessed a new filter we added based on demands of naming and addressing communities, that is ICANN and RIR. So, other than this, you are not [indiscernible] anybody else? Like [inaudible] will be there in that. Is that also being considered?

DAVID CONRAD: That’s using community in sort of a larger sense. We respond to the demands of the community to modify the data schemas as necessary to meet whatever the requirements are and the requirements have evolved over time. It used to be that prior -- back in the 70’s when the protocol was first developed, there were no fax machines. When faxes were created, they added a field to allow people to put in fax numbers.

Now, very few fax machines exist and it’s no longer a mandatory field. In the future there may be additional fields that are necessary for additional contact information or additional technical criteria or additional parameters that need to be specified, but they will be added on whether the technical community requires to make sure stuff works or what people require in order to facilitate using the WHOIS registration data in the way it was intended.

GÖRAN MARBY: Thank you. Was it interesting? I hear a lot of yes.

DAVID CONRAD: Just one thing. For anyone who's interested in RDAP, there is going to be an RDAP session tomorrow, an update on RDAP at 3:15 in room 114. I believe it's part of the tech day, the ccNSO tech day and I believe Francisco Arias of ICANN is going to be providing an update on Francisco implementation in that session.

GÖRAN MARBY: Thank you. And of course you can always reach out to David at any given point in time and ask more detailed technical questions about WHOIS. Now you actually know more about WHOIS than most of the population in the world. So thank you very much, and applauds to us all.

**[END OF TRANSCRIPTION]**