SAN JUAN – Domain Abuse Activity Reporting (DAAR)
Thursday, March 15, 2018 – 11:30 to 12:00 AST
ICANN61 | San Juan, Puerto Rico

UNIDENTIFIED MALE: Current session is Domain Abuse Activity Reporting, DAAR, running from 11:30 until noon on Thursday, March 15, 2018. Meeting Room 209-BC ccNSO.

UNIDENTIFIED MALE: Here we go. One, two, three, four, five. One, two, three, four, five. One, two, three, four, five. Second attempt 60 seconds later.

UNIDENTIFIED FEMALE: Good morning, everyone. Welcome to our DAAR session. We have John Crain, our Chief of SSR at ICANN OCTO department, and he's going to present for the session. Thank you.

JOHN CRAIN: Okay. I see many familiar faces, so many of you have seen some of these slides before, but I've shortened them. I'm not going to go into as much detail as I normally do for that very reason, but I am going to include some statistics that you've not seen before that may or may not be interesting. We shall see.

So, the standard thing. What is the Domain Abuse Activity Reporting system? It's basically a system for taking data feeds and associating what people are asserting as abuse against names, and we can do this by TLD quite easily by getting the zone files, etc. We're looking at all the gTLD registries and registrars that we can collect data for, zone and registration data.

We have a large set of reputation feeds. I did leave the list in because I know people would like to know what feeds we're using, and we can accommodate historical studies. We've been doing this for a year or so now, a bit over a year, and, of course, every day we do it, we have more data, so at some point we'll be able to say we've been doing this for 10 years maybe.

And we look at multiple threats. We've picked four different threats that have been mentioned in various places from the GAC and from the community. Phishing, botnet, malware, and spam. And the basic idea is to try and take a scientific and transparent approach so that we can use this for policy discussions.

So what are we using for the data? We're using DNS zone data. There's a process called CZDS, Centralized Zone Database System at ICANN, which allows you to get all of the gTLD zones. We're using some elements of WHOIS data, and we're using a

bunch of block lists or reputation lists that have different terminology. Some of them are free, some of them are commercial, so they're different types of lists and when you see the list of them, you'll be able to figure out which ones are which, I hope.

The idea is to publish data from DAAR, i.e. so not necessarily push out the data – give people access to DAAR but to publish data from DAAR into something we call the Open Data Initiative, which is a different initiative in the OCTO or the Office of the Chief Technology Officer at ICANN, and the kind of data that we publish there will be greatly dependent on discussions with the community.

It won't be, I can almost guarantee you, I can totally guarantee you, it won't be raw lists of domain names. Licensing and other reasons mean that's never going to happen.

So, what do we use it for? Well, there's a lot of things you could use it for. You could use it as reporting on threats at TLDs or registrar levels, study histories, helping operators understand or consider how to manage their own reputations. I really look at the data we get as really an indication of reputation rather than, per se, abuse. I mean, obviously, uses abuse, but these are lists that they use widely in the security industry, so it's more of a

what's the reputation of these TLDs, these registries, or these names with the reputation industry.

But the main purpose of DAAR is to provide data to support the community in both analytics and policy discussions. So, we collect all the zones. I told you this. When it comes down to spam, we only use names that… [inaudible]. I'm getting ahead of myself here. So, we only use names that appear in the zones. We don't say it's not all the registered names, right? There are many names that are registered that don't appear in a zone.

At the time they're not in the zone, they're not resolving, they're not a threat. At least not a threat at that very time. We've got about 1,240 gTLDs. I have exact numbers in later slides from the data set I'm using for this and about 195 million resolving domain names.

We use WHOIS but the only part of the WHOIS that we actually care about is the registrar ID. We're trying to figure out which registrar holds which name. Unfortunately, the WHOIS doesn't let you say, "Please just give me the registrar ID with the name." I wish it did but it doesn't. We could get this information probably for other ways than just querying the WHOIS, but we wanted to try and eat our own dog food, if you like. We wanted to try and do things the way people outside do, and it's hard.

Do the math. If we wanted to update the registrar ID every single day or imagine every single hour to make sure it's always currently at 195 million names, I don't think we're ever going to successfully do that. 195 million WHOIS queries every day is probably going to upset some people if we did that against their infrastructure.

Looking at different ways of getting that data. Not sure what it will be. I also don't know how GDPR will affect this. I don't think any of us do. But that's the only data we use. All the other data is not really relevant to what we're studying.

So, the names we count are unique names. The may appear in multiple lists but we only count them once, but we do store the data of which list they're in. So, if we see a name in our system, we can see that it was in three or four lists, but that's not what we count. And there's a lot of domain or URL lists that we use and I'm going to show that. And as I said, this really reflects how the external community, especially the security community, is viewing the industry.

So, these are the reputation lists. I'm sharing the slides. You can take the photos if you want, but I will be sharing the slides. Most of these, if you work in the security industry, then you probably know many of these lists. They're used in firewalls and all kinds of stuff. It doesn't really matter which set of lists we use or how

wide a set of lists we use. We'll never capture all the abuse. All right? It's like there is no list that has every single piece of abuse and even if we add in another 10 lists, we probably won't have everything. But we have, we believe, a good enough sample and as I said, we're using the lists that are used in firewalls and other devices, web browsers, etc.

Some of the lists are really big lists. You saw the lists. And some of them are small. We basically went out, looked at what people were using in their firewalls, looked at what people were referencing in academic documents, looked at their processes for adding and delegate names, and out of about 80 or 90 lists we looked at, these are the ones we came up with.

We can add, we can delete lists. We'd rather not because that will change the dataset, the history data, but as long as we put in a flag of some kind, so we can take note of the changes, it would be okay. So, we do have that capability.

So, let's look at some pretty pictures. You'll have seen some of these pictures and some of them you won't. So, about 10% of the names that resolve are currently in the new TLD space, which leaves about 90%, if my math is correct, in the legacy. Legacy is everything before this latest round of new TLDs. So, things like .info, etc. fall in the legacy in these statistics.

For the gTLDs that have at least one name in the abuse list, i.e. if you have zero abuse, we're not counting you in this chart. The new and this last six months of data has slightly more abuse names than the legacy. The new seems to be going down, the legacy seems to be going up, data gives you questions, not answers. I don't know. We will, if this patent keeps going, we will do some investigation. The whole point of DAAR is that we can see these kind of things.

When I did that the other day, I went, "Hmm. I wonder. We will figure this out." There is way more spam than anything else. That should not be a surprise to anybody. But the botnet command and control and the malware and the phishing, they are there in much lower numbers.

So, what kind of numbers? So, Alain Durand, I have to give him credit, who's doing our Internet Health Indicators Project, took DAAR data and munched it to find out what's been the averages across the resolving names. So, this is not registered names, once again, this is resolving names. And he came up with some averages.

So, out of every 10,000 names is about 4.28 names that are phishing, 3.28 that are malware, 2.89 that are botnet, and 86.73 that are spam. Right? That makes sense. I think anybody who's

studying this stuff probably see similar kind of statistics, depending on their dataset.

This is 1,143 TLDs with at least one abuse name in them and… I'm sorry, no, across all the TLDs and 1,952 registrars we're trying to collect data on. So, if you look at this dataset, you'll see that normally there's one or two registries that have more by number than others. Right? There are some really big zones out there and if you have a really big zone, you're likely to have more abuse names. That's logical. That's not saying anything about what kind of job that registry is doing, but if that TLD has got millions and millions of names, then you expect them to have at least a few more spam and other issues.

By the time you get up to sort of 90% of the traffic, you're down to not many registries, not many TLDs. For malware, 90% of it or actually more than 90% falls in 7. For phishing, it's 11. For botnets, it's 5. And for spam, it's 18. I don't know how he did these statistics. He's just really good at munching Excel spreadsheets or something, but they're pretty cool.

Now, this registrar data is affected by the fact that we can't keep all those registrar IDs up to date. Now, I've checked our data against other people's systems. There are other people, including registries that do this kind of data and I've checked our data against some of the registries' data. They've been kind

enough to show me. A lot of registries are coming to us and like saying, "Let's compare our data so we can see that we've got the same kind of thing."

We're fairly accurate on this. Numbers-wise, we're pretty accurate and the listing of the TLDs, i.e. of the registrars, the order, may be slightly off, but it's not much different, so that's interesting. If we could get better access to WHOIS data or better access to those registrar IDs, I suspect we'd be much more accurate on this, but it's the same kind of thing.

There's a few registrars that have more spam than everybody else, and you see it's the same kind of numbers. With phishing, it's spread a little bit more across the registrars than it is across the registries. Like I said, data just gives you questions.

This is just the same data in a nice pretty graph. GTL 50 is the 50% gTLD 90 is the 90%, we're just including them so you have the numbers. It's not really easy to see on the graphs. So, what I thought this morning when I woke up, because sometimes I do that when I wake up, is what else can I find out about this?

So, I decided to go and see what the data would tell me. So, we know the average from that previous slide that Alain did from the Internet Health Indicators is 4.28 phish names per 10,000. Now, there's a whole range of how many people have at the very top of the list has 190 names per 10,000. It doesn't sound like a

lot of names, but when you consider the average is 4.28, that may raise some questions.

If you look at one order of magnitude or 10 times M2.1, that sits within 10 TLDs or 10 registries. Interesting. Let's do the same for malware. 3.28 is your average. The absolute outlier is 417. That's a little bit higher than 3.28, I believe. And they are the only one.

Stop calling out names, people in the audience. I'm not allowed to mention names.

That's the only one that's above there. Right, 10 times above. There are a few that are sort of nine and eight times above but this guy's a little bit further than 10 times above. Ooh, questions.

Botnet and command and control. I will say this about the command and control stuff. Take it with a pinch of salt because these are names that are resolving from command and control and a lot of those names are actually sinkholes and things like that. Some of the lists do a really good job of removing the sinkhole names, some of them less so. So, there may be sinkhole names in here.

But still, there's one guy who's 71.46 instead of 2.89 and it's only one that is an order of magnitude bigger again. I want to go dig a lot more into this data because I wanted to see if I can actually clean out more of the sinkholed stuff. If anybody got ideas on

how to do that, come talk to me. But it's interesting and you also have to remember that if you've got a lot of command and control names, I'm not quite sure what you'd do about that. I mean, obviously, you might want to be able to clean them up but the things that seem to affect botnet command and control is the whim of the guy writing the code.

I've seen very strange choices by them in the names, so that's a little bit interesting, so I want to do a lot more digging into that data. And, of course, spam we know is a lot worse, right? So, 86.73, we know that's a lot higher number. The highest is 4,112, which I believe is bigger than 86.73 by a little bit. Just a little bit. It's at least twice as much or maybe more.

Oops. 15 TLDs in spam, so we know spam's pretty rampant and there's lots of different definitions of what it is, so there's a lot of names on there, right? We saw that in the earlier chart. So, I'm not surprised that there are 15 registries that are 10 times… so more than 867 names in their zone, at 10 times the average.

But 4,112. I only have one more slide because I only have 30 minutes, and I want to give some time for questions. So, somebody's going to ask me this, so I'm going to answer it before you ask. The top score is those like the 4,112 and they're not the same registry. They're not the same TLD. So, somebody's going to ask me that, so I thought I'd just get it out there. I'm not

going to tell you who they are but each one of them is a different registry.

Questions, not answers. And I'm going to dig in that.

The other thing I wanted to tell you about is that we've contracted two independent reviewers to look at the methodology we're using to make sure that we're not completely insane. The first of those reviews is in my e-mail box but I've not got around to opening it. The other one should be coming in soon. Our plan, which is in development as we think about this this week, as they came in, is to publish them and do some form of public comment. My suspicion is that we'll just use the standard public comment thing, but I'm neither a lawyer nor a policy person, so I'm going to get help from my colleagues, so we're going to open them up to comments. We're not going to edit them in any way. They're going to go out in the raw format that we got them in.

If there's enough interest, I'm considering inviting the reviewers to come and sit on a panel. If nobody's interested, then I won't, so that's down to the community folks to come and tell me.

I'm hoping, because I've not read these yet, that they turn around and say, "God, these guys are good. This is all perfect." Yes. That's what everybody does. Yes. I'm suspecting that we may make some changes based on what we're told, and that is

perfectly fine, so we're hoping that the two reviews are somewhat aligned and that they give us good indicators of where we can improve things. Hopefully, those indicators are also good for other people that are doing this kind of measurement, and then we will review it and then we will go on to the next discussion, which will be about exactly what we're going to publish. In the earlier slide sets that I didn't show all the slides of, we make a statement, which is it's better to do it right than do it fast, and we're still staying on that, so that's why we're not just publishing everything.

So, the next step is going through those reviews with the community, especially the people in here who have expertise and statistical analysis in the abuse industry. Please review them, please say nice things, be nice to me, but please also be very constructive. So, if you see an issue in there, let us know. Also, if you think some things are more important than other things that have been highlighted, please let us know. With that, I'm going to open up to questions, because I'm sure there are some, because I see certain people standing there.

CRYSTAL ONDO: Do we just go for it? Hi. Crystal Ondo, Donuts. I'm wondering if you are planning on giving contracted parties a view of where

their TLDs or their registrars sit, because it would be really interesting to see.

JOHN CRAIN: I will happily sit with a TLD registry and discuss their own registry data. I believe having a discuss with our suppliers that are contract, our licenses allow us to do that. We can't pass you the feeds, we can't get too far in the weeds, but we can absolutely discuss like this kind of statistics how they meet for your registries in private. Yeah. This is supposed to be a tool to help the industry, not anything else, not tease them, although… Go ahead, can you mention your name, as well, because we're being recorded.

UNIDENTIFIED MALE: [inaudible] .moscow. Is the way to subscribe to like some statistics on our own registrars or our own registries because without it, we cannot comment on your interpretation of results because in science, you need to be able to verify something and blind trust usually, it leads you to very dark places. Thanks.

JOHN CRAIN: So, we have published a lot of the methodology and the data, so you could rebuild it. I'm not saying you have to go rebuild this. What we publish will be a discussion with the community at

some point, so please say that when we get to that stage and say, "Give us more data." We are going to be bound by our licensing, so in the ideal world, you just pass of the feed to the affected registry and they'd be able to do what they wanted with it. But we have licensing agreements that prevent us doing that, so that's a discussion we need to have.

Our principle is we want to be as transparent with this data as we can and, obviously, we can be more transparent with a registry in person and show you what we have. Obviously, we don't have [inaudible] for .moscow. I'm happy to sit down and show you the .moscow stuff we have.

DAVID CONRAD: And if I can add, one of the goals of this particular project is to make sure that everything that we can do, everything that we do within the project is reproducible, so the whole point of this project is to allow for reproduction of our results and we would be – I can't tell you how ecstatic we would be if someone could take the feeds that we have, apply the same methodology that we documented, and come up with ideally the same answers or if they come up with different answers, publish that and then we can figure out who did things the different way. Oh, and I'm David Conrad, ICANN CTO.

DAVID HUGHES: David Hughes. So, first I want to say thank you. I want to point out that I believe this is the most important work being done in ICANN. I mean, GDPR [inaudible], but whatever, okay. But this is actually useful. A request in Panama, can we have more than 30 minutes to talk about this? I mean, I understand the limitations of what you can and can't say and it makes sense to make it short, but more time, if you can bring in independent guys, that would be awesome. And then my last question is the relationship or coordination with M3AAWG.

JOHN CRAIN: So, the 30 minutes wasn't necessarily our choice. We have weird schedules here. I will happily sit and talk all day about this with you because I think they're useful conversations. Note taken. We will talk to the scheduling people and say the community wants more time, and then the community will get more time.

DAVID CONRAD: Actually, my understanding is that the community sets the schedule, so if the SO/AC Chairs are, I believe, they get together and decide exactly what's going to be on the agenda. So if you believe this topic deserves more attention, more time, please relay through your SO/AC Chair.

JOHN CRAIN: M3AAWG? So, DAAR does [not of course] have a relationship but we have a relationship with M3AAWG. M3AAWG actually sends a representative to ICANN meetings. They've actually pointed a finger at somebody and made them the victim, so you go so ICANN meetings, they've been much more active in the discussion around these kind of issues and they're bringing that here. We send our staff to M3AAWG meetings. I believe we're actually… Are we a member of M3AAWG? I believe… Yep. We're a member of M3AAWG just like we are of APWG, so we are actively involved and the more they get involved here, the better, too.

Was there one more? I think there was. No? Thank you. Oh, yeah. Yes. Thank you.

KRISTINA ROSETTE: Kristina Rosette, Amazon Registry Legal. Question about the CZDS data. Obviously, ICANN has to request it just like any other requester, which means in turn that ICANN has to agree with the ICANN terms of use, which means that ICANN is contracting with itself to abide by the terms of use. And independent of the legal issues that that raises, I'm curious as to how exactly ICANN Legal would address an enforcement issue if a registry operator believed that in connection with DAAR, ICANN was misusing the data.

JOHN CRAIN: I am not a lawyer, so we have talked to Legal and they said we were okay to do what we were doing. I didn't ask them that specific question. Could you send me that in e-mail?

KRISTINA ROSETTE: Absolutely.

JOHN CRAIN: Because not being a lawyer, I only understood 90% of it and we will pass it on to our legal folks and ask.

UNIDENTIFIED MALE: I'm [Victor] [inaudible] Sigma. I've recently run into a novel sort of form of abuse, perhaps. I'm not sure where it will fit into your picture. It's something called spider webs, which are dense webs of SEO magnets that have started appearing in some interesting papers about this. I don't know if you have any interest or even ability to track these things if they show up on any list you can get feeds off. But I started running into them as part of my [DANE survey] running into these clusters of weird domains that are linked to each other.

JOHN CRAIN: I'd love to talk to you offline and figure out if there's even if it's not part of DAAR, if there's something funky happening in the namespace, I love to figure it out.

UNIDENTIFIED MALE: Okay.

UNIDENTIFIED MALE: Yeah. Just would want to add that the abuse that we're doing the analysis of in DAAR is largely driven by GAC, at the GAC Beijing communiqué and then the [inaudible] about communiqué. And my impression from those communiqués was that they were statements of these abuses such as suggesting that there would be – the abuse evolves over time and that if there were other abuses that are new and interesting, then that's something that the GAC would presumably want some sort of mechanism to address.

So, the whole system is supposedly designed to deal with evolution of abuse and the DAAR system itself has extensibilities of something new and interesting does come up, we can modify the system to take a look at that, as well, based on community demand.

JOHN CRAIN: Okay. So, we've reached our time but being as we're not bound by other people using this room or Adobe, we'll take questions.

SIMON: Simon [inaudible] from Spamhaus. Thank you for producing this aggregated data. I think it's going to be really useful. It helps back up some of the messages that we've been putting out there. Just for the people that don't know, Spamhaus.org does actually have a TLD naming and shaming page, so anyone who's interested, please go have a look at that. I'd love to see how our view compares with the meta view. I appreciate – we're probably never going to get to see that but it would be great to see that. Anyway, that's it.

JOHN CRAIN: I would not be able to comment on whether our lists matched at the moment, but I'm well aware, obviously, you're [inaudible] and those [inaudible]. I think other people are. One of the things we might want to do in the future is either a panel or a discussion about what kind of other things are out there, so that at least people in the industry are aware all the various lists.

KRISTINA ROSETTE: Thank you very much. Kristina Rosette Amazon Registry Legal. My question is that the data feed providers use their own criteria

for deciding what abuse is. How will ICANN resolve conflicts between what a feed provider says is abuse that does not necessarily align with the contracted parties' own definition of abuse?

JOHN CRAIN: This is why we are saying we are measuring the reputation of it. That's why I say that. This is what people are using in the security industry. I don't think you want ICANN going out and building any of these lists ourselves, so we use the widespread used… sorry, I'm tired of my language is a bit off here. So, we're using the lists that are used in safe browsing [in] firewalls.

Yes, there may be some distinction. We've looked for the lists that actually designate why those four characteristics. We've not put anything in that doesn't have that. So, we're kind of aware of that and we struggle with it, but that's a comment than when we have the discussion about what should and should not be in there as lists, we can have.

KRISTINA ROSETTE: Can I ask a quick follow-up?

JOHN CRAIN: Yeah.

KRISTINA ROSETTE: So, specifically I'm thinking about, for instance, the registry contract doesn't prohibit spam and spam is under your definition of abuse. So, if a registry or a registrar does not call abuse spam in their own registry or registrar agreement, how is it that ICANN can call that abuse and use that against a registry or registrar?

JOHN CRAIN: Number one. We're not using this against a registry or registrar in any way. We are providing data that's aggregated across a multiple set of abuse providers. This is information that is in use by network operators around the world right now and today to block registries outside of ICANN's control whatsoever, both registries and registrars.

So, what we are doing is we're aggregating data that is being used by network providers and those aggregated data, those aggregated data reflect operational realities. Now, the use of spam is actually an indicator of likely abuse in other areas. We're not using it in any way within ICANN to do anything with regards to any of the contracted parties in any way, so it's not quite accurate to say that we're using this against the registries.

FARZANEH BADII: Farzaneh Badii from Noncommercial Stakeholder Group. So this, your mandate is so broad and you do not tell us it's not… it's kind of like it's not limited and it worries us that you say you can just add to this list of abuse and another thing that worries us is that you take the GAC advice and just come up with a report. I think you need to consult with the community more and also keep your mandate limited with ICANN mission and also not to take action. I mean, it is reassuring that you are saying we won't take action but we are worried that what will happen in the future maybe one day you say community told us to take action or GAC, so we are concerned about this.

JOHN CRAIN: Okay. I will say that the three other forms are not just in the GAC communiqué, they're also in an element of one of the contracts. Yeah.

UNIDENTIFIED FEMALE: [inaudible].

MAX: Max [Mozo] of .moscow. I suggest remove references to the compliance from all maybe future presentations because without proof, a registry or registrar, they cannot do anything. Basically, we have contracts and in many cases, it's going to be a

violation of local administrative code to breach a contract without apparent reason because we cannot say, "Come on, guys, someone told us that you've been bad, so we decided to [come in] and do nothing." It's not a good thing.

And the second, you need to realize that it's not possible to use this kind of data in any compliance matters at all. The last question. Do I understand this right? The health indicators are just the interpretation of your output.

JOHN CRAIN: For this specific… I mean, there are lots of health indicators. The ones that Alain showed in the health indicators thing. Hang on. Let me go back and show you them. These ones, yes. Those are DAAR output. Derivatives, yeah.

DAVID CONRAD: One point to reiterate. The point of DAAR is to actually provide information that the community can use to modify and adjust policies as they see fit. It's entirely reasonable to make the assumption that the information that DAAR provides is not helpful in the context of modifying those policies, but unless you actually see the data, then it's hard to make those decisions.

It is true that there's nothing in DAAR data that would be fed into ICANN's contractual compliance other than here are some

outliers outside of the normal bounds, which is information that they already have from third parties that are constantly saying that these parties are outliers. So, the point of this is to try to provide a base level of information that the community can rely upon. Please remember that one of the reasons DAAR started is because after a number of trade industry reports of how the New gTLD Program was being abused with method from organizations that were selling products to combat abuse with undefined methodologies, a number of the contracted parties actually approached ICANN OCTO and said it would be really nice if there was one canonical listing of these types of reports, so that's what initially triggered DAAR, which is about two years ago, whenever the blue cat paper was published.

So, with that, we've already gone over and I think we appreciate the interest and if this is a topic that you think needs more time, please recommend it to the SO/AC Chairs. Thank you.

JOHN CRAIN: Okay. With that, thank you, everybody, and probably see you all in the Public Forum.

[END OF TRANSCRIPTION]