BARCELONA – Domain Abuse Activity Reporting (DAAR)
Wednesday, October 24, 2018 – 17:00 to 18:30 CEST
ICANN63 | Barcelona, Spain

UNIDENTIFIED FEMALE: Good afternoon, everyone. We're going to be starting in a couple of minutes. There are still some empty seats here in front, so please feel free to move over here. Thanks.

JOHN CRAIN: Good afternoon, everybody. We're going to be a little short on chairs, so I'm going to stand so other people can see. It's good to be here in Barcelona. This is the GDPR session. Everybody is in the right session? No, it's not. It's the session on – let's see if this actually works at a distance. Okay, my clicker doesn't work, so that's going to be good. Ta-da!

The Domain Abuse Activity Reporting tool. Who here has seen presentations on this before? Okay. I'm sorry, but there's a bunch of people who haven't. Many of you have seen these slides before. You can sleep for about the first 20 minutes and I'll wake you up when we get to the good stuff – or not.

So, this is a system we've been working on for a couple of years to gather data and collect information about how others see the abuse in the domain name space. It's called the Domain Abuse Activity Reporting System. There are many systems out there like this. This is not necessarily rocket science. You may have seen similar things before.

What we do a little bit differently is that we actually look at all of the generic TLDs that we can get zones for. Often, people that do this will just look at their own zone. They won't necessarily look across the ecosystem. And we use a lot of different data feeds, reputation block lists.

The other thing that we do is we actually published the methodology. So, we actually want other people to be able to go and do this.

The basic idea behind this was can we facilitate discussion around abuse in the domain name system and can we provide data to the community or to other practitioners that would help them understand the threat landscape that we all face?

So, all of the data we use is data that you should be able to get yourselves if you wanted it. Now, some of it is free, some of it you have to pay for, but it's all available. As I said, we wanted to publish the methodology so that other people could do this, so we don't use data that you can't get. So, if you did want to do this, you could. [inaudible] fees. We pay for them and they have licenses.

We're trying to publish as much of the data as we can wherever licenses permit. Has everybody here heard of something called the Open Data Initiative, or ODI? This week, we're calling it the Open Data Program. The reason for that is it's pretty much running and it's going to go to operations now. It's no longer an experiment in our group but it's now actually going to move towards our operations people and it's going to start running.

The data from this is not yet in there and we're still discussing whether or not or what we can actually put in there, but at some point, we're hoping to publish quite a bit of data in there.

So, what could you use a tool like this for? Well, it's reporting on threat activities. So, if you had your own system like this, you could do your own reporting. We see all kinds of things in there.

You can also look at things historically. We have about … Probably about two years' worth of data in there. The newer data got better and better. Slides I'll show later go back just to this year and we will be looking at older data to look for more trends. So, you can do trending and look at how the history goes.

We can help the operators. Who here is from a registry? So, we can help you guys by sharing data with you. We can't give you all the data that's in there because of the licensing, but we can come talk about your data with you. Some of you have had meetings with me in your offices or here at ICANN meetings where we've sort of gone over what we're seeing and it just helps inform. It helps inform the discussion. And of course we could do studies because now we've got lots of data.

But the main purpose, as I said again, is to sort of inform the community, so that when we're having discussions around abuse, we kind of know the scale and the threat landscape that goes with it.

So, DAAR collects, as I said, all the gTLD zones. It's all public data. It's only used as the names that appear in those zones. So, if you've registered in your TLD 10 million domain names but only a million of

them are actually resolving, we're only measuring the ones that resolve. The idea behind this is we're trying to look at the threat landscape and our thought was, well, if it's not resolving today, then it's not really a threat at the moment. So, only resolving names.

At the moment, there's about 1240 gTLDs approximately that we're collecting zone files for. We don't have all of them because some of the very old ones don't actually put their zone files into CCDS, the system that we use to collect them from, and we're looking at about 196 million names to date. So, it's a lot of names that are resolving and I'll show you some pretty pictures later because that's what most people like. So, it's got quite a lot of data in there.

We always get the GDPR question. Yes, it's GDPR compliant. We do use WHOIS data. We're not affected by the redaction because all we actually care about is the IANA ID, really. And we like some of the data like when this was created, etc. But we don't use any of the PII or the identifiable data that's in there.

It counts unique abuse names. So, we have a list of unique abuse names or numbers. A name can only appear in that once. Then we also flag them by the type of abuse they are. So, if you think about this as a database array, you could have one flag for it being an abuse name but you could also have another flag there saying it's spam and it's phish. But it's still only one unique abuse name. It could be used for different kinds of abuse. And we do actually measure that.

So, I'll actually show you a list of most of the block lists that we use. That may change over time. But we use multiple of them. And we're

using block lists that have some specific flags in them. We want block lists that flag names by the kind of abuse that we're looking at and we're looking at four specific types of abuse. We're looking at Botnet, Command and Control names. We're looking at malware names. We're looking at spam. And we're looking at phish. I'll come back at those. You know it's a database, so you can do everything with it that you can with databases and you can make pretty histograms. You can push out data, of course, into CSVs and do other stuff with it. So, it's kind of a cool tool.

And because it's using reputational block lists, we think of it more … We call it an abuse reporting tool but I often think of it more as a reputation reporting tool because it's showing how the industry that builds these block lists sees the domain space.

I have to say this because people keep asking. No, we do not build our own block lists. We do not say one way or the other whether a piece of … Whether a name has been used for spam or not. We use other people's data. We don't say whether it's a botnet command and control name or not. We just use other people's data. But this is the data that everybody else is using in their firewalls and other places.

So, as I said, they must have the threat classification that we won't. When we say evidence that they are trusted by the operational communities, that's people that are actually using them. All of the lists that we use are actually used in devices and blocking technologies out in the Internet. So, these are what people are using to protect their users on their network.

So, who here thinks they use a reputation block list? All those who didn't put your hand up got the answer wrong. We all use reputation block lists. If you get e-mail and you have spam filtering, you're using a block list. If you browse in web browsers, you're probably using block lists. If you have a firewall on your network, you're going to be using block lists. So, you may not know that you're using block lists, but you are. And more than likely, you're using the ones that are on the list that we use because they are pretty ubiquitously used. They are all over the place.

These are just some examples. Google Chrome uses the Anti-Phishing Working Group block list as part of its safe browsing. That's one of the lists we use. Akamai uses SURBL. Facebook composes and shares its own data through [inaudible] exchange. Everybody is doing this. Commercial firewalls use these lists. Enterprise mail messaging systems, third-party e-mail service providers, and the list goes on and on and on.

Does DAAR identify all of the abuse? No, it doesn't. We don't pretend it does. It has a specific set of RBLs or block lists that we use, so we only have the names that they have captured. There are many other block lists. They may not categorize them by the threat models that we're looking at, the particular threat categorization. So, we only look for those four types of threats. We only use certain lists. And we only look in the gTLD space.

So, no, it is not an image of every single piece of abuse or abusive name on the net. It is just a view of the threat landscape.

Here are some of the lists. We don't necessarily use all of these at the moment, but they are, for the greatest part, what we're using and we're probably going to adapt these over time as new lists come and as we have discussions about which lists we should use.

For example, we're trying to find more data on malware. So, we may add some malware reputation lists that we don't have at the moment. So, they will change. These slides are uploaded, so you don't need to take notes.

Partial lists of academic studies that we read when we were looking into doing this and that sort of talk about some of the science. The science about block lists, the science about identifying types of abuse, etc. Once again, the slides are there. Lots of great reading, if you're the kind of person that likes to sit around reading white papers. There's plenty of them here and we could probably give you a few hundred more if you want that kind of stuff.

So, why are we doing spam? Well, anybody who actually works in the sort of operational security realm actually knows that spam is sort of one of these delivery mechanisms for lots of other stuff. But also, these things were mentioned by the GAC. Some of the things are mentioned in contracts, etc. But it's sometimes a bit of a contentious thing about whether or not we should measure spam. At the moment, we are and I suspect we're going to continue. I was actually involved in this operation here called Operation Avalanche where spam was part of the problem for the delivery mechanisms for the malware, etc., behind it. So, there were lots of examples out there.

We are thinking about maybe putting together a white paper on how spam is used to abuse, etc., because I know people want to know more about this, so we're actually looking into whether or not we can actually produce extra material mainly for educational purposes about how spam is used to actually do bad things.

The other thing, of course, is even if we don't think that spam is an abuse problem, filtering people do. For the greatest part, that's one of the things that they filter on. That's where they look at TLDs and things and think about their reputation.

Who here has seen lists, for example, of top-level domains ranked by badness? A lot of that data they use is spam data. That affects the reputation of our industry. So, as I said, I see this a little bit as a canary in a coal mine kind of thing about what the reputation of the industry overall is and specific parts of the industry.

So, what's the project status? Which is probably the bit that those of you who have been here before want to know, so you can wake up now. We've been through the standard slides.

So, great news and bad news. DavidPiscitello, great friend and colleague, is still a great friend but no longer a colleague. He has moved on to do other work. But, on the bright side, he's still on the SSAC and he's still in this field and I suspect we'll see him at an upcoming ICANN meeting, and instead of being here with me being my friend and colleague, he'll be over there with you throwing things at me, which will be fun.

But, on the bright side, I have a new staff member. Her name is Samaneh Tajalizadehkhoob. I may have got that wrong. I think she's actually watching this. So, Samaneh, hi. These are all your best friends in this room now and they're all going to be nice to you if I bring you to Kobe, aren't you? Well, two or three of them said yes. We'll see how it goes.

So, Samaneh is a recently graduated PhD from the University of Delft, has written lots of white papers on identifier abuse, classifying that and how it is used. I recommend you Google her and read some of her white papers. One of the main reasons that she got the job is because I read all her white papers and she clearly knows more about this than I do. So, great. She's on the team. Her focus will be abuse research, and specifically she will be managing a lot of the DAAR project work. So, that's good.

So, we published some reviews. Who here read the reviews? Well, we did publish them. Not many people read them, apparently, in the room. Go read them. They are awesome reading if you're the kind of person who likes to read white papers.

So, we got two industry experts to review our methodology paper. We said to them, "Here's what we're doing. Here is the product. Have at it." We did pay them, so in that sense, they were working for us. But we gave them complete independence and they provided reviews and we gave them some input on the reviews. What we said is we don't understand what this means. I think you mean this other word here. They were not DNS people, so occasionally they would use terminology that confused

us and we clarified terminology, but that is all we did. So, they are 100% their reviews. So, we published them and then we asked for comments. And we used an e-mail address to do this and you could send them in.

I understand that some people would have preferred that we used the comments platform. We went round and round internally about that. What we got out of this … This all where you can go read it all by the way. We only got five comments. We got one that was published publicly by the Registry Stakeholders Group and then we got four from security practitioners.

We are going to publish the comments along with reviews of those comments and answers to them but some of those practitioners do not wish to be identified. If you work in the security field, you will understand why and it's one of the reasons that we decided to go this way rather than the other. Arguably we could have done both. I agree. But we are where we are.

Our plan is, by the first of December, to publish the reviews and our answers to those reviews and our own comments on … So, we will publish our comments on the reviews. We will publish our comments on the comments to the reviews. That's a bit of a [inaudible]. But, first of December we intend to have all of that out for public and we'll publish on the website where the DAAR papers are, most likely at somewhere near the same announcement, but we will publish the URL when we publish it.

We intend to start publishing monthly reports. Now, we've been intending to do this for a long time and every time we have this

discussion we say we intend to publish monthly reports but now the board has said we can publish monthly reports. They will be anonymized and they will include a lot of similar data as to some of the things I'm going to show you now with explanations, etc.

If people want to see different data as we publish these reports, let us know. Maybe we can do it in a different way as we do it. We're pretty open to this. This is for the community. So, if the data we're publishing isn't meeting your requirements or your needs, let us know and we'll talk about what we can and can't change.

We are still investigating publication into the Open Data Program. I still want to call it the Open Data Initiative but it's the Open Data Program. So, we're still discussing that. I don't know how long that will take. I don't know how that will go. Much of that discussion goes above my pay grade.

So, we've actually been using this in discussions. Last time I gave a presentation on this, I'm trying to remember where it was – it was a different part of the planet, but they spoke the same language. We were in Puerto Rico. I stood up and I said, "If you work for a registry and you'd like to learn more about the data related to your TLDs and only your TLDs, I'm happy to have a discussion." And lo and behold, a bunch of registries came and said, "Well, we'd actually love that."

So, I've actually been going around having, I think, excellent discussions with registry operators about what we see versus what they see and how we can measure that and how they can measure that. This is really been done on a peer-to-peer basis, as security operations

people. It's not a compliance thing. It's not an ICANN is coming to [inaudible] thing. And we get to sit down and have some really good discussions. I've had a few here at the ICANN meeting with registry operators where we've gone over what we're seeing and whether or not it maps with what they're seeing.

Interestingly, sometimes it maps exactly. Those tend to be registries that have been around longer or have more resources and maybe have a security person and an abuse person and they tend to map pretty close to what we have. Now, as a registry, you have more data than we have. For example, we still cannot map 100% all of the names to their registrar. We have 196 million names. In theory, if we wanted to be accurate to within, say, a 24-hour period, that means we would want to do 196 million WHOIS queries and I'm sure you will want us to do that, but not. So, we need to find a solution for that and you guys, of course, have that data.

So, what we find is that the numbers that the registries have are slightly different because they just have a closer view to the data than we have.

Of course, if you're from a registry and you still want to see your data, talk to me. You can write to daar@icann.org or you can write to me directly john.crain@icann.org and we will happily spend the time to look at that data, maybe even see if we can find some more data behind that. So, data I always tell people only gives you questions. It rarely gives you answers. But when you take the data out of DAAR that we have access to that we can't pass on because of our licenses, we can actually go and do some investigation around that and look at things

like, "Where is this badness? Where on the network is it? Are there patterns in this data? We can discuss that with you." We're here to help. So, if you want to come talk to us and have us help you go through these kinds of problems, please, please talk to us.

So, pretty pictures time. Everybody loves pretty pictures. That's a pretty picture. It's just a scatter diagram. Anybody can do these. You can probably do these in Microsoft products like Excel. Oh, that's where I did do this. So, this is pretty easy.

So, we're looking at phish. This is a point in time document and it's all of the registries, 1.6% of registrations. I'm looking at them by percent of registrations here. You can also, of course, look by number of abuse names, but I personally think the percentage of registrations is more interesting.

So, almost all of it … Now, some of these, of course, have almost no registrations down this end. This is how many registrations they have in the zone. So, somebody having … These guys all have less than ten names in their zone. You'd be surprised if they had too much badness.

Now, we do see people on occasion – that was once a day – that only have two or three names in the zone but one of them gets flagged. So, they would be up here somewhere to the left, up in the top, because they'd have almost no registrations but a large percentage of them. Not in this graph. Then, of course, as you go here, you get more and more.

So, really, when you read these kinds of graphs, you want to start looking sort of in this quarter for where the real outliers are. Well, 1.6 or

1.58 or something, I think that one was the outlier. I don't know if that's good. I don't know if that's bad. That is just the data.

Malware is … I think this one is Orion's Belt. I was never good at stars and things. These guys over here are interesting, but as you can see, most people are under the half a percent. These guys still below 3%. We couldn't see this kind of data two years ago, or maybe even a year ago, before we started doing DAAR so it's causing some really interesting questions to appear. Spam, glorious spam.

Most of the people, under 5%. Then, this little fountain up here with folks getting as high as 40% and a lot in the 30s and so on. So, sort of "what's going on here?" kind of questions. Go ahead.

UNIDENTIFIED FEMALE: Could you use the microphone, please, and give your name and affiliation?

ROB GOLDING: Rob Golding, Astutium. Are these domains you feel are specifically registered for these purposes or are these domains which are currently being used for these purposes?

JOHN CRAIN: These are names that are on [reputation] block lists. That is exactly what they are. What the reputation block list decides these are, we don't get into. Now, often, those block lists are things that are considered a threat at the time that they're measured.

For example, and I know some of the people in the industry are here from the block list industry, if it's not in the zone for more than a few days, then typically they'll pull them out of the block list. So, the fact that it's resolving and it's on a block list, you make your own decision about that. But the reputation guys have said this needs to be blocked. That's what it tells you.

You can dig into this and you can have conversations backwards and forwards forever about what it means and whether it's really a threat or not a threat, but it's on the block list, so people who are looking at these block list data, they think that's a problem.

RICHARD HILL: First of all, thank you very much. This is fantastic. If I understand correctly, dot-com is the one there at the bottom right and that actually has the lowest percent abuse. Is that correct?

JOHN CRAIN: I have no idea. I'm not allowed to name names. That is just somebody with a lot of registrations and a very low percentage. I don't know who it is. We've known each other long enough, Richard. You know I'm not going to fall for that.

But that's an interesting problem with data, right? We're not naming names, but you can look at things and say, "Well, I think I know who this is." But the data is what the data is.

And of course if you take the unique names, remember some of these unique names could be on more than one of these last slides, you're going to get a very similar pattern. So, if you say unique names on abuse lists, you've got the same pattern. You've got these outliers that are up in the 40% and then you've got that same scatter.

One of the things it does kind of tell you as an industry is if we could clean this up and we could hopefully make this unattractive, something is going to happen. Now, we don't know what's going to happen. We can produce a cause by pushing on this piece of string, but we don't know where it's going to go. Are they going to go somewhere else or are they going to disappear? We don't know. We can guess at some of this stuff, but if we don't push on the piece of string then we know what's going to happen – nothing.

So, if you rank the TLDs by percentage of abuse – and remember, there were some in there that were up in the 40s, and this is the unique abuse names, and you got then down into that below 1% average – so, we went and cleaned them up – if just the top five did that, they have about 15% of the abuse names in their space. So, if they managed to clean those up and they went away, you'd probably get rid of at least 14% of the abuse. Top 10-20, top 25-36. So, if you cleaned up the top 25 TLDs and we helped them stop being abused by these people … Because remember, as a registry, you're actually being abused here as well. Obviously, there's a lot of other victimization going on of people out on the Internet, depending on what kind of abuse it is. But they're also abusing the registry systems. That's a third of the abuse. That's quite interesting.

And once you go to about 65-ish on the list, everybody is sort of below 1%. So, it really is interesting that we have this very sharp tail where, at the top of it, there are few registries that are very heavily abused.

So, resolving domains. Not registered domains. Registered domains is a completely different thing. You can sell domains. Not everybody wants them to resolve and not every resolving domain is necessarily going to some business. There's landing pages. There's all kinds of stuff out there. It's going up. Can't really tell on this, but it's going up both here and here. These are new TLDs and it's dropped off the map. Down here it somewhere new TLDs. You'll have to trust me. I've got it down here. I did put the little chart in with what was what, because otherwise, my boss will tell me off if I didn't. We've got growth, slow but steady growth, of resolutions, about 7 million between new and what we call generic or the previous TLDs.

So, our definition of new are those ones that came in the last round of new gTLDs, just to be clear on that. So, 165.5 million, 68.5 million names roughly, and 23.5 million domains roughly in the new gTLDs. So, that's sort of how the market splits because we can measure that. Not registrations, but resolutions.

Unique abuse names. Now, I've been talking to the people that we work with this on. They're sitting there ignoring me at the moment, gentlemen. And they have longer data than I have and they say there's a lot of cyclical nature to this. We're going to get the two-year data and we might see more seasonal changes in this data.

But I will say one thing. It's actually going down. The number of names on the abuse list we look at is actually consistently going down over the last seven or eight months. I'm an optimist so I'm hoping that's not just some seasonal thing where it's going to go down and it's going to shoot back up for Christmas. I'm also a realist and I do know that spammers and phishers and that love Christmas, so we may see a spike there but we won't know until we get there.

Interestingly, they're moving. Although it's going down, the spammers and the phishers and the malware guys are actually moving the names they're using. They've slowly started moving them, as the registrations went up, from the older TLDs to the new TLDs. So, about 700,000 of those domains are in the older TLDs and about 900,000 names are in the new TLDs.

But there's been a decrease of about – not 4.5. That's the wrong number. About 450,000, sorry. Bad number. There's been about half a million decrease [inaudible] being there, of about .5 million listed abuse names over that period. I will change that before we publish them.

Phish had some kind of spike a little bit and have gone down. If I look at this, that looks like something that could become some sort of [inaudible] wave. You can see that being seasonal but we won't know until we've got more data. But we will have more data. Ten years from now, we may have a lot of data about this and we'll be able to tell more.

I don't know what happened in malware domains. Somebody cleaned up big right here. These are snapshots, so it's probably some kind of curve that could have happened anywhere in the middle. But between

May and June of this year, somebody in the old space cleaned something up, which is cool. And we've got a slight uptick in this last month in the new space. That's somebody probably doing a campaign.

Spam follows the abuse. Timeline, because as I said, it's most of it. So, it's kind of the same as the abuse domains. [inaudible].

Botnet command and control is really interesting. Botnets, or the malware that's behind the botnets often use something called domain generation algorithms and they generate tens or even thousands of names that could be used in the future, and these DJA lists can then of course end up on these block lists.

Now, if there's an action like Avalanche which took down I think it was 800,000 names, you would expect there to be 800,000 names on here. But, of course, security practitioners talk to each other. The idea of putting those names, however many you put into the zones – and they didn't put 800,000 in but they put quite a lot of names to resolve – is that people can get to them. And the reason you want people to get to them is so that you can identify them as a victim. You're trying to figure out what the infection base of a piece of malware is.

So, a lot of these block list providers, if you are involved in one of these and you go to them and say, "These are actually good guys," they take them out of the lists. So, most of the names that are registered by the good guys for what they call sinkholing don't actually appear in these lists.

So, something happened here. It could be that somebody phoned up one of the major RBL providers and said, "Hey, this is such-and-such law enforcement agency," or, "Such-and-such security body and these are our sinkhole names. Please take them out of your list." But I don't know. I'm just guessing there. But something happened.

Botnets is interesting. The interesting thing about this for me as somebody who works on a lot of these take-downs is this bottom line here. It used to be that they only used generic TLDs and country codes. Only earlier malware variance, that's what they did. And now, of course, just like they're moving their spam and things like that, they're also picking on new TLDs for botnet command and control names.

The bad news is there's pretty much nothing you can do about what they decide to do in their code. But we've seen that change. We've seen them start to use new TLDs in their code for their algorithms. So, that's interesting. You might get a phone call from me one day if you run one of those new gTLDs saying you're going to get a court order or something from a law enforcement agency because they're doing a botnet takedown and today is your lucky day and your TLD is being picked on by the bad guys.

So, that's the kind of data that we can produce from it, and of course, it's a database so we could produce any other kind of statistics you wanted and we have more statistics and if you prefer pie charts to pretty lines or to scatter graphs, we can do that, too. We can produce whatever you want. We'll put out the first report probably in a couple of months. I did ask Samaneh to automate that because she's one of these

clever engineer types and I thought that would take a few weeks and she messaged me this afternoon to say she was done. Oh, that's right. That's why you use code and programming. It really speeds things up. So, we may be earlier, but I'm still not promising. But we may actually get them out earlier.

So, where do we go from here? We've got plenty of time because I wanted to have discussions because last time we did this it was too short and we didn't get time to actually discuss any of this stuff.

So, number one is although we did the comment period, if you want to comment, you can still send us comments. If it's constrictive, useful comment that will make the product better, we'd like to hear it. We want this to be the best product it can be so that people in the industry can learn from it and that we can learn from it. So, if you have comments, feel free to send them in.

With that, I want to open the microphones and see if anybody has questions about the generics of it. I don't want to answer comments that are in the reviews, in the comments on the reviews because we're going to do those in writing but if you have questions or comments, please feel free to ask them now.

UNIDENTIFIED MALE: Hi, my name is [inaudible] reports or is it also for third-party services like a browser to warn to the user about the domain?

JOHN CRAIN: So, this is really statistical analysis and input into the policy discussions. It's not a tool that we feed to browser providers or anything like that. The input into this tool comes from the same things that the browser providers and the firewall guys use. It's a lot of the same data. This data doesn't push out to anyone. This is not our data. This other people's data. It's not some kind of rocket science where we have some magic switch that you can then push to other third parties. I mean, the licenses doesn't even allow us to do that. but even if we could, I don't think that people think that's ICANN's role.

You're very quiet today.

UNIDENTIFIED FEMALE: I have a question from Jim Prendergast online. "At the beginning of your presentation, you said that all of the data you will be using is also available to anyone on the street. With that in mind, on Monday, ICANN compliance will be sending audit notices to all new gTLD registry operators that are focused on abuse monitoring and mitigation efforts. Will DAAR, OCTO, or anyone else involved with DAAR have access to any or all of that data, either in aggregate or individually?"

JOHN CRAIN: So, aggregate everybody will have because we're going to publish some of that stuff. I didn't say anybody on the street because you kind of have to pay for some of this, but in theory, I guess they could. Compliance can have access to this data, I guess. I don't think it's – or at least if you talk to Jamie, you should really ask Jamie this question. I don't think

they intend to use it as some kind of push through compliance tool. But you'd have to ask Jamie that because I don't run compliance. We're not hiding it from them but we're also not … We're not going saying, "Look, here's a report. Go look at these 20,000 names that some registry had and do an audit on that." They have their own audit processes. We're not part of compliance. We talk to them. Obviously, if they have questions, they come to us about abuse and how abuse works and all that stuff, but they're a separate department.

So yeah, this data is out there. They have their own data sources, too. I actually don't really know how the compliance process works 100% because it's not my job. But I'm happy to find out for you, Jim. If you pop me an e-mail, you have my e-mail address. I can pass that on to Jamie. The lady in the corner has a question.

UNIDENTIFIED FEMALE: My name is [inaudible] registry. I have a question about what do you expect [registrars] to do to facilitate DAAR either now or in the future?

JOHN CRAIN: To facilitate DAAR. Okay. I don't think … So, there's a couple of questions. When we go through the comments process, there were some really good questions in there both from the Registry Stakeholder Group and from other people. We highlighted some problems that we had that I would love to have a discussion, not just with the registries, but with the ICANN community in general, about how we solve.

For example, we can't measure with high confidence against the registrars because we can't get that label. Maybe there's something we can do there. I don't know. That would be a good discussion.

The other thing you can do to facilitate it is join the discussions and tell us what kind of data you'd like to see out of it. For example, when I visited registries, they've said, "Oh, I wish you could give us this data on a regular basis, automated." Maybe we can. There are parts of this data that we can pass on because we talk about putting it into Open Data Initiative. Maybe we should send that data to the registries. If that's what the registries want to do, then please let us know. More importantly, let my boss know, so that we can spend the resources making that happen.

Personally, I've heard this from probably four or five different registry operators that they wish they could get this data on a daily or weekly basis. Maybe that's possible. But if you want us to do things like that, because the community drives a lot of where we spend our money, please go to microphones, maybe at the public comment or wherever or go talk to my boss and say, "We want this," because if our licenses allow it, we'll do it. So, the best thing you can do to facilitate the discussion around abuse, because DAAR is just a tool. The goal is to improve policy and to improve abilities and to know more about the abuse ecosystem together.

So, the best thing you can do to facilitate it is actually get involved in discussions like this. Actually suggest to us changes. Actually tell us that

you've gone and you've taken the methodology and built your own tool. That would be cool. I've seen tools at registries.

And by the way, some of these registries have better tools than we do. There's a lot of registries that have really similar tools and some of them are really cool and they've let me look at them but they've not let me steal the code and I don't blame them. So, just be part of the conversation. That's the best thing you can do for this.

UNIDENTIFIED MALE: [inaudible] dot-EU registry. I was wondering if you see a relationship between the abuse levels in TLDs offering free domain registrations and those where you need to pay.

JOHN CRAIN: Not in DAAR because DAAR doesn't measure that and that's a good part of the abuse discussion. There is anecdotal data and some people would say it was more anecdotal but I don't have the data. The pricing has an effect on who registers your names. I think we all know that's pretty obvious. If your names are free, then people who don't want to pay for names will [inaudible]. But we don't have data on that.

Pricing data is actually not … At least, maybe it's [inaudible] my engineering mind. It's not as available as you would think. I think conceptually maybe, but DAAR does not, and probably won't, because I wouldn't know where to get the data from. It does not measure impacts of pricing. There has been some talk about studies and there's

been a few studies out there about pricing but that's not this. I think this gentleman was first.

UNIDENTIFIED MALE: Sure. Thank you.

UNIDENTIFIED FEMALE: Sorry, could you provide your name and affiliation, please?

UNIDENTIFIED MALE: Sure, [inaudible], Neustar. So, on spam specifically – and I'm not a fan of it.

UNIDENTIFIED MALE: I thought you were.

UNIDENTIFIED MALE: No. I run a mail server on my own time. I'm not a fan of it. I'm curious your thoughts, because when you compare it to something like malware or phishing, one of the nice things about malware and phishing, especially if it's something really obvious is you can actually visit the site and see it. Phishing, particularly. You don't even have to be real technical to go and visit it.

One of the challenges with spam is oftentimes it's a compromised system, part of a botnet sending junk e-mail to recipients and maybe that junk e-mail is driving them to a domain name.

JOHN CRAIN: Okay. So on the spam, we're not taking the names out of the headers. This is kind of the landing pages where they're sending people and stuff. But, yes, some things are easier to identify than others. But malware can be on a compromised system just as much as spam can. Phish often points to compromised systems.

UNIDENTIFIED MALE: Sure. Yeah. And the reason I only mentioned it was because working in infrastructure one of the things you do is you want to verify [inaudible], especially if you're going to take some kind of action. It's often easier with malware and phishing because you can visit it and see the phishing page and then it's clear. If it was spam, a lot of times what will happen is a third party will provide you a data source that says, "Hey, we believe these domains are involved in spam." But often that's all you get. So, it's one of those things.

I'm just mentioning it. I just thought if you had any thoughts on it because you're someone who [inaudible].

JOHN CRAIN: So, I don't disagree with you, but the lists that we take the data from are the ones that are being used to block things.

UNIDENTIFIED MALE: When we do [inaudible] spam [inaudible] a lot of times they may not completely block it but they might score it really badly.

JOHN CRAIN: Yeah, exactly. So, these are the lists that people are using, whether they're blocking or they're giving them low scores and filtering them off to somewhere else, these are the lists that people are using. The fact that most of the products seem to use these same … Because there's lots of lists that don't get used as much, but the fact that these are in heavy use, to me at least, in the case that they've done their homework and that the producers of those security solutions put trust in their data.

Now, there's a difference when you're blocking something and trying to protect something than when you own that asset as a registry because you're making a risk decision and if you think that there is a 95% chance that something is going to be problematic, well, you block it, right? And you want to be 100% sure sometimes right before you take action. But that's just the nature of the game. But people are using these lists.

When you see the lists out there of who are the worst TLDs or the baddest TLDs or the spammiest TLDs, they're using these kinds of lists. When you see people blocking stuff and your customers getting blocked, it's because they got on one of these lists.

So, yeah, it's not 100% accurate all the time. There are false positives, but the data that you can find out there – and there's not much and this is another thing maybe we can do some research on – is that the lists we use tend to have low false positive rates, like less than 1%, some of them .0something but there is not much study out there. Maybe we should do some study.

UNIDENTIFIED MALE: For some of the work I've done, we do actually look at the lists and actually go and do further research. We see false positives and phishing and malware lists not because they were ever false positives. It's just that hosting providers sometimes subscribe to these same lists and they'll receive it first and they'll go and they'll change the permissions on a compromised website so people can't get to it, before anybody else does. So, it's a false positive because by the time someone else gets there, it's already been cleaned up. I thought it was interesting.

I guess I only had one other questions. On DJAs, I think you've probably worked on a lot of the same take-downs as I have, one thing that we're also seeing is DJAs will have a large selection of domains in a given day and the attacker only has to register a few of those domains because the compromised systems will try 1,000 DNS queries a day, and if one of them hits, then they've got [inaudible] control server. But, on the other side, on the registry side, you have this challenge of now you have 1,000 names or 100 names or 10 names a day that can potentially be abused.

JOHN CRAIN: Yeah. This Internet thing is really funky. You know, 800,000 names in Avalanche. I can't remember how many we did in [inaudible]. I think it was 50,000 names a day over 100 TLDs and they just need to get one, and if we're trying to prevent the badness, then we kind of need to block them all. It's just like DDoS and things like that which is actually enabled

by many of these things. We're kind of on the losing end of an arm's race here. Yeah, it's depressing. Go ahead. Could you use the mic?

UNIDENTIFIED FEMALE: Could you use the microphone, please?

ROBERT MARONEY: Yeah. Robert Maroney, DVP. You've been very clear about the limitations of your source of data and it's not under your control. I'm just curious, what's the latency once an issue might be resolved by a registrar or a registry? Could you just speculate – it's just speculation in your professional world – on how long it takes to clear, if you will, these blocking lists?

JOHN CRAIN: Who here who is more qualified than me wants to answer that? Because there's a bunch of people. You want to pass it?

RAYMOND: Hi, Raymond [inaudible], one of the SURBL guys here. Depending on the nature, usually less than a day. Can be even 12 hours. But, as soon as it drops from the zone file … So, as we see it disappearing, we take action on that. And that's also [inaudible] for registries and registrars to share also with us because we get zone files also from the same system as you guys. [inaudible], but if we get the zone files faster, we can also verify faster. So, there's always reasons to get that more speeded up. And it's

not always a limitation on our side, but we also have the same inputs as what you're using.

JOHN CRAIN: And I will say that we've had lots of conversations about this over the last year or so about what is the feedback loop between this industry and the blocking industry to make sure that, instead of talking across each other, which we sometimes do, that we actually talk to each other. Can we feedback data about the take-downs that we do to the RBL providers that announce them?

I imagine – and I know because I've talked to people – that, for example, if they got the zone file, it wasn't in or they got some kind of notification where they could do some kind of check to see that it's no longer resolving, they're happy to take it out of the block list. We've never, as an industry, really sat down and talked about how we interact better with the operational security community to actually make things better for both sides. Part of DAAR, the purpose of DAAR was to drive discussion and if these kinds of discussions come out of that, it's a win. Does that answer your question, Robert? You can see them better than I can.

CHRISTINE: Thanks. Christine, Amazon Registry. I had a question about the … You said that on December 1st you were going to publish the results of the public comment and that the Registry Stakeholder Group comment, which I helped write, so if you have questions, let me know.

JOHN CRAIN: Oh, thank you. That's good to know.

CHRISTINE: The four private researchers … So, in the interest of trying to learn from everybody, you said some of the researchers wish to remain private, so is it just their name you're not going to publish but you'll publish all of their questions and their comments?

JOHN CRAIN: So, once they realize the comments were going to be published, we got different [inaudible]. So, I talked to them and said, look, we need to publish the comments so that the community can see them and so we can take action on them. Some of them said, "Yeah, sure, go put my name on it," and some of them said, "Yeah, but take my name off." And at least one of them said, "Oh, let me just redact this and send you a slightly different version because there are things in there that will identify me personally and I don't want …

We're dealing with the security community. I work in that community. You think I'm paranoid, these guys are [inaudible]. I always say it's not paranoia if they're really out to get you. So, you have to be a little bit careful when you work like on botnet take-downs and things like this because you are actually … And people in the industry need to understand this. You're actually dealing with organized crime and bad people. So, you have to sort of give them a little bit of leeway.

They're not used to the ICANN transparency process where we just put everything out there. So, we're trying to accommodate them. Some of their comments are excellent. They just don't want to be associated with them because they do this kind of work on a daily basis. They look at these things. So, their comments will be out there. Their names will not be attached. You shouldn't be able to tell who it is from them and we will answer each comment they put in line by line. Well, not line by line, but comment by comment and we will do the same for the registry stakeholders. Now that I know who wrote it, we will …

We're actually quite a way through, going through them. With Dave leaving and Samaneh coming on, we've got a bit of a delay which I apologize for, but that's life. But we're going full steam ahead. So, we're going to … If we can get them out earlier, we will, but we're committing to December 1. Now Richard.

RICHARD HILL: Thanks. Richard Hill. From your previous question, I get the impression that you're not allowed or you don't wish to or whatever publish the actual names of the worst offenders, yet I think that would be – not the people. I mean the domain names that are the worst ones. Yet, that would appear to be useful. So, what would have to change so that you could actually publish those?

JOHN CRAIN: My board would have to tell me I can publish them. Like I said, it's above my pay grade. I completely understand why people are nervous about

that. Not just liability issues and stuff, but also reputation issues. So, it's not a decision that I get to make as a staff member and I suspect the board will only make that in consultation with the community.

To be honest, there are plenty of other people putting out these lists. Everybody knows who those names [inaudible]. If you're in the security industry in any way, shape, or form or even if you're in the DNS industry, you know who those are.

So, yes, name and shame is fun. But it might not be the most effective. What I want to see is I want to see [inaudible] graphs change. I want to see the registries who are being abused – and I see most of them as being abused – learn, take stuff back from this, and clean up. We can probably do that better by giving them the data than naming and shaming because the registries and the registrars are mostly good people trying to do a job. They're trying to turn a profit. And they're not security people. Most of the registrars and registries are not security people. So, if we can help by getting data to them, that may be more productive than naming and shaming.

Now, not everybody agrees with that and that's a discussion for the community to have.

UNIDENTIFIED MALE: Probably one thing that would also probably help is – and I know you mentioned data sources that have been used in filtering products. And one of the things, for example, I've used Spamhaus and SURBL stuff for specific purposes – those are largely built for those filtering

applications. So, the [inaudible] here's a domain name that we believe is [inaudible] of use. Great. I'll put it in my filter or I'll score it lower. But, richer data.

For example, if you tell me a domain is involved in phishing, if you give me the full URL of the phishing page – not just the domain name, the full URL. Because half the time it's a compromised WordPress blog.

UNIDENTIFIED MALE: Yeah, and they've buried it in WP-upload or WP-config, some sub-directory and it's not even a [inaudible] going to pick it up. It's like a custom landing page. So, if you get that, that's far more valuable. Other data sources, for example, have that. I think that would help.

The other question I really have on spam is, is it also a bit of like a squeezing a balloon problem? Because you deal with it here, but if spam were still able to make a profit doing what they're doing, I still think we should deal with it. But that's kind of a bigger meta conversation. How do you make the entire activity undesirable enough that people just simply stop doing it?

JOHN CRAIN: Yeah. Fully agree. There's going to be some of that squeezing the balloon but it doesn't necessarily mean we shouldn't squeeze it, right? That's always a dilemma. It's one of those things about the perfect being the enemy of the good. If we can lower the level, that's less victimization. If we could just shorten the time that they can abuse people.

UNIDENTIFIED MALE: And the only reason I mention this is one of the challenges is when data sources and [charts] get thrown up, people say, "Oh, this TLD has got this much." Sometimes it might be because the data set isn't rich enough or the challenge itself or the operational challenges that are working with it at a DNS level are different.

JOHN CRAIN: Yeah. That's why we went for the stuff that's out there in products instead of having a back-and-forth conversation with ourselves about which data is rich enough, which is not. We chose to go with what is actually in use out there. It's almost religion when you get into some of those levels. As I said earlier, and it's something I strongly believe in, we need to have those conversations with the people bringing the data and we need to have a feedback loop so that … Maybe we need richer data, but until we start talking to them, having those conversations, we're not going to get there.

UNIDENTIFIED MALE: Yeah. Actionable, really, is …

JOHN CRAIN: Yeah. Actionable is the word you're looking for most of the time. The folks from SURBL and Spamhous and that, they're very smart people, so if you talk to them and we can have a conversation, maybe we can improve the ecosystem a little bit. I'm all for improving the ecosystem.

ROB GOLDING: You touched briefly on something about the time and the time to take-down there which is of interest to us as a host. Are you now or are you planning to do trending to see how long something stays in the DNS while it's bad and whether that changes over time, to see whether or not this is having an improvement on the ecosystem?

JOHN CRAIN: So, other people are measuring that. It's one of the things we could measure. As we go through the review … As we go through the comments, there's all kinds of suggestions like different things we should measure. We've got a lot of the "you should use my data because my data is [cooler] than the data you're using" which we expected. You should measure different abuse types. All the stuff you would expect around something like this.

But we are … And this is why I hired a data person. We are looking at what are the different things we can measure, either inside or outside DAAR. DAAR is just one tool. Maybe there are other things we can do. But what I really want is for us all to have the conversations and that's starting.

So, if we were to turn of DAAR tomorrow, I'd be upset. But it's already a win because we're already talking about these things.

UNIDENTIFIED FEMALE: John, I have a follow-up question from Jim Prendergast to clarify his earlier question. “Will DAAR have access to the data gathered from the upcoming audit I referenced?” Thanks, John.

JOHN CRAIN: Oh, no. No. DAAR doesn’t use data that is unique to ICANN. People say to us, “Well, you could go get all the WHOIS data that ICANN has and you could take all the IANA IDs out of it and associate that way. Well, that kind of defeats the point of other people being able to do it.

Also, there’s a few lawyers in this room, I suspect. We have contracts around how we can handle that data and data that ICANN gets pretty much always is bound by purpose. So, even if we want to get the data – and trust me, I’ve tried. Even if we want to get the data, we can’t. We have good lawyers. Maybe not everybody agrees but I think they’re pretty good lawyers and they will not let me touch that data. So, no, we will not be getting that audit data into DAR. I can’t see any way that would happen.

UNIDENTIFIED MALE: One other thing about contracts is on specification 11.3B which is part of the registry agreement, phishing, malware, and botnets are called out but spam isn’t listed as an item, so I think that’s one of the … Registry operators are focused on phishing, I assume.

JOHN CRAIN: Yeah. I agree.

UNIDENTIFIED MALE: Phishing, botnet, and malware.

JOHN CRAIN: And that is a policy discussion that if we're having conversations around this, that's a policy discussion for whenever the next contracts get done. And maybe that's right that it's those and maybe it's not, but that's ….

UNIDENTIFIED MALE: And it's impacted years of operational activity, too, [inaudible] contracts versus a broad definition of use.

JOHN CRAIN: I am not a lawyer. I do not want to get involved in contracts negotiations, thank you very much. But, yes, I suspect that's where those discussions will be. And as you said, there were the three threat types that were listed in 11.3B. Richard?

RICHARD HILL: To that point, I don't know, but I suspect it's because their definition of span is not the same according to different legislations. So, there's no kind of universally agreed definition of spam.

UNIDENTIFIED MALE: Yeah. Historically, it's been commercial. It's a lot of commercial junk e-mail.

UNIDENTIFIED FEMALE: Sorry. Can you talk into the microphone, please? Thank you.

UNIDENTIFIED MALE: Oh, sorry, I thought I was. Historically, it's been a large number of commercial junk e-mail trying to sell products and some that contain malware. I don't know the breakdown on a given day or week.

JOHN CRAIN: There are lots of different classifications of spam. We didn't dig into that. Maybe we could. Maybe we can't. We'd have to look and see what the classifications are. But yeah. Most of these conversations are … There's never one side that it's clear that this is exactly what we should do and this is exactly what … Especially around spam [inaudible] is bad. Some people would say the stuff in the headers are where it was … The name which we all know can spoof is a [inaudible]. But I disagree.

UNIDENTIFIED MALE: Yeah. And technically it is possible to send a spam e-mail message referencing a domain that may technically not be involved in anything. I can forge an e-mail. SMTP protocol is simple.

JOHN CRAIN: Okay. I'm not seeing anymore questions. Oh, I got one in the back. Last minute. Is this the only microphone we've got? Oh, we've got another one. We've got one coming down. [inaudible] when he gets the microphone.

RICK: Thank you. My name is Rick [inaudible]. You said I think at the beginning that you don't use any country code domains data at all due to zone file problems I think is the main reason. But I was just thinking that the lists that you use probably do give you country code domains and the data they produce, so you could at least get sort of absolute numbers. [inaudible] maybe you might be missing out on some trends by not taking them into account, so you might say there's a reduction in malware over time, but they might have all just moved to cheap country code domain names. And you could capture that if you took the country code domain data from the lists you use.

JOHN CRAIN: Having those discussions. Oddly enough … So, I said I'm an optimist but I'm actually not. I'm a bit of a pessimist. So, I thought when we announced that we had this project and the data that everybody would scream and say, "No, no, no. Stop. You can't measure things." And some people did. And a lot of country code people said, "Can we give you our zone every day? Could we be measured?" We're looking at that. There's a whole bunch of things around that because you're only going to get some of the country codes. You're not going to get them all. We have

data, of course. Every country code that's on those lists. But how do you normalize that?

So, it might be a slightly different look than the current DAAR stuff and slightly different statistics. There are lots of other things we could measure, too. I think the country codes is an interesting area and it might not be in DAAR. We might do it by some other method but we don't know yet. I think that's in one of the comments and we'll answer it.

Okay. For the adverting behind me, who goes to the GDD Summits? So, after the one in Bangkok, we're going to take two days for something we call IDS which is a more technical conference. And one of those days is going to be on abuse. I am going to be managing that content. I'm going to be asking people in the industry to help. I think – and this is where you'll get to call me crazy – that we will dedicate half the day to how do you measure this stuff and the other half of the day to how are people tackling this stuff. So, how do we measure abuse and what are the people doing to tackle abuse?

One of the comments I get from the registries I visit is, "What's the best practice? What should we do here?" It would be good if industry players could share some of their knowledge. So, we're planning to do that in Bangkok.

Then, I have to advertise for our for our friends from DNS OARC. So, if you want to get completely technical, the Operations and Analytical Resource Center – I used to be on the board, I should know this – are

having theirs the two days afterwards. So, if you want to go spend an entire week in Bangkok, we've just given you a reason.

But I would like you to come to the abuse discussions because I think the strongest thing we can do together is actually have these conversations and try and find solutions together. Like I said, if the only DAAR has done is kicked off those conversations, that's a win for me.

If we have no more questions – and don't knock the microphones over. You have 15 minutes of your life back. Does everybody want to have another session in Kobe? Does anybody not want to have a session in Kobe? No? Okay. We'll do another session in Kobe. I'm going to be hanging around for the next 15 minutes if you want to come have private conversations.

UNIDENTIFIED FEMALE: Thank you, everyone. The slide materials that John just presented have been posted to the public schedule available to you. The recording for this session and transcripts will be posted to the public schedule as well within the week or so. Thank you.

[END OF TRANSCRIPTION]