ICANN73 | Virtual Community Forum – Tech Day (1 of 3)
Monday, March 7, 2022 – 09:00 to 10:00 AST

KIMBERLY CARLSON: Thank you. Hi, everyone. Happy ICANN 73 to you all. Welcome to Tech Day on Monday, March 7th. My name is Kim. Along with Kathy, we are the remote participation mangers for this session. Please note that this session is being recorded and follows the ICANN expected standards of behavior.

During this session, you may use the Q&A pod to submit your question or comment. We will read them aloud during the times set by the chair or moderator of this session. If you would like to ask your question or make your comment verbally, please raise your hand. When called upon, you will be given permission to unmute your microphone. Kindly unmute your microphone at this time to speak. All participants in this session may also comment in the chat. Please use the drop-down menu in the chat pod and select "respond to all panelists and attendees." This will allow everyone to view your comment.

Please note that private chats are only possible among panelists in the Zoom Webinar format. Any message sent by a panelist or a standard attendee to another standard attendee will also be seen by the session hosts, co-hosts, and other panelists.

This session includes automated real-time transcription. Please note that this transcript is not official or authoritative. To view the real-time transcript, click on the closed caption button in the Zoom toolbar.

And with that, thank you for joining and I'll turn the call over to Dr. Eberhard Lisse, chair of the ccNSO Tech Working Group.

EBERHARD LISSE: Welcome, everybody, from my lovely home office and not from Puerto Rico, which is a shame but what can we do. We hope that the next meeting in a year or so will be in Puerto Rico so that we all can meet in-person there. We will probably all meet in-person in The Hague in three months. But as we discussed among the panelists a little bit earlier, I am becoming a great fan of the hybrid Zoom format so that people—that we even extend the audience for people who can't make it there. We have always had access from virtual participants but we will push a little bit harder.

That said, I think we have got ourselves … As usual, I'll go through the agenda beforehand. And we have quite a decent program, I think. I will start with my opening remark that I'm busy with and then we have Graeme Bunton. Let me just quickly look. He just waved his hands. He's from the DNS Abuse Institute and will make a brief look at potential architectures for preventative methods of DNS abuse mitigation.

Then, Ed Lewis will follow up on his previous presentation about DNS Core Census, [CS code], another dataset that is available. That was a very interesting and good presentation last time so I was very keep on following this up.

Then we have our mandated break. And then, Andrew McConachie is going to speak about the root server data that root server operators

collect in terms of the RSSAC002 specifications. That link is clickable. So since you have here the agenda and we will publish it, if you want to have a look what's going on, you can just click.

Then we will have Craig Schwartz from fTLD. They do .bank and .insurance. He will talk about DMARC on the public service domain list—not on the second level as usual but on the top level. DMARC is a commonly used thing, which is a black hole to me. I do not understand it and I never get it right. I'll never be unable to send emails to my son when he uses his Gmail. Eventually, I'll figure out how that works. But it's always good to hear about stuff and advances in that area.

Then Gustavo Lozano will speak about the TLS client authorization using DANE in the MOSAPI website that they have. As we all know, they have an API. That way you can access data that is collected.

Then, always, we have our host presentation, or rather, we always offer the host a presentation. And Dr. Rodriguez is going to present. Last time he did, it was very interesting after the flooding and hurricane. That they had that was really very keen. So I'm quite anticipating to hear the follow-up on that.

Then we will have Paul Vixie and Kathleen Moriarty, or the other way around, speaking about going dark. That basically means using more, and more, and more, and more encryption. If people don't use that or are not aware that it's incoming, operators may find it difficult to interact with others.

That's actually something that we even see on our own little registry that we have for .na. We find it very difficult to convince people to use just basic GPG encryption if we want to give them login credentials. It's always a stress. Never mind doing virtual agreements, doing agreements electronically, which is currently not lawful. But there is a new act coming in which requires some form of doing this.

Then Owen and Graeme Bunton were very kind and interested in organizing a roundtable on DNS abuse. Dan is not going to speak but he and Graeme were organizing this table. I'm grateful for that. It's about an hour and I have no input or insight into the actual content. I like my roundtables always to be organized without editorial control or input.

Finally, we have probably got the hallmark presentation. Dmitry Kohmanyuk will speak about the recent experience that they have in .ua. As we all know, the Ukraine is under an aggression—attack, which is a violation of international law. In any case, they encountered in the weeks past some DDOS and now they encountered actual shelling. So it's going to be the hallmark presentation for us to hear what they did. We all can learn from this in a smaller way. We are not getting all into war situation but we have had disasters, natural and otherwise. So it's always good to know what you do, what you can do, how you can plan, and so forth.

Finally, I have volunteered Regis Masse from the .fr, French ccTLD. He's a member of the technical working group. I like to rotate the closing remarks. I don't like to do them myself so I want one of the others to

give a little bit of input or insight into the presentations. That said, I was to minutes faster than I intended. Graeme Bunton has the floor.

GRAEME BUNTON: Thank you, Eberhard. Good morning, everybody. It's an honor to go first. My name's Graeme. I'm from the DNS Abuse Institute. For those who don't know me, I've been around here for a while. I spent quite a long time at one of the largest registrars in the world, and as part of that, chair of the Registrars Stakeholder Group for a number of years.

But now I'm working at this thing called the DNS Abuse Institute, which is an initiative created by PIR, who run .org. PIR recognized that DNS abuse is an increasing problem for registries and registrars around the globe, gTLD as well a ccTLDs, and that it's a complicated global problem, and that really, solving or helping to improve the ecosystem on DNS abuse required some centralization function or really someone to step into it and coordinate activities around the world. So here we are. The DNS Abuse Institute exists.

It's got, briefly, three pillars—education, collaboration, and innovation. Innovation is where we're really looking at building tools and technologies to help with DNS abuse. I'll talk more on the roundtable about some of the other innovations that we're working on—a centralized abuse reporting tool and some intelligence initiatives. But today, I'm really previewing some of the thinking that I'm doing around preventative methods for mitigating DNS abuse and work that we're going to be beginning later this year—probably in Q3–Q4.

So I'll share my screen—I've got a presentation—in a sec. But briefly, what we're going to do here is I'm going to give you some context for how I think architecting preventative measures are going to work and what I mean by that. Then we're going to go into a bit of a dive into a retail registrar's domain registration flow and some of the opportunities for doing preventative mitigation in that flow. Then we can talk about what needs to be done to make this work really well.

So with that, I'm going to share my screen. There should be plenty of time for questions at the end of this presentation. And you all should see slides.

EBERHARD LISSE: We do.

GRAEME BUNTON: Great. Awesome. Okay. So here we go. Preventative versus reactive is the first thing to probably explain a little bit. These are what I consider two broad buckets for how we're going to reduce DNS abuse.

Reactive is much what it sounds like, where a registry or registrar is receiving a report of abuse. They're going to investigate and they're going to mitigate as they see fit. There is lots of room for improvement here. You can improve the speed of reports. You can improve the quality of reports. You can improve the speed at which a registry or registrar is going to act. Lots of work to be done around reactive and work that the institute has underway but really not what we're talking about today.

Today is about preventative methods. This is detecting potentially-malicious domains before either registration is completed or before the domain resolves. There's a couple key pieces here. One is that it's potential. The harm hasn't occurred yet. If you're looking at these from a very high level, boy, preventative is better in many ways. You want to stop a harm before it's happened. Reactive measures, no matter how quick, someone has detected that this domain name or website is up to something no good and something bad has happened. So you really want to get in front of that as much as possible.

But there's tradeoff here in cost and business impacts. Preventative measures are, by their nature, introducing some sort of friction in the registration process and that causes some problems for registries and registrars who, by and large, have spent the last 25 years trying to remove friction from the registration process. But long-term, I think it's pretty easy to argue that preventing abuse from happening in the first place is going to be cheaper. You don't have to employ as many abuse people to triage tickets and manage these things.

So we need to think a little bit about where we're going to try and do this preventative mitigation—this potentially-harmful domain detection. There's really two places. You can do it at the registry or you can do it at the registrar. The benefits of doing this sort of technology at the registry is that you can apply this sort of detection across an entire zone so you have consistency within a TLD.

The problem with doing it there is that registries have very limited information in comparison to, say, a registrar. So they're going to have

whatever WHOIS information they've got and the attributes of the domain name themselves. But they don't have any of the payment information, for example, or other really useful attributes of the transaction that's occurring.

Then also, there's a registrar as an intermediary. In most cases, I understand that that's not the same for all ccTLDs that might operate as both. But often, you are then trying to … You've suspended the domain name. You're then going back to the registrar to try and communicate with the registrant that something's not quite right.

At the registrar, you've got far more information about who's doing the transaction, the credit card details, who the customer is, what other domains they might have in their account—way more attributes that you could try and leverage to do this sort of potentially-harmful domain detection. The downside is that you end up, at the registry level, with a boatload of different implementations. There's not a lot of consistency here. So you need to make a choice about where you're going to try and do this.

Another key piece of this is incentives. Like it or not, this is a commercial industry and we can't just expect everybody to go and do everything out of the goodness of their heart. So as I mentioned earlier, preventative methods require friction in the registration process and that's a tall order.

Not only that but it requires scarce and expensive engineering resources. That means really getting people to write or integrate code in a thing that's not driving more revenue, that is maybe reducing your

costs long-term. The reality of the industry, especially in the gTLD space, is that most of the large registrars' codebases are very old. They're still dealing with a very complicated ecosystem. Dedicating dev resources to a potentially-not-critical task is a tall order.

So long-term, you need to be able to demonstrate that not only are you lowering your costs but you're having little or no impact on revenue. So you need to keep these things in mind as you're looking at what technologies that we can implement to try and reduce DNS abuse.

All right. That's hopefully some useful context for how to think about where to implement these sorts of technologies. Now I'm going to dive into a little bit of how to actually architect these things, or how I think the best place to architect these things is going to be.

This is a very high-level diagram of a domain registration process. We've got the registrant does a search. They're going to select a purchase. They can create an account and pay for the domain name. The registrar is going to submit the domain to the registry. The registry puts it in the zone and the registration is complete. It's resolving.

So as I've discussed, there's two places for preventative methods to be put in place. One is going to be somewhere around in here and this is before the domain is submitted to the registry. Then we have another opportunity. Whoa. I just opened up Lucidchart. There we go. Then we have another chance, at the registry level, before we put the domain in the zone.

But we're really going to be talking, for the rest of this chat, about detecting potentially-malicious domains before these things land at the registry. So in this context, I'm also really focusing on a retail domain registrar. In the wholesale model, this is going to be different because they're going to have access to different information. I saw there was a hand. I'm going to get through this and then I should have some time for questions. So that's a high level. Go ahead.

EBERHARD LISSE: Just a quick … No hands will be entertained during the talk. There is sufficient time for discussion afterwards so there is no reason to worry about it.

GRAEME BUNTON: Okay. So let's dig into a more detailed look into this domain registration flow at a retail registrar. Again, a registrant search. Registrar returns results. They're going to select a purchase. We're going to go into some sort of account creation method, where there's some personal information submitted, a payment method.

And I will say this is built based on a number of conversations with some large retail registrars and how they've architected this. I've turned this into an optimized version of this flow.

So they do a payment auth. This is not actually taking money. This is where they check that the credit card is good. If that's a no, it's usually, "Try again. Use a different payment method."

But the key bit that I want to talk about today is this, which is this fraud and abuse check. Just about every registrar has something like this in place right now. That's because it's a high transaction volume business, generally. And in order to do that high transaction volume, you need to be checking for fraud. Almost everybody is using some tooling like this as part of their relationship with their payment processor. Big payment processors are going to be like Stripe, Square, PayPal, Braintree. I think there's about five or six that are going to represent about 85–90% of the payment processor marketplace.

So what I think we need to do—and one of the two key points I'm going to make here today—is that we need to get better about using this. Those different payment processors have more than just credit card checks and that they have rules and tools that you can use to input more information about a transaction. And we can leverage those to both reduce fraud, which is where a registrar's self-interest lies, and reduce DNS abuse, which is where our collective interest lies.

So that means not only putting in t the payment method and perhaps things more advanced, like the IP address or the geographic location of the IP address, against the location on the credit card. But you can also put in things like the domain name itself. Does it include particular dangerous words like "support" or "login?" Does it have a lot of dashes? The attributes that, if you're building a machine learning model like the COMAR project did for SIDN, are identifying attributes of potentially malicious domain names.

So that we can leverage these particular checks without having to develop new technologies, without having to integrate other services, we just need to be better and smarter about leveraging the tools at our disposal already. I think there's a boatload of work to be done here and I think we can use it to really effectively reduce DNS abuse.

So in this payment flow, a domain goes into this check. More information is passed. And we have three results. We have a "no," we have a "review," and a "yes." The second component that I'm going to talk about is what happens between the yes and the review. So just to make sure that's clear, no is the fraud and abuse says, "This domain is just bad. You should end this transaction," and that's what happens.

The next bit is what happens is if it's a "yes" or a "review." "Yes" is pretty straightforward. This is where the check says the domain is good. It gets submitted to the registry. Payment is collected. Registration complete. Easy peasy.

The next bit is important because I've seen this done differently at every single registrar I've talked to. So a domain goes into a queue. And what should we do where it's in the review queue? The first thing I think we should do is that we submit it to the registry, which actually captures the domain. The domain is registered.

The next thing we do for a domain in review is to lock or suspend it so that that domain cannot do any harm between when it's registered and when the review occurs. I say this because I have seen registrars that don't run their manual abuse queue review 24/7. So they're putting domains to review but they're already being registered. And they're

seeing large upticks in abusive registrations Friday at 6:00 PM because bad guys know that they're going to get a full weekend of phishing out of a domain name before it gets nuked on Monday. And for them, that's plenty of time.

So if you're architecting this as a registrar, by registering but suspending your manual review queue domains, you're going to limit the potential for harm but you're also not going to lose the customer. The domain is there. You can send them a notification that says, "Hey. We've got your domain but we need to clarify some things. Please call our abuse team or we will follow up on Monday—" something like that.

This, I think, gets us to a nice happy place where you're not alienating customers too badly but you're also preventing harms from occurring over the course of a weekend. It allows you to reduce the number of hours for which your abuse mitigation team is running. So I'll try and wrap it up relatively quickly from here. If it's a "no," you're going to release that auth and delete the domain going forward.

This is the work that I think we need to do. We need to identify the transactional attributes so that we can get the most value out of what we're putting into these common payment processors, tune the weights of those attributes. We need to globalize these learnings. All the work done in the space right now is very Anglo-centric—very focused on English. These are global problems so we need to be better about that. Then we need to share these results. This is really going to work the more that we can share it with everybody around the world.

To sum all of that up into two bits, one is let's get the most out of our payment processors that the industry is using already and find that place for self-interest of reducing fraud. But also, we can reduce abuse. Then the second is to architect those domain registration flows to make sure that whatever is going into the manual queue that we're reviewing isn't able to be harmful on the internet for the amount of time it takes us to get to that review. So I'm going to stop there and I am happy to take questions. And I'm going to stop sharing my screen.

EBERHARD LISSE: Thank you very much. Interesting presentation. I forget, in my introduction, to welcome specifically the Fellows and the NextGen participants, which I apologize for, and here will do. Do we have any questions? Please raise them. Raise your hands. There is questions in the Q and answer. There is a question from Yoshiro Yoneya. "How often does manual review happen?"

GRAEME BUNTON: It happens all the time, especially for larger registrars. They have a constant queue of domains that are going into that manual review. Tuning it is very different. So you want as few domains going into that queue as possible and you want to make sure that … In a perfect world, that detection would be perfect. The domain is either good or it's bad and there is no manual review. The reality ends up being … And it's very different across registrars, depending on how well they've implemented these checks.

But it could be that something like 10% or 20% of their registrations are going into a queue to manage. If that 20% is 90% false positives, all you're doing is irritating customers. But if that manual queue is 90% abusive and you've architected appropriately, you've caused very little harm and you're doing pretty well. I think I see a hand from Stephen.

EBERHARD LISSE: Stephen, you have the floor.

STEPHEN DEERHAKE: Thank you, Eberhard. Thank you, Graeme, for the presentation. As a registry, we're pretty dependent on our registrars for doing the initial round of vetting. I'm just wondering if you have any thoughts on the balance of us, as registry, doing additional vetting on what's coming our way. Or are we really dependent on registrars for this? Thank you.

EBERHARD LISSE: Thank, Stephen. In most circumstances … And I understand it's not universal. In most circumstances, registrars own the customer relationship. They really should be, as much as possible, the pointy end of the stick on abuse. So they should be robust and vigorous in their anti-abuse stuff. That's part of the work of the institute that I'm focused on is making sure they have the tools and resources to do that.

But I also fully think it's within the power of the registry to keep their zone in the way that they want it. And I think there's a number of different mechanisms for doing that. You can do your checking on the

bits and pieces—the domains that are coming in. Then I think, broadly, there are ways to incentivize good registrar behavior.

They pay my bills but I also think it's an excellent program—the QPI, the Quality Performance Index from PIR. If you haven't looked at it, it's a really interesting thing where they incentivize registrars to have lower abuse. That appears to be meaningful and effective. I see a couple Q&A in the chat. Eberhard, I'm leaning on you to manage time.

EBERHARD LISSE: We'll manage the queue and the time. Don't worry about it. We have enough. Nick Wenban-Smith from Nominet. "A lot of malware arises from perfectly innocent registrations where the website is later compromised. Who is best place to mitigate this, registry versus registrar?"

GRAEME BUNTON: Thanks, Nick. So in the circumstance that I'm describing—this preventative approach—it's only going to be for malicious registrations. You're not really going to catch potentially harmful or potentially compromised sites because they haven't been compromised yet. They haven't been registered yet. They don't have an old version of WordPress on them yet. So all of this is really about detecting malicious registrations before they occur.

For a compromised website—and this is the subject of a panel later this week, a plenary session that I'm actually moderating—for compromised, it should be the host and the registrant first and then a

registrar should do a balance of harm test to see if the abuse is bad enough that they should suspend the domain. But that's a touchy subject we haven't done a lot of work on and I expect us to do a bunch more of that.

EBERHARD LISSE: John McCormac asks, "Would a registry-level blocklist for attempted registrations of problem domain names be a good idea? This would be kind of a blocklist bent on phishing and malware domain names. Confirmed bad registrations could improve its accuracy." Before I let you answer, I don't think is a very good idea because centralized lists are never up to date. But what is your opinion?

GRAEME BUNTON: My opinion is much like yours, Eberhard. I think RBLs, blocklists serve a purpose within the community—usually, more in the reactive approaches so that they're potentially useful for getting abuse reported to registries and registrars.

But the sort of work we're talking about here is going to be way more beneficial to identify the attributes of potentially-harmful domain names, which people have done. It's never perfect. This is going to be if you're doing north of 90% accuracy, I think you've done really well. But this is less about sharing a list and more about sharing the attributes of domain names. I think there's lots of room for this to be good based on that.

I saw a question in the chat from Romulo about, "Does the user have any actions against—"

EBERHARD LISSE: Can we do it in the right order, please?

GRAEME BUNTON: Sure.

EBERHARD LISSE: The first question was from Jose Gonzalez. "What must we do when there are domains in the SRS WHOIS system but the domain never goes into the DNS?" I don't think that's really an important question because if it doesn't get registered within two days or so, it gets deleted and then the money gets refunded. That's automatic in the gTLDs. For ccTLDs, it depends. We don't have this problem, maybe, in .na because we are too expensive and we don't refund. So the registrars can register whatever they want.

But the phishing industry lives off the industry of registering domains for a day or two and then getting their money back—using it for a day or two. So if it's in the WHOIS, it will say "suspended." If it then gets deleted, it gets deleted. That's basically the point.

Then I had a hand from Jacques Latour.

JACQUES LATOUR: Yes. Hi, Graeme. Good presentation. The question is if a domain goes pending review, what's the maximum time that it can held in that statue until a review is done?

GRAEME BUNTON: I don't think there is one. I think it benefits registrars to do that as quickly as possible. I think there's a natural review period based on the grace period of the registry, which I think is typically five days, if I'm not wrong, because that gives them the ability. You would want to do it in that amount of timeframe so that you can delete potentially malicious registrations and get your money back for that, especially if it's registered with a fraudulent credit card, so that you're not doing chargebacks and things like that.

JACQUES LATOUR: Perfect. Thank you.

EBERHARD LISSE: Okay. I'm closing the queue after Nick Wenban-Smith. So Romulo Cachin, a Fellow, welcome. "Does the user have any actions against 'no' decisions?" I don't think that's really a question for the presenter because that depends on every registry. They have their own appeal process. On ccTLDs, it's totally different because every ccTLD can do whatever it wants as long as it's consistent. I would think a "no" decision is not an issue for this presentation. That's more a legal process. Each registry, if you want to register a domain name and the registry refuses, what can you do about it?

GRAEME BUNTON: I think you're right, Eberhard, again, that this is going to be independent based on each different registrar's implementation of how they do this. If it's a no because you've failed their fraud check, they may fully just tell you to go elsewhere and that can be fine. But everybody's got their customer support queues and you can always try and reach out to your registrar to explain that that was wrong.

EBERHARD LISSE: My view on this is if somebody actually contacts the registry, that's a strong indication of not malicious because this is all automated high-volume things trying to overwhelm the registries. If the registry makes a mistake and you actually communicate with them, then they usually would not be too difficult. If the software decides something and then the abuse process—the manual check—takes it off and you say, "No. That is a mistake," then usually I would think that's good to go.

And finally, we had Nick Wenban-Smith's post into the question and answer. "ccTLDs are obviously out of scope for ICANN policies. So do you think that ccTLDs should individually mandate these sorts of checks at the point of registration? At .uk, we do it all at the registry level so that there is consistency for every registration regardless of the registrar chosen. And they have got 2,000 registrars."

We do this at the registrar level and we tell them, "If you register a domain, you're not getting the money back." And we are relatively expensive so we see very, very little of this. If we have the complaint

from the competition commission, but they go through an [inaudible] process, we haven't had to take a domain down. But we have discussed it with them. We would be a friendly defendant so they would sue us without trying to get our costs from us. That's the way we do it.

But each ccTLD should have their own way of doing it. If you have a functioning system like .uk has, I wouldn't need to change it. If you have nothing, then that's an obvious question.

GRAEME BUNTON: I think this is about that balance I was talking about at the top of this presentation between the consistency across a zone and the balance with the more information that a registrar has, especially in a context where you've got 2,000. Give or take, there's about 450 gTLD registrars out there—less than, I think, 80 with more than 1,000 names. So in the g space, I think this is easier.

But whether it's up to a ccTLD to mandate or not, I think, is up them for sure, of course. I'm going to be working on making sure that this information and these techniques are as available and socialized to ccTLDs, registry, registrar, everybody who wants it. That's the job of the institute is to develop these tools and technologies and make sure they're available across the ecosystem, regardless of cc or g. Then cc's or gTLD registries and registrars can do with that what they want. If it gets to a place where it's of quality enough that you think it's reliable to mandate, it could very well be extremely effective. Thank you.

EBERHARD LISSE: Okay. Thank you very much. There is one more question which I will ask the person to put into the chat so that we can follow this up directly and we are a little bit over time. Thank you very much. It was a very nice, interesting presentation.

As far as ccTLDs are concerned, if you run a registry that also runs a gTLD registry like Nominet does, then it's easy. You won't operate two different technologies for two different types of registries, so you just do it because as a ccTLD, you can do whatever you want. You can also adopt gTLD.

Anyway, thank you very much, nice presentation. I liked that. And Ed Lewis is now the next one and he has the floor.

You are muted. I can't hear you.

EDWARD LEWIS: Yeah, I was going to be nice and smooth here and share my screen right away. But then I realized I didn't know which window I wanted to use. Sorry. Yeah, this is the right window.

So, hello. Edward Lewis here from ICANN. I'm going to talk about some data sets that I have available. The title of the talk is DNS Core Census. I realize I gave a talk on the DNS Core Census last vTech Day and it was called an update then. So it's an update on the update but I have some more information coming up, too.

So previously on the vTech Day I talked about the concept of the DNS Core Census. And at the time I had really hoped to have the data

available. Besides the idea of creating this Core Census, I wanted people to have data in their hands so they could make use of the aggregation of the work that I had done.

The data is now available, which is the reason why I'm back on the agenda this time. And while I'm here, I'm also going to cover a second data set that's also been made available. It's available to everybody if they go to pretty much the same website. Now the reason for talking to this audience is that these datasets have information about TLDs across the board, ccTLDs and gTLDs.

And being that this is a ccTLD-somewhat focused audience—it's not exclusively—you've probably some interest here in what's in the data. There's some interesting information to have out there. Also want you to be aware of what's out there, not that there's anything that could be dangerous to be spread around but be aware at least that this data is out there and available.

Now, the other part of this is that I am doing this work kind of in an immature way. This is a first effort to produce data in this way, continuously updated all the time, updated every day actually. And in the format that I'm using, it's kind of experimental in some ways. There are so many choices that can be made. So I've brought this up in one other venue so far. They were not focused on TLDs. This venue is focused on TLDs so I'm hoping that people will dive into this and see whether they think the data is good or not, worthwhile of their time.

So in the talk I'll talk a little bit about the DNS Core Census. I covered that in more detail last time so I won't repeat that stuff. I'll introduce

the TLD Apex History in a quick fashion and also talk a little bit about how do you get the data and also what are the things I'm looking for in this data program. And I'm not going to read all my words on the slide. I want to be done in about 10 minutes so I have some time over for questions. And I may have to step out at the top of the hour.

So, in summary, there's a lot of information out there. It's focused on people and not scripts and that's been one of the frustrations I've had. I always have to pull down an HTML table, parse it, and then pull out the information. The computer has to pull information out that it just doesn't want to have to see, like white space, tabs, improper formatting, inconsistent formatting over time and so on.

There are a lot of sources of information about TLDs, so just getting it all into one place in one useful data structure has been a hassle. And it's one thing to do this for a short-term project. Presentations and dissertations would generally be one-offs where we do a lot of manual work, put the data together. Trying to do this repeatedly over time, which is where we get a lot of more information, is hard to repeat all the steps. So I'm trying to simplify life for a lot of people who are playing with data. And in operations long-term knowledge and history has also been very, very beneficial to have.

So the drivers for this, to get a little more focused on the application. I do a lot of work where I look at TLD activity, things that happen at the top of the DNS naming hierarchy. And a lot of times I have to divide between gTLDs and ccTLDs because the policies behind the two are

different. And so their behaviors tend to be different. And we look for trends that are different.

IDNs especially, there's interest in IDN activity and then again we want to divide that between gTLDs and ccTLDs. And there are a few other things in there. Geographical-regional groupings are something I'm asked to produce stuff for. If someone's giving a talk in South Asia, they want to know about South Asia, if they want East Asia and so on. And sometimes it's not very clear exactly what belongs in that region just by looking at the TLD names.

Now the data sets I'm producing are not judgements, ratings, or rankings. They're just collections of reference data. I do very little munging of data to create some fields. And I get the data from a lot of various ICANN sources primarily. But I also have some other external sources I lean on. And in a quick snapshot, these are the URLs that are of interest and I'm not going to go through this here. I'll talk more about this in the coming slides. But this is the menu of URLs that we're going to have information.

Licensing is something I'm working on. Don't have a final thing on that. Licensing was something I was asked about when I was trying to clear this data for publication. Across the board people don't have good licenses stated so I figured it's a good time to start this. The idea here is that we're trying to get the data as spread as possible, public or commercial. With attribution was one of the choices—some of the licenses we proposed.

On the other hand, this entire system is beta so it may go down or may be removed at any time. And we just want to say we're doing the best we can to give you the data and see what happens. I hope to have more formal terms published at some point just to make sure everything's clear.

For the Core Census, there's a version 010, a tag in there that's important because I have an older version that's available but I may not keep that around. There's a table directory which has a catalog. It's a file which is a catalog of all the other files in the directory. In my previous talk, there was a big hang-up about CSV versus JSON formats. So for the time being I have both out there.

I only have 35 days of the census out there. It's a rolling window. Again I'm a little constrained by disk space so for the beta program we're just limiting it to 35 days. And everything out there is compressed. There's a doc directory which is important. It has a data dictionary for humans, HTML and PDF so you can read it. It explains there's 161 different columns in the census spread over nine tables. I'll get to that.

And also I have some example code. There's actually two sample clients in that directory, one for pulling the CSV version and one pulls the JSON version and just prints out something. You can use the code for your entertainment. I don't have the census taking code in there, of course. This is just to pull the stuff.

For the TLD Apex History I have pretty much the same thing, a JSON and CSV file of the entire history. And you'll see why there's only one file there if you dive into it. There's a data dictionary. This dataset's a little

simpler. It's a little deeper but it has fewer columns. And then also sample code for pulling the CSV copy out and displaying fields to get some idea of what is in there.

For the Core Census, I started out with the question—and this is trying to motivate your interest in the data sets out there—is xn—example a ccTLD or a gTLD? It's not apparent. It's very hard to do this. So that's what's kicked off the entire work about two years ago.

Also, if you want to know by regional information, you can pull things out there. For South Asia, if you want to know how many ccTLDs are in India, India has double digits in IDNs because of all the scripts that are used there. And other places will have more than one script in places. Some may not have any.

So I pull information from a lot of places there, all over web pages IANA has, ICANN has. And I go out to other sources and I also pull in IP address information, IP address, ASN numbers, trying to find a mapping of the TLD's operations to the web. And it's a list of places there. Let's see, I'm going to speed up again because I want to make sure I have some time for questions.

The Core, it's the top of the—

EBERHARD LISSE: Edward, you're the last presenter before a break so we have no problems getting a little bit into a break. Take your time.

EDWARD LEWIS: Okay, well, actually I may have to take my son to school, so I have pressure on this side.

So the Core is the top of the DNS tree—TLDs, the delegations, down through in-addr.arpa, ip6.arpa, and subzones. And I go down to what I call the commercial registration boundary which is where TLDs in general are registering names on the behalf of somebody else, the registrars or other commercial activity.

The census pulls all this information I have available to me into one place. It assigns a category, whether it's a ccTLD or a gTLD or sub-ccTLD, sub-gTLD, and so on. And the jurisdiction, a two-letter ISO code that applies to the TLD. I use a worldwide one for the gTLDs, because they don't have a specific location. I also have—and I have it in italics here—what is considered to be useful in my eyes. That's an open question. More can be requested and some stuff can be dropped if it needs to be.

I look down through NS records. I look down at address records and so on. And I create nine tables. Of the nine tables, there's information here about what's inside of them once you go through them. Going through the data dictionary will help a lot to understand the columns again. I have compressed CSV and compressed JSON available out there because some researchers believe that we have to have JSON around where CSV is sometimes a little smaller. They compress the same way but CSV's a bit easier to ingest into other platforms.

The first version was 002. I did speak about that in some venues. It's available somewhere. I may try to backfit that data into 010 format if

there's interest in that. It makes using basically a year's worth of back history.

And I'll skip to the next slide. Now the data is from many sources that are in transition. This is why this is kind of in a beta program. Pulling this data is not straightforward. Some things have disappeared in the year or so that I've been collecting data and I had to replace the source. And it hasn't been that easy to just aggregate all this data. It seems like it would be very easy to do this but it's not.

Now I have an experimental arm of this work. I have two things in there, the determination of the commercial registration boundary. That is something which I have to look back at that to see whether it's doing a good enough job at it. I don't think it is quite yet. Also I have inclusion of whether going down to the ROA, the Routing Origin at a station, for BGP security information. I've been playing with that to see how detailed that can be.

It's a little immature. It's a lot of questions about this. I would appreciate a lot of feedback on this. And anything that comes back about … I want this to be easy to use and I don't know that I've been successful so far.

Now TLD Apex History, the other thing. There was long ago questions about how should we set up our DNSSEC. One of the big questions, how long should I cycle my keys? And the effort here is looking at the others who are doing this. What do they do? How do they roll their keys? How do they set their parameters?

And this is an observational platform which goes out and it looks at the top of each zone that I've been studying, only the TLDs, the top levels. I pulled out the records that are in the second half of this slide. And I'm able from that to determine a lot about how the DNSSEC is intended to work.

This is not an assessment of how it's worked. I don't check to make sure validation's working and all that. I am able to pull out enough detail to see what was pretty much intended by the actions of the operator, myself knowing the protocol and how it's put together.

This is a single table. It's updated every day. If a record appears in a day that's put in the table, it's not granular below that. It's not hourly, it's not to the minute. It doesn't see real-time changes to the zone.

Example entry, I have a thing in there where it has the owner, the type, the days it's seen. These are consecutive days that a record's been in there. So you might see a key show up in one state for a few days and then switch to another one for 90 days while it's being used for a [porter] and so on. Again, I do not cover validation or whether it's broken. It's in the system.

The data's been accumulating for some time. We've released the data going back to 2014 at this point. We've used the data to talk to operators in some of our meetings in the past. And it's been very helpful for an operator to see some history they may not have been keeping. It also gives you an idea of what their periodicity of work is. What's their clock? How have they been going about their business? And DNSSEC may just be one symptom of the overall operations of a registry.

And for protocol developers what I like to do with this is see how features are being used and when are they being deployed and at what rate. Is it good enough for operations is a good question.

My favorite toy is visualizing key life cycles and I think I've given a talk on that in vTech Day but maybe a year or two ago. I haven't updated that in a while but I will be coming out with more information about that at some point. And this really gives a good picture of how things have been put together over time.

So questions around here. I have a JSON format and this is my tickler to those who love JSON. Is this just a cool thing kids do? I understand that CSV can be harder to parse but my experience with using Python and the DataFrames and pandas, the two of them are just as easy as each other. But I'm open to being educated on that.

The resource records and fields, are they sufficient? What should be in there? And at what point do I go from making sure this has everything to just being overall scope creep?

So I would really appreciate people taking a look at the data. The intended benefits is to be a better reference for what's at the top of the DNS. Some historical recording of the state of the Core of the DNS would be nice to have. And I'm trying to take my data from as authoritative a source as I can or from direct observations, meaning that I haven't been as aggressive in getting ccTLD zone files involved. Because they're hit and miss, I don't want to just take the ones I have. I'm not being opportunistic here.

In terms of costs on your side, being in a beta period, I would really appreciate constructive criticism coming back. Be prepared for errors, things going wrong now and then. Be prepared for changes to the format, although on these last two I'm trying to be very careful, an operator mentality about that.

And if you have any feedback, I have information on this final slide, email address is in the middle there and that's right now the best contact for any feedback. And I will pull up here and let us see if there are questions, comments, and so on. And I will go back to the slide with all the URLs for the time being.

EBERHARD LISSE: Thank you very much. Again, very interesting to hear what you can do with available data. There is one question on the pod. "Is there any possibility of adding resolvers to the census? It would be very nice to have at least the quad resolvers, the big ISPs. It is it for the same reasons of observability, deployment of technologies, etc."

EDWARD LEWIS: Okay, so I have a couple of reactions to that. One is that I am reluctant to declare a roster of who are the appropriate resolvers to include. I would be willing to listen to other people's lists. I think it's a community decision there. What do we want to have included there? Because you get into a commercial area. I don't want to decide who is the worthy participant and who is not a worthy in that. So if someone has a good

list that they feel really should be looked at and it's a well-founded list, that would be the first step.

The second thing, though, is I don't have any specific insight into resolver operations. We could probably investigate that through some other way. What's going on here is I'm looking at the configuration information. This is not at all operational packet flow out there.

So depending on what's required, we'd have to decide who do we want to look at and then what are we trying to glean from information about them. I hope that's a proper answer, if that's the right context.

EBERHARD LISSE: All right. We have time for one more question, if there is any. All right, so then you are released to the school run.

EDWARD LEWIS: Okay, and I will be back.

EBERHARD LISSE: All right. Drive carefully.

EDWARD LEWIS: All right.

EBERHARD LISSE: All right, the next—

KIMBERLY CARLSON: Eberhard, I think you muted yourself.

EBERHARD LISSE: Oops, I'm sorry. Thank you very much. I muted myself on error. Next item on the list is a break mandated by the ICANN Meeting Committee. So we'll meet again at 14:30 UTC. That is about in half an hour and two minutes. And Andrew McConachie will take us through the root server analysis as on the agenda. All right, see you later.

KIMBERLY CARLSON: Thank you all. Please stop the recording.

[END OF TRANSCRIPTION]