ICANN72 | Virtual Annual General Meeting – Tech Day (1 of 3)
Monday, October 25, 2021 – 09:00 to 10:00 PDT

KIMBERLY CARLSON: Hi, everyone. Happy start to ICANN 72 to you all and welcome to Tech Day on the 25th of October. My name is Kim and along with Kathy, we are your participation managers for this session. Please note that this session is being recorded and follows the ICANN Expected Standards of Behavior.

During the session, questions or comments will only be read aloud if submitted within the Q&A pod. We will read them aloud during the time set by the chair of moderator of the 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 that time and speak.

All participants in this session may comment in the chat. Please use the dropdown 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 messages sent by a panelist or standard attendee to another standard attendee will also be seen by the session hosts, cohosts and other panelists.

This session also includes automated real-time transcription. Please note 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 again for joining and I'll turn the call over to Dr. Eberhard Lisse, chair of the ccNSO Tech Working Group.

EBERHARD LISSE: Good evening, good morning, or good day, everybody. Welcome to our fifth virtual Tech Day. This time, it's supposed to be in Seattle. I, myself, I am sitting in Germany at the moment, not in Namibia as usual, because I needed to visit some family and I wanted to test out the Starlink satellite dish that I got my sister on the roof. But due to some technical problem, we only get 10% of the speed. And they are very cool and very helpful in remediating and sending us a new dish and cables. Next slide, please.

We usually go through the agenda in the beginning. I'll do my opening remarks as usual. And then, Justin Chen from University of Taiwan and also working with Net-Chinese. That's the one. We'll talk about—now I lost my train of thought—a proof of concept of DNS identifiers. That's what it was. Then Ed Lewis will speak about the DNS Core Census.

And we put Justin in the beginning because he has the worst daytime. Garth Miller is later on—also has a bad time zone but he came very late into the game so he had to take what he could get. We still put him as early as possible but due to the lengths of the presentation planned for Ed Lewis and him, we couldn't put him before lunch—before the break.

Then Eduardo Alvarez will speak about—Garth will speak about an RDAP-enabled security initiative that CoCCA has been working on for a while. And Eduardo Alvarez from ICANN will speak about ICANN's RDAP

conformance tool. Then Roy Arends will speak about DNS magnitude, a way of measuring DNS—another way that he has come up with or some other colleagues of his and him. I don't know. Then Ed Lewis will talk about DNSSEC algorithm choices and Viktor Dukhovni will speak about DNS parameters for TLDs.

So as you can see, I've tried to group these things a little bit. And we actually got these last two together a little bit earlier so that they can choose which one is the better one for introduction.

Then we have another break. We are a bit flexible with the breaks. If we're under or over, we'll shorten or lengthen the breaks. It's not really that necessarily but we have to go with the flow of the ICANN meeting.

Greg Freeman will then speak about RPKI deployment experience at Lumen. Then Roy Arends will speak about the hyperlocal root zone service. I think we have heard all of this a little bit, that you can localize the root zones, which don't change much, so that less traffic comes with all this going up to the top. And then, Dmitry Kohmanyuk from the Ukraine will speak a little bit about ccTLD infrastructure. Do you want to do this yourself or do you want to farm this out?

And then, lastly, I have volunteered Jacques Latour, or rather he volunteered himself if nobody else did. He will then close the Tech Day as we usually do, giving a brief review of the presentations—what was interesting, what was not so interesting, what was helpful.

That means we can now, four minutes in front, start with the first presentation. Justin Chen from Taiwan has the floor.

CHI-YUAN CHEN: Thank you, chair. Let me share my screen first. My permission is disabled now.

EBERHARD LISSE: Kim, can you enable him again, please?

KIMBERLY CARLSON: It looks like he's still a co-host.

EBERHARD LISSE: Okay. Just try again. We're not in a hurry. We have a few minutes to get this working. There you go.

CHI-YUAN CHEN: Okay. Can you …?

EBERHARD LISSE: There you go. Now we've got it. Go ahead.

CHI-YUAN CHEN: Okay. Thank you. Hello, everyone. My name is Chi-Yuan Chen. I am a consultant from TTC in Taiwan. This talk is about a proof of concept of DNS identifiers for 5G mobile edge computing. And this project was funded by our government, MOTC.

Currently, in telecom, the numbering was based on the ITU Recommendation E.164. However, this numbering space is limited. This work is based on the concept of ICANN 5G Technology Report to use the DNS as an IoT identifier and the management of IoT devices is through the auto-location. We also designed two scenarios for the network slicing environment of next generation telecommunication network. This is our first-year result of our project and we focus on the 5G mobile edge computing only.

For our key innovations, first we showed, by using the DNS identifier, the management of fixed and moving IoT devices under 5G is more efficient. Second, based on the 5G network slicing, core network will send the APN, cell ID, IMSI, MSISDN, EMEI, IP to Gn DNS to give identifiers on IoT devices according to the RADIUS.

This is the structure of our proof of concept platform. We built a private core network according to the 3GPP Release 15. And we setup network slicing for eMBB, URLLC, and MIOT.

In all the proof of concept scenarios, we used 5G core network emulator to create two domains, the domain one and the domain two, and sent the required information by RADIUS protocol. We also implemented a dashboard to monitor the network status.

This is our equipment list. We used a 5G router and IPC for IoT devices. And we used an Amarisoft Callbox for the gNodeB. For the DNS server and the dashboard, we implemented a program to collect information in virtual machines.

In our 5G simulator, the red and blue squares [implement] the function elements. We use a broker program, Session Manager Function, SMF, to send the AAA information.

This is the IoT parameter list, including APN, IMEI for numbering, and NSSAI, SST, template, throughput for network slicing.

For identifier mechanism, we referred to the 3GPP and proposed the IMEI as the DNS identifier in the terminal layer for our private core network. IMEI is a unique number per device. This is a sample for our proof of concept.

In the registration process, the broker program will collect information from 5G Core API and forward UE information, including IP address, IMEI, cell ID, and the timestamp. In the DNS server, when it receives the information, it will generate a DNS A record, then serve the information to the dashboard.

This another sample for the DNS identifier generation. The fully qualified domain name for v4 and v6 were generated in step four.

There are two test cases in our first proof of concept scenario, the inter-domain and the cross-domain.

For the first test case, interdomain-the IoT devices were moved from server one to server two in the same domain. You will need to update information in the DNS record.

There are some data we collect from the Callbox and the DNS server during the registration process. The FQDNs were generated in step four, in the [py] server.

In the dashboard, we can see the [management] of IoT devices and domains and we can click on a button for the detailed information.

This table shows the status after our first registration process, the IP address received from the step one to the step two. And then the FQDN with cell one generated from step three to step four.

Later on, we moved the IoT device from the cell one to the cell two. There are some data we collected in the step five through the step 10.

In the dashboard, we can see the four IoT devices are moved to the cell two now.

For the step six through step 10, the FDQN will be updated with the new one in cell two. We can see the label.

This screen shows the changes in the dashboard when IoT moves from the cell one to the cell two.

For the second test case, cross-domain, the IoT devices were moved from the cell two to the cell three in a different domain. It will also need to update its information in the DNS record with different DNS server.

There are some data we also collected in the step five through the step nine. And now we can see that four IoT devices are moved to cell three in domain two now.

Also, the FQDNs were updated with the cell three and the domain two now.

This shows the changes in the dashboard when the IoT device is moved from the domain one to the domain two.

For the scenario one, we show IoT devices were given identifiers automatically. We can also trace the movement of IoT devices in the dashboard.

For the scenario two, we showed that large amount of IoT devices are given identifiers automatically and that later, 100 IoT devices are registered in different cell in the different domain.

This table shows the parameter list of our 100 IoT devices. They have different data than in the slicing [parameters].

From the dashboard, we can see that 100 IoT devices are registered in the two domains in the different cells. And in this scenario, we used the DNN and the IMEI as the identifier of IoT device.

In this work, we verified a mechanism and method of auto DNS identifier. We can apply it to the 5G slicing network and manage the IoT devices efficiently. In the next phase, we plan to do the proof of service on real and the public 5G network.

Especially thanks to the MOTC and ICANN and thanks to our panel.

EBERHARD LISSE: Thank you very much. Technically, well beyond me but very interesting, I think. IoT, we have had many presentations, or some, about IoT devices before. This is, in any case, where the future is. So probably, using the DNS to manage them is a good idea. I saw one hand, which I don't see anymore. Are there any questions? Please raise your hands and we will them open it, if there is one question. No more hands? One of the panelists has a question. Jacques Latour, you have the floor.

JACQUES LATOUR: Yes. Thank you. So I understand the RADIUS and the domain name assignment are to allow the IoT device to move along and get provision. So my question is are the IoT devices themselves aware of that identity or that's something separate?

CHI-YUAN CHEN: Okay. Can you hear my voice?

JACQUES LATOUR: Yes.

EBERHARD LISSE: Yes, we can.

CHI-YUAN CHEN: Okay. Thank you. Yes. IoT devices can aware this identifier from the DNS query.

JACQUES LATOUR: But is the IoT device aware of its own identity when it changes? Because presumably when it changes zones, like cells, the identifier changes, which means it changes the identity of the IoT device.

CHI-YUAN CHEN: I think both the IoT device and the core network can use the DNS query to update the newest identifier and the related information.

JACQUES LATOUR: Okay. Thank you.

EBERHARD LISSE: Thank you. Okay there are few questions in the Q&A. Yoshiro Yoneda asked, “What kind of IoT devices are you assuming?”

CHI-YUAN CHEN: Okay. Wait a minute. We used the 5G gateway and the industrial IoT—many industrial personal computers, as our IoT device.

EBERHARD LISSE: So it’s just a proof of concept but he wanted to know what types of IoT devices would you be using this for—were you assuming that are to be used? I saw there were parking meters, and cameras, and so on. That’s probably what he means.

CHI-YUAN CHEN: Oh. It's the DNN, data network name, for the IoT device. And we used this name for our network slicing tool to our proof of concept.

EBERHARD LISSE: Okay. Justin Mack from MarkMonitor asks, "What TTL, time to live, values are used for the A and AAAA records that are being generated?"

CHI-YUAN CHEN: Sorry. Time to live?

EBERHARD LISSE: Time to live, TTL value of the A and the AAAA records. Can you answer that? Do you understand the question? Did you understand the question or what is the problem? All right. I think we're having a communication issue here. Thank you very much. I will ask the people who asked this question to take it offline and send it directly to you on the e-mail address that was listed. Thank you very much.

Okay. Can we have the next presenter now, Ed Lewis? And un-share Justin's screen, please?

ED LEWIS: I think he stopped sharing. I can start.

EBERHARD LISSE: There you go.

ED LEWIS: There we go. All right. Everybody, can you see and hear?

EBERHARD LISSE: We can see and hear you very well.

ED LEWIS: All right. I'm going to have a talk here called an update on the DNS Core Census, which might be promoting a little higher than I probably should because many people have never heard of my work on this. It's a small project. So I'm going to go through, basically, a coverage of some work I've been doing for the past couple of months and this is the agenda. Go through the census. What's DNS Core? What's the census? Then talk about the sources of the data that I'm playing with and the assembly of the data that I'm doing and then some commentary about making this public at some point.

So in a nutshell, the work here, the DNS Core Census is a small project that I've been working on for about a year and a half. It started out as a module to help me do other data collections, where I look at data and I want to split the data into pieces to make sense of out what's happening along certain lines, whether it's geographic, whether it's some other type of information, whatever is going on there. And TLDs are where I've spent all my time studying. I have a lot of history data on this.

So I had an earlier version of this. I numbered it 002, arbitrarily. And I presented it twice at DNS-OARC in August of 2020, a year ago plus. And then I had another talk at LACTLD in September the same year. And I

have the 002 on the web. It's actually accessible but I'm not promoting the URL for that anymore because I'm about to make a move. And there is at least one person who's accessed it over time and I've been in contact one-on-one with updates on that. So sooner or later, it will move and it will be in a new place. I'm not going to promise when right now.

So I do have some feedback from that and some other communication with other researchers. And there were three things that I decided to do, that I wanted to start over again. The number-one reason for starting over again was actually my third sub-bullet there where I had to make the process more rugged. I found out when data feeds disappeared, my program wasn't handling it very well and would just be absent for a day or two.

But there were a few comments. One was to rearrange how the data was being stored internally. I'll get into that a little bit later on. And then another person suggested some more [instances] that I might include, which I've managed to do in newer version, and gave me the idea that maybe I need to get some outside input on what should be or shouldn't be in this over time. And finally, making this version public. With this version, being the 010, public, it's an ongoing effort. And it may be out there who knows when. Again, I don't like to make promises.

The origin of this work, to go back to where this is coming from, I have to decide, a lot of times, whether a TLD that I'm looking at is a ccTLD or a gTLD. The reason for this is that gTLDs and ccTLDs run under different sets of rules. They behave very differently. When you're looking at

behavior over time, you want to see does a certain policy seem to have a certain impact. Or if you see a lot of TLDs doing something, is it because they are independently thinking of that or are they all following a given set of rules.

It's not just the gTLD/ccTLD division that is interesting to me. There's some regional ones in there. IDN or not is another. And there's many other ones out there. To give you an example—not to harp on the data I'm showing here—but you see these pie charts. They're mostly red and green. And what's interesting in this is that the charts are different from place to place. This is divided from all TLDs, to all ccTLDs, to all gTLDs of the first division. That shows you a big difference between those two. And then, with the ccTLDs, I can divide again from the two-letter ccTLDs and the IDN ccTLDs.

So in the old day, the real problem I had was in the old days, it was simple to determine a TLD, whether it was gTLD or not. We had the two-letter codes everywhere. But when we go into IDN ccTLDs, it made it a much more complicated question to answer that. And I had to do that for many people, over and over again, internally, when they asked me to work on some analysis for them. That's what started me on the effort to build a census.

I saw there was a hand raised. I'm deciding whether we should take questions as we go. I think I'll let questions—

EBERHARD LISSE: No. We'll take questions at the end. I don't know to interrupt the presenters. It's not good for the flow.

ED LEWIS: All right. Okay. Cool. So the scope creep. Initially, I just wanted to know whether TLD was ccTLD or gTLD. But as you start looking at this, there are other factors. There is when did a TLD begin. How old is it? We have the legacy TLDs from before 2012. For a while, when I was doing this work, TLDs were very young. They might have been months old. Now they're usually a couple years old. We look at different phases of what goes on in them.

Another thing you can pull out of this information, if you dig deeply, is what TLDs share the same platform—at least share the same DNS. That's important because when you see a lot of TLDs suddenly adopt a new DNS feature, is it because a lot of people made the same decision—they ran off and said it was a really good idea—or did just one dominant player in DNS platforms decide, "I'm now going to assign everything." And with one flip of the switch, you see everything fly at one. And there's some examples of that. You can see that easily. In fact, my other talk will dive into that a bit.

Another one is geographical, geopolitical regions. I always get people going to a conference somewhere and saying, "I'd to have information about my audience's area of concern. They're from this part of the world. What do you know about that part of world or about that restriction?" and so on.

Also, you've got information about sizes. Are TLDs large, medium, small? What kind of cryptography have they chosen? There are many ways you want to carve this up for different types of analysis. And finally, too, where are the name servers on the network? The addresses, routes, autonomous system numbers are interesting.

One that inspired me to this work, too, is a lot of people like to divide the gTLDs into the new gTLDs and the old gTLDs. There are some finer distinctions to be made here so I really wanted to go in deeper and find out where I could split things up.

The novelty here is not so much getting this information. It's that this information is all over the place. It's scattered. I have a lot of sources that ICANN has published. And I wanted to simply put it in one place. And I also want to make it history so I can go back over time and do some of the historical analysis.

So, for example, of a different type of division here—and again, I'm not going to [inaudible] the underlying data. But these are pie charts representing one thing about DNSSEC across all the TLDs. And I have here the ICANN geographic divisions pulled apart. And you can see again, the pie charts, they all look pretty much different. They're red, and green, and one's larger in some area, and so on. The pattern is different depending on where you are in the world. Geography still rules, even on the Internet.

So coverage creep. Besides trying to put all this stuff together, how much of the DNS can we do? If we could have everything in the DNS in one place, that would be great but we don't have that. It's

unmanageable to have that much data. On the other hand, if you pick just the root zone and its delegations, which is where I spent a lot of my time early on, it's too small a sample set to get an idea of what's going on, on the Internet.

So I want to say it's more of a right-size thing out there. So what's sensible? What's interesting. And that's kind of an open question. When I was initially doing this work internally, there were a lot of people wondering what I meant by "sensible." So in a nutshell—and I'll go through this again—I settled on ccTLDs and gTLDs, the reverse map zones and the RIRs, and then the zones that tie these together.

Other possible cores out there that could be interesting to do this kind of work on, but you'd have to do a lot more work to assemble the lists, are either the high-volume or popular zones. I see a lot of research that goes into the Alexa lists or Cisco's umbrella. I think that was the name of the other one.

This isn't the popular zones—not the zones that get a lot of hits, necessarily. You can go with genres like social media, or public sector, or anything like that. You might want to look at how an industry is working on the Internet. Again, that's interesting but not what I could easily put a rope around.

Or technically-complex zones. In the DNS, the root zone is actually a pretty simple DNS zone. It doesn't get updated in very fancy ways. But as you go down the stack, features pop in because at the leaf edges of the DNS, we use it in many other ways, like for traffic engineering. We like to point things different ways. So there are many perspectives here.

A lot of them are worthy or membership but it's very subjective and the lines aren't drawn clearly.

So defining what I call a DNS Core … And I'm using the DNS Core as … I just threw a name out there. It's the elements of the DNS close to the root zone which primarily exist to delegate other zones. Top-level domains are a part of that, including sub-parts of TLDs out there.

The Regional Internet Registries, I throw them in there because they are registries also, and they run DNS, and they're all part of this magic that goes on. And then there are a bunch of other zones that fill in the cracks of those areas that people forget about but they're worth adding to see how they run.

Generally, these are zones which are under guidelines set by a community and that's a very loose definition. They see themselves as a primary service—not that DNS is not important to social media giants because DNS is very important to what they do. But their social media platform is their leader—the lead interest at what they're doing.

And also, a lot of the TLDs tend to stick closer to the applicable standards. They don't look for innovations in the DNs. They try to make sure things are run well and flat as an even platform for all others. They're not trying to support some other purpose-built application.

So the DNS Core starts with the root zone. I go down to something else that I've invented called the commercial registration boundary. And that's the area where, basically, the registrants start paying for delegations. For example, customer.example. That's the typical

example in the gTLD space. It could be a category below that. It could be a city province under a ccTLD, where payment may not be cash. It might be something else or some other qualification.

Or even if you go down to the reverse map, where do the service providers start delegating DNS for their PTR records, essentially. The concept is still experimental. Generally, the boundary is usually a third layer or it might be deeper.

So to add a little color here, this is a cartoon of the DNS Core, showing what I just covered—the root zone, the TLDs on top, the RIR component in the middle. Arpa and a lot of things that are under it would be in there, too. And sub-TLDs and sub-gTLDs. There are a few sub-gTLDs out there in the older TLDs.

So to give you an idea of what the core looks like—again, to add more color to my slides—the Core, there about 4,000 zones that I cover. 2,300 of them are sub-ccTLDs. That refers to the provinces or the regions within some of the ccTLDs that have really got very deep label usage.

There are about 1,300 gTLDs, 300 ccTLDs. That's what we're more commonly looking at when we look at the center of the DNS. There are about reverse map zones. That means some number .in-addr.arpa for IPv4 or the corresponding IPv6 delegations.

And I also have in these other zones these special use ones there because sometimes they pop up and you want to know that that was a special use. You shouldn't see it necessarily but you want to know that in research when you come across it because you will. I won't go

through the rest of it but the test was interesting because there are 11 test TLDs out there that are inactive but they're still on the record books. And you may come across those if you're doing some research and some historical.

The other things about the scattered—or the zone depth, most things are actually three levels deep. Two levels is popular. And you can get all the way down to nine levels. There's some ccTLDs that have gone nine levels down before they go out to their registrants for whatever purpose that would actually be ccTLDs. It's not the reverse map off the top of my head.

Other things. There are 4,000 zones, 6,000 name servers. There's a lot of sharing. A lot of zones share name servers. Only about 10,000 addresses and a lot of addresses are shared between name servers. And route origins, there's only about 2,000 of them, and so on down the line. The bottom four numbers are just counts of what I grab as I go along through getting some of the data.

So the census itself, how it's embodied is it's compiled daily. I run a script. It takes about an hour and a half to go through it and collect everything. I store it in two things I'll mention later. I have a database backend, which made it much—a tabular database. SQL-like is, I guess, the way to say that. I also have, for each run, nine CSV files. I have nine tables that correspond with CSV files. And also, I spit out nine equivalent JSON files, just to be cool with JSON again.

There's a little background to this. Why not lead with JSON? That's been a problem that's been at the back of my head. I had to put it on the slide

here. Initially, I did something much more JSON-like. Let me see my next slide. Yeah. This is on my next slide. I did something that was much more richly-typed. It was a very hierarchical thing. I had three JSON files that came out. There was the zones, the name servers, and the addresses because they were shared. And I had a lot of ancillary information out of that.

To flatten that, though, I couldn't flatten that easily. So I had to restructure this for table bases into nine of them. Six of the tables, basically, are just the details that are in the original JSON structure. And this came about—the comment that I should be looking more at things like pandas.DataFrames, which is a tabular data type within Python. It just seems to be much more acceptable across the field of numerical analysis or data analysis out there. We like to use tables out there, not rich data structures out there.

So I did a lot of work to spin this backward in my mind from the JSON world back to CSV. It just seems kind of odd but it seems like that's the way a lot of the analysis work is going these days, is they want to be in that space.

So to go into what it is that's in my census, I'm going to list the sources of data and walk through them that I have here—a couple pages of this. From IANA, I have access to the root zone database, which is out there. It's the database itself in XML. It's the history of what's been built up over time—not the history. It's actually a snapshot of what it is over time.

There is a Repository of IDN Practices that was requested. It's a page which shows IDN tables—the policies used by certain TLDs out there. And that Special-Use Domain Names is one that's a registry that IANA is hosting which is, I guess, more … I don't know who the proper operator would be but the IETF documents define what's in there. And again, that's for reference to make sure we account for names that are special-use, that might pop into the Internet. There's a lot of playing done with that.

From the ICANN sources here, I have a couple of websites and one other file in here. The ccTLD fast-track string evaluation gives a lot of metadata about the ccTLD IDNs that are in play. Registry agreement termination keeps track of the process to stop TLDs in running. So we have a history of why they may have been stopped. Sometimes, if the process itself doesn't succeed, we may actually have some gTLDs come back to life after the process is stopped. I found that interesting when I went through that page.

The ICANN geographic regions. The regions I've been using in the slides are the ICANN-assigned regions for some of the ccTLDs out there. There's a gTLD contract status page, which is known to some people that are involved with certificate authorities. It has a history back 10 years. It's not probably very well-visible. It's not an HMTL page. It's a JSON structure, which is very interesting.

And then there are various zone files that I can get my hands on, that most people can get out of CZDS. I'm sorry. In this line I'm including the zone files that ICANN publishes. If you go to that website, you'll see a

list of what we have on some name servers that you can grab. That's not the CZDS. That's actually different, although I do have access to those, too.

Another source of data from the UN. It was requested that I bring in the regional tags that are used by the UN because if you're not an ICANN person, the UN probably is the place to go for regional definitions. There's a bunch of different levels that are added there. That's just metadata from the other data.

From the RIRs, first I was using RIPE's RPK validator. And when I started doing the slides, I realized that they had discontinued that just about a month ago. And I've since replaced it by a lookalike service which is running a local Routinator instance. I don't have a lot of detail here because I just did that it's picked up the pace since then, giving me the data. I know there are other options out there and we might want to look at some better … I have to see whether it's reliable or not. The data that it gives me seems to be very much equivalent to what I was getting from the RPKI validator. So far, so good on that.

And then also, I pulled the zone files that the RIRs publish. Some of them through FTPs, some through AXFR. Most of them are generally public. Two of them, I requested and got allowed to pull them, to work on them.

Also, beyond ICANN and the RIRs and IANA and the UN, the next series, I went out to some other public sources out there. There's a cool tool from Team Cymru. You ask about an IP address and it will tell you what routing information is available for it, what AS Numbers originate the

route to that address. And also, it gives you some information about the name of the operator, which is cool to keep around.

I use the Public Suffix List but I don't use that as … We know that this is an interesting topic to get to the PSL. I use that as hints to help me figure out the commercial registration boundary. I don't republish that in whole. I just use it in some other work. I don't think I have any notes about the PSL membership of anything in the output here but I use that as an input to my work.

And then I do some DNS queries, of course. This is what takes some time. And this is done to fill in information that's not already covered—for example, the SOA, the DNS record and so on. And in particular, a lot of those I get from other sources but I sometimes want to get the DNSSEC information, whether it's available or not. And again, the DNS queries are chiefly to just get the commercial registration boundary poked at.

So this boundary, it's an estimation right now. I target a couple names that have been suspected of sub-elements. There's some out there that are pretty obvious. You might see .co or .ad or .gov are used across the board. I try those. And then I also through the Public Service List, which gives me another set of candidates. If they pop up to pass my little test, I include them as part of the boundary out there.

I'm not looking for the registrations in those areas. I'm not trying to discover what's outside the commercial registration boundary. I'm just trying to define where a registry leaves off and lets it go out to their registrations out there. It's experimental. There are a couple other ways

I could try to do this but I haven't spent whole lot of time with it. It's a little harder to do than it sounds.

So now, assembly of the data. What do I do with all this stuff? I grab all this stuff. It takes about … I'd say it's an hour and a half every day that it goes out and grabs everything. After my first attempt, I realized I didn't really want to generate much information. I just want to pass things through as much as possible in a straight-through phase. So you can always go back to the original source and get the original information and read on what it means. But just trying to collect it all in one place.

I use keywords. They hint at the source it comes from to help guide my way backwards but I also keep track of things come from. It's meant, ultimately, to be a utility for research, not a product of research. So it's not going to definitely tell you anything. The census doesn't tell you anything. You would use that as part of some research work and I've don't some examples of that. But again, being the one that wrote everything, I'm not really a good judge of whether I've been successful at that.

But there are these synthesized records. I have the start and end of the census, which of course, only I could come up with. I have the sources, which tells me where I got information from in a not-exactly-obvious way. It's more to help me debug things. Then I have a category, where I create my idea of gTLD, ccTLD and so on, and sub-gTLD to sub-ccTLD because in many places, it's just assumed that this list is a list of something. So I had to create the category.

The jurisdictions are the two-letter ISO codes. They most easily come from ccTLDs but then I assign a worldwide one for all the other elements that are gTLDs. I'm going to go through the next couple sides a little more quickly because I see there's some questions coming in and I want to leave time for that.

But for the zones, I have this information here. I've got the technical information, the zone cut information, all the bookkeeping you'd want to see, the root database, the gTLD, IDN metadata, agreement termination if it applies, and so on down the list there. There's just the list there.

The name servers, I have the addresses that are used for them, whether it's a registered address, meaning it's in the database somewhere. Cut-point, authoritative, and they should all agree but they don't always agree. And for cut-points, I also try to keep track of where the glue is coming from, what zone had that.

The glue information and authoritative is supposed to be the same. But as you go through the DNS core, it's not always true. There are some lingering differences over time. But for the most part, it's actually pretty clean.

Addresses, I have the address family. For some cases, I realized I had to keep track whether it was v4 or v6, even though some languages will let you do that automatically. Route origins, where is the route coming from? That's the BGP information, the CIDR notation of the network prefix and the AS Number. And then also, I keep track of the registered and glue values across the board. And also, I keep track of what name

servers claim the address. And on the same servers slide, the name servers also keep track of what zones claim them. So you can look for all the sharing out there.

Route origins, I have the information about the route origins. And the AS Number table pretty much is the name of the operator, as I'm pulling that from Team Cymru.

DNSKEY records, DS records, RRSIG records. These are three tables that each have columns that correspond to the fields of those records. And whenever I find a record of those types, I store that. This way we can look at what parameters are being used in places, like what kind of keys, what algorithms are out there, for example, how the DS record has been hashed and so on throughout that.

And finally, the IDN practices. This pretty much the most passed-through of all things I have here. I read the IANA webpage. I convert it from the human-readable HTML that you can see, and visualize, and click on, to computer-readable, which is just these columns of what's in that information there.

Now, I don't have much about ccTLDs. I have that they exist. I have their technical data. But I don't go deep into them. I don't into the size of the ccTLD. I don't have population information for that. I know there are a lot of other sources for that. But I chose not to add bits and pieces. I wanted to be clean about this.

I don't have the complete coverage of the commercial registration boundary, which is more interesting in the ccTLD space—maybe

enough to understand the depth and shape of ccTLD name spaces out there.

It would be great to have more coverage here but I'm trying to be mindful of being able to easily label this work—to say that it's just gTLDs with some ccTLD stuff out there. I know that some ccTLDs have name server zone files available. But until we have at least most of them, I don't really want to expand it into that area right now. I'd rather point to where they are or have them work with me in a different way. I don't want to exclude them. I just don't think it's convenient right now.

Presentation of the data. It's stored in an SQL-like database. It's a database, of course, which is not accessible to the public but it's built, ideally, for the Python data frames approach. I also have the data being generated into CSV and JSON text files but I haven't been able to get them in a place where anyone can play with it, at this point.

So my next steps. My next steps is to publish the dataset at some point. I do have documentation waiting to go, too, this time around. And I'm looking for feedback. Of course, you see my bottom. It says feedback requires published data. That will come along. In the meantime, expressions of interest would be appreciated.

But the feedback is, is this stuff worth the effort? Is this data worth the effort? I sometimes wonder about that because one thing I didn't put in the slides was that there's a lot of space taken up on disks to hold all this stuff. Maybe I don't want to do all that.

What other data would be helpful. I've gotten a few suggestions already in the past. I'm sure there's more to go. And what's the best way to make this usable by researchers. I'm guessing from what I do one way but I may not be the best at doing this.

So with that, I am done with my slides and I see that there's some raised hands and questions. I guess I should turn it back to the moderator to work on this. Or should I just—

EBERHARD LISSE: Thank you very much. I promised Viktor first crack. So, Viktor, can you unmute yourself, please?

VIKTOR DUKHOVNI: Sure. I could use data like this, especially about where the commercial registration cuts are. I currently attempt to get that data from PSL with lots of exceptions. But it's not accurate and lots of work. How do you figure out where the boundaries are? And can I get that data?

ED LEWIS: I'd be willing to share that. I have two things, right now, that I'm pulling. One is way back when I did a scrape of HTML, I did a hand-scrape of Wikipedia to find out what some ccTLDs advertise as their sub-elements. I threw all those names into a jumble and tried them across the board. That's one. I don't know that I have any statistics on whether that's a good strategy or not. But then the PSL is pretty much the other place I rely on. I have an idea of using some other passive DNS

information but there's so much of that, I haven't been able to make much progress of it.

But I'd be willing to share the … Right now, at this point, I could send you my list of usual suspects in an e-mail, Viktor. That's probably the best way to do that.

VIKTOR DUKHOVNI: All right. And I may have data for you. We should probably exchange some. There are places that are difficult. Like Poland and Norway have lots of city or other things like that, that are hard to track for me.

ED LEWIS: Yeah. I say I like to go naming no names. But you're not far from the truth on that. And they're not the only ones. There's a few others that are very deep. One TLD impresses me because they have the majority of the PSL names but they have it all in the same zone. It's amazing.

EBERHARD LISSE: All right. Thank you. I want to get Steve Crocker, who has got his hand up for a while, please. You may unmute yourself now. There you go.

STEVE CROCKER: Thank you. Ed, first of all, this is great work. I congratulate you on this. I particularly like the focus on the backends and the name servers. That will give some interesting perspective on the underlying organization—emergent organization of the operations.

I know you're fully aware but for the benefit of the audience, I've been involved for a long time and helped create the database and the maps that are shown in the DNSSEC Workshop. So on Wednesday, for example, of this week there'll be a presentation of that. That work is older than I can remember, even though we started it and then transferred it over to the Internet Society, which has been publishing these maps on a weekly basis.

What's interesting is that underneath those maps is a full-scale database that has a complete history of everything that's gone into it. So there is rather richer data there than might first appear. It doesn't cover all the things that you have, by ay means. It was very focused on just tracking DNSSEC deployment.

But we did look at some of the same organizational issues. How do you classify a TLD into CCs versus Gs and some things that didn't fit very cleanly into that? What do you do with IDNs and so forth? So there's some areas where we have related statistics. It will be interesting to try to reconcile the numbers that are in that database versus what's in yours and so forth.

The commercial registration boundary, I think, is very important. And the Public Suffix List is a mixture of several different things. So I agree with you. There's partial information there.

One comment about tracking things. I think you've emphasized, as best I can tell, doing everything in as automated a fashion as possible. Obviously, that's desirable. What we discovered is that much of the

data that we were interested in had to be collected by hand. That was the bad news.

The good news is, there isn't that much data. It doesn't change that often so that it's quite manageable. So for example, if you want to know what region, for example, a TLD is in, that's determined basically once. It doesn't have to be scanned, and rescanned, and rescanned over and over again. So I just wanted to mention all of those things.

I guess one additional comment is in looking at the backends, are you prepared for backends changing as they do once in a while, when the backend operation is transferred? A few examples come to mind—for example, .au has moved, I think, twice now, if I recall correctly.

Oh, yes. One more thing. For the ccTLDs, do you keep track of the governance structure and is that of interest? Because I think there are some effects related to that, that might be interesting to track over a period of time. Thank you very much.

ED LEWIS: To answer the first question I remember, in terms of backends changing, I've seen that. For the census, the census is just trying to track that. It's not impacted by that happening. Now, I have some analysis which goes over and does that, where it tries to discover what I call DNS houses and AS houses. I've run that for a time using the old census form. I need to rewrite it for the new census form.

It's interesting to watch that because you see when an operator may change. You can see in the newspaper that someone bought somebody

else. And then, months go by and you start seeing a change at the IANA database, saying that they've changed the backend database there. And then maybe a few weeks or months go by and you see the DNS RNAME change. So by looking at different elements, you can get a heuristic for who's actually the backend operator.

Also it made me very much aware of this that when we do transition TLDs from one place to another, we don't transition the technology always at the same time. It takes technologists a few months or weeks to get everything smoothly run over there. So the census actually highlights the process even more so. And with another layer of coding, you can start dividing who are the co-managers? Who's managing all of these TLDs together. There's 240, at the time, players out there.

Oh. The second question, the ccTLD governance. My first answer was so long I almost ran out my memory buffer here. The governance there, I'm not exactly sure how to interpret that. But I do record the backend operators of the TLDs out there. I don't know that I have the registry administration, which might give you an inference as to whether—who is it that's running this. Is it a company running it? Or is it being run by the jurisdiction that owns the ccTLD name and so on.

When you say "governance," I'm not sure if that … We can talk more offline about that. But we can figure out what we want to record in this.

STEVE CROCKER: Yeah. The question is whether it's being run by the government, being run independently, or something [creative] thing.

EBERHARD LISSE: Okay. Can we come to an end of the particular thread? Because we're running a little bit over time. There's one more question in the queue. Thank you very much, Steve. It was very interesting, though. Brett from Nominet, I think, asked, "Do you plan to make that data available to other interested parties? I understood you to say it's not research. It's a tool."

ED LEWIS: Right. Yes. I very much plan to make it available to interested parties and also uninterested ones, too, if they want. Yeah. I plan to make this publicly available on the network. I'm getting closer to that but I can't promise how close I am. But yeah. It definitely will be out there.

EBERHARD LISSE: Great stuff. Thank you very much. I'll look forward to the next … Sorry. I muted myself just now. We are going to break now for half an hour until 17:30 UTC—that's 25 minutes—and then Garth Miller will start. Thank you.

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

[END OF TRANSCRIPTION]