COPENHAGEN – Tech Day (Part 1)
Monday, March 13, 2017 – 11:00 to 12:45 CET
ICANN58 | Copenhagen, Denmark

EBERHARD LISSE: Good morning, everybody. If you could all settle down now. Okay. For all who don't know me, my name is Eberhard Lisse. I'm the Chair of the Technical Working Group and I am the .NA ccTLD Manager.

Today, we have got a nice program again, I think. We are struggling a little bit with fitting this into the way the ICANN meeting has been reorganized. I am on such a working group where I should pay close attention but it's sometimes difficult to get the good presenters or presentations or both into the slots that don't compete with others because they are all often double-booked. So, we can't really increase the audience by avoiding double-bookings. We have to basically live by the restrictions we are bound with.

That as it made, I think we have got some very good presentations. First, Alexander Mayrhofer from .AT Austria who can already come to the table, please. We'll talk about the analysis system, I think. Then we had to rearrange the schedule a little bit because one of our presenters was again double-booked. So, the host presentation which was supposed to be in

the afternoon will be second. Then Maarten Bosteels from the Netherlands will talk instead of Giovane Moura about Let's Encrypt which I call No Domain Left Behind. And then Dmitry will talk about some TLS statistics from .RU.

We have got a actually working clicker. After lunch, it will be Jay Daley. And then one of the two headlines I think is going to be Andrew Sullivan talking about the events of the 23rd of October. Sorry. The first one I meant was Maarten Aertsen, and not Maarten Bosteels.

Then .BE has recently moved their infrastructure into the Amazon Cloud and they will talk about their reasoning and experience so far. Then NL Labs will talk about DPRIVE. And then – I forgot his first name – Slaunwhite from CIRA will talk about their analysis software. And then another headline will be Bobby Flaim and Europol will talk a little bit about their initiatives to take down 200,000 domain names in one single fell swoop.

So, I hope we can all remain in the meeting. And it's very promising that almost every seat is taken. It usually has become a habit that more people come to take it into the ccNSO meeting in the same room the next day. And with that, I'll hand over to Alex.

ALEXANDER MAYRHOFER: Thank you, Eberhard. Good morning, everybody. I'm talking today about what –

EBERHARD LISSE: Speak closer to the microphone.

ALEXANDER MAYRHOFER: Yes. Sure. Is that better? Yes. I'm going to talk today about analysis based on the DNS big data infrastructure that we have. And we coined the name DNS Magnitude for it. And it's another answer to the question, "How popular is this domain?" And this is yet another, in that case, DNS based approach.

So, the motivation behind that, be it internal or external, at so many points in time, we get faced by the question. "So, hey, how popular is this domain name actually? Is it a really important domain name? Or is it just unused essentially?"

You could look at the queries and tell that someone that it had so many queries. But somebody like outside of the DNS industry will probably not realize whether that's a lot or that's nothing. So, I had the idea to create a single easy to understand popularity figure for each and every domain name within a TLD, in that case. That's based on the DNS statistics.

And I really like to copy the earthquake magnitude figure because essentially everybody hears on the news, there's an earthquake of magnitude 7 or something like that and everybody knows it's really bad. The question is whether a lot of DNS traffic is good or bad is a different story. But that's how the name of the DNS Magnitude came together.

So, we started with some DNS data exploration. We use the DNS query statistics because that's what we had. And so, the basis was I believe that the domain name is popular obviously if there are a lot of queries for it. Yes? So, the basic assumption was that the popular domain name would trigger higher query rate.

We started looking at single days of the DNS traffic for .AT. A single day of DNS traffic or .AT is between 420 and 680 million queries, something like that.

We initially removed the NXDOMAIN responses. We, later on, edit them again for analysis of popular NXDOMAINS, obviously. And what we saw is that we had queries for almost all existing domains. And even for all the domain names that used to exist at some point in the past will still get queries.

Looking at this, the problem with DNS queries is the extremely high disparity. I'm pretty sure that everybody else has something like that as well. So we have about 1.3 million

domains in our zone. So, if you look at the scale to the left, almost all of the domains have almost no queries. And the DX axis goes up to 20 million queries. That is queries per day. So, there are very few very popular domain names and a lot of essentially unused domain names I would say.

So, obviously, the linear scale is not really suitable for that. So, we went for logarithmic scale. That looks more natural. So, we get a center around four or something like that. And there's a long tail of very popular names that goes up to I think to 17 or something like that. There are very few names that actually get close to zero queries. So, most of the domain names actually in some way for some reason being queried.

Earthquake magnitudes use logarithmic scale as well because I think that the magnitude increased by one if the energy increases by ten. In that case, we use the natural logarithm. And now, so the maximum magnitude I got from that single day was 16.9 something or so. But I really wanted it to scale from 0 to 10.

The definition of magnitude 10 would mean that all queries in a single TLD would be for just a single domain name. So, that's actually not achievable in practice obviously. But what we get is that the logarithm of the domain names are between 0 and 16.91. So, if I scale that to the logarithm of the total queries, I get a nice scale from 0 to 10 if I multiply it by ten.

So, that's essentially the formula looks more complicated than it actually is. So, we take the logarithm of the number of queries for a certain domain, divide it through the logarithm of the number of all queries and multiply it by ten.

However, as you probably anticipate already, the number of queries is highly dependent on the TTL of mostly DNS records in the zone. And we don't have any control over the TTL that people set on their zones. So, the first domain that we have actually was the highest magnitude is an ISP. He doesn't have any really super popular names but he has a really low TTL on his NS records and also on the A-records of his name servers.

So, essentially, he has 22 million queries out of our 450 queries on that certain day. So, he has a huge impact on the query load on our name servers and therefore also high magnitude. And as you can also see, most of the names here are like infrastructure zones. So, zones with lot of name servers on which a lot of zones are actually delegated get the highest query rates.

I turned my head a little bit around this. But I came to the conclusion that actually those infrastructure queries are as important as a query for an A-record for a website because otherwise, the DNS wouldn't work.

But how do we get rid of the TTL problem? As everybody knows, I'm sure a TTL expiration would trigger a re-query from mostly the same source IP address for the same domain so that the name server would refresh the information.

So, my approach was rather than counting queries, we would count the number of unique IP addresses which query a certain name during the day. No matter if they query a domain once for a thousand. So essentially, it's like getting what percentage of the result of population is interested in a certain name.

So, the new basis that we use was the number of distinct source IP addresses per domain. And that's how the host space top ten looks better. It's still dominated by infrastructure domain. Infrastructure domains like univie.ac.at is one of the domains where all name servers are running.

Telekom.at is obviously a big ISP as are the others. But we also see a first few server or let me put it that way. Service operators like Google on spot number eight. And you can also see that the effect of the TTL is greatly reduced. So, it looks more natural to me.

That's the current working definition that we use. The magnitude of a domain name is the natural logarithm of the unique number of hosts for that domain name, divided by the

natural logarithm of the total number of unique hosts we see during the certain time period, multiply it by ten. So, it gives a zero to ten scale.

So, the question is can we also look into certain services? For example, what I did, I reduced the base of the data to A/AAAA records and www.% or origin which reduces the number of queries that get into that analysis from about 450 down to 44 million queries and about 400,000 hosts. So, it's already 10% of the base data that we have.

And then again, it looks quite reasonable. So, Google is number one in Austria. That's easy to see. So, it looks quite natural. However, as I said before, I turn my head around whether or not I would like make queries for the web more important than infrastructure queries. And then I decided no because they're like e-mail providers, they mostly have MX traffic I guess.

An infrastructure domain is really important if there's an outage in an infrastructure domain. Potentially, a large number of other domain names are affected. So, I decided to move further for the plain queries and not reduce them to a certain subset of a service or host name.

These are some examples in that case still web based. So, Amazon, for example, is reasonably popular in Austria. It gets

7.8. ORF. That's the local broadcasting station, gets 6.5. Google gets 9. Our own domain gets 6.1 mainly for the infrastructure reason. Our website is not that popular at all.

A small restaurant around the corner of our office gets 3.5. That's what the typically medium to small enterprise gets. So, it's actually quite nice. A domain name that I own that's essentially unused I think even one of the name server for testing purposes doesn't resolve, gets 0.6. And the Austrian post service gets 6.8.

So, I showed that to a couple of people within our company and asked them, "So is that useful? Does this tell you something?" And they said, "Yes. That's exactly what I want to know. If a customer calls me, I want to know whether that's really important domain name or not."

So, we went ahead and implemented that. In that case, not for the web service but for all services. And we added it into our internal BI panel, I would say. On the right side, you can see that each and every domain name has this small symbol. I didn't find a better ICANN than the RSS feed thingy. But it seems adequate.

And so you'll see immediately whether or not some domain name is really popular huge one or it's an unused one. And we also on the list of transactions that we have in our internal BI

panel, the DNS magnitude is listed next to each domain name. So, you can see that if a really popular domain name is being deleted, you can prepare yourself that the customer will call you in half an hour or something like that.

I also tried to look at the timeline of the magnitude of certain domain names. And to my disappointment actually, most of the domain names are really stable. So, they almost have the same magnitude across the whole year. That below, that stripe with the sparkline of the magnitude is from the domain Pokemongo.at. That was the only one domain name I could find that suddenly became popular. But still it is only a popularity of 3.7 so it's as much as a typical hotel or restaurant website.

So, this is still a work in progress. We also did that for NXDOMAINS. It's really useful. So, some of the NXDOMAINS are actually really popular. I bet that our registrars would love to have the data. But obviously, we are still discussing what we can actually do with that data.

Another example that I did, I looked at whether there's a relation between the DNS magnitude of the domain name and the Delete Propensity of a name, so whether or not the name is deleted. So that is a box plot of that. The interesting thing is I actually expected a higher correlation.

So, as you can see, on the left side is a plot of the DNS magnitude of the domain names that didn't get deleted in a certain point in time. And on the right is the set of domain names that actually got deleted. So, the correlation is not as super high as I would have guessed. On the other hand, no single domain name was deleted with a magnitude over 5.8 which in turn made me think that I could actually create even an alert in some way to the customer service that an important domain name go deleted, brace for impact.

Work in progress, I'm feeding in the DNS magnitude as a new parameter into a neural network that I run to predict whether a domain name is going to be deleted or not. I have early results and they are pretty promising but there is nothing I can share in this point in time.

What we are doing now is we are calculating this figure on a weekly basis for each of our 1.3 million domain names. And the tools we are using is we are using ENTRADA and Hadoop as storage for the DNS traffic that we get. On top of that, we are having Impala as the SQL query interface into Hadoop. I'm sure these are really familiar to you, guys.

For the statistics part, I actually used R for prototyping. And then I actually moved to PHP for the implementation of production service. Don't tell anybody but it actually works. The result

restored the result in Redis. We stored the DNS magnitude of each domain name in a single bite. So, the magnitude map of all 1.3 million domain names is about 1.2 megabyte per week. So that's really a nice data structure.

For the first time in our company, we used Airflow for orchestration. So, for all this coordination of [inaudible] Air, blah, blah, blah, blah. We used the Airflow for the first time. And that was a really nice that that we probably use for other stuff as well.

It's about 300 lines of code in total including all the orchestration stuff. So, it's really quite simple. It was about I would say one month of work until getting this done which is turning my head around a little bit between the working days. Oops, for the work. Yes.

I really want to look into refining the algorithm. For example, the lower end of the magnitude between 0 and 2, that's essentially very sparse. And that's because there are quite a number of A to Z queries from clients. So, there is a single client that actually does A to Z queries for 1.3 million domain names every week, even for domain names that have been deleted five years ago.

So, I really want to refine the algorithm to remove those clients from the calculation so that I only get natural DNS traffic, if you

will. And I'm also looking into changing the lower end of the scale to a linear scale rather than logarithmic because at that point, it's actually important whether you had queries from certified distinct client IP address during the week or 38. It makes an interesting difference. There's quite a tensed density of domain names.

I need to compare that work with the work from the NZRS. They did a linguistic based approach on domain popularity. I did some work on comparing it to the Alexa 1M list. It doesn't really compare very well because Alexa obviously only comes in as web traffic. It compares a little bit better to the Umbrella Top 1M list. That is the list that the open DNS folks actually create with the similar methodology as far as I understand.

As a follow-up study, I really want to do an empiric study on what happens if I change the TTL of a name service NS record in a delegated zone, how much does it affect the traffic? I also need to look at whether prefetching has an impact on all this. And mostly, the service based DNS magnitude might be heavily affected once QNAME minimization takes off.

Finally, I probably need to talk to an ISP to get access to data on the recursive resolver, maybe let them run my algorithm on their machine so that I compare whether their better vantage point gives different results.

So that's about it. Any questions? I'm happy to talk to anybody about the algorithm and also urge you to try it out.

EBERHARD LISSE: Okay. First is first. Wes Hardaker, I was going to abuse the prerogative of the Chair but I made some notes so I won't forget. No, go ahead. You're standing. Go ahead.

WES HARDAKER: You can go first if you like. Yes. So, Wes Hardaker, USC.

EBERHARD LISSE: And we have got 10 minutes for question time so nobody is going to rush.

WES HARDAKER: Wonderful work. I'm actually looking into doing a lot of research on similar things and looking at especially like the sparkline kind of stuff with attacks that occur over time. So, my one suggestion for follow-on work would be to actually look into defining a magnitude which would be fairly simple for a domain's typical traffic to suddenly spikes according to Pokemon Go sudden usage and say before then you'd have to refigure the algorithm

to say how big that jump was and scale the magnitude based on the jump rather than the total.

ALEXANDER MAYRHOFER: Okay. Thanks.

EBERHARD LISSE: Everybody speaking must please identify one's self so that the remote participants know who's speaking. That was Wes Hardaker. I must sometimes you get surprised. It's probably one of the harder to tap presentations because it is simple to do.

We had presentation here by Stephen Deerhake and myself in the past Singapore recall where we did very simple mechanism how even on a smaller ccTLD to grab the data of the wire put it into an SQL database.

I like to do R. As you know, I'm a gynecologist. So, I do this to track my claims with the medical funders who say they are supposed to pay me in 30 days and on average they take 80 days. And they hate that when I give them pictures and show them that. So, I have a little bit of insight into R. This is actually some stuff that I want to get if it is open source.

Some advice is that it's a dual court L-studio which is more an IDE. But the same people have got a way of easy web applet. So,

this is something that you might want to look at to make a nice simple web applet that then does that calculations directly if the data is too big. In this regards to ICANNs, Google is your friend. I found a nice one that I can send you.

ALEXANDER MAYRHOFER: Coming back to what you said, essentially, what I use R for was to plot the histograms and to do the first version of calculating the logarithm of a number. So, I used it as a very expensive pocket calculator in that case. There are other things that I use R that are really nice and especially the neural network stuff, that's really promising.

EBERHARD LISSE: And the graphics. The graphical analysis, R has some very, very cool tools. I even had to buy the book about this to figure out how this works. But they've got some very nice, easy to do is click. Once you figured out the learning curve, really cool tools.

ALEXANDER MAYRHOFER: And regarding what you asked whether it's open source or not, it's not open source because I'm a really bit shy about the code that we have currently but I'm more than happy to share it personally because it's so simple. It's like just the heavy-lifting is

more like getting the data from A to B. And getting everything right so that the timeline of working with the data, the data flow, the ETL orchestration is the harder part, not the algorithm. That's really simple.

EBERHARD LISSE: Why aren't? Can you turn the microphone on, please?

ALEXANDER MAYRHOFER: While you walk, one simple note. Obviously, this only considers the popularity of a domain name within the TLD or within the zone. So, for comparing different TLDs, we would need to find the different measure.

ROY ADAMS: I'm Roy Adams. I work in the office of the CTO in ICANN. Fantastic stuff. We've known each other for a while. And you know I love this stuff as well. What I'd like to do is to apply your technology on –

ALEXANDER MAYRHOFER: To the root.

ROY ADAMS: Yes. And initially just as the traffic that we see for domains ending in .AT. And as you know, we see cache misses. If the top-level domain is cached, then there's no reason for resolver to come to L-Root. If it's not cached, that's what I call a cache miss, we will see it as a root server, one of the root servers.

So even though we see the cache misses, overall, the magnitude number should be a reflection of what you see. Assume for a second that all cache misses are equal. So, I'd like to do that exercise and see what we can come up with with you guys.

ALEXANDER MAYRHOFER: Cool. Happily.

ROY ADAMS: Perfect. Thank you.

ALEXANDER MAYRHOFER: Thank you.

EBERHARD LISSE: Any more questions? Then I think it is of the hand. Thank you. Now, we come to the host presentation. So, Erwin Lansing and his colleagues can come to the floor. For you who haven't been here, as we usually ask the ccTLD hosting, preferably if it is a

ccTLD hosting, to give us a little bit of an oversight of how they're running their ccTLD and what special stuff or research they find at the moment interesting, attractive or worth talking about.

ERWIN LANSING: Thank you very much. So, I'm happy to be here, happy to work in Copenhagen. I think quite a lot of you have seen me around at these meetings for the last couple years. That's why I thought I'd bring in some reinforcements.

Unfortunately, our head of IT could not be here so I'm going to try to read his slides which I just got from him. But our Head of Development, Nikolaj is here. So, Johnny was going to talk about our new IT strategy. Nikolaj has something about our services and development processes. And whatever time is left, I'm going to talk about a very topical issue right now which is domain abuse mitigation which what the role of a registry could be in that.

So, this is what Johnny looked like if he could be here. He'd probably be around maybe later today or tomorrow. So, he started DK Hostmaster last summer. And we're going to change quite a lot of, not only our technical system but also how we do stuff.

Over the last couple of years, we changed most of our Fortran interfaces. We had a new self-service portal. We accredit our EPP platform quite significantly. And we're working on a registrar portal right now. So, now it's time to actually look into our backend systems and completely overhaul those. The reason to do that is not just because they are almost 20 years old and/or cannot be grown but also, we want to focus more on our customers and deliver what they want instead of focusing on what we can do with the technology we have and what we can't do with the technology we have right now.

This is also why we need to get ready for what we call a streamline project. And what's the streamline project? This is going to be the biggest project we've ever done in our almost 20-year history because like I said, we're not only going to change the whole backend system, not only just the database or business logic but also going to think about how we do things and how we do our processes or procedures, how we interact with the customers. So, it's quite a big project not just from a technical viewpoint.

So, like I said, we have our home-grown legacy system right now. It's mostly batch based. Most of you who are born in the '70s know what that means. We definitely want to get rid of those. We want to go do something that could be more or near

real time. Actually, for those who, makes sense, if you order a domain name, you don't want to wait an hour for the domain to be in the zone. That should be seconds, maybe minutes depending on what we can do.

Everything we're going to do right now, we have to change our perspective from ourselves looking at how we do stuff that's because that's how we've done it for almost 20 years. But is that also making sense for the customer? Maybe not. Maybe then we could change that. We have to take a look from the customer what do they want from us.

I think with that, I'm going to hand it over to you, Nikolaj.

NIKOLAJ RAVN HANSEN: Hello, everybody. I'm Nikolaj Ravn Hansen, Head of Development. As was Johnny, I've also fell anew in the wonderful world of domains. I joined DK Hostmaster six months ago. I spent the last 15 years in Ticketmaster. So, I come from a very customer focused industry. So, it's been a really interesting experience to see both ICANN as well as domain industry in general.

I just want to talk briefly about how we do things in our Research & Development department because historically, it's been very much focused on technology. And now, we are more

transitioning into meeting the customer requirements and what the customer wants.

So, it's been bit of a change from what's been done in the past, a lot of learning experience with that. But as you mentioned, Erwin, what we have and we have in-house development support our entire platform. What we're doing at the moment is we have two major development streams. So we have one focusing on bigger projects that could be we build a [inaudible] last year. We implement a new version of EPP. And then we have what we call our theme stream as well.

So, one thing is having focused on some very big project that take it long time. But at the same time, we want to make sure that we support all the other application we have. So that's why we split it up in two different parts. And since we're not that big a team, we are ten people in total including testers and products and project management. We have been forced to rethink our processes and our tools to make sure that we can do this.

First of all, it's an open source shop. We are running the majority of stuff on the Mojolicious platform which is Perl-based. And we are running that with the NGINX on top of that.

We are very focused on the agile approach so we're running a three-week sprints on both streams. This task require a lot from

the, you can say, the footwork, all the initial work. So that's what we've been spending a lot of time on getting better at understanding our customer needs because that can sometimes be challenging.

So what we've implemented is basically a lot of interaction with our customers being both internal, in the company as well as registrars. We have focus groups and are very much in-touch with them when we build new stuff. We focus a lot in getting stuff out of the door quickly for evaluation in sandbox environments and making sure that we collect a lot of feedback from these different parties.

But also, in order to do so, we need to have continuous integration. Basically, all of the stuff that relates to infrastructure, if you like, we need to ultimate that as much as possible. So we are relying heavily on continuous integration. We are relying heavily on automated testing. On our self-service portal, as an example, we have some 350 documented test cases. Of those, we run around 150 every night manually. So, it's not just unit testing. It's actually end-to-end testing that we want automated every night.

And then again, RERO, release early, release often. And with this whole scope we've set up, we're actually able at the moment to deploy somewhere between 10 to 15 feature and [inaudible]

releases every month across the entire stack of services we have. Just to give you a bit of idea what we have.

We have of course our legacy, .DK platform in the middle. But then we have a wide variety of different tools and application sitting on top of that. We are very much moving towards software service so we're providing more and more services, more and more APIs.

Just to mention a few of the key deliverables we are working on at the moment, we are in fact already live with the client, RDAP. We have an RDAP client available already. And get up if you want to check it out. And we expect to give live with the server side during this quarter.

Then as some of you might be aware, we have some fairly strict laws in Denmark regarding validation of when you want to register a domain name. So we do have quite a tedious validation process at the moment you need to do, go through a lot of steps to actually get a domain name, working hard on getting that process more streamlined, if you like, and make it more easy for the user.

A part of that is also that we have this requirement of using this public two-factor authentication scheme for Danish residents called NemID. Some of you might have heard of it. So that's

basically what's happening during the spring that all, everybody who wants to register a personal domain name in Denmark cannot do so if they're a Danish resident without using this two-factor authentication.

Then we're also looking at something that you, Erwin, will cover a bit more into later. We're trying to automate the screening process. We really want to be kind of front runners in trying to avoid abuse of faked websites and funny domain names that sends you all kind of web places.

So, we are looking at trying to build at least initial screening of these requests. We can see if we can see some patterns if we can at least catch the most obvious, you could say, the most obvious cases of potential abuse. And then we're also working on building a Registrar Self Service Portal. So, we want to move as much of the day-to-day work that the registrars currently has our customer service do for them, have them do it themselves.

And as I mentioned, the link below, we put all documentation of all our public APIs on Github, on this address. And we also, as I mentioned, have the RDAP client as well if you want to check it out.

ERWIN LANSING: So that's me. I think we can't get away from noticing that both cybercrime is increasing but especially the focus from other bodies interested in cybercrime is increasing. And they're increasingly looking at the registry what we can do when we hand out a domain name to someone.

So, the question is what can we do, what can't we do as a registry? In Denmark, we currently have a lot of focus on IPR violations. There is a lot of Nike shoes. If you want some cheap Nike shoes, come to me. I got a list of 400, 500 domains. I'm not sure if they will actually ship you some.

There's a news article here from last week where the police seized about 1000 domains this year to date. So, this is clearly an area where we have to look into what we can do to make it a more safe environment to be at.

However, as a registry, we can't just be the judge, jury and executioner especially for content. We have some rules in our terms and conditions where we can handle domain abuse. This is primarily used for typosquatting and malware hosting.

For disputes between third parties, there we have Independent Complaints Board that takes decision on that. And it's actually quite cheap to register a case there if you want to actually your money back. And for everything else, we have the Courts of Law.

We worked with them last year to get them a guidance on what kind of information they need to gather, take it to the judge then come to us. So, we right away, at the first time, have to write information for us to be able to seize the domain portal police. And we don't have to go back and forth several times.

So, what can we do as a registry? I think there's been a couple of years, many years actually, there's been a lot of talk about database accuracy. Two or three years ago, we had a new domain law in Denmark that require us as the registry to check Danish registrants against either the Civil Registry Database or the Central Business Registry.

We also do send a paper letter to make sure that the address actually exists and don't come back. The postman in Denmark is not allowed to put the letter in the box if the name doesn't fit with the name on the letter.

We currently do not actually check that is that registrant that's logged into the system or in the domain name. We just check that the registrant exists at that address. So, this will be before summer, we will be requiring those registrants to log in with the common Danish login system NemID to confirm the identity so we're also quite sure if that is the right registrant or a domain name.

This is easy in Denmark because we have those good databases. However, for the rest of the world, it makes it quite a lot harder. Within Europe, there might be some possibilities. Just talk about eIDAS to have those common national login systems talk to each other. There is the VIES business registry for all European VAT numbers. And to maybe other databases, we can link up to.

So, the better thing to do or have to go for the rest of the world is to take a risk-based approach. We can have some indicators that might say this is something that might be used for abuse or we have some other suspicion from maybe an existing domain name that is being abused for something and it is grounds for additional ID check.

What we're working on right now is to look into what clues those criterias be that might give it a hint of a domain name that is registered with the purpose of being abused or is being abused in other ways. Of course we have a list of registrants that previously had a domain name that was misused. There are some countries we know that are especially sending out bad domain names, free e-mail addresses, Gmail. 163.com is quite common for people that want to do something fishy.

There might be indicators where the phone number is not actually in the country where the address is. And we can also start looking into just like Alexander look into the DNS traffic

data, see what kind of fallacy you see there. And there's also external sources like blacklists and other databases we can link up to. I think it's hard to find all stuff but it could be an interesting thing to just start looking more into that.

And with that, I will just close up and say that the registry, we do have a role to play in fighting cybercrime. We can't do everything. We also have to work together with other parties like the courts and the police and also internationally what we can do and link up our own databases. There's no one size that fits all. There's no simple solution to just find everything and close it down. So, this will be a long project to come for all of us.

And that's it. Any questions?

EBERHARD LISSE: Thank you very much. I have one or two questions. NemID is something that every Danish citizen can access like in Germany, they have an electronic thing on your ID card and in the Netherlands, they have what's called DGD where your address with the municipality and they do a confirmation that you're the person, you are who you are.

Do you have a registry/registrant model or do the clients register their domain name directly with you? And if you have a registrar model, can they confirm the identity to the registrar and then

the registrar securely agrees with you or that means are you trusting those registrars or do they have to register with the registrar, then have to go and to confirm you? How much is the registration fee per domain name?

ERWIN LANSING: Forty five Danish krone including VAT. That's €7 I think-ish.

EBERHARD LISSE: And how many do you have?

ERWIN LANSING: One point three million. And 5% are signed by DNSSEC.

EBERHARD LISSE: Yes. That would be the question for Wednesday.

ERWIN LANSING: Yes. That will also show what algorithms they use. So, come on Wednesday to the DNSSEC workshop. So, the question to how does the validation work, it's interesting and quite complicated. So, let's try to give the short answer.

We do have registrars where we don't sell domain names directly so you have to go through a registrar. However, I think

we have 200 or 300 domain name registrars, it's quite hard for them to integrate with specific Danish NemID system. So that's what we do.

I'm not sure how NemID is going to work specifically. The current validation is that we do send the name and address to a civil or business registry. And if it just matches, the registrant doesn't have to do anything more. It just matches, it's okay, it's fine. If it doesn't match, then the registrant has to go back and come to one of our sites to somehow change the address so it does match or in the end, log in with NemID to confirm it.

Do you have an idea on how NemID is going to work?

NIKOLAJ RAVN HANSEN: Well, we will be using it primarily as a signature, a secure signature. NemID in itself does not contain any address information. So, you cannot just by entering your NemID, you can say, "I live there and my address is this and this." It's basically just your signature. So, we do as you have showed before, we hook up to both the Social Security number database as well as the official company database in Denmark.

With those two things in combination, we're actually able to see that all the specific NemID is actually belong to that specific Social Security number. And that way around, we actually are

able to validate the identity. That's how it's going to work. I think it's hard to see a model where we will move this responsibility to the registrars. But we are looking at making this entire workflow more simple and with a lot of few steps in the future.

EBERHARD LISSE: That would call for an EPP extension.

NIKOLAJ RAVN HANSEN: Both yes and no. We do a lot of things in EPP already. But I don't think we can complete about the fact that as a registrant, you'll need to have some kind of interaction with us to finalize this validation process.

EBERHARD LISSE: My point is not in the action. The point is that you automate it on your side that you don't have to have staff resources, human resources to look at how many registrations you have per day that you can automate it and only push the ones that don't pass the automated validation or to human. That's what I'm thinking about.

NIKOLAJ RAVN HANSEN: Yes, exactly. That's also part of our whole scheme for this risk assessment. We're looking at the model where we will have, let's say, green, yellow, red lights. All the ones with NemID of course there's green light because we know they're okay. And then based on this risk assessment, yellow, we could let it pass through. Red, we will definitely need some kind of manual intervention. Potentially, asking for some additional ID, maybe a copy of a passport or something like that. And that's primarily for the foreign registrants. But as much as possible, of course, automated.

EBERHARD LISSE: Okay. If there is no more question, thank you very much. There is one more. Please come to the mic. Just stand up and identify yourself to the microphone, please.

[SHO BADAM]: My name is [Sho Badam], I'm from India. Have you developed any tool which can categorize the domain names which are easily resolved or not resolving which are booked in the .DK?

ERWIN LANSING: Not yet but we're looking at both looking into our zone. We call it the zone quality index to just scan the zone and look how they

look if the domain actually resolves and how many have are resolved, how many have a newer server. I think about 10% of domains actually don't resolve for some reason and I'm looking into why that is. That's going to be an interesting project development.

[SHO BADAM]: Is the documentation process is affecting your number of registrations under .DK?

ERWIN LANSING: Not yet but that's something of course we have to look into. And that's something that could be [inaudible].

[SHO BADAM]: Okay. Thank you.

EBERHARD LISSE: Thank you very much. Give him a good hand. Next one is Maarten Aertsen who I just confused with Maarten Bosteels. He's going to talk a little bit about Let's Encrypt. I didn't really know much about it myself but when I heard about the topic, I started to look into it. And when Warren Kumari said that he's a big fan

of it, that I looked even more deeper into it. So, I'm quite sure we're going to enjoy this.

Chuck, you can turn this little clock off. We are not scrap for time.

MAARTEN AERTSEN: Thank you. So, good morning. Can I see show of hands who knows about Let's Encrypt? Okay. So that's easily half. So, who uses Let's Encrypt in some form or another? Cool. Thank you.

So, for the people who didn't raise their hands, I'll give you a very quick intro. Let's Encrypt is a certificate authority started by the Electronic Frontier Foundation, Mozilla Foundation and the University of Michigan.

Last fall, I decided to investigate how successful they were in their start, in their first year. And I wrote a research paper together with Maciej and Giovane. Let's see if this works.

So why look at this? Mozilla and Chrome have been tracking HTTPS adoption over the past couple of years. And we've seen a trend to see the adoption or the use of HTTPS on page loads surpass 50% in telemetry by both these organizations. There's a lot of people who are very happy about this. And I think that makes sense.

Yet, it also means that there's a lot of people who still do not have access to encryption. The thing with encryption is that on the web in HTTPS is that you need certificates. And these are pretty much a major hurdle for small businesses or perhaps individual users, people who just register a website or maybe your grandma or someone else. It would be nice if our end goal would be to get HTTPS all around the world to get these domains also covered. So, that's the title of our paper is No Domain Left Behind.

So why is it that certificates make this hard? First of all, there is a cost of purchase. For most CAs, it costs money to buy one to get your identity verified even if it's just the proof that you are owner of a certain domain. But also there's a cost to deployment and renewal. Maybe we can do a quick show of hands who has ever screwed up a certificate in a web hosting configuration. We have a very trained audience, I see. But anyway, so that happens. And if you do that scale, there's complexity involved especially when you need to renew lots of domains.

Now, Let's Encrypt, for the record, we're not associated with them. They aim to reduce these barriers by decreasing the cost of purchase and renewal. Essentially, it's free. And also, they created a protocol to automate request and issuance procedures which is called ACME.

Now, what this enables you to do is to run software which automatically requests and deploys certificates on the web. So, our question here is does this actually help to democratize encryption? And we defined democratizing as it is adopted also in the lower end of the markets. In the place where you would not expect people to use encryption on the web not just the bigger companies, maybe the technologically more advanced groups of people.

So, that's what we set out to answer. And our approach was to analyze all the issued certificates in the first year of Let's Encrypt. So, the period is September 2015 to September 2016 and we show the adoption trend from various perspectives.

Now, we're specifically interested in showing the coverage for the lower cost end of the market. That is specifically shared hosting. Now, our contribution is then to show that 98% of the certificates issued by Let's Encrypt correspond to domains outside the Alexa 1M. It's a little bit obvious because the Alexa top 1 million is just 1 million. And there have been millions and millions of certificates issued.

But it still shows you where these certificates are going. Yet issuance is not just restricted to the lower end. And I'll get back to that in a bit. We show that the more popular domain in terms of Alexa rankings, the more likely the organization is using Let's

Encrypt in some form or another, which is actually surprising because you would assume that more popular domains have money to spend on existing CAs.

So, the second contribution is that we show that the growth of Let's Encrypt is attributed to really big players specifically in web hosting. What they do is they deploy encryption for all of their hosted domains. And we'll get back to that.

But three hosting providers were, in September 2016, responsible for 47% of Let's Encrypt certified domains which is pretty big. That shows you the power of automation. That's not just interesting because these are big companies but it's interesting because the end-users of these hosting providers are maybe your grocery around the block or your local bakery instead of the big companies where you would expect people to use HTTPS.

The fourth contribution is that we analyze that the issuance, the total issuance is actually concentrated in the market for shared hosting. And finally, it would appear that it's not just people trying out this technology for the first time. No, they keep using it. So we show using survival analysis that the majority of certificates are renewed after they have been issued. The standard lifetime of Let's Encrypt certificate is just 90 days. So, if you can see that domain consistently reuses certificates by Let's

Encrypt, that shows you that they are actually engaged and are actively using.

That leads us to conclude that Let's Encrypt is actually starting to democratize encryption. And I personally believe that is no coincidence that in a period of time that we see a growth in the graphs by Firefox and Chrome that these guys came into existence. Although, I do not have a proof for this.

So, how do we do this? We looked at one year of Let's Encrypt certificate issuance. The data uses certificate transparency logs which is an abandoned only log of all certificates issued. So we have 100% coverage of all certificates. We used a passive DNS database by Farsight Security to be able to historically map domains to IP addresses and then mapped IP addresses to organizations using an older methodology by some of the co-authors.

Now, what's very important to realize here is that the way we talk about domains in this study is in terms of 2LD or 3LD level. So, for example, your certificate is issued for www.example.org, we then reduce that to example.org. So, when we say a domain is Let's Encrypt certified, what we mean to say is that within the domain, there is at least one fully qualified domain name with a certificate for it.

And that's important to realize because otherwise, the statistics for Let's Encrypt get weird really quickly especially if you want to compare to other CAs because other CAs have, for example, lifetime of two years. And in that timespan, if you have 90 days validity, you have something like five, six Let's Encrypt certs.

So that's about the methodology. Now, let's dive into some of the results we found.

So first of all, what this graph shows you is that 89% of certificates are issued outside the Alexa 1M because the purple line, by September 2016 and actually from the start of 2016, trails around 2% of all domains we know. And it's good to realize here that what we did. We don't know the full set of domains in the world. I don't think anyone knows. So, what we did in this research is we use the domains from DNSDB in particular month as a random sample of the full DNS space.

So, when we say 2% of those are in the Alexa 1M, it follows that the 98% which is not is outside that list. It's very hard to say something to say whether the DNSDB set is representative but we are not aware of a more representative sample. So, I think it's as good as we can do right now although we're open to suggestions. The other thing you can see here is that the smaller the list, the smaller the contribution, which makes sense I think.

What's more interesting is that when you look at the relative use within a ranking, so for example, the use of Let's Encrypt within the Alexa top 1K, so the top thousand domains according to Alexa, the use of the Let's Encrypt is a lot higher than the broader rankings. So in this case, 18% of domains, of the top 1000 domains is using Let's Encrypt. Whereas if you take a larger sample, for example, the whole set, it's just 2%. And you would usually assume that the richer a company, the more popular the domain, the more money there is to find you favorite CA and not the free one that just started. So I think this really speaks to the trust in Let's Encrypt that there's staff trying this out.

Let's see. Then we made a number of graphs which show the adoption within a month, showing what kind of organization uses Let's Encrypt. So I'll take a moment to explain this one. So what you see here on the lower axis is organizations, so using the DNSDB IP information and Maxmind GEOIP, we clustered certified domains by organization. And in November 2105, this way, we identified 60,000 organizations.

In November, on the left axis, on the Y axis, there were 14,000 domains certified. And DNSDB had 127 million domains. Now, what this graph shows is that the in the lower quadrant, that there's a lot of smaller organizations who have an equal share, more or less, of Let's Encrypt domain.

So, there's no big steps in this graph except for the middle. And it's marked by gray which means it's unknown to us. So we don't know what the organization is. So this graph shows you that in the beginning of November, it was lots of individual organizations trying out Let's Encrypt.

Now, switch to September 2016 and the graph changes a lot. You see lots of big step functions. What that means is that there are suddenly these organizations which have a very large amount of Let's Encrypt domains associated with them. And specifically these three steps are Shopify which deploy HTTPS using Let's Encrypt to all its customers. It's WorkBest.com, automatic. That's a big one. And it's OVH, the French hosting provider.

Now, you can really see this kind of bumps nicely in the graph. And this also allows you to quantify the total percentage of the full domain space certified by Let's Encrypt which is associated with these big providers.

Now, if you look at who is using these, these big providers give you a little bit of information because all these three are in the area of web hosting. But we can also actually examine using these organizations what kind of business they are in. And then you get into this graph which shows you that more than 60% of the domains associated with Let's Encrypt are used for web hosting. I don't think anyone thinks that's a surprise.

Other chunks are the CDN business. You can see that it starts off with something like 10% but then it grows and it gets smaller and smaller. So, web hosting is really dominant. The white parts in the bottom are organizations where we don't know what their business is. So that's basically a shortcoming in our understanding of the data. But at least 60% is web hosting.

Now, what's even more interesting is that in web hosting, it's almost all shared web hosting. I think this is really where Let's Encrypt shines compared to all existing CAs because think about it, if you look at a traditional CA, what percentage of the domain says certify would be on shared web hosting but it's certainly very low.

What you see here is that of the 60% of all domains certified by Let's Encrypt, more 90% is shared web hosting. So, these are all people who would not have had HTTPS with a certificate by Let's Encrypt were it not for them to start this. And that's of the lower cost end of the market.

Three minutes to go?

Final finding is we can see that the certificates renewed on a regular basis. So this graph shows you the survival rates of domains after they started using it. And you can see here that the line becomes flat after around between 180 and 270 days,

which really means that these people have learned to set all the major renewal up and after that, it just works which I think is a pretty nice thing. I think Let's Encrypt would like to have it even higher but this is a nice start.

So, in summary, we see that Let's Encrypt is used widely in the lower cost end of the market. Specifically, this sector would be unlikely to deploy HTTPS because of the barriers mentioned. It enables big hosting providers to issue and deploy certificates. And that way, they quickly and automatically enable encryption for large numbers of domains.

The people who have been trying Let's Encrypt seem to stick because 70% of the Let's Encrypt domains remain active. I think we would hope to see more big providers consider to adopt this kind of solution because it seems very effective.

Future work, we would like to extend the measurement period. What we did not do right now is to actually see whether this is deployed and also deployed correctly. We are not aware of good data sets, big data sets which include sub-domains so using TLS SNI that seems to be a thing which is not generally available. And also, this is bound to be used by malicious actors. So it's interesting to see how that shapes up.

For more information, please see your paper or send me an e-mail. I'm here all day. Thank you very much for your attention.

EBERHARD LISSE: Thank you. Thank you very much. First of all, it's very nice to see finally somebody who is using LaTeX and Beamer as the presentation software which is the same what I'm using.

Secondly, we have been setting up two new servers in Wintook because our two servers that we're using there are 20 years old and have never failed. The new blades are indirect already for two years and we haven't really come to do it.

Ubuntu has a very simple way of setting this app. I'm busy doing this file while I'm listening to the presentation. No big drama. Rick Lamb. I'm cutting the lines after the first question because we're already running slightly behind so we cut a little bit in the lunch time which is long enough for this. Sorry.

RICK LAMB: All right. I'll keep it short.

EBERHARD LISSE: No. The lunch break is long enough for this.

RICK LAMB: This is Rick Lamb, ICANN. You may or may not know the answer to this but I have a question. Wonderful work, very promising results, makes me very happy. Who pays for the audit for the web trust audit for these guys? Or is this a system where you've avoided or somehow gone around that process?

MAARTEN AERTSEN: I'm not associated with Let's Encrypt but I think I read on their website that they have a system of sponsors. So organizations sponsor Let's Encrypt. But I wouldn't know. I think it's probably though.

RICK LAMB: It's a stability, long term stability question. All right. Thank you.

WARREN KUMARI: Yes, I think Mozilla paid a fair bit towards the web trust. I have long been concerned about the CA's infrastructure and system sufficiently that I found that the DANE working group try and design an alternate system for this.

When I first heard about Let's Encrypt, I must admit it thought it was a really stupid idea. I didn't think there was any way that it would actually manage to launch and be deployed. And I'm

really, really, really happy that I was wrong. I think this is the best thing ever.

MAARTEN AERTSEN: I share that feeling.

NIGEL ROBERTS: Yeah. My name is Nigel Roberts from .GG and .JE regisitries. Couple of comments and a quick question. First of all, congratulations. I think it's very worthwhile. We've actually put a link on our front page to you, guys.

MAARTEN AERTSEN: Thank you.

NIGEL ROBERTS: We have a lot of people who ask us for secure server certificates. We have to tell them where the domain registry. So maybe if only a few of those go in your direction, it will help.

The other thing is we would be interested to send you a small amount of sponsorship. So, if there's some way that you take smaller sponsorship, we'd be interested in that. And certainly, do you do green bus stuff?

MAARTE AERTSEN: So, I'm not personally associated with Let's Encrypt. So we did independent research on their effectiveness. But I think they would be interested. So, I would invite you to get in touch with them. I don't know about the status of EV certificates. I've been following Let's Encrypt but I'm not with the group or with the organization.

JOHN LEVINE: I'm wondering, have you looked at free certificates on services other than HTTPS? I have Let's Encrypt search on my POP, IMAP submission and SMTP servers. And I'm trying to figure out am I weird or is this the way to the future?

MAARTEN AERTSEN: So, we did look at that because all the stats we did were regardless of which service they were deployed on. So, the stats are representative in that way. I also use them for other services so I don't think it's weird. So, the [inaudible] organizations and what business they are in doesn't tell us very much how they are using it. So that's why we arrived at web hosting. I wouldn't know personally what the stats are for this.

JOHN LEVINE: Did you actually probe the servers to see if the certificates are in use?

MAARTEN AERTSEN: No.

JOHN LEVINE: Okay.

MAARTEN AERTSEN: So that's future work. I think there was a paper in IMC doing something like this last fall. So, there are people doing this. We haven't yet.

JOHN LEVINE: Okay. All right. Thank you.

EBERHARD LISSE: There is a remote question. We always will take remote questions even if their lines have closed. No, because it's much more difficult for them. And we want to encourage remote participation. And in particular, the Namibians have promised they would look into this so I must behave myself.

UNIDENTIFIED FEMALE: Okay. The first question is from John McCormick, HosterStats.com. "For web masters, do search engines consider Let's Encrypt certificates as less trustworthy than a paid certificate from another big vendor? And has this affected uptake of Let's Encrypt certificates?"

MAARTEN AERTSE: First of all, I wouldn't know because I'm not associated with Let's Encrypt. But I have never read such a suggestion anywhere. I know that Google, for example, says they use of they score use HTTPS in a certain way in their search engine algorithm. I don't know whether they do it differently.

UNIDENTIFIED FEMALE: Okay. And just one more from [Tomslin Sammy Enlar]. "Can ACME protocol also automate non-Let's Encrypt certificates?"

MAARTEN AERTSEN: Yes. I think if other CAs decides to implement the same software or implement a protocol, why not? It would be actually really nice to see another independent party also supporting this. I haven't heard of anyone scheduling to do this but that would be very interesting indeed.

EBERHARD LISSE: Okay. No more question, thank you very much. Thank you very much. And next presentation, last, before our long lunch break is Dmitry Belyavsky. We have a long lunch break so he can have his full 20 minutes. Please remain seated until we're done.

DMITRY BELYAVSKY: Hi. I will tell you about TLS statistics regarding usage of the TLS protocol in the Russian domain space. Okay.

First brief history of TLS. There were five versions of TLS protocol. Two of them are deprecated. Three are in more or less active use. Here is the statistics of its popularity. And while we're waiting for the TLS 1.3 which should appear I hope this year.

Now, we live in a period of increasing value of encrypted traffic. More than 50% of traffic is encrypted by measurements of browser vendors. New protocols usually require encryptions by design. Some years ago, hosting providers began to enable TLS by default. The first was Glob fly if I'm not mistaken. And so, the last well used unencrypted protocol in fact is DNS but there are a lot of recommendations for providing encrypted solution for DNS, DNS privacy.

Some words about Russian domain space, we have three TLDs. .RU has more than 5 and a half million domains. .RF has more than 900,000. That is the most successful IDN domain in the

world. .SU has about 120,000 domains. We have also two new gTLDs. We have third-level domains.

We have our statistic platform's name, the Stardom. So, it's a project of our registry. And our registry provider the Technical Center of Internet. It's based on Registry data. We analyze our domains. We can analyze foreign domains. The platform allows different variance of access to data.

So, some slides about visualizations and reports, not very interesting.

About one and a half year ago, we started to collect our TLS related data. Here is a very brief description of our methodology. We iterate our zone files. We go to port 443, collect certificates. We build chains of trust to browser roots. And after that, we provide the different metrics.

There is a more complete description. The link is provided. So first of all, we started collecting data in July 2015. When we started, there we're about 30,000 domains with valid certificates. And now, in January, we had more than 200,000 domains. So the total number of domains or having certificates has increased six times. The statistics is slightly different for certificates and for websites because some certificates include more than one names in .RU. Some was about CA distribution.

The most popular CA is Let's Encrypt. It's not a surprise to anybody. It has about 46% of Russian certificates market. The second position is Cloudflare with 59.5%, then cPanel and GlobalSign.

Before the appearance of Let's Encrypt, before March 2016, Cloudflare was a leader on the market with about 30%. But when Let's Encrypt appeared, we see the significant growth of certificates and very significant growth of Let's Encrypt certificates.

The migration to Let's Encrypt is [RSA] limited. But people prefer to migrate to more cheaper CA to get the certificates cheaper or for free or in bundle. So, total migration seems to be between 5% and 10% per period.

When we compare which domains have certificates now and will have their certificate a month later, we see that it's usually more than 90%. So, it's a good sign that we have a growth in market of certificates.

Some was about algorithms or SHA1 had significant share in 2015. Now, we have only 116 certificates in all .RU space using SHA1. The most popular algorithm for digital signature is RSA. The most popular K length is 2048 bits. The elliptic curves algorithms are not very popular. They get only 15% of their

market. It had up to 32% before Let's Encrypt appeared. So some interesting facts that almost all elliptic curve certificates are from Cloudflare.

We have something about just the ten certificates with elliptic curves which are got from other CAs. We have about 70% of certificates that are either for free or parts of this or that bundle. We have only 600 EV certificates. And we expect that there are some more EV certificates on the third level. But we can't measure it.

We don't see any correlation between EV certificates and using DNSSEC. And our last research regarding using TLS in MX servers, we see that 70% of IP address, of domains IP address is MX are protected with STARTTLS.

So, some words about what do users think about TLS. It's a problem that people trust to Green lock. And now when we have cheaper or certificates for free, it can be a problem because it's very cheap to register domain names looking similar like, for example, PayPal or famous banks, get certificate for free and provide phishing. So, the only solution that can be suggested is using EV certificates which are much more expensive. And the most difficult problem is to explain it to clients.

Some words about what we are to worry about. We have to worry about mobile application. Now, a lot of services use mobile application for the purposes. In Russia, it's something like must-have. And we don't have a research regarding Russian-based mobile application. But the worldwide research show that there are a lot of certificate validation errors since such application. In the end of 2106, there was a study of VPN solutions for Android which showed that five or six VPN solutions though not ensure necessary protection.

We should worry about different forms of TLS termination because now most protected software are browsers. And so that was a very interesting study regarding using TLS proxies which are used for TLS termination that show that TLS proxies often don't provide enough security.

The last slide.

We have a problem by design of bigger infrastructure that any CA can issue a certificate for any domain. We have a limited set of solutions until now. DANE, certificate transparency by Google, certificate pinning. But none of them is comprehensive enough. So I think we will see problems in this area more and more. Thank you.

EBERHARD LISSE: Thank you very much. Any questions? All right. Then I can release all of you to lunch. We will be here. Let me quickly look. I think it was 1:30 or 12:45? 1:45? Okay. Kim, you can leave this on like this, yes?

[END OF TRANSCRIPTION]