ABU DHABI – Internet Technologies Health Indicators
Wednesday, November 1, 2017 – 10:30 to 12:00 GST
ICANN60 | Abu Dhabi, United Arab Emirates

ALAIN DURAND: Good morning. This is going to be the ITHI session. We are going to start in just a couple of minutes, waiting one more person to show up. So let's start. Welcome to the ITHI session. ITHI stands for Identifier Technology Health Indicators. We have two components mainly on the project, one about names and one about numbers. We just got Alan Barrett who's going to talk about the current status of the number effort on ITHI. Just in time.

UNIDENTIFIED FEMALE: Do you have the microphone?

ALAIN DURAND: Yes.

ALAN BARRETT: Thank you. Good morning. My name is Alan Barrett. I'm going to speak about what the Internet number registries, the RIRs are doing to help the ITHI initiative. I'm not really John Curran, I don't look a lot like him. John was unable to be here.

Note: The following is the output resulting from transcribing an audio file into a word/text document. Although the transcription is largely accurate, in some cases may be incomplete or inaccurate due to inaudible passages and grammatical corrections. It is posted as an aid to the original audio file, but should not be treated as an authoritative record.

Is there a clicker or something?

UNIDENTIFIED FEMALE: Yes.

ALAN BARRETT: Thank you. Okay. So I'm sure you've already heard from [Alain] the goals of the ITHI initiative. Essentially, ICANN asked the RIRs to work with them on measuring on the health of the number resources, and so the Regional Internet Registries agreed to develop some metrics. We've been doing some work over the past six months to a year now. I think it's close to a year.

So staff at all five RIRs, the staff that I involved with the number registries got together and thought about what we could measure to contribute to this effort. So we've drafted a strategy, drafted some metrics, and we are in the process of community consultation. So I think on the next slide – no. Yes, okay, so in about three slides' time, you'll see the URLs for the proposal which has been developed and which is currently out for public consultation. So we expect that to be completed towards the end of this year.

And then after the public consultation, of course, we'll revise the document taking into account any new developments and any public comments, and we expect to release our next version of the metrics document towards the beginning of 2018. And then we're not sure how long it will take to implement. Each of the five RIRs will have to do their own work towards implementation. So although we collaborate, we each work on our own part of the system.

The scope is we want to measure all the Internet number resources which means IPv4 addresses, IPv6 addresses, and autonomous system numbers. These are the number resources that the RIRs deal with. So we allocate or assign these resources to our members who are usually Internet service providers or other businesses who have a need for significant amount of Internet number resources.

We're only going to consider those with parts of the resources which the registries deal with. There are certain parts of the address range which are reserved, so in some cases the IETF reserves certain parts of the number ranges for reserved for future use or reserved for experiments, or reserved for some special purpose. These are out of scope. We're not trying to measure the health of those parts of the resource space.

Okay, so I promised some URLs. It might be a little too small to read, but these slides should be available somewhere for download from the ICANN website. If you go to the agenda and click around, I'm sure you can find a way to download the presentation.

But if you go to the NRO website, nro.net, you should be able to find the information on the ITHI project. There's a call for comments which is open until the 12th of December. There's a document you can download, and if I have time later, perhaps I'll be able to display part of it. I'm not sure if the Tech Team here in this room is able to go to that URL for the proposal, there's a PDF. And if we have time later, perhaps we can talk about that in more detail.

We're holding consultations in each of the five Regional Internet Registries in our public meetings. As I'm sure most of you know, the five RIRs do their policy outside ICANN. We have our own meetings. Each of the five RIRs meets twice a year to discuss policies, and as well as policies, we'll also be discussing the ITHI project. It says this fall, I think it means around the end of this – yes, the second half of this year.

Okay, we do have a mailing list which you should be able to find on the NRO website. It's called ITHI Consult. There's not been much activity on the mailing list, but the archive is accessible from that URL. You can go there both to subscribe to the mailing list if you wish to receive messages in future or to send any messages, and you can also read the archive of all past messages that other people have already sent as part of this discussion process.

Okay, so thank you. I'm not sure how much more time we have. I think we do have time. Yes? So then could I ask the Tech Team, somebody to put up the PDF which is in the middle link in about the second to last slide in the presentation? Yes, that one. That PDF. Okay, thanks. Sorry, I should have provided this in advance. I thought having a link would make it easy to find.

Okay, so this is the document which hu can download, and if we scroll to the next few pages, there should be a table. Let's scroll or page down until we find a table.

UNIDENTIFIED FEMALE: [inaudible]

ALAN BARRETT: No, I'm not doing it. Can I do it? Does this do it?

UNIDENTIFIED FEMALE: Yes.

ALAN BARRETT: Oh, okay. Great. So here we go.

UNIDENTIFIED MALE: [inaudible]

ALAN BARRETT: The text of course is too small to read, so I'm going to have to stand really close here. Okay. So we started by trying to identify what metrics we could use to tell whether or not we're measuring the right thing. We wanted to know, is the information complete? Do we have – aha, there we go. So, is everything complete and unique? So we found some metrics here like, do we have everything registered in the database, or is something missing? For uniqueness, we think we can measure whether there are duplicate records, perhaps more than one RIR might claim to have jurisdiction over the same part of the address space. And in the ideal case, each part of the address space would be handled by exactly one RIR.

Can we scroll down? I'm not going to go through all of this, but I'm just going to highlight a few of the points.

So, is it comprehensive? Which means, is everything there and is it all unique? Is it correct? So we want to ask things like, does the registration listed in the database match the official sources? So, for example, we can look at the legal name of the organization. If some part of the address space is registered to a company, is the company name in the RIR's database the same as the company name in the database perhaps held by the legal authorities in that country? That could be one aspect of correctness that we think we can measure. Is it up to date? So we can measure how recently has the information been updated, and even if there's no change, how recently have we checked to see whether the information is accurate? So we can give people a way of saying, "Yes, I've checked this and it looks right," and we can measure how often that takes place. So that can be one of our metrics.

So there are several here, there are a few pages in the document which you can download from the NRO website, and you can comment on that. I don't think I want to go into more detail now. The point was simply to show you that we've produced this matrix with all kinds of different metrics, and it's still a work in progress. And even after comments, in the RIRs we will need to write software to go into our database and measure these things. Okay. Thank you.

ALAIN DURAND: Thank you, Alan. Are there any questions for Alan? No? Well, I have one for you. We are being asked from time to time about abuse metrics, and one of our metrics is the abuse on the domain name, we're going to talk about this data. There was a discussion yesterday, for example, on the DAAR project where people were wondering about abuse on the number space.

Should we redirect them to you, or should we try to do something different?

ALAN BARRETT: Okay. I think abuse of the number space would be within scope of what the RIRs are trying to do here, but I'd like to distinguish between abuse of the numbers and abuse that's simply used as the numbers to facilitate some other kinds of abuse. For example, sending spam e-mail certainly is a kind of abuse, but I think it would be out of scope, whereas hijacking addresses would be within scope.

ALAIN DURAND: Okay. Thank you very much. Thank you for your presentation. Questions? Dave.

DAVE PISCITELLO: Alan, I like your distinction, and I think it's great that the RIRs are pursuing hijacking, squatting and theft of accounts that [can cede] large blocks of addresses. But you left the other part on the table, and so if we want to include things like IP addresses that are associated with spam in our DAAR project, I'm assuming that the RIRs would have no objection.

ALAN BARRETT: We don't have a consensus position on that, so I can't say whether or not the RIRs would have an objection. We have a process where we need to reach consensus before we can take positions like that.

DAVE PISCITELLO: Okay. Let me just make a clarification. What we might want to experiment with is not only pulling in reputation block lists that are domain block lists, but also using reputation block lists that would help us identify IP allocations or autonomous systems where there's a concentration of abuse, because that's a corresponding way to take a look at the overall threats that we examine.

There's actually a very fine product that you can look at where I'd hope we'll get data, it's called Seclytics.com, and you get some amazing views of the threat landscape at the IP and ASN level. So one reason I bring this up is because it would be very useful to sort of see, especially if we were to be successful in mitigating use of domain names, see if they would return to just primarily using IP addresses or focus on specific autonomous system numbers.

ALAN BARRETT: Thanks. You raise some very interesting ideas, and I think it's very useful to track all of that. But speaking for myself as part of AfriNIC, I don't think we have the resources to do it. If somebody else wants to, I would have no objection, and I'd like to see the results.

DAVE PISCITELLO: Okay. Thank you.

ALAIN DURAND: Okay, so questions?

WANG WEI: Wang Wei from KNET. I know that five RIRs are promoting the deployment of RPKI. Will this research help to [format] your repository of the IP addresses of RPKI in the future?

ALAN BARRETT: In that document, there is something about RPKI. There are measures about how much of the address space is covered by RPKI and whether the certificates are valid or if there are errors in the chain of certification. I'm not sure if that really answered your question.

WANG WEI: Well, it helps, really. [inaudible]

ALAN BARRETT: Okay. Thanks.

ALAIN DURAND: Well, thank you all. If there are any further questions about this effort, I would encourage people to go and comment on the mailing list and provide feedback to the RIR on this project. We had a tremendous effort in the last year to put all this together and really wanted to thank you for that. And we see [outer] movement with this other part that Dave was mentioning earlier. Thank you very much.

So the next segment is about ITHI for names. I'm going to make a very quick overview, and then I will hand the microphone to Christian who is going to dig into some of those metrics.

The process we want to use in the namespace is to access some data sources that may either have private identifying information or that may be under contract and that cannot be made necessarily public, but that contact information that are quite interesting for us, thinking in particular of some of the information about the anti-abuse that Dave Piscitello was talking about earlier, but there are others.

This for us will be raw data that we may want to access but we can't really publish, so we want to extract from this something that will be shareable with the community, some kind of an aggregate view of this, and start to perform some analysis, maybe calculate some average numbers or prepare some graph, and make all of this then available to the community through the ODI project. So there will be a session tomorrow on ODI which his the Open Data Initiative, and if you have any question exactly on how this is going to happen, then we can talk about this.

So this is the pipeline of information. What is really important here is this barrier here with the dotted line where we want to isolate as much as possible from any place where there is raw data that contains PII or contains contractual data. Next slide, please.

We have a number of high level metrics that we have defined, so I'm going to go quickly through all of them. The first one is about data accuracy, that maps very much with what Alan Barrett was showing us earlier on the number space. One major difference is that the RIR have access to all the registration data. In the IANA namespace, ICANN does not. The registrars have this information, we don't. So we cannot go and measure directly this.

So we are looking at indirect measurements through the level of what we call validated complaints per million registrations. Validated complaints is something that has been received by the ICANN Compliance department and that someone in the ICANN Compliance department has judged that, "Yes, this sounds like a real complaint, we're going to investigate." Not necessarily take action, but at least really investigate. So try to remove all the spurious complaints, frivolous complaints. This is not necessarily the most satisfying thing, but that's where we are planning to go for now. Next slide.

Next slide is about abuse, and we are going to be a customer of the DAAR project. There was a big presentation yesterday afternoon from Dave Piscitello on the DAAR project, and we would like to keep track of spam, phishing, malware, and botnet independently. So Dave, if you'd like to say a word about the project, maybe.

DAVE PISCITELLO: Sure. Briefly, we are collecting the zone files for all the generic top-level domains for which we can obtain zone data through either the CZDS program or directly through legacy arrangements. We then obtain WHOIS for the sponsoring registrar field only, and the IANA [ID] to associate the domain names with the sponsoring registrar, and then we have about 15 to 18 reputation block lists, and what we do is we calculate or count the numbers of spam, phishing, malware and botnet, and associate those with each of the gTLDs and each of the registrars.

So we have data from January 1, 2017 at this point, and we have reporting on the registry level with very high confidence since the beginning of May. And so we will be feeding information to the ITHI project, and the current status of that is we're actually doing manual transfers right now, but we're building in automations where that information will be made available in some pull fashion for Alain to populate the M2 metric.

ALAIN DURAND: Thank you, Dave. One of the challenges here is how much information we are going to make public. This is a discussion that is happening within the DAAR project, because we don't want necessarily to go into the naming and shaming game. And we will do essentially what he community ask us to do. Now, remember that we will be a customer of the DAAR project, so if the DAAR project and the community reach an agreement that they'll publish up to a certain level, we will match that level. we will not go any further than that, of course. Is that a correct statement, Dave?

DAVE PISCITELLO: Yes.

ALAIN DURAND: Thank you. Next slide, please. So M3, Christian is going to talk a lot more about this. This is about overhead traffic to the root. Next slide.

M4 is about usage and leakage as seen as recursive servers. What's the difference between those two? One is seen at the root, and the one will be seen at recursive servers. Next slide.

Still talking about recursive servers, we had a number of anecdotal evidence of problems of the customers of those recursive servers. We have heard things about recursive servers handing out wrong information. For example, you ask for a certain business server, and you are being directed to a competition. That's not necessarily what you expect. Now, those are anecdotal evidences. We don't know how widespread it is. But we have heard enough of those anecdotal reports that we would like to find a way to measure this.

Another thing that we have heard is sometimes not the recursive server itself, but the service provider operating recursive resolver will redirect port 53 like DNS requests towards its own server. So if you were trying to say for example, say, an open resolver like the google 8888 and you think you go there, actually, you're not going there. You're being redirected to the ISP resolver, and that may fall back into the previous type of problems. Again, we don't know how prevalent those problems are, that's why we'd like to track them. So we're going to work with Geoff Huston from APNIC and the system of Google [ads] that he has that may help us to shed some light on this. So I expect by the next ICANN meeting, we'll have some details on this one. Next. Next.

The next set of metrics are more related to protocol parameters and how are those things actually being used. The first one is about the DNS protocol parameters directly. So the IANA registry contains a number of registries that have registered parameters related to the usage of either DNS or DNSSEC and others, and we want to see how much of them are actually being used. So Christian will also talk into more details about this. The next one.

And the last one is M7, this is keeping track of DNSSEC and simply counting the number of top-level domains that have been signed, so it's a much simpler one. So Christian, I would like to invite you to [inaudible] dig a little bit further into some of those metrics.

CHRISTIAN HUITEMA: Thank you, Alain.

UNIDENTIFIED FEMALE: [inaudible]

CHRISTIAN HUITEMA: Yes, please. Good morning. Okay, we can go to the next slide. Oh, yes, I'll use a clicker.

So Alain gave a brief overview of all the metrics that are envisaged in the ITHI indicators there, and I will go in details into three of those that are related to the DNS, either because we are looking at the DNS root or looking at the DNS leakage of names, or looking at DNS parameters.

Our methodology there is to look at the traces of traffic for which we want the cooperation of recursive resolvers, and we analyze those traces of traffic to count the occurrence of various issues. So I'll see that here. First, let's look at the overhead in the root traffic. You may have seen graphs like that in previous reports on the state of the root. The root traffic, if we look at it, a lot of the traffic at the root is not really necessary. About 60% of the traffic of the root, of the queries sent to the root, results in "no such domain" responses. And that's something which is interesting.

Then even the traffic that is correct, that basically there is really a domain like that, if we look at the TTL, what we observe is that the TTL may be like six hours, but we see a query for the same resource a few seconds later, a few seconds again and a few seconds again. So this is a big source of override for being not TTL compliant.

So we propose to look at that with three metrics: first, measure the ratio of "no such domain" response to the total queries, then look at the ratio of non TTL compliant queries over the total queries, and then another metric is to look at these NX domain responses and try to find out why we have so much of that.

Remember that if we [add] perfect recursive resolvers, they will kind of know the root, they will cache, they will do negative caches and things like that, and we'll see very few of that. But we looked at [addresses,] we don't have any data that we can trust to the point that we can publish them. But we know that globally, those "no such domain" response fall in four categories.

The first category are the special use names defined in RFC 6761. These are names like .local for example that have been reserved by the IETF for special use in special protocols. Normally, we should not see them arrive at the root, but yet we do, and we want to see how much of that there is.

The second is the frequently leaked names. These are names that are not reserved by the IETF, not registered as top-level domain, and yet we see them in many root queries. An example would be the .home domain, which doesn't exist, it's just a string, but we see it happening. Then we look at names that don't seem to make any sense. They look like automatic generation, like this string. I see that it's actually taken from one of the traces. It's obviously not meant to be resolved. And then we have all others. So we want to look at that.

For the RFC 6761 names, well, that's really easy to measure. We know how many there are because they are registered, so we just have to count how many times we see them and compute a percentage of this – I mean how much of the overhead is accounted for by these special names. So that's basically relatively simple, and that will be the metric 3.3.1.

Then we have the overhead with the frequent names. This is a bit more complex, because those are not registered, and so we cannot say, "Hey, we are going to count how many times we see .home or .coop or something like that." Of course, we could do that, but we would miss the new ones and new leaks. So we have to develop a system in which we look at the most frequent strings like that that we see, and we count them. And what we will do is that we will retain all of the strings that account for more than 0.1% of the leaks so that we can have something that is manageable. And we'll publish that as the metric 3.3.2.

The metric 3.3.3 is interesting. It's the pattern. We are very intrigued by this business of having automatic name generations and would like to understand how much of the overhead is caused by these automatic name generations, and what kind of them. There are several kinds. So each kind of name generation corresponds to a pattern. A pattern might be defined by something like, "Is it a numeric string of three characters?" Or, "Is it an automatically generated domain of 10 characters?" Or something like that. And this is clearly going to be a work in progress, but our goal is to understand this phenomenon, and then once we understand it, measure it and see whether we can do something about it all. Excuse me. We are not going to do anything about it. We are just publishing the data. But [someone] might. Yes, Dave.

DAVE PISCITELLO: I believe that algorithmically generated domains for botnets would fall into this category.

CHRISTIAN HUITEMA: Yes.

DAVE PISCITELLO: It would be interesting to have a conversation with some of the people who do the malware reverse engineering and identify the algorithm, because it may be useful to be able to pattern match against malware.

CHRISTIAN HUITEMA: Yes.

DAVE PISCITELLO: And when the root sees a new set of domains that do not match known TGA patterns, that information might be extremely valuable to trigger some investigation, especially on the IPs that are initially announcing them, because that would be a way for our community to help identify perhaps at the onset or birth of a new botnet.

CHRISTIAN HUITEMA: Yes. So there is a little problem with what you say, is that for privacy reasons, we don't have to have IP addresses in our metric files. But we could definitely have some kind of parallel system that does that.

DAVE PISCITELLO: If we aren't doing individual hosts, but you are doing the sider allocation, would you keep that?

CHRISTIAN HUITEMA: I would have to ask.

DAVE PISCITELLO: I know that everyone gets nervous about an individual host-assigned IP address, but even knowing what the longest sider string is, subnet string is, would probably be very useful for us to work with the malware [inaudible]

CHRISTIAN HUITEMA: There’s definitely a need to cooperate, if only to observe the patterns and observe which ones are growing and which ones are decreasing, and correlate that with botnets and see what we can do.

DAVE PISCITELLO: Right.

CHRISTIAN HUITEMA: And then you could have another system which is not meant for publication like this one that actually does look at the IP address in some controlled fashion.

DAVE PISCITELLO: Okay. I think it also will be interesting to see how the whole notion of situational disclosure and collection plays out with

GDPR, because in the case of trying to have accountability for harm, there's an argument that that kind of information would actually be appropriately collected and appropriately curated.

UNIDENTIFIED MALE: Okay.

CHRISTIAN HUITEMA: Okay.

ALAIN DURAND: I think that's a really good point. We need to separate what is for the archives, what is the long term and the trend that you're trying to observe in ITHI, versus what is situational and could have security and stability impact.

CHRISTIAN HUITEMA: And the fourth metric for the overhead is basically all the stuff we don't understand yet. That's probably what in that category that are the new patterns that we're speaking about, and we'll probably mine the data there to create new patterns and put them back, and try to identify them. So that was the metric M3. M3 is about the overhead at the root, trying to understand why there are only so few of the root queries that are actually useful.

The metric M4 is about usage of TLDs. We want to understand basically how TLDs are used, and we want to do that by looking at what users are doing. In fact, we want to do that, as Alain said, by looking at the queries that are directly sent by users. So the idea is to [take] a sampling of traffic that originated from actual users, and look at the TLDs that are queried in that traffic.

And we will get four kinds of metrics out of that. The first one is just TLD usage. Out of one million queries, how many go to this TLD and that TLD? And that's interesting per se just to see how TLDs are used, whether traffic equalizes between different TLDs, or is it very imbalanced like it is right now?

The second metric is about leakage of special use RFC 6761 names, and to see how much of that is happening directly from users. Now, I will go back to that later, but you must realize that what we will see by looking at user traffic is not what we see by looking at root traffic, because between the user and the root, we have resolvers that are doing caching, that are doing things like privacy enhancements, and so the statistics of the user traffic are not the same as the root traffic. So we're going to see what are the statistics of the root of the user traffic, like how many times do we see .local or .example in user traffic.

The third one is the leakage of nondelegated strings, that is a string like .home or .coop. And again, we want to look at that directly in the sure traffic, because it will be saying like out of one million user queries, we see so many quires for this name and that name. The methodology for measuring will be pretty much the same kind of statistics that we can do at the root, but the raw data is different. And because the raw data is different, we'll get a different signal.

And the fourth one is the leakage of other strings, which is basically to try to find out what's happening there, and again, something that we might investigate later. So that's the usage of TLDs, and I believe that if we have that, we'll get a good idea on how TLDs are used in practice and how we can see them evolving.

The next set of metrics is about IANA, and specifically about DNS parameters in DNS queries. If we look at the IANA process, the IANA process typically starts in the IETF with a working group creating a protocol specification and saying, "Hey, for my protocol, I need as table of parameters managed by IANA." So they contact IANA, IANA creates a table of parameters, there are generally some initial values set at that point, then IANA sets a parameter of registry which is managed with the process specified in the RFC. There are several options.

Developers who are doing applications use those parameters. Hopefully, they use the registered parameters, but sometimes they do squatting by inventing their own parameters. And then there is some practical usage that again we can see. Now, the reason we look at DNS parameters rather than, say, IP protocols and things like that, is because DNS parameters are easy to observe when analyzing DNS traffic, and also because we care more about the DNS. Conceptually, we could do that for other kinds of IANA parameters, but [IANA seemed] more traffic.

Now, registries, we have many registries like that for the DNS alone, things like [R types] or EDNS options, and DNSSEC also has its set of parameters for [algorithms] meant for keys, then as it set of parameter for certificate types, there is a very large number of parameters of registries. And for those registries, we have two questions that we want to answer. One is, okay, we defined parameters. Are they actually used? Does someone actually use them somewhere, or is all that mechanic for no good reason, just some kind of overhead?

The second one is, do we observe squatting? So they're kind of two opposite. "Are the parameters used?" Is basically, is this registry useful? "Do we observe squatting?" is, do we see usage of unregistered parameters because people could not be bothered going through the process?

I'll give you an example about how we think about that. Suppose that we have –

UNIDENTIFIED FEMALE: [inaudible]

CHRISTIAN HUITEMA: No. I was trying to look at the laser pointer there.

UNIDENTIFIED FEMALE: [inaudible]

CHRISTIAN HUITEMA: No, okay. This one?

UNIDENTIFIED MALE: Yes.

CHRISTIAN HUITEMA: Oh, yes. So basically, what I have here, I have typically the outcome of the statistical analysis. I've taken a fictitious example of some fictitious parameter that would have 16 possible values. And of those 16 values, 10 have been registered at the registry and the other six are not registered and in theory should not be used.

The statistic gives us – that's the bar you see there – the number of times we see this parameter in the traffic. With that, we create two numbers. First, we say, “Okay, do we see anything about that parameter in the traffic? Is it used at all?” If it’s used at least one, or if we have a very large dataset, we might take a threshold, if it is used more than a threshold we consider it being used. And then we can compute here the ratio of how many of those parameters are used over how many are registered, and that gives us an idea of basically the usage of the parameter.

The next metric is to look at those nonregistered values, which are what we call the squatting part. And for that, we sum these little bars there. We sum the red bars, and they sum to eight in my example, and we sum all the bars, the blue and the red, and they sum to 86 in my example, so that is the volume of nonregistered traffic over the total volume of traffic, and that gives us the squatting metrics. How much squatting do we observe on that parameter? And we believe that with these two metrics, we will have some handle about how the IANA registries are used.

So we have a little problem of naming, because there are many registries. We’ll do a naming that includes the registry class, something like DNS or DNSSEC or [DANE], the class of registry for IANA. We’ll take for each registry three metrics: the usage metric that I described in the previous slide, the squatting metric, and also the volume of each registrant value. The volume of each registrant value is interesting for us to see things like take

DNSSEC algorithm, looking at how many times we see people using RSA or [inaudible] or something like that tells us about the evolution of the usage. It's not directly for the health of the parameter, but it's very useful for analysis.

So take the example of one example for the RR types. The RR types is a second registry in the DNS class if you go to the IANA website, so it would be the registry number two that gives us the metric with this funny name like M6.DNS.2.1 for the first usage metric, .2.2 for the squatting metrics, and .2.3.the value such as the value 28 for example for the record type 28 gives us how much of that we see. And so that gives us the M6 metric.

Here is the list of all the registry parameters as we'll be talking. We might want to add more registries in the [first of] release when we find out ways to basically track the traffic there, and that will give us – oh, thanks. So that was the presentation of the metrics themselves. The next question is, how are we going to measure that?

The proposed methodology is to do effectively sampling. We are not going to look at all the traffic on the Internet and extract the numbers. We are going to do sampling. And we are going to do sampling by having collection points in which we take copies of the [traces] and analyze them.

Those collection points, at each collection point we'll take collection every day, typically at a random time in the day so that we have a spread of time of day, and we are not going to collect huge files. We think of collecting about one million transactions each time we do a collection. One million transactions is not that large if you look at the volume of DNS traffic. And that gives us enough statistical sampling that we have good precision. If you think about it, we are interested in events that happen at least 0.1% of the time, so 0.1% of the time each of those events, things like that would happen a dozen times, so the deviation would be quite low and the precision would be good.

The data will be analyzed at the collection point. It will produce a summary, and the summary will be sent to ICANN. So that's the separation. Alain mentioned that there is a slide saying, "Hey, we don't want to look at the raw data." The raw data is like the [traces,] there is all kinds of information in them. We want to look instead at just the summaries, and from the summaries we compute the metric. So we'll take the summary for many collection points, and we'll aggregate them to compute the metrics.

In order to do that, we developed a small tool that will – it's a small program that can read a PCAP file and produce the summary file, and then can read several summary files and produce the metrics. So if I look at that in a graphic way, what you see there is that at each collection point, we assume that people will be using some kind of monitoring tool that will produce a [catch all] file. The monitoring tool can be something like DNSCAP, and with the tool that we have developed, we can do extraction of a summary file.

Now, many capture points will do the summary files we will put on a network share, and then at ICANN, we'll take those summary files from the network share, compute a merged summary and compute the metrics. The point here is to have a clear separation.

I briefly touched on that in the description of M4, we realized very quickly that we cannot just do these measurements by looking at root data. We can do the M3 measurement by looking at root data. Indeed, we must, because if you want to measure the overhead at the root, you have to look at the root. But if you want to look at usage, you want to look at the usage just before the recursive resolver.

Because between the recursive resolver and the root, we'll have caching. If the resolver were perfect, they'll cache everything and won't see anything. Suppose they do, for example, NSEC3 aggressive. They know exactly how many TLDs are at the root, so we'll see an occasional query for refreshing a TLD value, but we'll never see any wrong query, no NX domain, nothing. And we'll not be able to do statistics.

The same thing goes for the DNS parameters. The DNS parameter in the user traffic, suppose that the resolver do QName minimization, which is the idea that they only send a need-to-know query at the root. Then we'll only see requests for TLDs and for the NS record of the TLD, and we won't get any information about the other parameters. That means that in order to get that information, we have to be located at the resolver and get the [user] traffic directly. And that's of course a problem, because I don't operate a recursive resolver. Well, I do, but I'm the only customer so it won't be very representative.

So that's why we have to get cooperation from a recursive resolver in order to get that data. Now, if we mention cooperation, we have to look at issues like GDPR. If I ask a resolver, "Hey, can I get your [traces]?" The answer is, "No, I cannot legally do that. I'll get in trouble." So what we want to see is remove the PII traffic. We don't look at the IP address of the users, don't look at the domain name that they are looking for, definitely don't look at the pattern of usage queries. But it turns out that none of those metrics that we have there require PII of any kind. We do not need the IP address, we do not need the queried name, we just need a TLD.

So if we summarize those data, we don't have a piracy issue, and we believe that we can convince some recursive resolvers at least to say, "Yes, I can do that." The summaries themselves are pretty small. A typical summary is 8-16 KB, so you can inspect what we are sending back to IANA and you can verify that we are not getting statistics about what the users are doing.

The tool design, the tool that we have is meant to ensure that privacy is respected. So we have developed a single tool which has three functions. Parse a capture file, merge several summaries, and compute the metrics. The tool is open source, it's on my GitHub server, it's available with a classic open source license, and it's written in C++ for Window and Linux.

One thing that we have [to] make sure, having it open source means that anybody can verify that we are not pulling shenanigans in the code, but what we have also done is make sure that the code can run in a sandbox. It doesn't need any Internet access when it is running, so it's isolated, there is no leakage of data. And the summaries don't have to be – the tool doesn't have to write directly on the network share. It can write on the local file and then use a script to put it on the network share. So it's all controlled.

So hopefully, we have solved the privacy issues, and we can at that point start asking for help, getting volunteers who say, "Yes, I think it's good to have those metrics for the good of the Internet. I think the capture methodology is safe, there won't be any issue." So if you believe that, please contact us and we can start organizing some collection. Thank you.

ALAIN DURAND: I wanted to add to this that we are planning to work with another third party that is going to do some kind of an audit code review of the code to make sure that it's okay, there's no problem with that. And we are starting discussions with two very large recursive server operators. Two is good, but we would like to get more, so if you are interested, again, please talk to us.

Now, I would like to open the floor for any questions for Christian, or any question for me or for Alan, or any other question that you might have on the project.

Well, if no further question, then maybe we can release you for an early lunch. Sounds promising? Alright, well, thank you all very much. We're going to post all of this on the web, of course, and we'll send some of that via ITHI mailing list. Let me remind you that there's a call for comments on the number side, and you have seen the slides from Alan where the URL is on the NRO website. Please participate. We've done some great work and would like some feedback on that. Thank you all very much, and see you next time.

[END OF TRANSCRIPTION]