JAMES GALVIN: Okay. Thanks, everyone. This is the NCAP discussion group on 27th of May in 2020. Thanks very much, everyone, for taking the time to join us today. I am James Galvin, one of your co-chairs, and I will be monitoring the call today. Of course, that doesn't excuse anyone from jumping in and speaking whenever they feel the need to add something here.

Let's start with updates to statements of interest, follow our normal ICANN protocol. Anyone have any updates to their SOI that they want to announce? I'm not hearing anything, so we'll move on here. We don't have any new members as of yet here in this group, so I think that that's fine, but please do tell your neighbors, colleagues, countrymen. It's always good to get some new and fresh voices in this discussion, although I suspect the stuff that we do later as we [inaudible] study two in earnest will really strike some more interest in folks.

So jumping on ahead, we had the definition of name collision. We really do want to restructure that document and get all that in front of this group, but we haven't gotten that to a place where it's useful to put it in front of the group just yet, so didn't want you to think we were forgetting that action item. We're just going to jump in here and continue to move forward with our study two analysis notes.

As we had discussed last time, we had gotten through question three and rather than doing four, five and six right away which seemed to jump right away into mitigation, we thought it might be a better follow on to start with question seven and jump in there as we have announced and we have now also up on the screen here for folks to look at. And we would talk about the criteria for how to determine whether or not a string is a collision string, what kind of data [inaudible] of things, information might be relevant for determining whether or not a string is a collision string?

Now, I appreciate there's some sensitivity to whether we know exactly what the definition of name collision is, but I think for the purposes of this discussion, we all have a sense of that and at least we can suggest things here for consideration and inclusion that will later get revised, enhanced, certainly more deliberately and carefully finalized before we have any final work products here.

But what I want to do for context, just in case folks have not looked ahead to see everything that’s here, question seven really talks only about the question of whether something is a collision string, to use the terminology that the board used in the resolution that it had provided. So whether or not it is a string that would manifest name collisions.

And it really only wants the criteria to know if something should be in that space. I want to call out the fact that if you jump ahead and look at question eight, that’s where we get into a discussion of, okay, now we have the various criteria for identifying potential collision strings. Now, how do we evaluate those criteria? How do we use them to determine whether or not something should be delegated or not and what advice can we provide to the board in that category? So I just want to call out that there's a careful distinction to be made here between criteria that might identify a collision string and separate from that is the criteria of how we would evaluate those metrics and hat information that we gather to determine whether or not something should be delegated.

And similarly, question nine actually gets into the question of evaluating what is the harm and the downside, what are the ways in which you can mitigate and protect again collision strings. So this would get back to questions four, five and six, kind of groups a little more naturally with that. So we're going to focus here on seven, potentially eight. I'm not opposed if people want to jump in and make notes about eight, although I'm going to focus primarily on seven to start with.

So we'll do that at first. Let met just pause there, see if anybody has any questions or comments about that sort of order and the path that we're going to take here. I'm not seeing any hands. Let me make sure I've got this scrolled properly so I can properly see hands. Okay, we're all set to go.

All right, there is some data that has been put here with respect to what are the possible ways in which we might identify a string as a collision string. Matt Thomas, one of the other co-chairs, at some point actually in the past here quite a while ago when we were first looking at some of this had pulled out some text that he had from some of the work that he’d done and put it here.

The entries that he put here and the text that he put here—and I'm assuming that everyone has looked at this and has read it—really focuses on actual data and actual elements that one might see. For example, looking at root server traffic, or even resolver traffic. So what are the kinds of things that you might get looking at DNS query data? What are the kinds of things that you could pull out that would provide you some indication that something is likely to manifest a collision string and therefore should at least be in this set of collision strings so that it needs to be more carefully evaluated as to whether or not it should be delegated.

To some extent, this is all fairly obvious stuff in the sense that this is kind of what the JAS report pulled out. There's a little more detail here about ways in which we might look at that data. But I think that we need to be very open minded about this question. In addition to DNS query data, is there anything else that might be interesting to go look at in determining whether or not something is potential collision string. It’s not just about looking at the DNS query data itself. Is there anything else that we can look at?

Oh, I should point out here, I guess, at the bottom of his list here, he does talk about certificates and legacy use. So there are a couple of examples down there at the bottom about other things besides just DNS query data which might provide some information that needs to be gathered as part of the board's decision process. we can imagine that an application will come forward, the board will probably look to the staff, I would imagine, if I were to invent some process here, to gather up whatever kinds of things we put here, whatever kinds of things we ultimately decide belong here. they'll probably look for staff to gather all that data and put it together as part of the package that gets submitted to the board when it's making its decision about whether or not to delegate.

So open question. I'm not seeing any hands go up, so I'm just kind of [inaudible] at the moment here. I think I've pretty well summarized what's here. the question is, do folks have any questions about the kind of stuff that's here? Do you want to expand on any of these descriptions? Do we want to be more specific about the kinds of questions that are asked? Is there anything else that you can think of that's not here that we might learn?

I think part of what I hope we will ultimately get as we get a chance to look at things like the name collision reports that we have gotten—there have been a few, certainly not a great many of them—and we'll be able to dig into those to see what they show. Maybe as part of looking at that, we'll be able to look at something, some other kinds of applications or data that might be interesting that we could add here. In fact, actually, I'll make a note here at the bottom of all this that suggests that we might learn of some criteria as we review existing reports of name collisions.

Jeff, go ahead, please.

JEFF NEUMAN: I don't know if this falls into this category, but at one point we were talking about potential—this would obviously be after applications are submitted, but at one point, we talked about potentially having ICANN “delegate” the strings just to collect certain data as to be able to assess whether if delegated to a third party, there would be some sort of risk. I don't know if that falls in this category or that falls into the mitigation, because I know this one says determining whether an undelegated string would be considered, and technically, that would be delegated even if it was ICANN delegating it to itself. But I don't know where that kind of thing would fit.

JAMES GALVIN: Thank you for that. That actually is where I was going to jump to when I was thinking about other things that we might put here.

We might also consider, as we get into looking at mitigation opportunities, perhaps something that we might consider is, as you say, let’s collect data and maybe we do—well, to make it concrete, right now we do controlled interruption and that seems to be a mechanism that we use to evaluate the presence of and the severity of collisions that actually might exist, might manifest if we don’t know about them in advance. It might be that as part of the review process, we want to take an opportunity. We might want to suggest that as something to be considered. And maybe there's some criteria around that. That’s something worth considering too.

I was going to ask Kim, if you can scroll to the bottom of question seven here, then you can see the note that I'm trying to type in this file at the same time. Folks can see that on the screen, and just looking at that, yeah, that’s perfect, thank you.

But I think that’s at least something that we need to think about here as we dig in later on and we get to look at the collisions that have happened and we review them. As we think about mitigation strategies, we want to add to this list, could ICANN delegate the collision string to gather some data about the collisions, or even just to see if collisions happen. Is that an option worth considering?

I can imagine something along the lines of maybe it’s not something you do all the time, but if something has significant collisions that already exist and already present in the data, that might be a criteria that’s useful in order to make a judgment or get some information to help influence the judgment of the severity of the collision so that you might consider whether or not you can delegate it anyway. That’s one of the questions that we ultimately have to answer: is it possible to get on and off the collision string list, and is it possible to be delegated even if you're on the collision string list?

Any other thoughts? Yeah, Rubens is just pointing out, we know that the public reports about collisions are limited. And even though in a technical sense, the small number of reports might be considered anecdotal, nonetheless, one of the things that we had always said in the project proposal in its original form is we really should dig into those. We should take the time to see what we can learn from what's there. I realize it'll end up being a bit of a tabletop exercise for us in this discussion group to try and extrapolate some principles out of that, because it is limited dataset, but nonetheless, we should do that and see what we can learn from it, whatever that might be, as limited is all of that is.

But thanks for this, Jeff. I tried to add that point down there at the bottom. Anything else from anyone that you want to suggest? Especially given this list I already have here. It was a fairly complete list to start with.

I think that these questions that we’re asking here about DNS query data, the thing which is not what's important here is this only talks about DNS query data, but it doesn’t actually call out—well, it does, actually. I'm sorry. So it does talk about query data from multiple sources, and because you are going to have root query data, there are going to be resolvers.

I suspect we should take into account in some way what if some other element to the infrastructure comes around. And I guess, could there be something, just as global public resolvers have come around since the 2012 round, who knows what the infrastructure might look like in the future? And it’s worth thinking about whether or not we want to offer some comments or observations about how you might look for data from other sources as they come along in the future. Danny, go ahead, please.

DANNY MCPHERSON: Hey Jim. Hello. I was just going to mention, in the sort of two-pager, whatever that document was Matt Thomas and I sent on root query data Verisign has, I think in the bottom of that, there's a paragraph that talks about probably six or eight things, QNAME minimization to local root to aggressive [incident] caching that have implications on visibility at the root, and anything that any sort of framework potentially relies on just query data could make problematic. Given that for example QNAME minimization is approaching 50% of all queries at the root, you don't see the full FQDN, and so you don't see things like DNS service discovery packets further down in a label base or something like that.

So I think that to your question, are there other considerations, even if we had a perfect framework for looking at just the queries in determining this, the efficacy of that is diminishing given a bunch of different things that are occurring. And that might be worth saying something about, I believe, if we get to this Work Stream or work area.

JAMES GALVIN: Thanks for this, Danny. You're absolutely right. I will point out that I think, yeah, it's just visible on the screen there too and I highlighted it. Hopefully, it should show up. The [impaired] observations is an entry that Matt Thomas had added here.

DANNY MCPHERSON: I totally missed that. There you go. Thank you.

JAMES GALVIN: Yeah. But that was also part of our technical gap brief that even this discussion group had written before as we considered the status of study two was observing that we really do need to see what kind of effect the [impaired] observation issue has on the ability to answer question seven. It's all well and good to think that these sort of obvious technical questions can be asked [with] DNS query data, and they were quite sensible in 2012 and they might still be sensible except for the [impaired] observations we are going to have to give some consideration to is what effect [inaudible].

And just to remind folks, this is one of the reasons why even if we end up just repeating the experiments that were done or the data collection that was done in 2012, part of our technical gap brief that we had written a few weeks ago was really all about observing that because the infrastructure is different now than it was then, it's important to look at those same experiments again and confirm and actually look and see what the data looks like now as compared to then, what effect have these impaired observations had in our ability to get data? Because if it's going to change our ability to be able to identify collision strings, that’s important to know and useful to know. That'll certainly affect the board’s ability to make a decision about whether or not a string can be delegated.

Okay, aside from there's coupe of things here that we've added, I don’t have any other things that jump out at me to look at, and I thought that this original list that we had here was actually a fairly complete list of the potential things to look at in DNS data and observing all the various sources of DNS data that is useful to grab onto and look at. So there are some additional points rather than just root server data, looking in some additional places to ask these kinds of questions and then pull all that together as part of our suggestions to the board on collecting information.

Okay, if we don’t have any other suggestions to add to it, then I'm going to suggest that we jump ahead to question eight here. We actually don’t have anything to start with in question eight to start at the moment. This is where things start to get a little bit interesting. This is about once I have something, what criteria, what kinds of things should we look at for whether something should be delegated? Or conversely, what kinds of things can we look for to demonstrate that a string which was in the collision string set could be removed off that set?

This is probably real meat of what we have to get into here. I see a hand up, so let me just pause there for a moment and call on Jeff. Go ahead, please.

JEFF NEUMAN: Thanks. I didn't mean to cut off the intro, I was thinking this relates to the topic we were talking about the last time, of how this could be a function of time. So there could be different criteria that we could apply now to keep it out of the entire new gTLD process versus criteria after application where the balance of harms may be more tilted towards and applicant that’s paid money and has gone through the process and done all this investment.

So I don’t think we’re going to come up a definitive set of criteria but more kind of a sliding scale and factors to consider. And that’ll also be as to whether it can be mitigated. If you know now, two years in advance, let’s say, of potential issues, you can do outreach, in theory. Maybe you can do outreach to the party that has the collisions and get them to fix their systems. So I think we’re not going to come up with anything static. But thanks.

JAMES GALVIN: Let me continue with my introduction too. I'll respond to your query here. maybe I'll call on Warren before I get into the rest of the introduction. But I tried to capture the point that you're making. Sure, you're right, I can imagine that the criteria might actually be structured and they might be more applicable at different points in time as a collision string or a potential collision string moves to being considered a string, not being considered a string, to being delegated, the application process. I just kind of gave three little bullets there about applications submitted, while it‘s being evaluated, after board delegation. There might be others. we don’t have to make this the complete list at the moment.

But sure, I can see that that might be something that we’ll need to be considered as we look at some of these criteria. It may be more applicable at different points of time than others. Warren, please go ahead.

WARREN KUMARI: There's also, as Jeff was saying, the possibility to mitigate some of this. There's also the privacy concerns or the lack of visibility that an applicant might have. If the majority of the queries for a potential string are coming from one organization, how is the applicant or whoever supported to find out about that, and how are they supposed to contact them and get visibility into who is making the queries? Currently, I don't know if there is a way. And then again, the [inaudible] making sure that whatever the criteria are are not gameable, like Coke causing queries for .pepsi to stop Pepsi being able to get their brand name.

JAMES GALVIN: Yes. So what I'm capturing here is, how does privacy of the data collected affect the ability to share it with the applicant who might want to propose mitigation options? Certainly, being able to mitigate any potential name collisions is something that we expect the application process to allow. We do have questions here where we have to look at the potential mitigation options. That’s up there in four, five and six. But I wanted to capture the point here in this discussion rather than trying to jump round. But yeah, so I captured your point. Thank you for that.

The rest of my introduction was really to—and we've kind of answered it actually ideally here in this case. I'm looking for kind of the concepts, the principles, the ideas that we want to evaluate the specifics as in I wanted to steer us away from the idea that, okay, in this discussion we’re going to decide right now that if root server data shows X amount of NXDOMAIN queries, then it’s automatically collision string. That’s not the kind of discussion that I want to be having here right now.

In fact, it’s not even clear yet in my mind that we even need to have a discussion about what that number might be. We might just talk about some of the criteria here and what all of that looks like and how we want to evaluate that and how we might propose that be evaluated. Maybe there is no explicit number, just something to think about.

Anyway, so we have the criteria up here. What other ways can we use to evaluate it? I think what we’re talking about here is a function of time is one thing that’s important. When we’re evaluating the data, can we share it or not? So whatever we collect in question seven, is there an option for us to be able to share that? Are there any limits to that? That might affect the ability to go forward in what matters and how it’s related.

We obviously need to consider—I will make a note here about the obvious, which is, are there limits or ranges that affect the interpretation of the data collected in question seven? That's sort of the obvious thing. I think there was obviously a lot of emphasis on that in the 2012 round. Okay, so I see some hands. Jeff, please go ahead.

JEFF NEUMAN: This is going to be another obvious thing, but what was the dividing line in the mind of—I don't remember if it was Interisle or JAS that originally classified .corp, .home and .mail as collision strings. I think .mail was the least of the three. So, what was the demarcation point between .mail and the next string, at least in the minds of whether it was Interisle or JAS or whoever?

But we should use that criteria, or at least use it to throw against the wall and see if that works. And Jeff's got his hand raised, so cool, I'll shut up and let the experts talk. Thanks.

JAMES GALVIN: Thanks. I actually don't remember where the line was. I remember that it was sort of obvious when looked at the data where to draw a line here, because you sort of get this long tail, this huge drop off and a long tail. But maybe we'll jump ahead here. jeff, you have your hand up.

JEFF SCHMIDT: Yeah. Thanks. And I think this is the right question for us to be discussing. It’s not a number, it’s a story, a criteria. The last call that I was on, two calls ago, we talked about this. Every collision scenario is a unique flower in and of itself, and you really need to dig in on a per-string basis before you can make a determination of whether something is “dangerous” or not.

It’s not just a number, there's more to it than that. So here's what we did, JAS, and our thinking and what led to our recommendation that .corp, .home and .mail not be delegated.

.corp was obvious in retrospect but not obvious when we look at it. In fact, other people including Interisle looked at it in advance and saw it was a large number but didn't understand why. We figured out why and then realized this is really bad. Only after very specific research on that very specific string.

.mail, again, a large number, different queries than what we were seeing for the other strings, but looking into the why, we found that it was hardcoded into a bunch of early Sendmail example configuration files.

So after understanding—and our opinion may have changed in the last six years, but back then, just looking at the queries, the volume, the geographic distribution, where they were coming from, we made the call that this was widespread and would be dangerous.

.home was several ISPs, same sort of situation where the numbers were large but it wasn’t just one particular geography or provider, but it was fairly widespread and we made the call that that would have been more dangerous or serious than others.

What's interesting is some of the ones that we didn't put on the list, [.fritz] being one of the most interesting. Anybody that’s looked at the data sees the .fritz and the [.fritzbox] numbers are fairly high, but the geography and the scenario in which it was occurring was actually fairly narrow. So it was limited to one ISP in one country with one set of hardware. We contacted those folks and learned that the hardware was being phased out.

So it was a very case-by-case situation, and at the end of the day, it came down to evaluating each case for broadness of application, geography, scenarios, who would likely be harmed, whether it was individual situations or larger enterprises or something that could be an infrastructure or an ISP scenario. That was why we separated out .mail for example. Home users weren’t the ones installing Sendmail, it was enterprises and ISPs that were more affected.

And based on the individual analysis, we made those calls. If anybody ever wants to put anything in the root again, I think you need to do the same thing, per string analysis o an case by case basis, understand exactly what it is, and find the unicorns if there are any.

JAMES GALVIN: That’s great, Jeff. Thanks very much. I tried to capture what you said, and made some notes there, but you created a careful reminder here. I'm going to do it this way. We really need to look back at the JAS analysis and review the criteria that were there. So thanks so much, you did a great job of summarizing all of that here.

I tried to capture it. I think the principle observation that you were making and how you did it even in your report was really, it’s not just a function of data, meaning the absolute numbers. There's a wider element to this. It really is a case by case thing and you kind of have to look at why and evaluate why those numbers are there. That’s an important part of the decision process. And we’ll have to give some thought to that too as we try to capture the generalities so that we can give some solid advice to the board on how to use that going forward.

“Or do we ignore?” Jeff Neuman says. “Do we remove unsanctioned use?” I'm sorry, I jumped into the middle of a chat thing. Never mind. Warren, go ahead, please.

WARREN KUMARI: Thank you. I do think that it’s important to note in here as well that whatever the criteria are, are things where gaming needs to be considered and taken into account. I do think that gaming is going to or has the potential to be incredibly harmful, both to people trying to apply for a string but also to the Internet as a whole.

If there's a new round started up and everybody starts to game the system by spoofing queries for strings that they don’t want to exist, we could end up with a substantial load of simply gaming queries. So it’s not only harm to potential applicants, it’s also harm to the underlying system itself.

JAMES GALVIN: Yeah. And we do know explicitly, just to call out one of the big points I think that you made there about gaming the system, we do have to be aware that whatever criteria we list here and we make some decisions about, we are going to have to consider how to provide advice or what advice we can provide so that the board can actually determine, or in some kind of best effort way, have some kind of indication of whether the negative criteria, if you will, the criteria that get you on the collision string list, were gamed, inflicted on the string. And that’s an important thing to watch out for, generally called gaming. So there are people that are going to do that, and we kind of have to look for that. We have to provide some ways to effectively evaluate it.

I think in general, that gets deeply into this “why” question that Jeff Schmidt was bringing up earlier in what JAS had done back in 2012. Even if you have raw data which would suggest that something is a potential collision string, you really do have to look at the source of that data, you really do have to give some consideration and some evaluation into why that data is present. That's critical to avoid gaming and things related to that. Jeff Neuman, go ahead, please.

JEFF NEUMAN: Yeah, and first thing is I'm just going to repeat what Danny said on the chat, because I think that that's important. Looking at it from a longevity perspective or how long these collisions or person, entity, whatever that was using this string, how long they have been using it. If it's something that they just started a month before the application period opens up, that may be very different than those that may have been using it for ten years or something like that. So I think that that's worth adding as part of the criteria.

And I don't really see this, but maybe it's also the ability to mitigate or reduce the harm. So I think it was Warren that basically said, it was either the last call or the call before that, a couple calls to the person or entity that was having issues and they resolved it right then and there even though there may have been a lot of queries to the root.

JEFF NEUMAN: I'm sorry, I have to ask a clarifying question. You started off by saying you wanted to refute what Danny said in the chat room—

JAMES GALVIN: No, repeat.

JEFF NEUMAN: Oh, repeat. Okay. Thank you.

JAMES GALVIN: In support.

JEFF NEUMAN: You seemed to be confirming what he said.

JAMES GALVIN: Yes.

JEFF NEUMAN: And I do want to point out with respect to what he had said, up here in question seven, we do actually have included in the data that would be collected suggested criteria. It's not just about traffic volume but it also would include some longitudinal trend that's actually listed there as an element in the first bullet under question seven. So [its use will be there.]

Although I did note here, under this, scale's an important fact to consider when evaluating data because you have to worry about gaming the system. And a longitudinal review of the data also matters here too. So this gets back to time, data also having a function of time associated with it. Was there data that was present before the application? Did it only start appearing after the application? That kind of thing. These are all things that have to be looked at when trying to judge the quality of the data that's influencing whether something becomes a collision string or not. Danny, go ahead, please.

DANNY MCPHERSON: Yeah, just wanted to make two quick comments. One was that, to the point of longitudinal analysis and Jeff Schmidt's point, I completely agree. I've seen a graph of traffic for about five or six years to A and J root for .mail or any subdomain thereof, and you can see for example a huge amount of 8.8.8.8 queries. I suspect there bay be Happy Eyeballs or something in IPv6. But that illustrates the changes over time that you can look at for given strings. It might be helpful, just to reinforce the longitudinal analysis point.

I also agree certainly when people stat looking at names and testing, it's super easy to send a query to the root for anything you want. So certainly influence things. So I think longitudinal analysis should inform that. the other thing I will say is that there are a lot of examples—and hopefully, we'll publish something related to this, and I've mentioned this already but we have been looking at the data and the top query domains that we see at the root. And instead of just talking about name collision mitigations, actually doing some outreach, and there are a lot of examples where a single e-mail has resulted in 50 million queries a day for example stopping at the root via notifying a security team at a big carrier, a big enterprise, that kind of thing. We hope to publish that soon, which is just like what we did with the .cba analysis that we published in 2014. And certainly, we talked to Warren about this and worked with him and a few other folks in the community on some aspects of this as well. Just to generally clean up some of the—what are the top new entrants or any anomalies at the root? And then are there cases where somebody wants to apply for, say, .brand?

And I think one of the biggest factors is sort of spatial analysis, and you have already captured this, is how many source IPs or ASNs, because if it’s a small number, it’s probably a specific configuration in that environment as Jeff said earlier, versus something really broad like .corp, .home and .mail for example. A really broad distribution is way harder to clean because outreach is just really difficult.

Anyway, we’ll provide some data on that, hopefully, before this working group wraps up. But we've already probably mitigated on the 10 or 12 strings at the root that were in the top 20 or so. We’re going to iterate through the top 100 that we see and do the outreach, just ask anyway. And probably to the tune of 500 million queries a day or something, which is a lot. And I think that that ought to be an option for people when they apply for something, to say, “Well, what amount of outreach would make this problem go away?” And if they want to fund that or do that, or ICANN does that or somebody else, then maybe that’s an option. Just wanted to put that out there.

JAMES GALVIN: That's great, Danny. Thanks very much. Certainly very much appreciate all the additional analysis that you and your team are doing and continue to do. I know that all of that is going to inform our work here substantially.

One of the reasons why I especially appreciate having Matt Thomas as a co-chair here, as part of our group—and I'm pretty sure Patrik would agree since he's really directly part of all of this and right in the middle of it. Thanks very much for that. Warren, go ahead, please.

WARREN KUMARI: Thank you. I had it up and then I took it down because Danny covered much of what I wanted to say about—it's also the [fan out] of the queries. One person doing a million queries is very different form a million people doing one query.

But I think there's also something which is even harder for applicants to see, and that's not just only the number of queries, but do the queries seem to look normal? And by that, I mean, is it a reasonable distribution of subdomains under the queries? Is there a reasonable distribution of TCP versus UDP? Did the answers seem to be cached? Etc.

And there's a large number of queries that end up at the root where the response never seems to be cached, so even though it's a large number of queries, they don't look normal and [inaudible] they're being created by regular nameservers. Whether that means that they're more or less concerning, I think, is still an open question. But there is a lot more than just the volume and the fan out, It’s also the type, the catchability, the UDP versus TCP distribution, geographics, potentially time of day or clustering. There are a lot more metrics. And unfortunately, a lot of it is stuff which is difficult for external parties to look at if they don’t have direct access to either recursive or root data.

JAMES GALVIN: Thanks for this, Warren. I agree with you. I made a note in the bottom of question eight, but I also wanted to remind people back up to question seven. One of the bullet items was unusual query attributes, which I think is something which takes into account—there's also other things like source address diversity, subdomain diversity and subdomain types. All of those fall into that category of what is the quality of the queries that are there. What do they really look like? And we’ll have to give some thought ourselves here to what it means to ask that question and what kinds of things we’re looking for. I'm not quite sure how to frame all that at the moment, but I think we at least for the moment have a sense of what we’re thinking about, and we’ll have to expand on that as part of our guidance to the board in its decision process. Warren.

WARREN KUMARI: Between Christmas and New Year, I got bored and made—I think I've deleted it now because I ran out of credits, but I made a machine learning model where I just tossed root data and said, “Classify this based upon the actual string, the number of queries, and the second level thing, and then show me anything that looks as though it doesn’t fit this expanded classification.” And that actually turned out to be relatively trivial and it showed a lot of interesting stuff. But then I got bored and deleted it in a fit of pique, but should not be hard to recreate. I'll look and see if I even still have it somewhere.

JAMES GALVIN: Okay. Certainly appreciate any insight that you have there that we can add to our list. That would be awesome. Thank you. That’s very helpful.

So Jeff Schmidt, you have your hand up. Go ahead, please.

JEFF SCHMIDT: Yeah, It’s been a long time, and so I don't know how many people on here actually remember or dug into our second report, the one that was 3000 pages long. But the whole last part of it from about page 40 on is actually the exact sort of pattern analysis that we’re talking about here. And for every string, we have a visual regular expression, kind of exhibition that shows patterns in the string. We break it down by query type and number of queries for every applied for string.

The patterns, as Warren mentioned, are fascinating when you start doing this sort of categorical analysis. But then you can see what's going on. I really like the visual regular expression representations that we put together in our report. It’s kind of a little bubble chart, because you can really see what's going on. And I think that sort of digging into strings individually is really important to figure out what's dangerous and what's not. But a lot of this work has been done. And for folks that haven't looked at our report in a while, I would encourage you to take a look at it, and maybe there's some inspirations on what we should do in the future.

JAMES GALVIN: Thanks for this, Jeff. You're absolutely right, and that’s a good reminder for us here to go back and pull that out as we really dig into this, we want to look at that previous work that was done. And it might be interesting to think about part of an application package might be to have some of that analysis done. So maybe part of the answer to this question for us to think about is, is some of that analysis done on everything that’s applied for? Is it only done under certain circumstances? We need to think about those kinds of questions as we consider how to answer this particular question and we’ll definitely want to do that.

JEFF SCHMIDT: Absolutely. Sorry, to throw in one more thing, we've got an XY binned scatter plot of source diversity versus SLD diversity. And that’s also fascinating, again, pursuant to what we've been talking about, the diversity of where queries are coming from and the diversity of what's being queried for in higher levels in every QNAME is really important in understanding whether something is an isolated incidence or a more broad incidence. And the way that we chose to visualize that are these binned scatter plots that the plot was two things against each other. Again, it’s just illustrative as we’re thinking about how to analyze these things.

JAMES GALVIN: Okay. Thank you. Excellent. And Warren is giving a link to a report in the chat room there saying that the global advisors report—yes, from 2015. Thanks for that, Warren. I think that that’s all part of the bibliography we already have, but we should certainly make sure. Sometimes it’s hard to remember the details of where stuff is or isn't, you see it so many times.

So let’s see here. On the one hand, I see a hand count of one at the top, but—oh, that’s because—

WARREN KUMARI: That was Brantly but it looks weird.

JAMES GALVIN: Yeah. I don't know if the attendees can type in the chat room and get their question in there and do that so we can look at it. Okay.

KIM CARLSON: I can give permission for attendees to speak, if you wish.

JAMES GALVIN: I'm not opposed at the moment, I guess. So, sure. if you want to give Brantly the opportunity to talk. Brantly, if you want to ask your question, go ahead, please.

BRANTLY MILLEGAN: Hi. My question actually was about question seven. You had the criteria legacy, and I'm wondering, is that better to find somewhere else, what you mean by that? How do you distinguish between alternative root projects versus new technological experimentation versus commercial squatting? Do you have more information about that?

JAMES GALVIN: Thanks for the question, Brantly. Actually, in fact, the point of that bullet item there is just calling out that there might have been legacy usage and we should consider what that is and how we want to evaluate that. So it’s really just a criterion to look at in looking at these things. The questions that you're asking are exactly the questions that we’ll have to dig into as we get into the analysis here of what we’re going to do with that information. This is just calling out the fact that we should look for any legacy usage and then we have to consider what to do with that information.

BRANTLY MILLEGAN: Okay. Where will that further analysis for that point be happening?

JAMES GALVIN: In our discussions. We have to get there. We’re not even there yet.

BRANTLY MILLEGAN: Okay.

JAMES GALVIN: So that’s future work. So, thank you for the question. Okay, [inaudible] look at the time, we’re actually at five minutes to the hour. We've kind of come to the end here, which is interesting. I'm thinking that rather than jumping ahead here to the next question, I think we've now nicely done seven and eight today at the moment. Certainly, these are—as always, this is not final and not definitive. We’re collecting discussion points and things and we can always add to them. And then of course, as we get into some of the details, we’ll get ourselves to a place where we might even start removing things at some point.

But I think now that we have started by looking at the first few questions where we were just trying to understand name collisions and their harm and queries and what that kind of stuff means, we've now talked about elements of what could be a collision string and what could not be.

So I think at this point, what I would suggest is we would pick up next week with question four. Let's go back now and let's start talking about mitigation a little bit and get some real notes on actions that might mitigate harm, ways that that can be done. We'll start with question four with that, and then question five and six also get into what happens when mitigation is present or could be present. So we'll then do that.

And at that point, we will have finished our full complement of questions here and we'll take a step back and see what our next steps can be. So we'll start next week with question four and pick up with that. Let me do a quick check here for Any Other Business or any other last comments from anyone on any particular topic or a thing that we had here.

Not seeing any hands, I'll just do a quick reminder to people that this group will not meet the week of ICANN 68. In case you haven't been tracking detail, ICANN 68 is going to be a virtual meeting and as a virtual meeting, it will be held according to Kuala Lumpur since that's the time that the actual physical meeting would have been held in. So they'll hold it during that time period, which will offset it for most people by quite a bit as compared to any ICANN meetings.

But we will not meet because this meeting slot will turn out to be pretty much in the middle of the night. Well, it turned out to be in the middle of that slot. Yeah, Kuala Lumpur time. And so for those who are going to be attending the ICANN meeting, it would be very challenging to attend this meeting in this time slot. And we also will not meet during this week, so we’re not going to have our full day before the ICANN meeting that we had targeted that we will ordinarily have.

So I'm sorry, that was a longwinded way of saying we’re not meeting the week of ICANN 68 which is that third week of June. But our next meeting will be next week on June 3rd. Any Other Business, last chance. Hearing no words, seeing no hands, thanks, everyone. Appreciate your time. We’re adjourned.

[END OF TRANSCRIPTION]