BARCELONA – Open Data Initiative
Tuesday, October 23, 2018 – 15:15 to 16:45 CEST
ICANN63 | Barcelona, Spain

CATHY PETERSEN: …for a few more people. Thanks.

MATT LARSON: Hello, everyone. Why don't we keep going? My name is Matt Larson. I'm VP of Research in the Office of the CTO and I'm here to give you an update on open data at ICANN.

So, here's the progress that we've made since Panama. We've completed an RFP to get a Software as a Service open data platform and went through the entire standard ICANN procurement process to run a very thorough RFP, considered multiple possible vendors and wound up with OpenDataSoft. I'm actually going to give you a brief demo here of the platform to give you an idea for what it's like and what it's capable of.

We've also completed the initial version of what's called the Data Asset Inventory. This is a list of all the data sets at ICANN, not necessarily all the data sets that we'll be able to be published but the idea is to do an inventory and find all the data so that we know where we're starting from and then from there figure out what can be published, what would need changes to be published, and so on.

That was a part of our public comment. So we showed that to the community and got feedback on if the list of data sets was complete, if

anyone thought anything was missing. And we also asked for the community's priority in terms of what data sets you'd like to see in what order, since depending on the nature of the data set, there might be more or less work involved in getting that data set published to the open data platform. So we want to prioritize according to community priorities as well as the level of difficulty so that we get the data published in an order that works for everyone.

So the open data platform is going to include what we're calling a pipeline for automated data ingestion. This pipeline will extract, transform and load the data sets from the various systems of record in ICANN, whether it be databases or Salesforce. In the research department in OCTO, we have some data sets. There could be all different kinds of systems of record and different pipelines will transform those data sets and send them to a staging server. And the idea is that we only publish clean data, ready to go, [it's had] any retractions/redactions necessary. So it's only public facing data. There's nothing private on the staging server. And then the open data platform from OpenDataSoft reaches in and pulls the data.

So now let me give you a demo. This is a live demo with everything that involves, so we just don't know what will happen. I'm hopeful. Alright. So this is the open data platform. This is OpenDataSoft. This is the production version of the platform that will be available to the public eventually. Right now we're in a testing/demo mode so it's not open as of yet but you can see it's branded to look just like the ICANN website. This follows all the official branding guidelines.

What we're looking here at the data tab that showing the various data sets right now. We've loaded 20 data set just again for demo purposes. We don't have any of the official pipelines working yet. We have an architecture for the pipeline that we suspect will be transforming much of the data that will go to open data. But at this point, the data that's on here are sample data sets.

So let me start with one from the Identifier Technology's Health Initiative (ITHI). This project actually intersects very nicely with open data in that it is a project that is if you're not familiar with it, over time going to be recording metrics related to the health of the unique identifier systems that ICANN helps coordinate and the ideas that over time we can look at these metrics and see how they're changing and evaluate the health of the identifiers. So with all these ITHI metrics, we need a place to publish them and this is an example of an ideal sort of data that can then be published in the open data platform.

So let's look at this particular metric right here. So when I click on it, we look at the particular data set. So here's the table, and of course you look at that and you go, "What is that? I have no idea." So you can go to the information tab and get a description, both a written description here as well as some standard tags of metadata describing this data set. So all of this particular metric – let's make this just a little bigger. All of this particular metric relates to the usage of the DNS root servers.

So as I read this, for example, the first part of this metric is the fraction of the request to the root for non-existent TLDs. So let's look at that. If

I go back to this table full of data and then I filter on just metric 3.1, it's actually a relatively small data set. It's only got nine rows here but it's one per month. I can go to the analyze tab here and I can make a graph.

So if I plot on the X axis data and break it down by month, and in this case we want – actually, I believe it's the same for this data set. Yes. So here we have over time, we can see the fraction of queries to the DNS root – at least to the ICANN Managed Root Server, L-root – how many are not valid TLDs or non-existent TLDs. So if we go back here, what else is there interesting in this data set? Well, along the lines of non-existent TLDs, we could look at the request for the most common non-existent TLDs that are queried. So that's 3.3.2. So if I go back to the table of data and clear this and then go down to 3.3.2, here's all the data, but again this might be helpful to analyze. In this case, I would want it plotted by TLD and here then we can see the most popular non-existent TLDs over the entire – I beg your pardon. We need to average. Let's average all the data points here. So that would be across the time span that I have which is 2018 but I could say, "What if I want to look for let's say one month?" How about June of 2018? Then I can redraw the graph accordingly.

I could also decide, what about .home? This is clearly the largest one. How is that doing over time? So I could say, "On the X axis, I want TLDs but I want let's say just .home." Let me start and clear the dates. Let's do this for all dates and then let's do just .home. I have TLD on the X axis but I want to break this down by date. And if I change the type of

graph to this then I can see over time how the fraction of non-existent queries for .home is at the root.

So this is just an example of how given the data set, you can do some quick graphs to analyze. Now, of course you can also then, using just HTML and CSS, you can tell a story. So this is an example of what we've done. We've created this page that has pre-set graphs that shows this a little more clearly. This could be a standalone web page on the platform that would show again an explanation of what this metric is. And then this is one of the visualizations I already showed you the percent of request to the root for non-existent TLDs.

Here's another one of the metrics in this data set. This is the fraction of unnecessary request. So if we look at all of the traffic coming in, taking into account caching … this is a calculation of all the queries that we think shouldn't have been sent to the root here because they were invalid because of a non-existent domain name or unnecessary because someone was apparently not caching accordingly. But the point is it's in the data set and we can graph it.

We also as part of this data set categorized the kinds of non-existing TLDs in here. Here are the four categories and then you can see how they're plotted over time by category. So this is an example of how you can combine visualizations and text to tell a little story. But the other thing you can do is you could go – this is interesting but I have an idea for Python script. I want to slice and dice this data myself. Well, you can go to the export tab and you can say, "I'd like to just export that as a CSV file," and then lo and behold, there's my CSV file, and

that's the entire data set for me to use as I like. Because that is the idea behind open data after all.

Alright. Let me show you one more thing. The home view of this platform will not be all the data sets. There actually is a homepage that we can curate to be a little more friendly to get people started. These are examples of those data stories like the one I was just showing you.

Let's look at a visualization corresponding to another data set. This is an OCTO data set. Since the end of 2011, we've been looking at gTLD zone files once a week to see how many delegations there are and then how many DNSSEC signed delegations there are. So if we look – this is the entire data set and every gTLD for all time. You can see the total number of all gTLD delegations back here, at the start of the data set, it was 134.5 million and then today is as you can see almost 192 million and that's the timeline. You can see the growth curve of the delegations overall and you can see there's a steeper curve here for signed delegations. And then, a shallower curve for non-signed delegations.

Then we can do things like filter, for example. Let's look at say .bank. .bank is an example of a TLD where everything has to be signed. So you can see here the number of delegations total and the signed delegations are the same. And up here we have boxes that you can – and this is a real world demo, not quite working as we've intended here but you'd supposed to be able to click here and change the boxes up here, and that's not quite working yet. But this gives you the idea of

the kind of interactive visualizations that you can do on the data that we can make part of the platform to make it easier to get people started so they don't just come in and see something just list after list of data sets.

So we've made pretty good progress on this and we definitely have a really good framework. We're very happy with how the RFP went. We're very happy with how OpenDataSoft responded and the future set of the platform. And we think we've got a really good start on that

That is what I wanted to show for the demo. In terms of next steps, we have some policy documents in draft. We have a draft Open Data Policy that's going to talk about what the ICANN org will do and what you can expect related to the implementing policies and practices for open data. We also have a Data Governance Model underway and we're developing the process just to determine how we're going to prioritize data sets for publication. With that, I would like to turn the final slide over to Susanna.

SUSANNA BENNETT: Hi, everyone. I'm Susanna Bennett, Chief Operating Officer. This Open Data Initiative is going to be transitioned to what we call Open Data Program. What that means is that we are going to operationalize the initiative, i.e. internally we've worked with many teams especially Engineering and IT Team and the Legal Team and Communications Team and many of the functional teams who have this data. And the work on the prioritization of building the data source. Feeding the data source into open data so we can share with the community. Of

course, we'll also work closely with the community to understand the party needs and make sure that we will get the data set or load it for the community. And also whole sessions to review them and to share, make sure that we answer the questions to make useful for the community. Thank you.

MATT LARSON: Okay. Thank you, Susanna, for the update. So that is the end of the prepared portion of the presentation. Our update for you today is pretty brief. We'd be happy to take any questions that anyone has. We have a mic up front here. Any questions in the room?

ROLAND LAPLANTE: Hi, my name is Roland LaPlante. I'm with Afilias. I've been tracking this all along. You guys have made a huge amount of progress. This is awesome.

I have a question, two comments. The question is: you talked at the beginning about having only clean data in this and I'm wondering exactly what that means.

MATT LARSON: Well, that means data that's been stripped of PII or otherwise legally can't be exported or data that would be sensitive that we wouldn't want to explore. So, for example, sensitive personal data or something like that. The idea is that only the data that gets to the staging server is data that ICANN intends and is comfortable with to publish.

ROLAND LAPLANTE: Okay. I've been looking at the Registry Operator Reports and your definition is completely different from what my question is. My question is really about the accuracy of the data. So I'm looking at the Registry Operator Reports and I'm looking at various aspects of it. And it's clear for me that somebody's reports have no data in some fields. They have anomalous data in some fields. I'm wondering if that's going to be – I think of that as cleaning it up but are you going to do anything with things like that that are obviously errors?

MATT LARSON: Well, one of the advantages of open data is that we can make all this data very accessible to the community so that people can analyze it and give us feedback like that that says, "Hey, we see issues with this particular data set."

UNIDENTIFIED MALE: Just to add, it wouldn't be ICANN org's responsibility to clean up that data. We can make anomalies or lack of data aware to the contributors of that data and it might be Contractual Compliance's job to go and slap people upside the head if they're not providing that data correctly. But within the context of open data, the open data platforms role is to actually make that data available not to modify it that way.

ROLAND LAPLANTE: I think that's a good policy. But if we identify particular data points or data sets that look inaccurate and the pieces of it that look inaccurate, is it possible to get that stuff corrected?

UNIDENTIFIED MALE: Again, it would probably be something to raise with Compliance like the Registry Reports and that sort of stuff or the contributor of the data, identifying where that data source was from and saying, "If it happened to be…" and trying to identify whether it's a bug in the data pipeline or if it's a bug in the original source of the data.

ROLAND LAPLANTE: Okay, great. Thanks. My two comments are: can somebody from marketing look at the labels on the data sets that you're presenting? Because as a community, I don't know if we want to talk about WHOIS inaccuracy. We might want to talk about WHOIS accuracy. I don't want to give anybody any more ammunition than they already have. So, that's one comment.

The second comment is that the bottom of each chart, whenever data is actually presented out of this, can we get the name of the data source that was used and maybe a link to it so that if we're looking at something that we want to understand better what the underlying data set is, we can go back to the original source and get an answer to the question.

MATT LARSON: Each of the data sets in the platform has a whole bunch of metadata associated with it including where it came from, who the owner is.

ROLAND LAPLANTE: So the viewer will be able to figure that out?

MATT LARSON: That's the idea, yes.

ROLAND LAPLANTE: Okay. Perfect.

UNIDENTIFIED MALE: Thank you very much. Good job.

MICHAEL KARANICOLAS: Hi, my name is Michael Karanicolas. I'm with the NCUC and I run a NGO corporation coalition that works in transparency issues. Congratulations on this. It looks really great. I have a few questions.

The first is whether or not – are you planning on releasing this as a beta for the community to experiment with, play around with and to offer feedback on?

MATT LARSON: Since this is transitioning over to becoming Open Data Program, I will defer to Susanna on that question.

SUSANNA BENNETT: Thank you, Matt. Great idea. Definitely, if that's workable together with the OCTO Team, we can make it happen. Thank you.

MICHAEL KARANICOLAS: Particularly because different types of users will have very different approaches to the data, so an academic researcher might approach it very differently from a commercial researcher and so it's important to get those different perspectives.

I also wanted to ask just as an idea and maybe building off what the previous speaker said, I think it's very important to be neutral in how you present the data and to aim for accuracy as opposed to spinning it, but one thing that I know that certain Open Data Programs do is they allow people to comment on data sets or they create a functionality to chat related to the data set that's there. So, I'm not sure if there's challenges to implementing that but that might be an avenue towards allowing people to – if they see an issue with a particular data or if they want to comment on a particular data, fostering that kind of conversation can be beneficial.

It also can be useful to solicit community feedback in terms of prioritization which I know you've done to a certain degree in the last public comment period. I'll put in my personal pitch for any financial information or contracting information if and when that becomes available. I know there's commercial sensitivities around that but to

the extent that it can be offered I think that would be also a good potential.

MATT LARSON: Yeah. My anticipation is that Open Data Program moving forward will be sort of a living program that'll continue to evolve as community requirements come up as suggestions arise. The idea of having a forum associated with it from a high level makes sense. We obviously don't understand the implementation details associated with it. I suspect that as with anything in ICANN implementation, it will be challenging. But the idea of having community involvement and prioritization and the descriptions and formatting of the data, I think makes a lot of sense. Thank you.

MICHAEL KARANICOLAS: Finally, I noticed across the top in your demo that there were a bunch of different languages. Can I ask about what you envision? Is it going to be total functionality across the different languages so everything will be translated or are there going to be degrees? Because I know that that can be very challenging. I'm from Canada, I know the Canadian government struggles with language, challenges in getting things out in both languages. ICANN is working with a lot of different languages. What are you envisioning for that approach?

MATT LARSON: I don't think we know the answer to that completely yet. The languages you saw at the top were an artifact of this being branded in

following the [inaudible] guidelines exactly at this point. That's a standard thing that's part of the guidelines. So, I honestly don't know the answer to that yet.

MICHAEL KARANICOLAS: Alright. Thank you.

MATT LARSON: Okay. I know we have some questions from the chat room. So, Cathy, if you could please read them off?

CATHY PETERSEN: We have several questions in the chat room. First one from Chokri Ben Romdhane. He's IT Senior Analyst at CNUDST Tunisia, ISOC Board member. His first question: "Why have you chosen to adopt the project Open Data Metadata Schema version 1.0?

MATT LARSON: I would like to ask Jay Daley to help answer that question. Within OCTO, we simply didn't have the expertise that we needed nor the resources to implement an Open Data Program, so we engaged to excellent subject matter experts to help us and they've done great work for us. One of them is Jay Daley, and so could I ask Jay if you could – at the risk of putting you on the spot – take a stab at that please?

JAY DALEY: Hello, this is Jay Daley, introduced as by Matt. There are a number of open data metadata standards and they vary in their complexity. In particular, they vary in the complexity of other standards that they need to reference in order to produce a complete set of data. Because our industry is a relatively immature industry compared to other industries, how do we choose one of the more complex types of metadata that required multiple different references, we would've struggled to find those references.

So, for our needs, the project Open Data Standard which was used by the U.S. government very effectively when they still had the strong Open Data Program, fit very well and it's simplified the issue of references to other areas. It may be that at some time in the future, possibly some years away, a more complex standard is needed but what I would imagine is that standards within the domain name industry would develop such as the work of the registry/registrar data group taking place in Europe. That would then provide additional taxonomies and other standard metadata sets that could then be referenced by the metadata standards we choose. That's a very technical answer but that's the main reason.

MATT LARSON: Okay. Thank you, Jay. I think we still have some more.

CATHY PETERSEN: Jay, you may want to stay over there. We have several more questions for you.

Second question also from Chokri: “Did the platform expose ICANN data sets for harvesting? If so, in which format?”

MATT LARSON: I believe that question is asking, “Can one export data from the platform?” That’s my interpretation of it. And the answer is yes, you can in CSV format, in JSON format. I believe there may be other one I’m not … Excel. They're not recalling. Yes, the idea is the data is indeed open. And as I showed in the demo, you can get the data out of the platform to slice and dice with your own tools and ideas if you like.

CATHY PETERSEN: Another question online from Chokri: “Will the community be engaged in the policy development process?”

MATT LARSON: I would say yes. We’ve already had the community involved this part of the public comment to help give us feedback as I said on the completeness of the Data Asset Inventory as well as on helping us prioritize the data sets. From what I understand of Open Data Programs, they're most successful when the consumers of the data are actively involved and it’s not just pushing data to the community but also hearing the community’s feedback. So I think as much as that’s a general question, I’ll give the general answer of yes. Susanna, any further?

SUSANNA BENNETT: Yes, definitely, Matt. This is for the community and we need to listen to the community to the right thing going forward, yes.

CATHY PETERSEN: We also have a comment online from Judith Hellerstein: "At-Large has developed its own open data tool, the Stakeholder tool which allows anyone to identify without are our stakeholders in a country and in a region and what are the gaps in a stakeholder representation in a country and in a region. The direct link to the Stakeholder tool can be found at (she put the link) http://bitly.com/ICANNST."

MATT LARSON: So there are no more questions in the Adobe Room. Are there any more questions in the room here?

SARAH INGLE: Hi, my name is Sarah Ingle. I'm one of the NextGen ambassadors and a member of NCUC. I'm just wondering if you can elaborate a little bit more on the Data Governance Model and what that might look like as well as what the role of the community might be in participating in that model? Thank you.

MATT LARSON: Well, I don't know how much we're able to elaborate now, honestly. Some of these processes are, frankly, under development. Also, as we've been advised by our subject matter experts who have done open data for other organizations, ICANN turns out to be unique

compared to some other organizations that they've done. So, to a certain extent, we're treading new ground here as they're helping us develop some of these policies. So I don't think I have a lot of specifics I can offer you at this point.

SUSANNA BENNETT: Thanks, Matt. Good question. Thank you. The operation plans to of course involve the community. How do we involve the community? We talked about advisory group for the community, so we start thinking through – help us to form an effective community advisory group. Thank you.

MATT LARSON: Any other questions in the room?

JUSTINE CHEW: Hi, this is Justine Chew from At-Large. Could you just remind us on the timeline of what happens now between the production and moving the ODI into a data program in terms of when it will be made available to the community to have a play around with and moving forward from there? Thank you.

SUSANNA BENNETT: Thank you. We're working on that plan. We have a draft plan right now reviewing with the team. As you know, there's quite a bit to it, so we want to make sure that we wrote all correctly and not disappoint the

community. So, it's in progress. At certain stage, we'll definitely share with the community. Thank you.

MATT LARSON: Okay. I'm not seeing … oh, wait. Another question in the room.

CATHY PETERSEN: It's actually a comment. Judith Hellerstein from ALAC just shared a link which everybody can see is pretty long so I'm not going to read it out loud, but this recording will be posted to the public schedule within a week or so. "Dev Anand Teelucksingh presented this tool today and the presentation can be found…" and the link is in the chat room for this session.

MATT LARSON: Alright. Seeing no further questions or comments, I'd like to thank everyone for coming. We'll give you some time back in your day. Thank you very much.

[END OF TRANSCRIPTION]