WHOIS Registrant Identification Study Webinar
TRANSCRIPTION
6th March 2013 19:00 UTC

Note: The following is the output of transcribing from an audio. Although the transcription is largely accurate, in some cases it is incomplete or inaccurate due to inaudible passages or transcription errors. It is posted as an aid to understanding the proceedings at the meeting, but should not be treated as an authoritative record.

Coordinator:

Coordinator: I'd like to remind all participants this conference is being recorded. If you have any objections you may disconnect at this time. You may begin.

Barbara Roseman: Thank you. This is Barbara Roseman. And I'm currently the coordinator of the Whois Studies Project. These studies were undertaken by the GNSO a few years ago in order to provide more factual data about the Whois, how it's used and what kind of data is available through the Whois.

The GNSO asked that these studies be done through ICANN. And we have chosen several different research groups to work on the various studies. The Whois Registrant Identification Study is being handled by the National Opinion Research Center of the University of Chicago. And NORC has been working on this for about 18-20 months I believe.

The report we're getting today is a summary of the report - the draft report that's been put on the Website for public comment. And Edward Mulrow, who is the head of the project, is going to be delivering the presentation and he has two of his project participants online with him also to answer questions as we move through - or I'm sorry, at the end of the presentation.

If you have questions that are not answered today or that you think of after the presentation is done please feel free to put them in the public comment section. We will be reviewing the questions there for additional inclusion in the final report and for clarification.

So, Ed, if I can ask you to start. I will remind everybody that the lines are currently muted and that we will open them at the end of the presentation to allow for questions. Thank you.

Edward Mulrow: Hi. Thank you, Barbara. I'll start here by just going through an outline of what we'll talk about during this Webinar. We'll start with the goals of the project.

We'll move from there into a brief description of how we went about sampling the domains that we looked at in this study and then we'll talk about how we collected data from the domains and how we coded that - the data from the various sources which were from the Whois records and then downloaded content for domain users and the domain content.

We'll also, as we go through, provide some answers to some specific questions that the GAC had posed as part of the study. And we'll end things by giving a brief overview of lessons learned while we conducted the study.

So the goals of the project, one was to do an exploratory examination of Whois data for a representative sample of the top five ICANN gTLDs. We wanted to understand registrants and the domain users, the types of entities that use the services and the kind of activities for the domains.

Our main focused areas were registrants and whether - what type of entity they were be it natural persons, legal persons or privacy proxy services. We also looked - were focusing on domain users trying to distinguish them between natural persons and legal persons. And another key focus was the type - if there was possibly commercial activity taking place at the domain.

There were some other things going on but these are the three focus areas that we'll have for today.

Now the Government Advisory Committee, the GAC, had posed these questions, which we will provide some answers to as we go through the information.

One was, what is the percentage of registrants that are natural versus legal persons? What is the percentage of domain name uses that are commercial versus noncommercial? What is the relative percentage of privacy proxy use among legal person users? And what is the relative percentage of privacy proxy use among domains with commercial use?

All right so to answer those questions and also look at other relationships between users and registrants we selected a sample of 1600 gTLDs from the five most common ones that are coordinated by ICANN.

As of June 2011 the top five gTLDs represented 98.5% of all of the domains coordinated by ICANN. And so those are DotCom, DotNet, DotOrg, DotInfo and DotBiz.

The sample size of 1600 was chosen to have good statistical properties but - and more so to the point for the types of estimates we're doing here, which are generally going to be percentages of - within various types of groups, in general a sample of 400 will provide estimates with good statistical properties and sort of a worst case scenario.

And so we knew that we would be wanting to concentrate on subgroups of domains with various properties. From a prior study we knew that there were approximately 25% of domains that were registered by privacy proxy services so we took - since that would be roughly 1/4 of the population we multiplied the 400 by 4 to get our sample size of 1600.

And as we'll see as we go through this we get reasonably good estimates for most of the subgroups that we're interested in. There are some where we don't find enough of them to do much further.

In the data collection part we constructed what we call the NORC bot, which is a multithreaded application using the Python programming language. It was an automated tool that would gather Whois data, publicly accessible http, https and ftp files as well as response codes from DNS blacklists and white lists.

This information is not static and can easily change over a short period of time so we attempted to do this in a simultaneous fashion or as near simultaneous as we could.

To get the Whois data in good formatted records we used the Whois ATI service. While that service was good for the study it did not always return complete Whois records.

So ICANN staff ran a separate Whois extraction process in parallel, sort of on the same day that we ran the NORC bot, and then these two Whois extractions were merged and compared. And it gave us a fairly complete set of records at least from the Whois perspective. This data collection was completed in March of 2012.

In terms of extracting the Web and ftp content only the www and ww2 subdomains were searched. Domains may have content on other subdomains and no attempt was made to look for that content.

We also put a download quota of 100 megabytes to ensure that extremely large files weren't downloaded from some domains. That restriction turned out to be not too strict.

After all that then we had a set of information that we would then code into variables. We had three broad classes of variables; that's the - what we called the Whois variables. Those were coded just on the Whois information. Domain users is information or variables related to the beneficial user of the domain and that's coded on downloaded content.

And then there's also domain content and variables where that concentrates on the activities that might be taking place at the domain. Once again that's based on downloaded content.

I'm going to do a little bit closer look at one variable from each type so apparent registrant type from the Whois category, the apparent domain user type from the domain user category and potentially commercial activity from domain content.

I do have some slides at the end of this presentation which may - which address some of these other areas so if there is interest in looking more closely at something like allegedly illegal or harmful activities I do have some slides that if you just ask a question about it we can talk a little bit more about it.

Okay so let's take a look at the apparent registrant type. Here we were looking to categorize the entity that registered the domain into one of the three categories, that would be natural person, legal person or privacy proxy service.

So for a natural person we tried to see if the Whois data, mainly the registrant name and the registrant organization, if that appeared to identify a real living individual. That's opposed to, for legal persons, it appeared to identify a company, business, partnership, any type of organized group of people that were not privacy proxy service providers. In this we included multiple domain name holders which we identified using reverse Whois email accounts.

For privacy proxy services we constructed a list of known providers. We started using a list from the study on the prevalence of domain names registered using a privacy or proxy service. And we added to that list as best we could.

So if the registrant information was found on our list of privacy proxy services then we coded that as a - that the registrant for that domain as a privacy proxy service.

Throughout all this there were some records that we could not make a determination on; we call those unclassified or unknown records. This would include when the data is completely missing or if we know that the registrant name or organization is just patently false or it was incomplete. And it could also include domains pending reactivation or deletion.

The next slide gives us a summary of what we found once we finished categorizing the apparent registrant type. And it also provides an answer to one of the GAC questions which is what is the percentage of registrants that are natural persons versus legal persons?

So we see that in our representative sample of 1600 domains 39% were - we classified as legal persons, another 33% as natural persons and approximately 20% were privacy proxy services. And that left 8% where we could not make a determination.

Okay let's move on now to the apparent domain user type. This would be coded based on domain content. So once again we were trying to distinguish between the beneficial user being a natural person or legal person; same basic definitions for those two it's just based on domain content.

There were a number of times where we could not make that determination but the reasons for that - for those situations were a little bit different. In some situation the domain had no usable online content so that means there is

either no content available or there was minimal html code that was insufficient to determine a user type.

Similar to that but different was - were the parked domains. These domains - the domain landing page had minimal html content but that was consistent with typical domain parking content.

And as we analyzed the data especially in an area sort of like potentially commercial activity we see that there is a difference between parked domains and those we said had no usable online content. The parked domains tended to have a larger proportion of potentially commercial activity taking place.

Even after taking care of that there were still some domains left over where there was available content but we could not determine the user type as a natural or a legal person.

The next slide here gives a breakdown of what we found in our sample of 1600 domains. The 37% of the users were legal persons, another 20% had no usable online content, 21% were - we determined to be parked domains, 12% of all the domains we looked at we could not determine a user type so it was an unknown user type and 5% we determined to be natural persons.

Now on the legal person user side, as I said, there's 37% in our sample. That is - that was actually 586 domains in our sample we determined as legal person users. And now we can answer another one of the GAC questions, which is what is the relative percentage of privacy proxy use among legal person users?

So we see here the breakdown just for the 586 domains that we classified as having legal person users and we see that 15% were registered using a privacy proxy service. This is a little bit lower than the overall percentage of 20% throughout the whole sample. So there's a slight shrinkage here. It's borderline whether that is a statistically significant difference. But

nevertheless there were less privacy proxy registrants within the legal person user category.

As we look at other categories the parked domains is where there was an increase in privacy proxy registrants. In our report you would see that in Exhibit 16 which we won't show here but it - the information is available in our report.

Now let's give a look briefly at what we did about potentially commercial activity. Here we attempted to categorize all observed monetary activities that in some countries might be legally considered as commercial. So things like ecommerce, collection of membership dues, promotional material, banner ads and pay per click ads are what we tried to - we looked for at the domains in our sample.

Next slide shows a breakdown of what we found. The five main categories, promotional content, pay per click ads, banner ads, ecommerce and membership dues, are there as well as within some of those like promotional content we - there were three different types of things we looked for. Note that a domain could show evidence of more than one of these activities at a time so they're not mutually exclusive.

We move on this gives us information now to answer a third - the third of the GAC questions which is, what is the percentage of domain name uses that are commercial versus noncommercial?

So when we include pay per click ads as potentially commercial activity we found at least one of the five activities detected in 905 of the 1600 sample domains so that's 58% of the domains in our sample. If we don't consider pay per click ads as potentially commercial activity then the number of domains with at least one of the activities dropped to 717 or 45%.

Now if we restrict ourselves to the 905 sample domains with detected potentially commercial activity - and that includes the pay per click ads - we can divide that up - divide that by the registrant type and we find an answer to the fourth of the GAC questions, which is what is the relative percentage of privacy proxy use among domains with commercial use.

And here we see that 23% of domains with commercial use were registered by a privacy proxy service. And that is slightly above the overall percentage of 20%. And again that is not a statistically significant difference so it's marginally higher than what we expected.

Okay so we'll conclude this first part of the Webinar then with going over a summary of the lessons we learned. Aside from finding out the relationships between our coded variables we also found out some things we'd like to share and the difficulties in collecting this data.

So collecting it in a nearly simultaneous manner is difficult but with a good multithreaded application such as the NORC bot it is a feasible task. Our summary report does contain summary information on NORC bot features in particular I believe in Section 4 of that report, which is the lessons learned section, we go into a little bit more detail on the various aspects of how we collected the Whois data, how we extracted domain content and how we searched the blacklists.

There were some subjective challenges that came about in our data coding. Things are not as clear cut as we'd always like to think they are when looking at something like data that is out on the Internet.

If, in attempting to impose standard codes on a (unintelligible) varieties of unique Websites it was not always easy to make a determination between some of these categories. For example very hard to distinguish, say, between parked domains versus domains that are there for reselling.

In terms of the variables we looked at the domain user relationship to the registrant was difficult to determine. Fifty-five percent of the domains we looked at we could not determine what the user/registrant relationship was.

However that is highly correlated with the domains without content. So if a domain had no usable content we could not really classify the user. In that situation if we couldn't classify the user we could not really determine a relationship with the registrant.

Another variable that was very difficult to discern was business structure. We had 65% of our domains classified as unknown. Originally we thought that this type of variable might provide additional insight into the registrant/user relationship but when you can't code such a high percentage of them it turns out to be not a very useful variable.

So just in summary this was an exploratory study that was a first step in ICANN's process to learn about domain name registrants and their relationships to domain users and the ways in which domains are used.

In many cases the classifications and characteristics and activities were difficult to discern and often had to be coded as unknown. But even with that we were able to find a large enough number of domains where we could code them so that important relationships were uncovered.

This is all summarized in our draft report which also includes an appendix which has much more detailed information about the data. That can be found at the link that I put in the slides.

ICANN, as Barbara mentioned at the beginning, is seeking comment from the public on this. So you can go to the same Website. The comment period will close at the end of this month. So with that thank you for listening. There are additional materials past this slide. And all of the material in the Webinar today is included in the Registrant Identification Study Draft Report.

So at this time I think we can open it up to questions.

Barbara Roseman: Thank you, Ed. There have been some questions in the chat room. And (Steven) and (Michael) have been answering them as they go along. (Steven), is there anything that you'd like to expand on from your answers?

(Steven): I'd actually like to open it up to the questioners to see if their questions have been answered fully. I know I've deferred Avri's question to this Q&A because it was a more detailed question that would entail a longer response.

Barbara Roseman: Okay. In that case why don't we open the floor to questions? And if you're in the chat room and can use the raise hand thing we'll be able to maintain a queue I believe.

Coordinator: Thank you.

Barbara Roseman: Otherwise, Avri, why don't you go ahead and go first since you had already started your question?

Gisella Gruber-White: Barbara, this is Gisella. Just to let you know that the lines have been unmuted for those on the audio bridge. And I don't currently see Avri on the audio bridge. Thank you.

Barbara Roseman: Oh okay.

(Michael Yogovich): This is (Michael Yogovich). I can address Avri's question. So the question is, "Knowing NORC data from a previous lifetime I am wondering whether the analysis could have differentiated between commercial and noncommercial legal persons."

Well so we'll answer this part of the question first. There's really two parts to this, legal persons - are you referring to the domain user or the registrant

legal person? For the domain user legal person we not only collected Whois data as we did in a previous study, in addition to survey data, but we also collected the html content associated with the domain.

So using that content and a variety of other tools such as Accurint, Google searches, LinkedIn searches, we were able to look at what was the domain user actually doing with the Website. And that - and if we made a determination that the domain user's function was being a noncommercial legal person that's how we could differentiate between noncommercial and commercial.

(Steven): I can still hear the audio feed.

(Michael Yogovich): Can people - can...

((Crosstalk))

Avri Doria: This is Avri. So a follow up question on that...

Barbara Roseman: Okay go ahead, Avri.

Avri Doria: (Unintelligible) question on that (unintelligible) oh on this one I'm seeing it, okay. Never mind.

Barbara Roseman: Okay thank you.

((Crosstalk))

(Michael Yogovich): To the second part of your question while we did - so we did collect data that would be able to make the distinction for public services. That was not a variable that we coded with - to that level of detail.

Edward Mulrow: I would just guess that those types of organizations, if we did see them we'd - I'm guessing we would probably say they were nonprofit users if we're talking about the domain user? And I put up the slide where we at least look at the apparent business structure. We see there is a small percentage of 2% were found to be nonprofit users. Again, there is this very large proportion, about 65% where the business structure was unclear though.

Avri Doria: Okay thank you. But that means that you could, in further analysis, differentiate (unintelligible) cases if you so desired.

(Michael Yogovich): Well we...

((Crosstalk))

(Michael Yogovich): Sorry, Ed. Go ahead. Okay we would need to first do additional work to code whether or not we detected those types of services. And then if we - once we had those in a coded data set we could then do additional analysis, yes.

Edward Mulrow: And I'd say what this study tells us if we were interested in that we would need to do a much larger sample in order to find enough of them to do any further analysis with it. As it is with only 2% being nonprofits in general if we then were targeting to have it as women shelters or human rights advocacy, you know, we'd need a much, much larger sample to really start finding enough of them.

Avri Doria: Understood. Thank you.

Barbara Roseman: Does anyone else have a question? I noticed that Steve Metalitz is saying that his audio on the Adobe Connect is not functioning properly so he may have had a question also.

(Lisa): Yeah, this is (Lisa). I'll jump in and ask Steve's question. He had one about whether if there was a natural person who seemed to hold a number of domains whether they would be - have been classified as a legal person or a natural person.

(Michael Yogovich): Could you repeat the - this is (Michael). Could you repeat the question please?

(Lisa): The question was if you found someone who appeared to be a natural person who had multiple domains whether they would have been classified as a natural person or a legal person registrant?

(Michael Yogovich): So I think a - this is (Michael Yogovich) speaking. I had addressed that previously in the chat window and I'll just restate my answer. When you - so it kind of - it would depend on the individual circumstances. We did use the reverse Whois tool to gather how many domains were registered to a particular individual or organization.

If that number was very large - I can't state off the top of my head but I believe, based on my recollection that there was a high correlation that exists between large numbers and - of registered domains and legal persons. However that was not the first variable that we checked to conduct the coding.

Barbara Roseman: So it would not have been automatic reverse Whois that you would do to find out if a given registrant or a given domain user had a number of other domains? That was something that you only checked after looking at both of those variables?

(Michael Yogovich): Correct. The - this is (Michael) speaking. The reverse Whois was - we discovered it towards the end of the study and we conducted in all domains and used it as a validation tool.

Barbara Roseman: I see, okay. (Lisa), did you have any other questions that you wanted to raise?

(Steven): (John) posted a question.

Barbara Roseman: Oh I beg your pardon.

(Steven): And I think (Michael) is best equipped to answer it.

Barbara Roseman: Were registrar and hoster coming soon pages differentiated from (PPC) parked pages?

(Michael Yogovich): So the pay per click - I think we do have a slide on this. We did - we attempted to differentiate between a parked domain and a domain reseller. However that distinction may not have always been made. We use - if there was enough information provided where we could make that distinction we would classify it as a domain reseller. However by default we first coded it as a parked domain.

Barbara Roseman: And ones that were parked with clear (PPC) then you counted those separately or not parked but ones that had a lot of (PPC) links? Those were counted separately?

(Michael Yogovich): Well so I - it depends. (PPC) was a variable that was classified for every domain whether or not we detected pay per click advertisements on that page. So if you wanted to differentiate between the domain resellers and the parked pages you would have to do a cross tab on the (PPC) variable and the domain resellers and the parked domains.

Barbara Roseman: Okay.

(Michael Yogovich): Does that answer the question?

Barbara Roseman: I think it does.

(Lisa): The pay per click ads you actually did count separately whether the ads were hosted or offsite correct?

(Michael Yogovich): We attempted to classify hosted pay per click advertising if we were able to have enough - if there was enough information to make that distinction. Again follow the same classification strategy where first if was identified as pay per click if we were able to then identify it as the pay per click ad was placed on the page via a hosting service the further classification was made that it was a hosted. Those two variables are mutually exclusive though.

Barbara Roseman: Okay are there any other questions?

Woman: I think we had one as we went along about whether a domain that had the name of a small business would have been classified as a legal person or a natural person. It might be worth giving that answer to the whole group.

(Steven): (Michael), you answered that question in the chat.

(Michael Yogovich): Yes, I'm trying to...

((Crosstalk))

Barbara Roseman: ...midway through the chat transcript so Chris Chaplow is who asked the question.

(Michael Yogovich): Oh so for the - is this the question that says - I was just trying to locate...

((Crosstalk))

Barbara Roseman: ...micro-business considered legal person...

(Michael Yogovich): Yes.

Barbara Roseman: ...or natural person.

(Michael Yogovich): So this is (Michael) speaking. And the micro name since it would refer to a business we classified all businesses as legal persons.

Barbara Roseman: So even if it wasn't officially incorporated you didn't like look for that information?

(Michael Yogovich): We attempted to do a further classification of whether or not the business was incorporated and the different types of incorporation. And that was coded under a separate variable.

Barbara Roseman: Okay great.

Edward Mulrow: Yes, I'd say that, (Steven), you can correct me if I'm wrong but these more detailed breakdowns like this are more likely to be in Appendix A to our report than in the sort of the main text which is just summarizing everything. But for more of these detailed breakdowns - if we thought that there was a sufficient enough number that were counted would appear in Appendix A.

(Steven): That's correct.

Barbara Roseman: Okay so if there are no further questions in the chat room or from the attendees then I think we'll say that this concludes the call. As Ed and (Steven) and (Michael) have pointed out there's a lot of information in the appendices to the document that is online. And the appendices are included in that document so there's no separate document to download.

And I encourage all of you to take whatever questions you may have remaining or that you thought of later to the public comment forum and

please enter them there. We will do our best to be responsive and include any significant answers in the final report.

I see (Steven) is still typing in the chat room so I'll let that...

(Steven): No, I was just going to thank everyone for attending.

Barbara Roseman: Oh okay great. Okay thank you very much. The call is concluded now. And, Nathalie, if there's anything else that you have to do to - or Gisella - to round up the call then I'll let you go ahead and do that now.

Gisella Gruber-White: This is Gisella. I would just like to thank everyone for participating and, no, just wish you good evening. Thank you.

Barbara Roseman: All right, thank you very much for attending, everyone. Good-bye.

Edward Mulrow: Bye.

END