Whois Registrant Identification Study Webinar

TRANSCRIPTION

6th March 2013 12:00 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: Thank you. The call is now being recorded. Please go ahead.

Nathalie Peregrine: Thank you very much, (Sam). Welcome to this Webinar. A few housekeeping rules before we start. Please be aware that there will be audio streaming in Adobe Connect room. If you wish to ask a question you may do so in Adobe Connect Chat or dial in to the audio bridge.

On the audio bridge all participant lines will be muted during the presentation. They will be opened at the end for questions. The recordings of this Webinar will be available on the GNSO calendar page shortly after this call.

Thank you very much and over to you, Barbara.

Barbara Roseman: Thank you. This is Barbara Roseman and I'm currently the Coordinator of the Whois Studies that were originated a few years ago by the GNSO. The Whois Registrant Identification Study was initiated to a better profile of who the various users of domain names are.

And this is intended - the work has been intended to further our understanding of how Whois functions and what its uses are so that we can better develop policy moving forward.

The National Opinion Research Center, NORC, was - at the University of Chicago, was selected to perform this particular study. And they've been working on it for the past two years. And their results have finally been put forward in this draft that is now currently available for public comment.

We'd like to present these findings today so that you can get them in a more summary form but all of the details that are presented today are also included in the report that's online.

So I'm going to turn it over now to Edward Mulrow, the Project Director and two of his participants are online as well for being able to answer questions afterwards. So, Ed, would you like to begin?

Edward Mulrow: Yes, thank you Barbara. So we'll be going through just some of the highlights from the Registrant Identification Study. And as we do so the outline for today - we'll go over some of the goals of the project as well as some of the key questions that we were asked to answer.

In doing so we'll go through and give a brief view of the sample design that we used to select the domains that were used in our study. We'll also talk a little bit about how we went about collecting the data and then we'll spend some time talking about how the data was coded after the data was collected and in doing so we will present some answers to the GAC questions that we were given. And finally to wrap things up we'll go over a few lessons learned.

So we'll start now with the goals of the project. So primarily this was an exploratory examination of the Whois data for a representative sample from gTLDs in the - for the top five - from the top five gTLDs coordinated by ICANN.

The intent was to understand the registrants and domain users looking at the types of entities using the services and the kinds of activities for the domains. Our focus area were on the registrants trying to categorize them as natural persons, legal persons or those who registered domains using a privacy or proxy service.

We also wanted to look at the domain users to determine between natural persons and legal persons. And also to look at the activities in particular the commercial activities that could be taking place at a domain.

In looking over this we also were given four sort of key questions to start with, although we went further than just answering these questions. But, the questions were 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?

So in order to answer this we collected - we did a sample of 1600 domains from the five most common gTLDs. And when we started this study - the planning for the study back in June of 2011 those top five gTLDs represented 98.5% of all domains - all gTLDs coordinated by ICANN.

We decided we should stratify by gTLD, that was the best stratifier that we were able to use. And then we allocated the sample based on the percentage of the gTLD in the population although for DotBiz and DotInfo we increased the sample size in those up to 100 to make sure that we had enough to examine within those gTLD groups.

Once we had the sample we began our data collection. We developed a - what we call the NORC box, that's a multithreaded application developed in Python. It was - it's an automated tool for gathering Whois data, publicly accessible files, web content and also response code from blacklists and white lists.

This information is not static; it can change at any point in time. So we attempted to collect this information simultaneously from all three sources. To collect the Whois data we used the Whois API service which we found to be a good way to get nicely formatted records from the Whois sources. However it didn't always return the information.

So as a backup plan ICANN staff also ran their own Whois extraction process that was done within the same timeframe that we ran NORC bot. And those two simultaneous extractions were merged and compared so that we could get as complete a possible set of Whois records for the 1600 domains in the sample. That data collection took place in March of 2012.

In terms of extracting the Web and ftp content we only looked at the www and ww2 subdomains. Domains may have had content on other subdomains but no attempt was made to look for that content. We also had a download quota of 100 megabytes to ensure that extremely large sites hosting gigabytes of data were not downloaded. This was not a very restrictive condition as most domains did not have that much information that we downloaded.

Now once we had finished the extraction process we went through and coded the variables. We classified the variables into three types. There's the Whois type, domain user type and domain content. For the Whois these are coded based on Whois information.

And it may have backed up - been backed up by some independent searches of public databases but we did not use a domain's own sites for any information so we tried to keep this just at the Whois information only.

The variables we looked at there were apparent registrant type, the country or region of the world under which associated with the registrant and the registrar also. And in a moment I'll talk a little bit more about the details behind the apparent registrant type.

For the domain user we also looked at the domain - the apparent domain user type. We tried to discern the relationship between the user and the registrant as well as the user's business structure. The coding for this was based on our downloaded content.

And then the content of the domain was related to the activities that might be taking place at the domain. So primarily we looked at the potentially commercial activity but we also looked for allegedly illegal or harmful activities and explicit sexual imagery. And once again these were coded just based upon the content that was found when we - from our downloads.

So now we'll look at three of the key variables that I'll talk more about today so the apparent registered type; that was a Whois variable. And we tried to classify three types of registrants, that would be the natural person, legal person or a privacy proxy service.

So for a natural person when we looked at the Whois data, in particular the registrant name and the registrant organization, tried to determine if it appeared to be a real living individual. Legal persons were those that appeared to - where the name and organization appeared to identify a company, business, partnership, etcetera, some group or legal entity.

That would include multiple domain name holders but it does not include privacy proxy service providers. Reverse Whois email accounts were used to help determine multiple domain name holders.

Privacy proxy services were determined by developing a list of known providers. We started with a list that was developed for the study on the prevalence of domain name registered - domain names registered using a privacy or proxy service; that was our guide.

We looked to enhance that more but once we had that list we looked for those - if we could match those providers to the information in the Whois record we categorized it as a privacy proxy service.

There were some records where we could not classify the apparent type based on the Whois data. This would include records where the name and organization were completely missing or were patently false or incomplete and that could include domains pending reactivation or deletions.

Now on the next slide we see how this broke out for our sample of 1600. And it also provides an answer to the GAC question of what is the percentage of registrants that are natural versus legal persons. So we see that approximately 39% of the domains registered in the top five gTLDs we categorized as legal person registrants. Another 33% were natural person registrants. Twenty percent turned out to be privacy proxy services and 8% were unknown.

So now moving on as we look at the domain user variables in particular the apparent domain user type based on domain content we tried to determine whether it was a natural person or a legal person. Definitions similar to what we used for the registrant type.

In this case though there were times where we could not classify quite a few of the domain user types but the reasons could be different. There were some times where there was no usable online content. By that we mean either no content was available or there was minimal html code that was not sufficient to determine the user type.

Parked domains, in some sense, were similar to no usable online content but the domain landing page - the html content that was there was consistent with typical domain parking content.

And as in our report we will see - you can see that in particular where commercial - potentially commercial activity is involved parked domains have much more commercial - potentially commercial activity taking place than domains where there is no usable online content.

Even after separating out those there are still some domains where we had available content but we could not determine the user type, that is whether it was a natural or a legal person.

So here's the breakout for domain user type. We had 37% were legal persons out of the 1,600 domains in our sample. Twenty-six percent had no usable online content, another 21% were parked. Twelve percent were domains where we could not determine the user type so those were an unknown user type. And 5% were natural person users.

So of the 1,600 domains 37% or 586 were determined by NORC to be legal person users. And this brings us to answering another one of the GAC questions which is, what is the relative percentage of privacy proxy use among legal person users?

So if we just looked within the 586 that we classified as legal person users we see that 15% were privacy proxy service registrants. This is slightly lower than the 20% that we found overall.

Statistically speaking this is borderline whether there is a statistically significant difference between the two or not but there is - there was a slightly lower percentage within the group of legal person users for the privacy proxy registrant.

And incidentally if you are looking for this information in our report Exhibit 16 shows much more than this; it will go through and show how the registrant breakout for other types of domain users. And particular there if we look at the parked domains we see that the privacy proxy registrants increase above the 20% when - for parked domains. So it lowered some for legal person users, it increased for the - for parked domains.

And now we'll move on to the coding of the activities at the domains and in particular here we'll look at the potentially commercial activity. So this was an attempt to categorize all observed monetary activities that in some countries might be legally considered commercial activities.

So the thing we looked for ecommerce, collection of membership dues for online or offline content, promotional material content, banner ads and pay per click ads.

So the table on our next slide shows a breakout of what we - how we tried to code these things. We had five main categories which we called promotional content, pay per click ads, banner ads, ecommerce and membership dues. There were subcategories there to help us in the coding process.

Now these are not mutually exclusive categories. A domain could show evidence of more than one of these activities. So in our analysis we just concentrated on whether or not any of these activities was taking place at a domain.

Which brings us to our next slide on potentially commercial activity, the GAC question which is what is the percentage of domain name uses that are commercial versus noncommercial?

So when we include pay per click ads at least one of the five activities was detected in 905 of the 1,600 sampled domains so that's approximately 57% of domains with potentially commercial activity.

If we don't include pay per click ads as the potentially commercial activity the number of domains dropped to 717 or 45% of all domains with potentially commercial activity.

Now there was one further GAC question which was what is the relative percentage of privacy proxy use among domains with commercial use. So including pay per click ads, as I said, we have 905 sample domains with detected potentially commercial activity and among those there were 23% which we determined were registered through the use of a privacy proxy service. That is slightly above the overall of 20%. Once again I would not consider that a statistically significant difference.

Okay so that's the summary. We have much more information included in our report. And now we'll just move on to some lessons learned. This was an exploratory study not only to find out some information about the registrants but we were exploring how to collect this type of data.

So collecting the Whois domain content and DNS DL information in a nearly simultaneous manner is difficult but if you use a multithreaded application such as the NORC bot the task is feasible. We include a summary of some of the features of the NORC bot in our draft report. In particular in Section 4 of the report, Lessons Learned, we go into some details about some of the features. It's still at a summary level but you can learn more about it there.

Data coding was challenging; it was subjective based on the rules that we put in place. There is some inherent ambiguity in Internet data so it sometimes is - might be apparent to us to classify things one way and other people might see it a different way.

If you attempt to impose some standard codes on a huge variety of unique Websites you'll find that sometimes it's not always possible to classify things

in such a standard way. For example trying to distinguish between parked domains and domains for reselling was fairly difficult.

In terms of some of the variables we looked at trying to discern the user relationship to the registrant that was difficult. We classified 50% of the relationships as unknown. That is highly related with domains without content. If we have a domain with no online content or it's a parked domain there is not much there to tell us what the relationship between the user and the registrant is.

Also we tried to look at the business structure of the user. That was very difficult to discern, 65% were classified as unknown. Initially we thought that this would be a good variable to look at because it might provide additional insight into the registrant user relationship but given the high percentage of unknowns that really didn't work out.

So just to summarize things before we open it up for questions, this was an exploratory study that's the first step in ICANN's process to learn about domain name registrants and their relationships to domain users and the way in which the domains are used.

In many cases classification of the characteristics and activities were difficult to discern and often had to be coded as unknown. Large number of domains were able to be coded so that important relationships were uncovered.

Our draft report is available at the link I provided in the slide. And as Barbara noted ICANN is seeking comments from the public and they can be submitted at the same Website. The comment period closes on March 31 of this year.

So with that said I thank you for listening. There are additional slides in this presentation that follow; I will not go over them. They come from the Registrant Identification Study Report. And so now I think we'll just go ahead and open it up for questions.

Barbara Roseman: Ed, thank you very much. There've been a couple of questions in the discussion chat and wanted to draw your attention to those. The main questions have to do with how categorization for legal person versus natural person was done. And you addressed this somewhat in the slides but if you could maybe go into that a little bit.

Edward Mulrow: Yes. I'll ask - (Michael Yogovich), are you prepared to give a little more background on that?

((Crosstalk))

(Michael Yogovich): Yes I'm prepared to give some more background on it. So first of all when we looked at the classification of a natural person we were looking as a single existing person, a human being, and a legal person was more of any type of business structure, any type of organization or anything that was not a primary individual that could - was not...

((Crosstalk))

Barbara Roseman: Somebody's line is open and they're having a separate conversation. If you could please be aware that all the lines are currently open for...

((Crosstalk))

(Michael Yogovich): Okay so I just see a chat line here in the chat. My question was not on legal person but on categorization of elicit activities.

Woman: There were two different questions, (Michael).

((Crosstalk))

(Michael Yogovich): Okay.

Barbara Roseman: ...one basically was on how legal persons were determined and the other was on how allegedly illegal activities were categorized.

(Michael Yogovich): Okay.

Barbara Roseman: (Steven) says that he can take that question on illegal activity.

(Michael Yogovich): Okay so I will - I will continue with the natural versus legal person. So we use a combination of manual and automated techniques.

(Michael Yogovich): So primarily when we were looking at the legal persons we were looking at, as I was saying, to business structure and - not so much the business - like what was the entity that was owning the domain doing. And so when we were looking at it we would do a review of both the domain. Was the question for domain user legal classification hosting? Yes.

Woman: I believe the question was actually registrant type, legal person versus natural person.

(Michael Yogovich): Oh registrant type. Yes, so then we would look at the Whois information to make this determination. And this was a - done, like I said, using a variety of natural or manual and automated review.

We used the previous studies that we conducted - some of the data - when we did the Whois classification to look up - to do a portion of that coding work automatically primarily to do privacy and proxy coding providers. If we detected any matches on natural person which I think we did not find many we would also have coded that.

And then - and we also then did the rest of the coding manually. Primarily, you know, if there's a field (in) the business structure we could identify a business structure or a business through doing Google searches or a variety

of other techniques after reviewing it then it was classified as a legal person. A natural person was classified if we could identify the individual as only using the domain for personal use.

Edward Mulrow: Just to add to that, (Michael). I brought up a slide that shows the Whois country or region by registrant type. And if we go down to the unknown registrant types you see we - there is kind of a spike there for China. And so it's possible that - we don't know for sure but just using only the registrant name and organization it's possible that we really couldn't - we didn't understand, say, the Chinese name to be able to give a classification between the natural person or the legal person.

I'm not saying that that did exactly happen but you see that in China we kind of have an extra spike higher than expected percentage where we don't know the registrant type.

I think, (Steven), you were going to answer the question about the alleged illegal activity?

(Steve): Yeah, so along with our main report there's also an appendix that contains a lot of analysis with not as much interpretation. But Section H does include the allegedly illegal or harmful activities that we attempted to code from the data set.

And so just reading from Table H1 the types of activities that we were looking for would be spam activity or advanced fee fraud Websites, phishing, cyber squatting, typo squatting, counterfeit merchandise, trademark infringement, malware. And some things that we actually didn't find included intellectual property theft, child sexual images, identity theft Websites and money laundering Websites.

And all told out of the 1,600 we only classified 18 of the 1,600 as having allegedly illegal or harmful activities.

Edward Mulrow: And there are a few slides at the very end of the presentation deck that we've made available that talk about this. And when you talk about that those were - our manual coding process we looked for those activities. We also sort of kept separate domains that we found on blacklists as well as white lists. And there's some minimal information there. And I believe (Steven), in the appendix, you have a much more extensive table on the types of - what we found based on the types of different blacklists.

(Steven): Yeah, there's a whole section on blacklist variables and there's a whole section on white list variables as well.

Woman: We've had another question here about the - your sample size relative to the size of the general population. Can you comment on how you chose the sample size?

(Steven): Yes, I would be happy to. So the sample of 1,600 domain names is obviously a small percentage of the domain names that are registered through ICANN on these five generic top level domains.

But it is a representative sample and therefore all the percentages that we give we also give standard errors. And those are based on sampling error. So you can see that those standard errors are not that large.

The problem with the sample size really is related to subgroups and whether or not certain subgroups have enough domains in our sample to be analyzed. And I would point out that for the natural person registrants we only have 78 or 80...

Edward Mulrow: Domain users.

(Steven): Oh, it's domain users where we only have a few...

((Crosstalk))

Edward Mulrow: ...person...

(Steven): ...natural persons?

Edward Mulrow: Yeah, the apparent domain user type - we had 5% were coded as natural persons so that was about 87%.

(Steven): So when we were looking to compare the legal persons and the natural persons among the domain users we didn't have very much power for that comparison. Also specific countries beyond the ones that were listed we put other Asia into a separate group because outside of China we did not have enough domains to do any specific analysis. Specifically the question was about the Philippines.

And so really it's the subgroups - certain subgroups that we don't have enough for analysis. But one key reason for the sample size is so that we would have enough privacy proxy registrants to analyze.

Edward Mulrow: Right, statistically speaking if we have a group of around 400, in this case domains, then we know that we'll have good statistical properties. That would be a worst case scenario.

So for instance we had - with privacy proxy domains we ended up with, what, around 320 in the sample which was more than enough to make good judgments about what was going on with the privacy proxy group. And as it turns out for many of the other categories the - what we found was large enough.

If we wanted to study more things like allegedly illegal activities or harmful activities there we, at least from our manual review, we really didn't find enough to do much more than report that we found, you know, a small

percentage. Any analysis breaking that group down further would not have good statistical properties. So we would need a much larger sample in 1600 if we wanted to explore that subgroup some more.

Woman: Yeah, and if I could just comment on that. We actually do have a separate study commissioned to look specifically at domain names involves - at least allegedly involved in the activities that (Steven) enumerated and analyze those.

Edward Mulrow: Did we have other questions?

Barbara Roseman: I think that this has been very helpful, Ed and (Michael) and (Steven). If there are no further questions then we can go ahead and close the Webinar at this time. This will be posted, Nathalie, on the GNSO calendar page, is that correct?

Nathalie Peregrine: Yes that is correct.

Barbara Roseman: And we will be holding another session of the Webinar later today. So if there are questions that you have that were not answered or that you think of afterwards please feel free to put them in the comment section of the public comment area for the report. Questions that are raised there will be addressed before moving forward with the final report.

So again thank you very much to the NORC participants and to those who called in today. Nathalie, would you like to close the call?

Nathalie Peregrine: Thank you very much for that. (Sam), you may now stop the recordings. This call is now closed. Thank you, everybody. Good day.

Barbara Roseman: Thank you.

END