From: Garth Bruen, KnujOn.com
Date: 07.29.2009
Subject: Review of NORC Proposal

I have detailed my concerns below relating to the NORC proposal (http://icann.org/en/compliance/norc-whois-accuracy-study-design-04jun09-en.pdf).

General Concerns about the Methodology

1. **Approach:** NORC is fully qualified to conduct research and survey analysis, but I'm not sure if the author(s) of this proposal understand the WHOIS problem at a depth that allows them focus on the right level of data. I believe it should be acknowledged first that the WHOIS accuracy problem is not a problem of the WHOLE record set. Rather, we are concerned the rate of WHOIS inaccuracy among domain names that are abused. Anyone who complains about WHOIS inaccuracy is usually doing so only after attempting to contact a registrant when a problem has occurred: spam, malware, illicit content, trademark infringements, etc. This is often compounded by poor secondary response from Registrars, ISPs, technical operators, and government. For the vast number of domains that exist without incident the validity of their WHOIS is not really in question. By not focusing on the general population of domain owners, ICANN can sidestep concerns about privacy.

2. **Size of survey data:** 2,400 is too small a set to reveal meaningful survey results. By their own recoding there are 102 million domains names in use, 2400 is 0.002% of the whole. Results from this less-than-one percentage cannot tell about the whole, but only the thousandth of a percent studied. It's like drawing conclusions about a shopping mall by only looking at one shelf.

3. **Source of survey data:** On page 2 of the proposal it is stated: *"ICANN drew and delivered to NORC a proportionate sample for these five domains…"* Citing ICANN for the source of the data invalidates the survey results. If the goal of this study is to satisfy ICANN/WHOIS critics, this statement will ruin that. It will be seized upon by anyone suspicious of the study in the first place and used as evidence

against it. A more respected survey would obtain the data independently or through a neutral third party.

4. **Mixing of data sets:** .COM, .NET, .ORG, .INFO, and .BIZ are really different animals, each with their own registries, volume, policies and standards. Therefore each should be the subject of a separate survey. Because of the overwhelming size of the .COM space, analysis of any other gTLD should be excluded. As we all know, the conditions for obtaining a .NET or .ORG were originally distinct, meaning a registrant had to be a network operator or a non-profit organization, respectively. Regardless of why the policies changed, the growth and consumer appetite for other gTLDs is remarkably different from .COM and requires a different understanding and approach.

5. **Schrödinger's cat**: With the survey conductors actually contacting registrants we may end up with behavioral changes in some registrants and possibly unscientific results. To be more effective, there should really be two studies: one without any registrant contact and another with a registrant questionnaire.

Are we trying to fix a problem, if so which problem?

It is not plainly stated, but I assume that the proposed study is intended to reveal information about *registrant* behavior and no other portion of the infrastructure. If that is the case we have to consider how WHOIS accuracy is impacted by other parts of the DNS.

WHOIS Access and Presentation

On page 3 of the proposal it is stated: *“For the *.org, *.info, and *.biz gTLDs, the Whois information…is standardized and easy to work with … For the *.com and *.net gTLDs, it is much more difficult to obtain, with many domains needing to be parsed by hand.”*

Right away there is an acknowledged problem that exists before the problem being studied. Not only are WHOIS access and standardization problems that should be studied before the general accuracy problem, but the differences in data access and credibility call for separate surveys. I am not questioning the integrity of

NORC staff, but saying one data set is unaltered and the other data set is subject to manual processing error means they cannot be mixed.

This concern is reinforced on page 14 where it is stated that: “…getting all WHOIS information for a sample taken from many registrars into a consistent data structure requires considerable work.”

So here we acknowledge that the inconsistent data formats is not just a problem of the gTLD registry handling but Registrars also. Different Registrar policies can severely influence WHOIS accuracy, therefore the WHOIS database is not really a single record, but a collection of many records. More information on this next.

Registrars Accept Double-Sets of Data

It is no secret that Registrars have two sets of client records: one for payment processing and another for ownership presentation. In most industries this would in fact be illegal. The ability for a registrant to enter different data in a WHOIS record further supports a research focus on the Registrars and not the registrants. Without the opportunity to create two sets of records, some Registrars may in fact have no inaccurate records whatsoever and should not be included in a study meant to determine registrant behavior.

Bulk Registrations

Most cases of false WHOIS data we have analyzed involved large sets of domains with the same inaccurate information. For example we have one registrant in our database with over 10,000 spammed domains and all of his WHOIS records are false. Compare this to the average domain owner who at most has fewer than ten registrations. The point is that there are some very distinct populations in the WHOIS space that are going to behave in different ways. The person or organization buying hundreds of domains at a time has a very different experience and intent than someone purchasing one domain for a specific reason and is unlikely to buy more. A survey that mixes these populations will reveal questionable data. To clarify, WHOIS record analysis of registrants who have specific domains for specific purpose and registrants collecting domains for speculation, parking, etc… should be studied separately.

Recommendations

I could give you many recommendations for improving this research, but I'll start with these recommendations:

-Conduct completely separate surveys for different gTLDs.

-Break results up by Registrar since there are varying data formats and collection policies.

-Study a much larger percentages of domains for more accurate results.

-Focus on WHOIS accuracy research for domains that have been abused.

-Do not mix survey results for registrants with fewer than 10 domains with registrants with large bulk holdings.