Contents

Terms of Reference for WHOIS Misuse Studies

These studies will examine the extent, nature, and impact of WHOIS public data misuse – that is, harmful actions taken using contact information obtained from WHOIS – by (1) analyzing reported misuse incidents and (2) conducting experiments to measure misuse and the effectiveness of anti-harvesting measures.

1. Objective

These studies will analyze reported and recorded harmful acts such as spam, phishing, identity theft, and stalking which Registrants believe were sent using WHOIS contact information. Specifically, these studies will attempt to prove or disprove the following hypothesis:

> **Public access to WHOIS data leads to a measurable degree of misuse – that is, to actions that cause actual harm, are illegal or illegitimate, or otherwise contrary to the stated legitimate purpose.**

Study results are intended to help the ICANN community determine whether WHOIS misuse is significant enough to warrant action and which combinations of anti-harvesting measures appear to reduce the most prevalent and damaging types of misuse.

In this study, WHOIS misuse refers to harmful acts that exploit contact information obtained from WHOIS. Those harmful acts [1] may include generation of spam, abuse of personal data, intellectual property theft, loss of reputation or identity theft, loss of data, phishing and other cybercrime related exploits, harassment, stalking, or other activity with negative personal or economic consequences.

It is not feasible to determine the percentage of actual WHOIS queries that lead to harmful acts, or to compare the frequency of harmful acts that exploit WHOIS data to those using names and addresses obtained from other sources. Rather, these studies will attempt to measure how often WHOIS misuse is reported and the impact on Registrants. By analyzing different types of WHOIS misuse (e.g., spam, phishing, identity theft, data theft), these studies will determine which occur most often and which are most impactful.

Finally, reported and recorded incidents will be correlated with anti-harvesting measures that some Registrars and Registries apply to WHOIS queries (e.g., rate limiting, CAPTCHA). Given WHOIS data storage and access path diversity (e.g., thick vs. thin Registries [11], query vs. bulk access [10], third-party WHOIS operators), misuse frequency may be affected by many factors. While this study cannot definitively attribute any observed decrease to a specific anti-harvesting measure, it can try to identify combinations and circumstances that appear to have the most impact.

2. Approach

This hypothesis can be tested using two fundamentally different yet complementary approaches: **Descriptive** and **Experimental** studies. The descriptive study documents and analyzes WHOIS misuse incidents (harmful acts) that have already occurred. The experimental study stimulates and records misuse to measure more reliably the impact of making WHOIS data public and WHOIS query filters applied to deter data harvesting.

Both kinds of studies are defined here because conducting both (either sequentially or in parallel) is one way to compensate for limitations inherent to each research method. If both kinds of Misuse studies are in fact conducted, terminology, inputs, and outputs for each should be defined consistently to facilitate reuse, integration, and correlation.

2.1 Descriptive Study

To conduct a descriptive study, sources will be surveyed to learn about harmful acts attributed to misuse of WHOIS public data. Inputs will be gathered from a representative set of sources, aggregated, and analyzed to categorize misuse by type (kind of harmful act), severity (impact of act on Registrant), and applicable anti-harvesting measures (WHOIS query/response filters). Input data sources include the following.

a) **Registrants:** As proposed by [2], survey a representative sample of Registrants that own domains in the top five gTLDs about specific harmful acts they have experienced which they believe were sent using WHOIS contact information.

b) **Registrars and Registries:** Pursuant to proposals [3][4], survey Registrars and Registries in several regions/countries to obtain contextual information about how WHOIS data can be queried for the above-sampled domains (e.g., supported query vectors, applied anti-harvesting measures, known harvesting attacks).

c) **Cybercrime Researchers:** To put WHOIS misuse into broader perspective, contact a representative set of independent industry research organizations that track related cybercrime activities (e.g., phishing, spam, identity theft) to gather examples and statistics regarding harmful acts occurring in many different regions/countries. Examples include the Anti Phishing Working Group, the Privacy Rights Clearing House, and the Online Trust Alliance (AOTA).

d) **Consumer Protection, Regulatory, and Law Enforcement Organizations:** To further put WHOIS misuse into broader perspective, contact a representative set of organizations that victims contact to report cybercrimes to gather examples and statistics regarding harmful acts occurring in many different regions/countries. Examples include the U.S. Federal Trade Commission, the FBI/NWCC Internet Crime Complaint Center (IC3), and the Identity Theft Assistance Center (ITAC).

Sources a) and b) can describe harmful acts attributed to misuse of WHOIS public data and/or the circumstances surrounding incidents first-hand; this primary research allows more qualitative analysis (e.g., actual impact on Registrants, effectiveness of WHOIS anti-harvesting measures). Sources c) and d) can only provide examples and aggregated statistics, but incorporating this secondary research may facilitate more quantitative analysis (e.g., which kinds of harmful acts are most frequently reported worldwide, to what extent are these acts attributed to WHOIS misuse).

A representative sample of Registrants for survey a) may be obtained by randomly selecting "n" domain names from the top five gTLDs (.org, .net, .com, .info, .biz), where "n" is calculated for each TLD to generate results with a 95% confidence interval. To enable analysis of global and region-specific misuse, sample design must also consider the Registrant's country/region to ensure that a representative set of countries are covered.

To obtain cost and consistency benefits, this study should build upon the foundation laid by the WHOIS Accuracy Study [6], as follows.

- **Sample Design:** The Accuracy Study started with a proportionate "microcosm" sample of 2400 domains from the top five gTLDs, without geographic limitation. However, because conducting telephone surveys in hundreds of countries is cost-prohibitive, that sample was refined to create a sub-sample of domains registered in just 16 countries. Industry standard "clustering" for studies covering large geographic areas was used to select countries with small, medium, and large domain populations, ensuring proportional representation in the sub-sample. The resulting geographically-clustered "verification" sample contained approximately 1400 domain names, sufficient to meet that study's 95% confidence interval objective.

- **Sample Cleaning and Coding:** WHOIS data for every domain name must include certain mandatory values (e.g., Registrant Name), but there is no RFC-standard record format or even a single global database from which WHOIS data can be obtained. The Accuracy Study therefore started with a "microcosm" domain name sample generated by ICANN. That sample was cleaned to eliminate parsing errors, mapped to Registrant country code and name, and then sorted by Regional Internet Registry. Only at that point could design parameters be applied to generate the cleaned and coded subsample required to perform Registrant Name and Address verification (the objective of the Accuracy Study).

Given differences in timeframe, the Accuracy Study's verification sample cannot be directly reused by WHOIS Misuse Studies. However, researchers are strongly encouraged to apply the same domain sample design, cleaning, and coding process to reduce cost and promote consistency across all WHOIS studies.

The sample used for survey a) also provides the list of Registrars and Registries for survey b). During survey b), researchers should not inquire about specific incidents reported by Registrants in survey a). Doing so would add considerable effort without much value; Registrars and Registries do not routinely hear about or track harmful acts experienced by Registrants. Instead, survey b) will obtain details regarding WHOIS operation that may have applied to queries performed on domains sampled by survey a).

Due to the diversity and complexity of WHOIS storage and access, survey b) must account for differences between "thick" and "thin" Registries and the impact of resellers, affiliates, and third-party WHOIS operators. For example, to identify the anti-harvesting measures (e.g., port 43 rate limiting, web query CAPTCHA, image-based responses) applied to queries about a given domain, this survey must learn about any measures applied by the Registry and the Registrar. However, misused data could also have been obtained in bulk form or from a reseller or a third-party WHOIS operator. It is not feasible to identify all possible vectors through which a victimized domain's WHOIS data could have been obtained, but survey b) will at least examine the primary vectors.

Surveys c) and d) add value by putting Registrant survey results into context using cybercrime examples and statistics covering larger, broader, or different sets of Registrants, Registrars, and domains. Surveys c) and d) should therefore include Cybercrime Research and Consumer Protection, Regulatory, and Law Enforcement Organizations that cover TLDs and countries beyond those examined by surveys a) and b). Note that some Registrants, Registrars, and Registries cited in c) and d) examples may also appear in a) and b) samples. However, to avoid over-representation, Registrants who reported WHOIS misuse to these kinds of organizations should not be specifically included in (nor excluded from) survey a) and b) samples.

2.2 Experimental Study

Harmful acts attributed to misuse of WHOIS public data can also be measured in a more reliable manner by conducting an experiment that monitors a set of test domains, registered through a representative sample of Registrars, distributed proportionally across the top five gTLDs.

To reduce cost and promote consistency, this sample should be a random subset of the Registrars identified by Descriptive Study b). (If that study is not performed, a suitable sample should be generated as described in section 2.1.)

A control set of unpublished Registrant addresses will be established, distinct from addresses published in WHOIS data for these test domains. Harmful acts against published vs. unpublished addresses will then be recorded, compared, and categorized by type (kind of harmful act), severity (impact on Registrant), and applicable anti-harvesting measures (WHOIS query/response filters).

These experiments should record harmful acts against test domains over a 90-day period. Network and/or client-based Internet defenses (e.g., anti-spam, anti-virus) may be used to detect harmful messages and defend research systems. However, evaluating the effectiveness of various Internet defenses is not the goal of this study. These experiments should therefore measure harmful acts *before* Internet defenses are applied by recipients, and the type and order of any such defenses must be documented. By correlating misuse frequency with WHOIS query vectors for each test domain, these experiments can also look for factors that contribute to or deter misuse.

These experiments can build upon some of the techniques used by an earlier SSAC WHOIS Spam experiment [8] – notably random generation of names in published / unpublished addresses. However, to meet the Misuse Study's objectives, researchers must start with a broader domain sample and design experiments to measure harmful acts that go beyond spam, including phishing, identity theft, and other cybercrime-related exploits.

The experiments summarized below should be used as a starting point for refinement. Researchers are expected to develop rigorous, repeatable test methodologies and formal test plans to conduct these and any other experiments they believe can meet study goals. All test plans must be reviewed and approved prior to study start.

a) **Email Spam:** As proposed by [5], compare the volume of unsolicited bulk email sent to WHOIS-published addresses vs. unpublished addresses. To differentiate between types of misuse, received messages must be divided into at least three categories: solicited email, phishing email (see below), and all other (unsolicited bulk) email – that is, spam.

b) **Postal and Telephone Spam:** Measure the volume of postal mail delivered to each Registrant's published address and calls received by each Registrant's published telephone number. Here again, unsolicited bulk mail and telemarketing calls would be differentiated from apparent attempts to "phish" for identities and all other postal / telephone communication.

c) **Phishing:** Categorize a subset of the email, postal, and/or telephone contacts received in the spam test cases as attempted phishing attacks requiring further analysis (e.g., impact assessment). These may include both mass-mail phishing attacks and spear-phishing attacks specifically addressed to the Registrant.

d) **Abuse of Personal Data and Identity Theft:** To detect and measure abuse of personal data in identity theft attempts, further analyze the *content* of email, postal, and/or telephone calls addressed to the Registrant. For example, letters from banks or merchants denying a credit application or purchase could signal attempted identity thefts, while email carrying a key-logger trojan (or a URL that leads to one) could represent attempted identity theft.

Some kinds of misuse simply cannot be studied through experimentation in a meaningful way and are thus considered out of scope. For example:

- **Harassment and stalking:** This study will not attempt to solicit or measure these kinds of harmful acts due the extreme difficulty of correlating such behavior to misuse of WHOIS public data for fictional Registrants.

- **Intellectual property theft and loss of data:** Cyber-criminals do not usually attempt to steal intellectual property without an interesting or high-value target in mind; a fictional Registrant is unlikely to be targeted by these harmful acts.

A key objective of these experiments is to identify factors that increase or decrease WHOIS misuse. To enable this, WHOIS query and response practices applied to each test domain should be examined, looking for relationships between anti-harvesting measures and the frequency harmful acts against published addresses. However, these experiments will only examine measures applied to WHOIS queries. Analyzing the impact of Privacy/Proxy registration services that prevent addresses from being published in WHOIS are beyond this study's scope.

As noted in Section 2.1 study b), anti-harvesting measures applied to WHOIS queries include limiting on port 43 and/or web WHOIS queries, CAPTCHA challenge on web queries, and image-based response formats. Furthermore, each test domain's WHOIS data may accessible through multiple vectors, including queries sent to the Registry's WHOIS server, the Registrant's WHOIS server, a reseller's WHOIS server, and a third-party WHOIS operator. Not only may each WHOIS server apply different anti-harvesting measures, but they may return different data elements (e.g., "thin" vs. "thick" Registry responses).

All of these factors must be considered when examining the processes applied to WHOIS queries about each test domain. If the Registrar/Registry survey proposed by study b) has already been done, this context may be available for reuse. Otherwise, this information must be requested from each Registrar and Registry used by test domains.

3. Inputs

Different sources are unlikely to supply the same information about each WHOIS misuse incident or represent those data elements consistently. Inputs obtained from all sources must therefore be normalized to enable aggregation, comparison, and statistical analysis.

Section 2.1 study a) and Section 2.2 experiments should gather the following data elements for each incident (i.e., reported or observed harmful act).

- Domain name *
- Type of Registrant (legal person or natural person) *
- Registrar (at time of misuse incident) *
- Complete WHOIS data (at time of misuse incident) *
- Misused WHOIS data (e.g., contact type, affected fields) *
- Could misused data have been obtained from other public sources? *
- Type(s) of misuse (e.g., spam, unsolicited phone call, harassment) *

- Description of misuse incident *
- Name and Type of entity that misused the WHOIS data
- National law or regulation that was violated by misuse incident
- Adverse consequences to Registrant arising from the misuse incident
- Purpose of domain (e.g., commercial or non-commercial), according to Registrant

Section 2.1 study b) and Section 2.2 experiments should also gather the following data elements for each affected domain.

- Registry type (thick or thin) *
- Registry's WHOIS server and query interfaces *
- Registry's applied WHOIS anti-harvesting measures, if any *
- Registry's documented WHOIS harvesting attacks, if any *
- Registrar's WHOIS server and query interfaces *
- Registrar's applied WHOIS anti-harvesting measures, if any *
- Registrar's documented WHOIS harvesting attacks, if any *
- Resellers/Affiliates relevant to affected domains

Because all data elements may not be available from every source, a minimum set of elements must be required for any reported incident to be included in study results. Proposed mandatory data elements are denoted with an asterisk (*) above. For example, surveyed Registrants may not know of specific laws that were violated, while experiments using fictitious Registrants may not have any adverse consequences.

It is essential that every incident denote whether misused information was published exclusively by WHOIS – this element is not only mandatory, but criteria for inclusion in study results. However, surveyed Registrants may not reliably know this, so Internet search engines should be used to verify that misused addresses are in fact NOT readily-available from sources beyond WHOIS.

All of the above data elements must be explained so that survey participants – even Registrants unfamiliar with WHOIS – understand what is being asked. Simple unambiguous definitions must be given for legal vs. natural persons and personal vs. commercial use; these should be consistent with those developed by the WHOIS Misrepresentation Study [9]. Types of misuse must also be clearly enumerated and defined, but must include an "Other" choice that can be used to describe incidents that do not fit into a previously-defined type. Focus group(s) should be used to assess how well participants understand draft survey questions so that definitions can be refined prior to final survey launch.

Survey questions must be phrased to encourage authentic responses and discourage over/under-reporting. For example, Registrants that have experienced misuse are more likely to respond to this survey, while Registrars may not have incentive to respond at all. Here again, survey wording should be vetted with focus group(s) to assess likelihood of participant response. Sample sizes must also assume a realistic level of non-response (e.g., 30% non-response is assumed by study [6]).

Finally, protecting sensitive data is critical to promote participation. Survey forms must acknowledge the privacy concerns that many misuse victims have and explain how responses will be protected and used. Individual responses must be safeguarded against unauthorized access or distribution. Cryptographic protection of stored/transmitted responses is not required, given that these studies involve WHOIS public data. However, participation is likely to be higher if only aggregate anonymous data is published.

4. Outputs

These studies will produce empirical results that characterize the extent, nature, and impact of WHOIS public data misuse. Specifically, study outputs should be designed to help the ICANN community determine whether WHOIS anti-harvesting measures are warranted and the combinations of measures that best reduce the most prevalent types of misuse.

The following raw data will be produced by the Descriptive Study:

- Total # of misuse incidents recorded by Descriptive Study
- List of Registrants surveyed/responding (categorized by TLD and region)
- For each Registrar and Registry surveyed/responding (categorized by TLD and region), available WHOIS access vectors and applied anti-harvesting measures
- For each Cybercrime Research, Consumer Protection, Regulatory, and Law Enforcement Organization, harmful act examples and statistics

The following raw data will be produced by the Experimental Study:

- Total # of misuse incidents recorded by Experimental Study
- List of registered Test Domains (categorized by TLD and region)
- List of Published and Unpublished Addresses per Test Domain
- # of incidents for each Address (categorized by type
  – e.g., email spam, phone phishing, postal identity theft)
- For each relevant Registrar and Registry (categorized by TLD and region), available WHOIS access vectors and applied anti-harvesting measures

Analysis that should be performed, based on this raw data, includes the following:

- Misuse incidents, categorized by gTLD
- Misuse incidents, categorized by region/county
- Misuse incidents, categorized by type of harmful act
- Misuse incidents, categorized by anti-harvesting measure
- Misuse incidents, categorized by severity (impact on Registrant)

In particular, analysis should attempt to answer the following key questions:

- Are some gTLDs or regions/countries especially prone to WHOIS misuse?
- Within the realm of WHOIS misuse, which kind of misuse is most frequent?

- Within the realm of WHOIS misuse, which kind of misuse is most harmful?
- Which WHOIS anti-harvesting measures best reduce frequency of misuse?
- What other factors (if any) appear to increase or decrease frequency of misuse?

5. References

[1] Working Definitions for Key Terms that May be Used in Future WHOIS Studies, GNSO Drafting Team, 18 February 2009

[2] Study Suggestion Number 1, Document misuse of WHOIS data, Steve Del Bianco

[3] Study Suggestion Number 21, To what extent is WHOIS data being misused, Kathryn Kleiman

[4] GAC Data Set 2, GAC Recommendations for WHOIS Studies, 16 April 2008

[5] Study Suggestion Number 14, Quantify the extent to which the WHOIS database is used to facilitate illegal or undesirable activities, Steve DelBianco

[6] Proposed Design for a Study of the Accuracy of Whois Registrant Contact Information (6558,6636), NORC, June 3, 2009

[7] [left blank/not used]

[8] SAC 023: Is the WHOIS Service a Source for email Addresses for Spammers, October 2007

[9] Terms of Reference for WHOIS Misrepresentation Studies, ICANN, September 2009

[10] Registrar Accreditation Agreement (RAA), ICANN, 21 May 2009

[11] Explanatory Memorandum: Thick vs. Thin Whois for New gTLDs, ICANN, May 2009