Contents

Terms of Reference for WHOIS Registrant Identification Studies

This study will measure the extent to which domains used by legal persons or for commercial purposes (1) are not clearly identified as such in WHOIS Registrant data and (2) are correlated to use of Privacy and Proxy registration services.

1. Objective

This study will examine WHOIS data for a representative sample of domain names, looking for Registrant Names and Organizations that are either patently false or appear to identify a natural person, an organization commonly engaged in non-commercial activities, or a Privacy or Proxy registration service. By analyzing Internet content associated with each of these domains, this study will attempt to measure how often Registrants who are in fact legal persons or engaged in commercial activities fail to clearly and directly identify their domain name's true ownership or purpose in WHOIS, including the degree to which Privacy and Proxy services are involved.

Specifically, these studies will attempt to prove or disprove the following hypothesis:

> **A significant percentage of domain names registered using Privacy or Proxy services are associated with WHOIS Registrant data that identifies a legal person as a natural person and/or obscures the domain's commercial purpose.**

In this study, a *natural person* [1] is a real, living individual, as opposed to a *legal person* which may be a company, business, partnership, non-profit entity, or trade association.

Commercial purpose [1] refers to the bona fide use or intent to use a domain name (or any content, software, materials, graphics or other information thereon), to legally exchange (or facilitate the exchange of) goods, services, or property of any kind in the ordinary course of trade or business. These differ from *non-commercial uses* [2] commonly associated with individuals and civil society organizations involved in education, community networking, public policy advocacy, promotion of the arts, children's welfare, religion, consumer protection, scientific research, and human rights activities. According to [3], registering a domain name solely for the purposes of selling, trading, or leasing that name does *not* constitute a "bona fide business or commercial use;" therefore, such domains are considered to have non-commercial purpose.

As defined by [1], Proxy and Privacy registration services provide anonymity and privacy protection for domain name users. *Privacy* services hide certain user details from WHOIS by offering alternate contact information and mail forwarding services while not actually shielding the user's identity. *Proxy* services have a third-party register domain names on the user's behalf and then license the use of the domain name so that a third-party's contact information (and not the licensee's) is published in WHOIS. According to ICANN's Compliance Department, obtaining the actual user's identity during a study of Privacy/Proxy registrations would be likely for Privacy registrations, but domains registered by a Proxy provider largely conceal the identity of the licensee.

Privacy and Proxy services are often used by natural persons and organizations like human rights groups that have well-known reasons for preserving anonymity and concealing personal identifying information. However, these services are available for use with .com and .biz domain registrations. In fact, ICANN's Registrar Accreditation Agreement [12] does not prohibit Privacy/Proxy use by legal persons and does not require Registrants of any type to identify domain purpose. As a result, this study cannot categorize such uses as impermissible. Instead, this study can help the ICANN community better understand how often and why such uses occur, providing the empirical input needed to consider any related policy changes.

2. Approach

This hypothesis will be tested by conducting a descriptive study which first classifies a representative sample of Registrants by apparent type. Domains clearly registered to legal persons require no further analysis. All other domains, including those using Privacy or Proxy services, will be further analyzed based upon the content of websites and other Internet data associated with each domain. Based upon proposals [7][8][9], domain names subjected to this content analysis will then be categorized according to actual ownership (natural vs. legal person), actual purpose (commercial vs. non-commercial use), and distribution by gTLD, country/region, and Privacy/Proxy use.

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

For cost and consistency benefits, this study should build upon the foundation laid by the WHOIS Accuracy Study [4] and WHOIS Privacy/Proxy Prevalence Study [5] 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 and translate non-ASCII characters, 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 used to verify Registrant Name and Address (the objective of the Accuracy Study).

- **Registrant Type Classification:** Next, based on WHOIS Registrant Name and Organization values, each domain name in the "verification" subsample was assigned an Apparent Registrant Type (e.g., natural person, registered business). All domain names apparently registered using a Privacy/Proxy service were then passed to ICANN for coding and confirmation by the WHOIS Privacy/Proxy Prevalence study (established to measure the prevalence of domain names registered using a Privacy or Proxy service among the top 5 gTLDs).

Given timeframe differences, the Accuracy Study's sample, vetted by the Privacy/Proxy Prevalence Study, cannot be directly reused by WHOIS Registrant Identification Studies. However, researchers are strongly encouraged to apply the same sample design, cleaning, coding, and classification process to reduce cost and promote consistency across all WHOIS studies. In particular, ICANN's help may be needed to efficiently confirm apparent Privacy/Proxy use and request Registrant identities from Privacy services.

This cleaned, coded, and classified sample will then be examined to identify and categorize domains that might potentially have been registered by legal persons or used for commercial purposes but not clearly identified as such in WHOIS Registrant data. To determine this, two questions will be considered:

- **Ownership:** For domains used by *legal persons*, how often does WHOIS Registrant fail to clearly and directly identify that legal person? This occurs whenever a domain being used by a legal person (e.g., a business, partnership, non-profit entity, trade association) to engage in Internet activities of any sort is associated with a WHOIS Registrant Name/Organization that (a) is patently false, (b) identifies a natural person's Name not accompanied by a Registrant Organization, or (c) identifies a Privacy/Proxy service's Name/Organization.

- **Purpose:** For domains used for *commercial purposes*, how often does WHOIS Registrant identify an entity more commonly engaged in non-commercial activites? This occurs whenever a domain being used to engage in commercial Internet activities (i.e., solicit the exchange of goods, services, or property) is associated with a WHOIS Registrant Name/Organization that (a) identifies a natural person, (b) identifies a commonly non-commercial user (e.g., a human rights organization, a community group), or (c) identifies a Privacy/Proxy service. Note that commercial activity includes revenue-generating ads posted on "parking pages" explicitly linked to registered but otherwise unused domain names.

These two tests are related but distinctly different. For example:

a) A domain with Registrant Name=John Doe (a natural person) that resolves to a website used by ABC Corp (a legal person) to describe products sold by ABC (a commercial use) fails to clearly identify both ownership and purpose in WHOIS.

b) A domain with Registrant Organization=Community XYZ (an implicitly non-commercial organization) that resolves to a website used by that community for on-line fund-raising activities (a commercial use) fails to clearly identify purpose only.

c) A domain with Registrant Name=John Doe (a natural person) that resolves to a website used by Human Rights 123 (an implicitly non-commercial organization) with no commercial content fails to clearly identify ownership only.

d) A domain with Registrant Name= Proxy123 (a Proxy service) that resolves to a website used by ABC Corp. (a legal person) to host pay-for-click ads (a commercial use) fails to clearly identify both ownership and purpose.

Note that the Registrant's *intent* is important but not readily or reliably discernable. For example, John Doe may have registered a domain using his own name or a Privacy/Proxy service because he owns ABC Corp or authored a website for ABC Corp or participates in Community XYZ or volunteers for Human Rights 123 – or John may be fictitious. This study finds all to be domain names where ownership and/or purpose were not clearly identified, but they clearly have different implications for the ICANN community.

To deliver useful results without subjectively guessing the Registrant's intent, this study will categorize domain names based upon information obtained from multiple sources (e.g., websites, spam URL lists, Internet search engines, individual and business directories). For example, domains that fail to clearly identify legal person ownership might be grouped into categories like Patently-False, Developer-Registered, Principal-Registered, PrivacyService-Registered, or ProxyService-Registered. Actual categories and mapping criteria are not specified here, but should be developed during the study to clearly describe and unambiguously differentiate between common scenarios.

3. Inputs

The first step in this study is to assign an **Apparent Registrant Type** to each sampled domain name using the following classes defined by the WHOIS Accuracy Study [4]:

1. Registrant Name and Organization are completely missing
2. Registrant Name and Organization look to be patently false (e.g., “99999”)
3. Registrant Name appears to be a natural person; no organization is named
4. Registrant Organization appears to be a registered business; no person is named
5. Registrant Organization appears to be a registered business; person is also named
6. Registrant Organization present but not clearly a business; person is also named
7. Registrant Organization present but not clearly a business; no person named
8. Registrant Organization appears to be a Privacy/Proxy registration service

Domains placed into classes 4 and 5 (registered businesses) are within the scope of this study but do not require content analysis because WHOIS Registrant clearly and directly identifies a legal person's ownership and implied commercial purpose.

Domains initially placed into classes 1 and 2 (missing or patently false) fail to clearly identify both ownership and purpose, but require content analysis to determine what type of entity appears to actually be using the domain and for what purpose.

Domains initially placed into class 3 (natural persons) and classes 6 and 7 (other non-business organizations) may or may not fail to clearly identify ownership and/or purpose; these domain names require content analysis to determine whether Registrant Name and Organization accurately reflect the domain's actual user and purpose.

Domains initially placed in class 8 (Privacy/Proxy) must be confirmed or reclassified using the methodology defined by the WHOIS Privacy/Proxy Prevalence Study [5]. All confirmed Privacy/Proxy-registered domains require content analysis to determine what type of entity appears to be actually using the domain and for what purpose. In this study, any Privacy/Proxy-registered domain that appears to be used by a legal person (for any purpose) fails to clearly and directly identify ownership, while any Privacy/Proxy-registered domain that appears to support commercial activity (by any person) fails to clearly and directly identify purpose. While these domains may in fact comply with existing registration policies, they are being studied to better understand how often and why such Privacy/Proxy uses occur.

After Apparent Registrant Type classification (including Privacy/Proxy confirmation) has been completed, the following input data will be available for each sampled domain:

- Domain name
- Registrant Name
- Registrant Organization
- Full WHOIS record for the domain
- Apparent Registrant Country Code/Name

- Apparent Registrant Type
- Privacy/Proxy Service confirmation (for domains in class 8)
- Protected Name/Organization (for domains registered via Privacy service)

The remainder of this study uses this input data to investigate, quantify, and categorize all domain names that might not clearly identify legal person ownership and/or commercial purpose, generating the output described in the Section 4.

4. Outputs

Study results will provide a breakdown of domains registered to natural vs. legal persons and domains used for commercial vs. non-commercial purpose, distributed by gTLD, geographic region, and Privacy/Proxy use. For domains apparently used by legal persons, this study will quantify how often WHOIS Registrant fails to clearly and directly identify that legal person. For domains apparently used for commercial purpose, this study will quantify how often WHOIS Registrant identifies an entity more commonly engaged in non-commercial activities. Incidents will be categorized and illustrated to help the ICANN community understand how often and why ownership and purpose are not clearly identified in WHOIS and correlation to Privacy/Proxy use.

To deliver these empirical results, this study will attempt to find and analyze Internet content associated with all sampled domains that have missing, fictitious, ambiguous, misleading, or obscured WHOIS Registrant Organization values. Domains shall be found to have commercial purpose if any analyzed content (e.g., website) is engaged in commercial activity. Domains shall be found to be used by a legal person if any of that analyzed content is published (i.e., posted or emailed) by a legal person.

As suggested by proposal [7], this study will use DNS to resolve each domain name, locate publicly-addressed Internet server(s) within the domain, and analyze public content posted there. In particular, first-level public website(s) associated with each domain will be visited to categorize observed commercial activity as follows.

- Any website offering commercial transactions (i.e., purchases), whether benefiting a for-profit entity, a non-profit charitable entity, or a natural person, is engaged in a form of commercial activity. Most on-line storefronts and auction websites fall into this category.

- Any website without commercial transactions which promotes for-profit goods or services is engaged in another form of commercial activity. Most registered business websites fall into this category, including privacy-sensitive businesses like medical clinics. In contrast, personal websites and many civil organization websites usually do not promote for-profit goods or services.

- Any website without commercial transactions or for-profit promotion which generates significant revenue by advertising other sites where transactions are made or goods or services are promoted is engaged in another form of commercial activity. Most ad-supported on-line publications and search engines fall into this category.

- Any website engaged in largely non-commercial activity but containing a modest ad such as a Google ads frequently posted on personal websites is still engaged in commercial activity – albeit sufficiently different to be assigned its own category.

- Any website explicitly linked to a pay-per-click ad page is engaged in a form of commercial activity, no matter who supplies that page (e.g., the domain owner, a domain reseller, a web hosting company, an ISP). However, this category shall NOT include cases where a third-party DNS server simply redirects all unresolved domain names to a generic ad page, since that content is not explicitly linked to the domain.

- Any website that offers no tangible content (e.g., a generic domain parking page, a custom "under construction" page, or an unresolved link) must be categorized using another method or (in the absence of available Internet content) treated as non-commercial. This category includes domains purchased for resale or lease but not (currently) linked to Internet content of any sort.

These categories are proposed as a starting point, to be refined during the study, keeping in mind that commercial purpose [1] does not depend on type of entity but rather the type of activity. Although web content classification is likely to remain somewhat subjective, category criteria and examples should be documented with sufficient rigor to enable consistent repetition. Sites that prove especially difficult to classify may be assigned to an "other" category for further study. Additionally, given that websites are updated over time, the key content upon which classification decisions are made should be recorded (along with a date/time stamp) for future reference.

Website content will also be used to identify the domain's actual user, which may be the Registrant itself, a Proxy-registered domain licensee, or a third-party using the domain with or without the Registrant's knowledge or permission. In many cases, the actual user (individual and/or organization) will be explicitly identified in a copyright statement or an "About us" or "Contact us" page posted on the website.

As noted in Section 2, discrepancies between WHOIS Registrant Name and Organization and the domain's actual user will be catalogued and categorized, with the goal of finding domains being used by legal persons without clear ownership identification. For example, a domain used by John Doe's business (a legal person) may be registered to John Doe (a natural person) or a Privacy service. This study treats these as different categories of ownership to give the ICANN community empirical data as to where and why such uses occur. However, it is beyond this study's scope to determine whether some legal persons have legitimate reasons for obscuring ownership (e.g., websites promoting free speech).

Any domain that can be categorized as used by a legal person and/or engaged in commercial activities based upon its website(s) alone does not require further content analysis. This is expected to be the case for most registered domain names – including parking pages that ISPs typically provide for domains registered by natural persons.

However, domains not associated with any website may still be used by legal persons to originate electronic communication, including unsolicited commercial email (spam). Thus, any domains for which actual user or purpose cannot be analyzed by website inspection should be investigated by other methods. Methods are not fully specified here and should be developed and documented during the study, based upon analysis of domains that prove difficult to categorize. Suggested methods include the following.

- Using third-party directory sites like SiteAdvisor, DomainTools, and AboutUs to obtain information about a domain name that has no currently-active website might reveal its actual user and purpose. For example, directories may return contact information, logos, and thumbnail images of related websites, captured at an earlier date. These directories may also be useful to safely preview domain websites prior to web content analysis. However, directories that simply display contact information obtained from WHOIS cannot be used to verify the domain's actual user.

- Using Internet search engines and databases to look for traces of email activity might reveal the domain’s purpose. For example, a domain found in a database like Spamhaus that tracks reported spam (unsolicited commercial bulk email) might be categorized as commercial, while a domain that only appears in email messages with no apparent commercial purpose may be treated as non-commercial.

- Using social networking sites to search for domain names might reveal whether the entity using that domain is a natural or legal person. For example, content found on LinkedIn or Facebook may be associated with individuals, organizations, or businesses, and may be accompanied by text descriptions and links that further reflect the domain's purpose (e.g., domains used by natural persons for personal email).

- Using DNS results to query other (non-web) public servers and associated IP address blocks might reveal whether the entity using the domain is legal person. For example, natural persons rarely own Class A or B IP address blocks, but many natural persons use a dynamic DNS service to associate a personal website with a single broadband service provider-owned IP address. Many public-facing business file or VPN servers return banner text at connect time identifying the server's owner. However, such queries shall not include probing the content of private networks or servers.

- Access to certain websites (e.g., drive-by malware sites, adult content sites) may be blocked or filtered by firewalls used to defend research systems from retrieving harmful or illegal content. In such cases, firewall logs, redirection responses, and alternative analysis methods may be used to learn more about the offending website and its content. For example, URL filtering databases usually classify blocked sites by type and identity. Malicious sites that have been taken down might be found by searching a directory like PhishTank to view earlier website images. However, care must be taken to avoid filtering out otherwise-harmless ads required to correctly categorize domain purpose.

After Internet content has been inspected and categorized for each sample domain requiring further analysis, the following raw data will have been produced:

- Sample set of domains (filtered by gTLD, country, and apparent type)
- For each domain requiring content analysis, the following additional outputs:
    - Actual Domain User Name and Organization
    - Does the actual user of the domain appear to be a legal person?
    - Does the actual purpose of the domain appear to be commercial?
    - Was legal ownership or commercial purpose ruled out?
    - If actual user or purpose could not be determined, why? (e.g., offline)
    - Category of Legal Person Ownership (if any)
    - Category of Commercial Purpose (if any)
    - List of analyzed Internet content (e.g., websites, servers, directories)
    - Critical data or images archived from these content sources which played a key role in confirming and categorizing domain ownership and purpose

This raw data will be used to create the following statistical summaries, including the following metrics:

- Percentage of Registrants apparently classified as natural persons vs. legal persons, categorized by gTLD, region/county, and Privacy/Proxy use
    - Percentage of domains actually used by legal persons without clear identification in WHOIS, categorized by gTLD, region/county, Privacy/Proxy use, and type of owner
- Percentage of domains apparently registered for commercial vs. non-commercial purposes, categorized by gTLD, region/county, and Privacy/Proxy use
    - Percentage of domains actually used for commercial purposes without clear identification in WHOIS, categorized by gTLD, region/county, Privacy/Proxy use, and type of purpose

Note: This study addresses questions posed by GAC data sets [8][9][10][11] and proposed studies [6][7]. However, this study does NOT focus exclusively on domains registered through Privacy/Proxy services as proposed by [7] – examining those domains are a primary goal of this study, but not the only goal. Nor does this study examine the general accuracy of WHOIS Registrant data or overall frequency of Privacy/Proxy service use, as those questions were already studied by [4] and [5]. However, Privacy/Proxy prevalence should be considered when determining sample size to ensure that enough Privacy/Proxy-registered domains will be included for statistical relevance.

5. References

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

[2] Noncommercial Users Constituency (NCUC) Charter, NCUC, August, 2003

[3] .BIZ Agreement: Appendix 11, Registration Restrictions, ICANN, December 8, 2006

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

[5] ICANN's Study on the Prevalence of Domain Names Registered using a Privacy or Proxy Service among the top 5 gTLDs, ICANN, September 28, 2009

[6] Study Suggestion Number 13a, Measure growth of proxy/privacy services vis-à-vis all registrations, Laura Mather

[7] Study Suggestion Number Study 18, Measure percentage of domains registered using proxy/privacy services that are natural/legal persons, or used for a commercial purpose, Claudio DiGangi

[8] GAC Data Set 5, Measure percentage of registrations who are natural vs. legal persons, GAC Recommendations for WHOIS Studies, 16 April 2008

[9] GAC Data Set 6, Measure percentage of registrations used for a commercial vs. non-commercial purpose, GAC Recommendations for WHOIS Studies, 16 April 2008

[10] GAC Data Set 9, Relative percentages of legal persons and natural persons that are gTLD Registrants utilizing proxy or privacy services, GAC Recommendations for WHOIS Studies, 16 April 2008

[11] GAC Data Set 10, Relative percentages of domain names used for commercial versus non-commercial purposes registered using proxy or privacy services, GAC Recommendations for WHOIS Studies, 16 April 2008

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

[13] Terms of Reference for WHOIS Misuse Studies, ICANN, September 2009