Contents

1. Objective 1
2. Approach 2
3. Inputs 5
4. Outputs 6
5. References 10

Terms of Reference for WHOIS Registrant Identification Studies

This exploratory study will measure and classify:

(1) The types of entities that register domain names, as identified in WHOIS,
(2) The types of entities that appear to be using those domain names,
(3) The types of commercial activities associated with each domain name, and
(4) How these types correlate to use of Privacy and Proxy registration services.

1. Objective

This exploratory study will examine WHOIS data for a representative sample of gTLD domain names, using WHOIS Registrant Name and Registrant Organization values to classify the types of entities that register domains, including natural persons, various kinds of legal persons, and Privacy and Proxy service providers.

This study will then analyze available Internet content associated with each sampled domain name to classify the types of entities that appear to be using those domains and the various types of potentially commercial activities associated with them.

Finally, this study will analyze inter-relationships between these categories, seeking to provide a foundation for answering the following questions posed by GAC:

- Percentage of registrants that are natural versus legal persons [8]
- Percentage of domain name uses that are commercial versus non-commercial [9]
- Relative percentage of Privacy/Proxy use among legal persons [10]
- Relative percentage of Privacy/Proxy use among domains with commercial use [11]

Entity and commercial activity classifications will be further developed during the study, based on sampled data, to help the ICANN community better understand the wide variety of possible correlations that may emerge and their potential implications on policy.

For example, reference [1] defines a *natural person* as a real, living individual, as opposed to a *legal person* which may be a company, business, partnership, non-profit entity, or trade association. Legal persons may have a wide variety of reasons for registering domains name or utilizing Privacy and Proxy services. Mapping Registrant types into more granular categories will thus better inform policy decisions.

Reference [1] defines *commercial use* as 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* cited by [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. Furthermore, according to [3], registering a domain name solely for the purposes of selling, trading, or leasing that name does *not* constitute "bona fide business or commercial use."

Given this wide variety of potentially commercial activities, as well as local differences in the formal legal definition of "commercial activity", mapping sampled domain uses into more granular categories such as those described above may uncover relationships between specific activities, entity types, and Privacy/Proxy registrations that would be lost if all domain names associated with any of these activities were treated as a single category by this study. Instead, this study will supply sufficiently granular data to support post-study interpretation of "commercial activity" in each country.

Some domain names use Proxy and Privacy registration services [1] to provide anonymity and privacy protection for domain name users. *Privacy* services offer alternate WHOIS contact information and mail forwarding services while not actually shielding the Registered Name Holder's identity. *Proxy* services register domain names on a third party's behalf and then license their use so that the provider's identity and contact information (and not the licensee's) is published in WHOIS. According to ICANN's Compliance Department, obtaining the actual user's identity during any study would be likely for Privacy registrations, but not for domains registered by a Proxy provider.

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 privacy of contact information. However, ICANN's Registrar Accreditation Agreement [12] does not restrict Privacy/Proxy service use to certain types of Registrants, nor does it require Registrants to identify domain purpose. This study can help the ICANN community better understand how often various types of Registrants use Privacy/Proxy services and why, providing the empirical input and context needed to consider related policy changes.

2. Approach

This empirical data will be obtained by conducting an exploratory study which starts by classifying a representative sample of gTLD domain names by apparent Registrant type. All sampled domains, including those using Privacy or Proxy services, will then be further analyzed based upon the content of websites and other Internet data associated with each domain. To address the questions raised by proposals [6][7] and GAC data sets [8][9][10][11], all domain names subjected to this content analysis will then be categorized according to entity type, commercial activity, and Privacy/Proxy use (distribute by gTLD and country/region).

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 Name Type (e.g., person, organization, multiple domain holder, Privacy/Proxy service). 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 (which measured 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 Registrant 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 further categorize domain names based upon the types of entities that appear to be using them and the types of potentially commercial activities associated with them:

- **Type of Entity:** WHOIS Registrant Name/Organization does not always clearly or directly identify the entity using a domain. For example, WHOIS Registrant data may be missing or identify a Privacy/Proxy service's Name/Organization. A domain name may also be registered by a web site developer or a small business owner or another third party. As such, this study will attempt to identify not only the type of entity that *registered* the domain, but the type of entity that appears to be *using* each sampled domain.

- **Type of Activity:** A domain name may be used to engage in a wide variety of potentially commercial Internet activities (i.e., solicit the exchange of goods, services, or property). This study will attempt to classify potentially commercial activities (if any) associated with each sampled domain so that inter-relationships can be examined between those uses and WHOIS Registrant types (including Privacy/Proxy service providers). Note that this includes revenue-generating ads posted on "parking pages" explicitly linked to registered but otherwise unused domain names. Given local differences in the legal definition of "commercial activity," this study will attempt to document all activities that might be considered commercial in some countries.

As detailed in section 4 (Outputs), these classifications are related but distinctly different. For example:

a) A domain with Registrant Name=John Doe (a natural person) may resolve to a website used by ABC Corp (another type of entity – a corporation) to describe products sold by ABC (one type of activity).

b) A domain with Registrant Organization=Community XYZ (a non-profit organization) may resolve to a website used by that same entity for many purposes, including on-line fund-raising activities (a type of activity).

c) A domain with Registrant Name=John Doe (a natural person) may resolve to a website used by Human Rights 123 (another type of entity) that has no apparent commercial content.

d) A domain with Registrant Name= Proxy123 (a Proxy service provider) may resolve to a website used by ABC Corp. (another type of entity – a corporation) to host pay-for-click ads (a type of activity).

Note that the Registrant's *intent* is not readily or reliably discernable from WHOIS data. For example, John Doe may have registered a domain under his own Registrant 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 must identify and classify all of these possibilities (and more) because they may have different implications for the ICANN community and WHOIS policy.

To deliver useful results without subjectively guessing the Registrant's intent, this study will categorize domain name entities and uses based upon information obtained from multiple sources (e.g., websites, spam URL lists, Internet search engines, individual and business directories). Actual categories and mapping criteria are not specified here but will be developed during the study to clearly describe and unambiguously differentiate between common scenarios and inter-relationships encountered in sampled data.

3. Inputs

The first step in this study is to assign an **Apparent Registrant Type** to each sampled domain name using the classes defined by the final 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 and Organization are incomplete, unable to classify
4. Registrant Organization appears to be a Privacy/Proxy registration service
5. Registrant Organization appears to be a multiple domain name holder
6. Registrant Organization is specified; person is also named
7. Registrant Organization is specified; no person is named
8. Registrant Name appears to be a natural person; no organization is named

Domains placed into classes 1, 2, or 3 (missing, patently false, incomplete) may not have a discernable WHOIS Registrant type but still require content analysis to determine what type of entity appears to be using the domain and for what potentially commercial activities (if any).

Domains initially placed in class 4 (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 using the domain and for what potentially commercial activities (if any). While these domains may in fact comply with existing registration policies, they are being studied to understand how often and why Privacy/Proxy services are used.

Domains initially placed into class 8 (natural persons) and classes 5, 6, and 7 (legal persons) also require content analysis to determine whether Registrant Name and Organization match the domain's actual user and what potentially commercial activities (if any) are associated with these domains.

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 4)
- Protected Name/Organization (for domains registered via Privacy service)

The remainder of this study uses this input data to investigate, quantify, and categorize all sampled domain names based upon the type of entity apparently using the domain and type(s) of potentially commercial activities, generating the outputs described in the Section 4.

4. Outputs

Study results will provide a breakdown of domains registered to and used by various types of entities (including Privacy/Proxy services), for various types of potentially commercial activities, distributed by gTLD and geographic region. For domains apparently used by various types of legal persons or associated with various types of potentially commercial activities, this study will also quantify how often they are registered using Privacy/Proxy services. Furthermore, inter-relationships between all categories will be analyzed and illustrated to help the ICANN community better understand how domains are identified in WHOIS Registrant data, including the percentage registered to legal persons [8], the percentage associated with potentially commercial uses [9], and relative frequency of Privacy/Proxy service use [10][11].

To deliver these empirical results, this study will attempt to find and analyze Internet content associated with all sampled domains, looking for data that might be used to classify potentially commercial activity (if any) and the identity of the actual domain user (which may or may not be the domain name Registrant).

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 all observed monetary activities that in some countries might be legally considered "commercial activities." The goal is to document a relatively broad range of potentially commercial activities, at a level of granularity sufficient to enable multiple post-study interpretations that apply varied legal definitions. Examples of activities that should be documented by this study include the following:

- Monetary transactions (i.e., purchases) offered by a website, without regard to the type of entity registering the domain or using the website. For example, most on-line

storefronts and auction websites let visitors purchase goods or services. Many non-profit websites let visitors pay membership dues or make a donation. Each of these activities should be separately documented as a distinct category.

- Promotion of for-profit goods or services by a website. For example, most registered business websites promote their own goods or services that are paid for off-line. These too shall be separately documented as distinct categories.

- Paid advertisement of third-party websites where transactions are offered or goods or services are promoted. For example, most ad-supported on-line publications and search engines exhibit this activity. Although researchers cannot quantify the revenue generated by ads, results should differentiate between sites that display an occasional third party ad and those that continually present extensive third party ads.

- Modest paid advertisements posted on what otherwise appear to be personal websites, such as the Google ads often displayed by personal blogs. This shall be documented as yet another type of potentially commercial activity – albeit sufficiently different to be assigned its own category.

- A pay-per-click ad page linked to a website, no matter who supplies the content of that page (e.g., the domain owner, a domain reseller, a web hosting company, an ISP). However, this activity shall be differentiated from 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 analyzed using another method or (in the absence of available Internet content) found to have no discernable commercial activity. Examples include 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 based on sampled data, keeping in mind that commercial uses [1] do not depend on type of entity but rather 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 and with sufficient granularity to support varied legal interpretations of "commercial activity." 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, the type of entity using each domain will be catalogued and categorized, along with possible relationship to the entity identified by WHOIS Registrant Name and Organization. For example, a domain used by John Doe's business (a type of legal person) may be registered to proprietor John Doe (a natural person) or a Privacy service. This study maps these entities into different categories to give the ICANN community empirical data about how Registrants and domain users are identified in WHOIS. It is beyond this study's scope to determine whether various entities have legitimate reasons for using Privacy or Proxy services.

Note that domains not associated with any website may still be used to originate electronic communication, including unsolicited commercial email (spam). Thus, any domains for which actual user or activity 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 such as SiteAdvisor, DomainTools, and AboutUs to obtain information about a domain name that has no currently-active website might reveal its actual user and association with commercial activities (if any). 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 analysis. However, directories that simply display contact information from WHOIS cannot be used to identify the domain's actual user.

- Using Internet search engines and databases to look for traces of email activity might reveal potentially commercial activities associated with a domain. For example, a domain found in a database such as Spamhaus that tracks reported spam (unsolicited commercial bulk email) might be categorized as exhibiting this type of potentially commercial third party advertisement activity.

- Using social networking sites to search for domain names might reveal the type of entity using that domain. 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 how the domain is being used (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 likely a natural 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, this study shall not query private data stored on 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 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 such as PhishTank to view earlier website images. However, care must be taken to avoid filtering out ads required to correctly identify and categorize potentially commercial activities (if any) associated with such domains.

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

- Sample set of domains (filtered by gTLD, country, and apparent Registrant type)
- For each sampled domain name, the following additional outputs:
  - Actual Domain User Name/Organization, Type of Entity, and possible relationship(s) to Apparent Registrant
  - If actual user could not be determined, why? (e.g., offline)
  - Type(s) of Activities associated with domain (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 categorizing entity type and commercial activities

This raw data will be used to create statistical summaries that quantify gTLD domain registration distribution by type of Registrant and actual user, types of potentially commercial activity, and relative frequency of Privacy/Proxy use. These summaries should attempt to address the specific questions posed by GAC data sets [8][9][10][11] regarding legal persons and commercial use. However, they should also dig deeper by breaking these types into more granular categories and analyzing possible correlations between various types of entities and commercial activities and how those domain Registrants are identified in WHOIS. In particular, it is beyond the scope of this study for researchers to determine what does or does not constitute "commercial activity" across the globe. Rather, researchers shall deliver the data needed by policy makers to apply varied legal interpretations of "commercial activity" that are applicable in each country.

Note that 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] Draft Report for the Study of the Accuracy of WHOIS Registrant Contact Information (6558,6636), NORC, January 17, 2010

[5] ICANN's Study on the Prevalence of Domain Names Registered using a Privacy or Proxy Service among the top 5 gTLDs, ICANN, September 14, 2010

[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