Draft Report for the Study of the Accuracy of WHOIS Registrant Contact Information

Developed by NORC at the University of Chicago
for ICANN

NORC Project Reference: 6558, 6636

17 January 2010

Contents

Executive Summary

WHOIS services are intended to provide free public access to information about the registrants of domain names. The information displayed is that obtained from the registrant at the time they registered the site, or the latest update of that information that they have provided to the registrar of their domain name.

There have been concerns about the accuracy of the information in WHOIS for some time, although the actual extent of the problems is not known. In 2005, GAO conducted a study which looked at the prevalence of missing or patently false information, and found that nearly 5% of WHOIS records in the top three gTLDs (.com.org. .net) had missing or patently false information in the registrant name and address fields. The extent to which information which *appeared* complete but was in fact inaccurate was not addressed.

This study was commissioned by ICANN in order to get a baseline measurement of what proportion of WHOIS records are accurate. The scope was limited to the quality of the information provided about the registrant (as opposed to the administrative or technical contact), since it is the registrant who has entered into a legal arrangement with the registrar for the domain name.

Under Registrar Accreditation Agreement Section 3.3.1.6, an accurate name and postal address of the registered name holder means there is reasonable evidence that the registrant data consists of the correct name and a valid postal mailing address for the current registered name holder. Adapting this for the study, there were three criteria to be met for any WHOIS record to be considered accurate:

1. Was the address of the registrant a valid mailing address?
2. Was the registrant named associated in some way with the given address?
3. When contacted, would the named registrant acknowledge that they were indeed the registrant of the domain name, and confirm all details given as correct and current?

An internationally representative sample of 1419 records was drawn from the top five generic top level domains (gTLDs, covering .com, .org, .net. .info and .biz). The address for each selected case was checked against postal records and mapping data for deliverability, searches were conducted in phone listings and other records unrelated to WHOIS for a linkage between name and address, and contact was attempted with the named registrant using phone numbers obtained during the association process.

Using strict application of the criteria, only 23% of records were fully accurate, but twice that number meet a slightly relaxed version of the criteria (allowing successful contact with the registrant to imply association, and requiring only that ownership of the site be confirmed, as opposed to confirmation of both ownership and the currency/correctness of all detail). Eight percent of records failed outright with obvious errors. The table on the following page gives more detail, the findings on the remainder, and limitations.

There is no question that there are people who register domains without disclosing their full or real identity. While we didn't find any cases where an identity had been stolen (that is, among the persons we contacted who had domains registered in their name, none denied having registered the domain), it would seem that, given the latitude that people have in choosing what information to provide when registering a domain name, identity theft may not be necessary; it is all too easy for registrants to enter any or no name, along with an unreliable or undeliverable address.

Most of the barriers to accuracy found (concerns about privacy, confusion about information needed, lack of clarity in the standard to which information should be entered, no requirement for proof of identity or address, the structure of WHOIS itself) can be addressed by the internet community. However the cost of ensuring accuracy will escalate with the level of accuracy sought, and ultimately the cost of increased accuracy would be passed through to the registrants in the fees they pay to register a domain. Cooperation among all registrants and other ICANN constituents will be needed to eliminate any commercial disadvantage accruing from enforcing greater accuracy.

Accuracy group	**Description of accuracy (1),(2)**	Unweighted frequency counts (3)	Population estimates	Estimated percentage	Margin of error (4)
No failure	Met all three criteria fully - deliverable address, name linked to address, and registrant confirmed ownership and correctness of all details during interview	353	23,117,442	22.8%	1.4%
Minimal failure	All criteria met but minor fault noted by registrant during interview	17	1,101,176	1.1%	0.2%
	Name unable to be linked to address, but able to locate registrant and confirm ownership	312	23,024,007	22.7%	2.2%
Limited failure	Deliverable address, name linked and/or located, but unable to interview registrant to obtain confirmation.	365	24,893,476	24.6%	1.7%
Substantial failure	Undeliverable address and/or unlinkable name, however registrant located. Unable to interview registrant to obtain confirmation.	109	7,202,472	7.1%	0.9%
	Deliverable address, but unable to link or even locate the registrant, removing any chance of interview.	177	13,949,721	13.8%	2.2%
Full failure	Failed on all criteria - undeliverable address and unlinkable, missing, or patently false name, unable to locate to interview	86	7,937,694	7.8%	1.8%
All domain names in top five gTLDs		**1,419**	**101,225,988**	**100%**	

(1) Definitions:

- *Unable to link: means unable to find any independent association between name and address, or name and/or address missing*
- *Unable to locate: means unable to get confirmed current phone contact information for named registrant*

(2) Limitations:

- *Failure on the linkage criteria could be caused by a concern with privacy (e.g. by having an unlisted phone number and not having name and address listed together in any readily accessible sources other than WHOIS)*
- *Failure on the confirmation criteria could be caused by refusal or inability to cooperate with the survey for reasons unrelated to the accuracy of their WHOIS record.*

(3) Each record is listed only once, against the most severe failing for that record.

(4) Margin of error is calculated on the basis of a 95% confidence interval, which is approximately the estimated percentage plus or minus the margin of error.

Introduction

The following explanation of WHOIS is provided on the ICANN website:

> WHOIS services provide public access to data on registered domain names, which currently includes contact information for Registered Name Holders.
>
> The extent of registration data collected at the time of registration of a domain name, and the ways such data can be accessed, are specified in agreements established by ICANN for domain names registered in generic top-level domains (gTLDs).
>
> For example, ICANN requires accredited registrars to collect and provide free public access to the name of the registered domain name and its nameservers and registrar, the date the domain was created and when its registration expires, and the contact information for the Registered Name Holder, the technical contact, and the administrative contact.

There has however been concern about the accuracy of the data for some time, and as a result ICANN commissioned NORC to design a study to assess the accuracy of WHOIS entries.

In 2005, the GAO conducted a study related to accuracy by determining the prevalence of "patently false" or incomplete contact data in the WHOIS service for the three largest gTLDs: .org, .net, and .com. While we have replicated part of this study in the conduct of the current one (the results are shown in the appendix), the GAO study involved only coding from the data as displayed in WHOIS, and picked up only the most obvious errors. A false name or address can appear compete, but it is only on checking it against other listings and in attempting to make contact that it might be revealed as false.

This aspect is the key difference between the GAO study and the current one. This study seeks to go several steps further in checking the accuracy of registrant information, including contacting the named registrant to confirm that they were indeed the registrant, and not, for example, a victim of identity theft.

Sample Design

A sample of 1419 domain names clustered among 16 countries was used in this study. Appendix 1 describes the sample design in detail, and key elements are repeated here.

According to the April 2008 Registry Operator Monthly Reports (Jan-Mar, 2008 for .aero) at http://www.icann.org/en/tlds/monthly-reports/, there were 15 global Top-Level Domains (gTLDs) with active Domains.

Table 1 below shows the total number of domains among all 15 gTLDs. Excluded from these gTLDs are .edu, .mil, and .gov, which were deemed out of scope due to the higher level of control (and thus accuracy) involved in registration of domains within those three gTLDs

Table 1. Summary of global Top-Level Domains (gTLDs) of interest to ICANN

Rank	Top-Level Domain	Total Domains	Percentage of Domains	Cumulative Percentage	Included in WHOIS Accuracy Project?
1	**.com**	**75,785,462**	**73.7%**	**73.7%**	**Yes**
2	**.net**	**11,478,837**	**11.2%**	**84.9%**	**Yes**
3	**.org**	**6,840,493**	**6.7%**	**91.5%**	**Yes**
4	**.info**	**5,092,053**	**5.0%**	**96.4%**	**Yes**
5	**.biz**	**2,029,143**	**2.0%**	**98.4%**	**Yes**
6	.mobi	903,941	0.9%	99.3%	No
7	.name	287,442	0.3%	99.6%	No
8	.travel	201,047	0.2%	99.8%	No
9	.asia	159,682	0.2%	99.9%	No
10	.cat	29,230	0.0%	100.0%	No
11	.jobs	13,279	0.0%	100.0%	No
12	.pro	7,994	0.0%	100.0%	No
13	.coop	5,861	0.0%	100.0%	No
14	.aero	5,414	0.0%	100.0%	No
15	.museum	528	0.0%	100.0%	No
	TOTAL	102,840,406	100.0%		

Because this study was to involve direct contact with the sampled registrants, and registrants in this universe come from all countries on the globe, it would have been cost prohibitive to use a totally unrestricted sample. Instead, the sample was clustered by country so the number of countries (and languages and systems) involved would be restricted to a manageable count.

However the country of the registrant is not readily available for the purposes of sampling[1], and thus a systematic random sample of 2400 records was pulled ("the microcosm"), the corresponding WHOIS records were extracted for each of these 2400 domains, and the country of the registrant coded for each directly from the WHOIS record.

Because of the systematic way the 2400 records for the microcosm were selected, the numbers of domains that came from each country in the microcosm is closely proportional to the number of domains that country has among the top five gTLDs. The countries were then ordered by number of domains, and four strata formed, with allocation to strata based on the number of domains. Countries were then sampled from within the strata, and domains sampled from within selected countries.

For cost efficiency, the sample was structured to pull slightly more domains from the strata which contained the countries with the most numbers of domains, with the final results weighted to adjust for this (e.g. in the final results, domains from countries in the "small" stratum carry a higher weight than domains from the "certainty" stratum). The final counts by country and strata are shown in table 2.

[1] We explored with ICANN the possibility of using IP address and Maxmind.com to assign countries to all domains prior to sample selection in order to avoid this additional stage of sampling, however once we checked the country assignments using those sources against actual country as recorded in a sample of WHOIS records we found too many inaccuracies.

Table 2. Countries selected in the sample

Country	Strata	Countries represented	Domains in sample
United States	Certainty	Themselves only	928
Canada			77
United Kingdom			71
Germany			61
China			49
France	Large (17-52 domains each in the microcosm)	Themselves plus Italy, South Korea, India, Portugal	35
Australia			35
Netherlands			35
Japan			35
Spain			31
Turkey			23
Sweden	Medium (6-16 domains each in the microcosm)	14 others, including Brazil, Switzerland, Hong Kong, Saudi Arabia, Poland	13
Russian Federation			10
Malaysia			8
Singapore	Small (1-5 domains each in the microcosm)	All others	5
Israel			3
Total			1419

Because gTLD is an intrinsic component of all domain names, the population was able to be stratified according to gTLD prior to sample selection, and so maintain a strictly proportional relationship across the five gTLDs included in the study. This can be seen in Table 3:

Table 3. Distribution by gTLD

gTLD	Domain counts			Percentages		
	Universe	Microcosm	Sample	Universe	Microcosm	Sample
.com	75,785,462	1801	1066	75%	75%	75%
.net	11,478,837	273	162	11%	11%	11%
.org	6,840,493	167	102	7%	7%	7%
.info	5,092,053	114	64	5%	5%	5%
.biz	2,029,143	45	25	2%	2%	2%
Total	101,225,988	2,400	1,419	100%	100%	100%

Accuracy definition and assessment

To allow verification to be done to a consistent standard, ICANN determined the following definition for Accuracy of a WHOIS entry, based on the contractual requirements between Registrar and Registrant.

> Under Registrar Accreditation Agreement Section 3.3.1.6, an accurate name and postal address of the registered name holder means there is reasonable evidence that the registrant data consists of the **correct name** and a **valid postal mailing address** for the current registered name holder.

Further clarification of this definition was also provided:

- The name of the Registered Name Holder is "correct" if the WHOIS data identifies the actual organization or individual that has consented to and entered a registration agreement with the registrar (even though the registration might have been arranged by or created for the benefit of a third party)
- The postal mailing address is "valid" if it accurately identifies a functioning destination or postal mail that has been designated by the Registered Name Holder. There is no requirement that the address be the primary residence of an individual or the headquarters of an organization. A valid mailing address could be a post office box or the address of a mail forwarding service arranged by the registrant or the registrar of a third party. The elements and format of the mailing address may vary by country and territory, but they should at least be sufficient to be used as an international address and must comply with the recommendations of the postal authority of the country of the registrants designated address.

For a WHOIS entry to be deemed completely accurate under this definition, the following measurable criteria must be met:

1. The address must be found to be deliverable.
2. The name of the registrant must be known or associated at the given address
3. When contacted, the registrant confirms they consented to the registration

The following steps were undertaken to assess these criteria.

Criteria 1: Deliverability of the mailing address

The address given for the registrant was first coded for type, using the following categories:

Address type	Unweighted frequency counts	Population estimates	Estimated percentage	Margin of error
1_Address completely missing	14	940,491	0.9%	0.3%
2_Address patently false	8	674,086	0.7%	0.6%
3_Partial - no detail below city or state	29	2,695,264	2.7%	2.2%
4_Street or physical address	1,187	84,941,486	83.9%	4.2%
5_Postal service address	181	11,974,662	11.8%	4.0%
Total domain names	1,419	101,225,988	100%	

Missing addresses included those who gave only an email address and not a postal address, as well as those where "N/A" or similar had been written in the address field.

Patently false addresses were obviously false – for example:

1 Mucky Road, Mucksville, Muckland MU11CK UK
PRIVATE, XXXXX, XXX, 99999

Addresses where no detail was included below the city/town level (for example, no street name) were coded as Partial, with the exception of those from such small towns that it could be feasible that no finer level of detail was needed.

Missing, false and partial addresses were directly coded as undeliverable. The remainder went through to additional checking for deliverability. For addresses within the US, we used the Smartmailer software which checks against USPS records of deliverable addresses, including presence of apartment number or similar detail for apartment buildings, valid ranges for street numbers, and all valid street-city-state-zip combinations. For other countries we used online address confirmation systems and mapping systems to confirm the existence of the address and the consistency of street-city-state-postcode combinations.

Depending on the presence and type of error found, we coded into Deliverable, Potentially deliverable, or Undeliverable. Anything with minor errors but which might be easily resolved (e.g. a missing apartment number, or a mismatching zip for an otherwise valid street-city-state combination, or a very easily corrected spelling error) we coded into Potentially deliverable. The cases that were coded as Undeliverable (apart from those with missing, partial or patently false detail) were in most cases ones where there was an outright mismatch between street and town, or the street number given was outside the range of the street.

Address Deliverability (Ordinal)	Unweighted frequency counts	Population estimates	Population percentage	Margin of error
1_Deliverable as given	1,163	79,087,682	78.1%	5.1%
2_Potentially deliverable	109	8,716,870	8.6%	2.3%
3_Undeliverable	146	13,421,437	13.3%	4.2%
Total domain names	1,419	101,225,988	100%	

To create a binary classification, we grouped "potentially deliverable" with "deliverable", on the basis that small errors are often corrected along the way by post offices, depending on the policies of the particular postal service, the volume of mail being handled at the time, and the individual staff handling the mail.

Address Deliverability (Binary)	Unweighted frequency counts	Population estimates	Population percentage	Margin of error
Deliverable	1,273	87,804,551	86.7%	4.2%
Undeliverable	146	13,421,437	13.3%	4.2%
Total domain names	1,419	101,225,988	100%	

We also created a single dimension combination of type and deliverability:

Address type and deliverability group	Unweighted frequency counts	Population estimates	Population percentage	Margin of error
1_address missing, partial or false	51	4,309,841	4.3%	2.4%
2_address complete but undeliverable	107	9,920,696	9.8%	4.1%
3_Potentially deliverable street address	98	7,872,661	7.8%	2.4%
4_Deliverable street address	997	68,111,031	67.3%	5.0%
5_Deliverable PO address	166	11,011,759	10.9%	3.8%
Total domain names	1,419	101,225,988	100%	

The proportion of PO boxes being fully deliverable is likely to be a slight overstatement from the true situation, because although there may indeed be a postal service at the address given, whether or not there is a particular box in that location of the number given cannot be checked without contacting every post service named directly. However should the ratio of deliverable to potentially deliverable postal service addresses mirror that of street addresses, we would reduce the percentage of "fully deliverable" PO boxes by only one percentage point to 10%.

Criteria 2: Association of name and address

Before we attempted association of name and address, we coded for name type. This was to assist identification of appropriate search sources in which to find an association, and to distinguish between lack of association due to an inability to even attempt an association (as would be the case with a missing or patently false name), as opposed to a lack of association due to other reasons.

Name type	Unweighted frequency counts	Population estimates	Population percentage	Margin of error
1_Name completely missing	16	1,035,321	1.0%	0.2%
2_Name patently false	10	641,860	0.6%	0.4%
3_Partial or unable to classify	18	1,844,138	1.8%	1.8%
4_Privacy/proxy service	224	14,852,653	14.7%	1.9%
5_Multiple domain name holder	125	9,064,126	9.0%	2.8%
6_Organization, person named	148	10,351,567	10.2%	2.2%
7_Organization, other	371	24,646,890	24.3%	2.1%
8_Person	507	38,789,433	38.3%	3.6%
Total domain names	1,419	101,225,988	100%	

Many cases changed classification over the course of the study, based on what we found as we tried to locate and interview registrants. For example, what might appear to be the name of a person turned out to be the name of an organization when we searched for contact information, or what appeared to be patently false was found to be an unlikely, but genuine, business name.

Specific notes about the coding of these categories follows:

- *Completely missing* – includes where the name field(s) are completely blank, or contain only a dash or a N/A annotation.

- *Patently false* – includes entries such as : ?; 9 9; self; domain admin; private; citizen; business; (where not found to be represented by a privacy or proxy service); muckimarie; (the latter when in conjunction with a patently false address, given that many businesses had similarly odd names).

- *Partial or unable to classify* – where they gave their first name only (and we were unable to rule it out as a business name, by for example, the structure of their email address or the lack of any record of any business of that name). Some of these were borderline "patently false".

- *Privacy or proxy service* – where the registrant name was that of a confirmed privacy service. To identify these cases, we took all cases from the sample which had duplicate name or addresses within the sample, had "privacy" or "proxy" or similar indicators in their name, or which had the same address as a known privacy or proxy service, then identified all the potential service providers among this group and attempted contact with each service provider to establish if they did offer a privacy/proxy service. More detail is given in appendix 3.

- *Multiple domain name holder* – any registrant with multiple domain names within the sample, or a name suggestive of a hosting or other internet service and evidence from WHOIS that they own multiple names. Unlike the privacy and proxy services, they are more likely to be the beneficial owner of the domain name (such as domain name investors). If we were unable to distinguish whether a registrant was a Multiple domain name holder or a privacy/proxy service provider (and we were unable to get a response from them to tell us either way), we classified them as a Multiple domain name holder.

- *Organization, person named* – any registrant with the name of a business or other organization (school, church) in the registrant name section, in addition to a person's name. We distinguished organizations which included a person's name from those who didn't because it gave us an additional possible linkage point. Not all organizations are legal entities, although many are.

- *Organization, other* – any registrant with the name of a business or other organization (school, church) in the registrant name section, but with no additional information such as a person's name within the business.

- *Person* – in most cases this is a two part name, or at least a surname which we found some other link to confirm it was indeed the name of a person as opposed to that of a business.

Following name classification we sought to find an independent association between the name and the address as given for the registrant. The ideal association was for a phone book listing or an equivalent standard, such as a business directory listing. However, given the prevalence of missing and partial entries in names and addresses, and the fact that PO Boxes are often a privacy shield, only a subset of the sampled cases were even candidates for such matching.

The definition of association however only requires that the person *be known* at the address – not that they *have a formal documented link* to the address. The phone listing approach has the inherent limitation that phone listings are often done in the name of one person only in the household, and many households contain members with different surnames. If Joe Smith is the registrant, but his home phone is listed in the name of his wife Mary Jones, we will not find Joe Smith in the phone book, but we will find a phone number for his address, and when we ring that address there is a good chance that Joe Smith himself will answer. Therefore we have an independent association – albeit not a direct one.

The association can even be made with a partial name. If in the previous example, Joe Smith had given only "Joe" as the registrant name, we still would have found him by ringing the number linked to his stated address. The Registrar Accreditation Agreement Section 3.3.1.6 refers to the need for an *accurate* name, but it does not require that it be a legal entity, so arguably even just the initials "JS" would be sufficient to identify this registrant at the given address.

Things get more tenuous when dealing with organizations, since there is less structure to the names. For example, one case that was initially on our "partial/potentially false" list was a registrant by the name of XMG[2], which turned out to be the initials of a registered business. We were able to establish that only by tracing through the address. The person answering the phone for that business knew what XMG stood for, even if no official business listing referred to it, and a Google search turned up hundreds of possible, but unrelated, leads.

There were other cases however for which we were *not* able to establish any satisfactory association between name and address, even though we found a phone number by which to contact the registrant interview them. Examples of such cases:

- One where the name was patently false and the address undeliverable, although not through attempts to hide but more carelessness in completing WHOIS (the domain name itself contained their full name, so under "registrant" they wrote "self"). A search on the name in the domain name to find someone at a similar enough address gave us the phone number, and on contact the person confirmed ownership.

- Registrants using a postal service address specifically for privacy reasons. There will be no link between the name and the address in such cases (unlike businesses using postal services for convenience, where a link is often apparent from them including the PO box in their official literature). However we were able to interview several of these by finding a phone number associated with the name in a close enough geographic range to the PO box.

In terms of assessing the quality of WHOIS data, cases where we were able to locate (i.e. find current contact details for) the registrant, even if we couldn't establish an association between the name and the address, would still mean the WHOIS data is arguably of better quality than the cases where we couldn't even establish if the registrant existed beyond the entry in WHOIS. As a result we ended up using a three-level classification for this second criterion:

[2] The letters are changed to protect confidentiality of the sample member.

Association of name and address	Unweighted frequency counts	Population estimates	Estimated percentage	Margin of error
Independent association found	753	50,301,048	49.7%	2.1%
No association, but able to locate registrant	424	30,648,240	30.3%	2.0%
Neither	242	20,276,701	20.0%	2.2%
Total domain names	1,419	101,225,988	100%	

Criteria 3: Registrant acknowledgement

We had several different processes for registrant acknowledgement, depending on whether the registrant was classified as:

- an individual registrant
- an organization which included a person's name
- an organization or other listing without any person's name (including those with completely missing or patently false names)
- registrants who potentially were privacy/proxy services

For the individual registrants, we attempted contact by phone, using phone numbers found through the association/locating work, as opposed to any number given in WHOIS. Persons who had a strong association found (such as a phone book listing) were often the most straight forward cases, but even among these there were challenges, such as people who had moved from the address listed in WHOIS since registering their domain. Multiple phone calls were made to establish contact, with more calls made the more certain we were that we had the correct phone number for a registrant (as might be confirmed by the wording on a voicemail message). For cases where we were less certain about the phone number, we would try it a maximum of six times, and then recommence the search for new contact details for that person. Around half the person sample was finalized on the first phone number found for them, and ten percent of the person sample took four or more phone numbers before we either found them, or deemed that we had exhausted all possible leads.

For organizations which included a person's name, we would first try any phone numbers associated with the address given, but ask for the named person on contact. Should that person no longer be with the organization, we would ask to speak with their replacement, or failing that, we would ask to speak with whoever would be responsible for registering domain names for that organization. If we hit a dead end trying to make contact through the organization or the address, we would commence searching for contact information on the person named, looking for someone of the same name in some proximity to the original address.

For large organizations with no name mentioned, we would start with the main number publically listed for the operation at the given address, and ask to speak with whoever would be responsible for registering domain names at the organization. With smaller organizations, or registrants where the name was missing, we would try any phone number we could find through a reverse search on the address.

The script used for all but the potentially privacy/proxy cases is shown in appendix 2, and basically seeks to establish whether they acknowledge registering the domain name, and :

- if ownership is acknowledged, the type of entity they are, the type of address, the accuracy of the address, and their familiarity with WHOIS;

- if ownership is denied or uncertain, the circumstances under which they think the domain name may have come to be registered in their name.

In total, 1,068 cases among the 1,419 sample were treated as an individual registrant or organization to determine Registrant Acknowledgement.

Among these 1,068, we were able to locate 797 cases – that is, we were able to have a chance of interviewing them because we found what we believe to be current working phone numbers for them, as indicated by the source of the number, the content of the voicemail message associated with the number, the caller ID, or information provided to us by someone else who answered the phone. Among these 797 cases, we were able to speak directly with the registrant in 529 cases (66%), and among these registrants, none denied ownership of the domain names.

Only a handful (35) of the located but un-interviewed cases refused outright to answer any questions. The remaining cases, all non-contacts after multiple attempts to contact them, are most likely to be either genuinely very busy people, or just generally disinclined to take a phone call from a survey company. There was nothing in these cases that led us to believe the refusals or non-contacts were related to the topic as much to a general disinterest in surveys[3].

Name type (final classification)	Domain registration determined	Able to locate but not successful in interviewing	Unable to locate to even attempt interview	Total cases handled as individual registrants
1_Name completely missing	3	12	1	16
2_Name patently false	3	3	-	6
3_Partial or unable to classify	3	11	3	17
4_Privacy/proxy service	3	1	-	4
5_Multiple domain name holder	6	6	7	19
6_Organization, person named	76	38	32	146
7_Organization, other	186	75	93	354
8_Person	249	122	135	506
Total domain names	529	268	271	1,068

All registrants who were deemed to be potentially privacy/proxy services were approached differently, in light of a separate study into the prevalence of privacy/proxy services, and because in most such cases there were multiple domain names associated with a single respondent so a "bulk processing" approach would be more time efficient for them. More information on the handling of these cases is given in appendix 3.

[3] To put this in context, in the survey industry it is generally acknowledged that most telephone surveys get more refusals than successful interviews, with most studies into the reasons for lack of cooperation finding it to be almost independent of topic and much more related to reluctance to spend time participating in a survey.

All contacts were made between June and October 2009, using experienced interviewers at NORC's offices in Chicago. International interviews were conducted with interviewers fluent in the appropriate language, with the exception of Japan and Turkey, where we brought translators on site to assist the interviewing staff.

Results

If we apply all three criteria strictly, i.e.:

1. address must be deliverable
2. an independent linkage between name and address must be found
3. the respondent must acknowledge ownership, AND confirm that all details are current and correct,

then by the strictest interpretation, only 23% of WHOIS records can be considered fully accurate:

Number of criteria being strictly satisfied	Unweighted frequency counts	Population estimates	Estimated percentage	Margin of error	Cumulative percentage
All three criteria	353	23,117,442	22.8%	1.4%	22.8%
Two out of three	625	42,804,131	42.3%	3.4%	65.1%
One out of three	355	27,366,722	27.1%	3.1%	92.2%
None	86	7,937,694	7.8%	1.8%	100.0%
Total domain names	1,419	101,225,988	100%		

However as discussed in the previous section, the requirement of *independent* linkage between name and address does not capture all degrees of association, and it may be appropriate to include as associated those cases where we were able to track down the registrant (by finding someone admitting to be the registrant, or finding some other contact information for the named registrant which makes us reasonably certain that we have found the right person (such as a voicemail indicating the registrant name from a telephone number associated with the address).

Moreover, the errors that some registrants admitted on contact were arguably trivial in nature, and certainly not so severe to have prevented us from finding them.

And so with a slight relaxation of the criteria to:

1. address must be deliverable
2. an independent linkage between name and address exists, or the WHOIS information enables us to track down the respondent, even if we cannot otherwise confirm a link between name and address
3. the respondent must acknowledge ownership,

then the proportion of WHOIS records which are accurate more than doubles, to 46%, and only 6% fail on all three:

Number of criteria being at least partially satisfied	Unweighted frequency counts	Population estimates	Estimated percentage	Margin of error	Cumulative percentage
All three criteria	704	47,073,062	46.5%	4.2%	46.5%
Two out of three	442	31,590,046	31.2%	2.9%	77.7%
One out of three	208	16,235,901	16.0%	2.3%	93.7%
None	65	6,326,979	6.3%	1.3%	100.0%
Total domain names	1,419	101,225,988	100%		

Not all failures of accuracy are equally serious; there is a qualitative difference between the registrant who has given their full name and address in WHOIS but who refused to participate in the survey component because they dislike surveys, and the registrant who gives a name which is so common that on its own provides little identification value, at an address that exists but to which we cannot independently link to the name, for which a phone number cannot be found. The following table classifies records by their primary failure (if any) against the criteria; while not all of the 29% of registrants shown with a full or substantial failure will have deliberately misrepresented their information, it is among this group that those with questionable intent are most likely to be found.

Accuracy group	**Description of accuracy (1) (2)**	Unweighted frequency counts	Population estimates	Estimated percentage	Margin of error
No failure	Met all three criteria fully - deliverable address, name linked to address, and registrant confirmed ownership and correctness of all details during interview	353	23,117,442	22.8%	1.4%
Minimal failure	All criteria met but minor fault noted by registrant during interview	17	1,101,176	1.1%	0.2%
	Name unable to be linked to address, but able to locate registrant and confirm ownership	312	23,024,007	22.7%	2.2%
Limited failure	Deliverable address, name linked and/or located, but unable to interview registrant to obtain confirmation.	365	24,893,476	24.6%	1.7%
Substantial failure	Undeliverable address and/or unlinkable name, however registrant located. Unable to interview registrant to obtain confirmation.	109	7,202,472	7.1%	0.9%
	Deliverable address, but unable to link or even locate the registrant, removing any chance of interview.	177	13,949,721	13.8%	2.2%
Full failure	Failed on all criteria - undeliverable address and unlinkable, missing, or patently false name, unable to locate to interview	86	7,937,694	7.8%	1.8%
All domain names in top five gTLDs		**1,419**	**101,225,988**	**100%**	

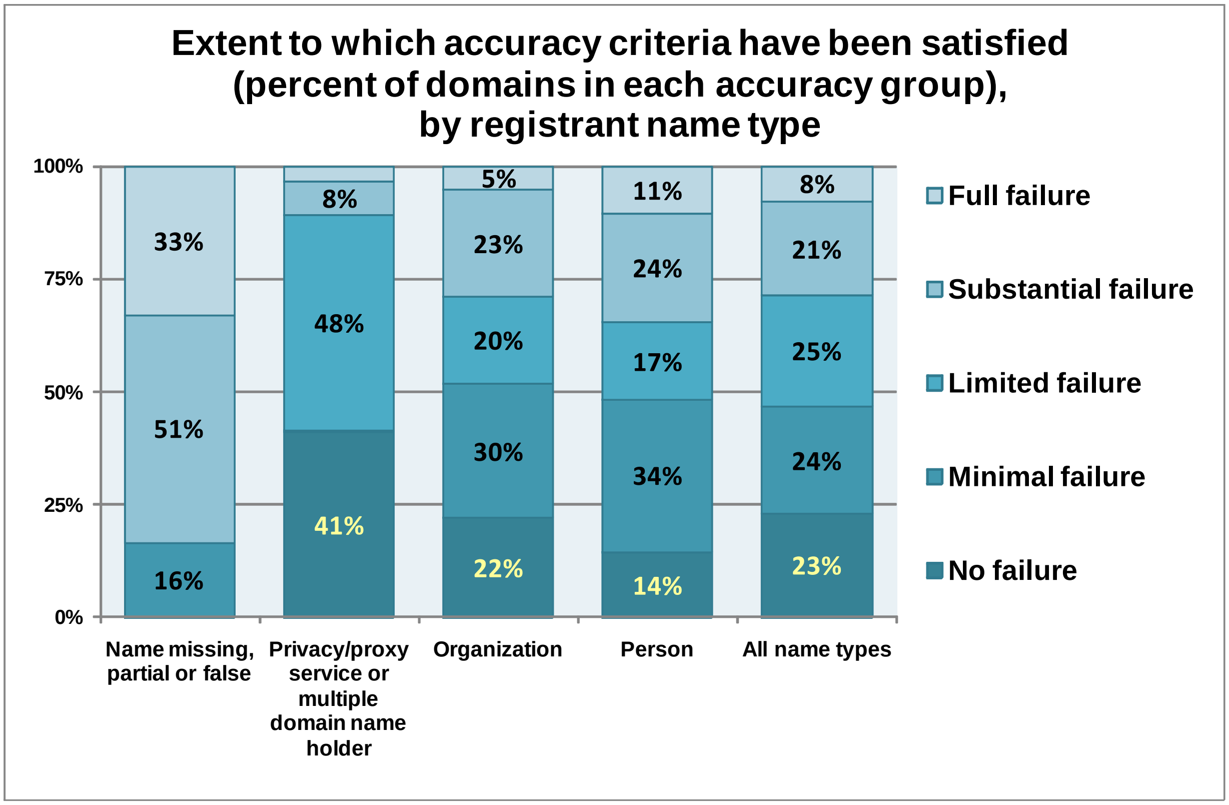
(1) Unable to link: means unable to find any independent association between name and address, or name and/or address missing

(2) Unable to locate: means unable to get confirmed current phone contact information for named registrant

The highest proportion of accurate WHOIS records was found among the domains registered through a Privacy/Proxy service, which would be expected as generally such services have no motivation to obscure either the name or address of the service. The majority of criteria failures among entries associated with Privacy/Proxy services were because we were unable to get a response from the service confirming that they did indeed represent all domain names listed for them.

Entries where the name or address is missing, partial or false by definition have some inaccuracy associated with them (thus none among this group are "fully accurate). However, the extent to which we were able to locate and in many cases interview the registrants of domains with such faulty WHOIS entries indicates the extent to which such situations are less the result of deliberate obfuscation as carelessness or confusion in completing the information fields that will feed into WHOIS.

As might be expected, organizations registering domains have a slightly higher overall accuracy in the associated WHOIS records than natural persons.



Barriers to accuracy

The majority of errors found in this study were ones that could feasibly be eliminated if several barriers to accuracy were addressed. The barriers occur at several points, the largest being at the point of data entry.

Barriers to accuracy at the point of data entry - from the registrant

Most of the errors found were related to the registrant's attitude towards domain names and/or WHOIS. There were two predominant themes:

1. Concern regarding privacy, and

2. Carelessness, and/or little perceived value in domain ownership.

Privacy concerns

WHOIS can in some ways be considered the equivalent of a telephone directory. However, unlike telephone directories which allow people to remain unlisted, WHOIS does not give that option. This creates the motivation to give partial or obscured details for those who do not want their information so publically displayed.

This is a different motivation to desire to remain completely anonymous, as evidenced by the fact we were able identify and contact some registrants who had given only partial or obscured information. It is also evident in the growing proportion of registrants who use privacy/proxy services who presumably are providing good data to the service provider (at least a valid credit card with consistent information), while responding to the very overt cautions about how WHOIS information is displayed when the service is offered during the course of registration.

In most registry type systems which could reveal information about a person's identity and address (motor vehicle registries, telephone directories, property ownership, credit status, medical records), there is an inherent tradeoff between the accuracy of the information and the degree of unrestricted and/or undocumented access.

Value perception

Many registrants do not perceive any adverse consequences to having incorrect information about them in WHOIS, or even domains incorrectly registered in their name. This leads to lack of attention or care in what they enter, for example

- one registrant gave the address of the person who owned the computer with which she used to register her domain,
- others acknowledged typographic errors in their information,

and to reduced motivation to keep their WHOIS information current:

- the most common error reported was tardiness in keeping the information updated following a change of address
- others were simply waiting for the domain to expire, having sold or otherwise abandoned the business name they registered the domain in.

Several registrants when contacted were not sure initially of whether or not they did register the domain listed for them, but none were concerned about their details being associated with the domain. They each reported having so many domain names they didn't really keep track of them all, and didn't see the need to. The sense was that domain names are so cheap and easily acquired, there is little value in them except where the name itself carries commercial value.

Barriers to accuracy at the point of data entry - from the requirements

No proof of identity or address is required when registering a domain name, which removes many barriers to entering inaccurate information. Requiring that the registrant name and address at least match that of the credit card which was used to pay for the account would go some way towards addressing this, given that reasonably stringent proof of identity and address is usually required to obtain a credit card. This however would still not be a complete removal of this barrier, because there is a large market for stolen credit card details and a determined thief could easily organize for registration to occur in the short window of time between theft and card shutdown. The cheapness of the domains relative to credit card charges for the vendor make the repeated checking against credit card details for continuity of registration unlikely.

Related to the lack of proof of identity is the broad scope within which registrants can choose to interpret the requirements for name and address. A full legal name is not required, nor is the address required to be one officially associated with the name. It was common for persons to give their work address for personal domain names, and for organizations to use the home address of someone associated with the organization.

Basic edit checks, if used consistently by all registrars, could eliminate some of the missing data issues, although in the absence of checks of identity, intentionally blank fields would just be replaced by plausible if incorrect information. The relative scarcity of undeliverable addresses (relative to other errors) indicates that most registrars are using various address checking software, but again, finding a valid address is no guarantee that it is indeed the address of the registrant.

Finally, some errors were strictly ones of respondent confusion at the point of data entry. A significant proportion of registrants interviewed – over 20% - were completely unaware of WHOIS, and consequently would have limited understanding of the information requirements. The pattern of responses for some cases indicated a confusion between the roles of registrant, administrative contact, and technical contact. For example, by writing "self" as registrant, or leaving the registrant field blank, while providing full and complete details about themselves in the administrative contact field. When asked to complete name and address information four times in the course of registering a site (one each for registrant, administrative contact, technical contact, and billing address), it is easy to see how these errors could arise.

Finally, only in late 2009 were changes made to support the entry and maintenance of non-ASCII character sets. Some apparent inaccuracies arose from script translation problems. For example, someone in China would need to translate their name and address into ASCII characters, but depending on where they were in China there may have been be no standard translation, and so once the WHOIS data is received a back translation is required in order to check addresses against listings that are maintained only in Chinese characters. Translating from A to B and then back again rarely re-reproduces A accurately, it is more likely that C will result, producing an inherent inaccuracy.

Barriers to accuracy in maintenance of accurate data

Even if information can be made accurate at the point of data entry, the maintenance of accuracy requires the registrant to keep the information current. Currently, the only penalty for a registrant for letting information get out of date is a communication from their registrar that they need to update it or their domain name will be suspended and possibly their ownership revoked. Even this is not a significant concern for many registrants when only a small proportion of domain names lead to web sites that the registrant has a vested interest in maintaining uninterrupted access to, and only a tiny fraction of domain names have intrinsic commercial significance.

However, even an improvement of the registrant motivation for keeping data accurate and current will not address the lack of provision for centralized checking. At the moment, only the registrars themselves are in a position to use efficient electronic checks of the data, from basic field completion checks through to cross checks against address deliverability databases and other databases by which identity might be confirmed. The process of combining WHOIS information from many different registrars and servers for the current process highlighted the near impossibility of a cost efficient centralized checking process, since different registrars used different fields in different ways, and mapping everyone successfully into a consistent set of fields ultimately required a large degree of manual work. A centralized database would, by virtue of being a larger data repository, make pattern based checking (such as credit card companies use to flag possible fraud activity) more powerful. However, like the removal of all other barriers already discussed, there would be costs involved in doing so which ultimately would need to be borne by the registrants in their fees.

Conclusion

There is no question that there are people who register domains without disclosing their full or real identity. While we didn’t find any cases where an identity had been stolen (that is, among the persons we contacted who had domains registered in their name, none denied having registered the domain), it would seem that, given the latitude that people have in choosing what information to provide when registering a domain name, identity theft may not be necessary; it is all too easy to enter any or no name, along with an unreliable or undeliverable address.

Most of the barriers to accuracy found (concerns about privacy, confusion about information needed, lack of clarity in the standard to which information should be entered, no requirement for proof of identity or address, the structure of WHOIS itself) can be addressed by the internet community. However, the cost of ensuring accuracy will escalate with the level of accuracy sought, and ultimately the cost of increased accuracy would be passed through to the registrants in the fees they pay to register a domain. Cooperation among all registrants and other ICANN constituents will be needed to eliminate any commercial disadvantage accruing from enforcing greater accuracy.

…

Appendix 1: Sample Design in Detail

Project Objective and Overview

For Phase I of this study, NORC selected a representative sample of domain names from five gTLDs (*.com, *.net, *.org, *.info, *.biz) that allows us to estimate the percentage of domain names that are "accurate" with a +/- 5 percent margin of error at the 95% confidence level.

This sample is an equal-probability sample so that every in-scope domain had an equal chance of selection. However, to reduce costs, we did not choose a simple random sample of domains. This would require selecting domain names scattered across the whole world. Instead, we follow the industry standard (used for worldwide surveys as well as nationally representative surveys in the United States) to select a multi-stage sample (or "cluster" sample). For the WHOIS Accuracy Study, the first stage "clusters" are countries. At the second stage, we selected domain names within each selected country. This is designed to minimize cost, but does not compromise the representativeness of the sample because every domain name (worldwide) had an equal probability of being selected.

Cluster samples are the industry standard for in-person studies because it is too costly to send interviewers to every county in the United States. Similarly, it would be too expensive to collect data on domain names in every country in the world. Prominent national area-probability studies (area-probability is the industry term for multi-stage cluster samples) done by NORC are the General Social Survey (done every 2 years), the Survey of Consumer Finances (every 3 years), and the National Longitudinal Survey of Youths (every year).

In-Scope Universe

According to the April 2008 Registry Operator Monthly Reports (Jan-Mar, 2008 for .aero) at http://www.icann.org/en/tlds/monthly-reports/, there are 15 global Top-Level Domains (gTLDs) with active Domains. Table 1 below shows the total number of domains among all 15 gTLDs. Excluded from these gTLDs are .edu, .mil, and .gov, which are out of scope.

The sample was restricted to the top five gTLDs only because they collectively represent 98.4% of the universe, and this restriction would simplify the selection and WHOIS extraction process. Any error arising from the exclusion of the remaining 1.6% of domains would be less than the sampling error of this study.

Table 1. Summary of global Top-Level Domains (gTLDs) of interest to ICANN

Rank	Top-Level Domain	Total Domains	Percentage of Domains	Cumulative Percentage	Included in WHOIS Accuracy Project?
1	.com	75,785,462	73.7%	73.7%	Yes
2	.net	11,478,837	11.2%	84.9%	Yes
3	.org	6,840,493	6.7%	91.5%	Yes
4	.info	5,092,053	5.0%	96.4%	Yes
5	.biz	2,029,143	2.0%	98.4%	Yes
6	.mobi	903,941	0.9%	99.3%	No
7	.name	287,442	0.3%	99.6%	No
8	.travel	201,047	0.2%	99.8%	No
9	.asia	159,682	0.2%	99.9%	No
10	.cat	29,230	0.0%	100.0%	No
11	.jobs	13,279	0.0%	100.0%	No
12	.pro	7,994	0.0%	100.0%	No
13	.coop	5,861	0.0%	100.0%	No
14	.aero	5,414	0.0%	100.0%	No
15	.museum	528	0.0%	100.0%	No
	TOTAL	102,840,406			

Frame Used for NORC Sampling

In April 2009, under instruction from NORC, ICANN drew and delivered to NORC a "proportionate" sample for these five domains of 2,400 total records. Each of the gTLDs were represented in their proper proportions. This is the frame NORC used to draw our revised sample of domain names for data collection.

First Stage of Selection: Assigning Country to Domain Names

In order to select countries, we need to know the country of the registrant for each domain name. For the *.org, *.info, and *.biz gTLDs, the WHOIS information (which includes the registrant address and country 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. Of the 2,400 selected domains, the country was identified for all but 54. Rarer countries might or might not be in Kent's sample, but countries with at least 0.04 percent (1 out of every 2,400) of the world's domains have a good chance of appearing in Kent's sample of 2,400 records.   The table in Appendix 3 shows the distribution of countries by country.

Determining the Number of Countries

Our main decision was how many countries to include in the sample. If we selected too many, the costs would be high because we would attempt to investigate only a few domain names in many countries. If we selected too few, the additional clustering increases the design effect and the necessary sample size to achieve the goal. We found the best compromise by selecting 16 countries.

Every country had a positive probability (based on the number of domain names in our frame) to be selected for inclusion and we have selected a representative sample of countries.

The five countries with the largest number of records (United States, Canada, United Kingdom, Germany, and China) all would have had a probability of more than 100%, so they enter as certainty countries (their selection probability is 1) and are allocated their proportional share of the sample. For example, the United States contains over 59 percent of the domain names, so it has received over 59 percent (928) of the total 1,571 domain names that will be selected.

The other eleven countries were selected proportionate to their number of domain names in three groups. The first group consisted of countries with at least 17 domain names in the frame of 2,400 domains, which corresponded to having at least a 31 percent chance of being selected. This group was sorted by their Regional Internet Registry, and the European and Asian nations were sorted further by location (e.g., Iberia, Western, or Central Europe). The second (consisting of countries with at least 6 domain names in the frame of 2,400 domains, which corresponded to having at least a 10 percent chance of selection) and third groups were also sorted by Regional Internet Registry, and further by location.

We refer to these three groups of non-certainty countries as Large (> 16 domains), Medium (> 5 domains), and Small.

Determining the Sample Size of Domain Names

ICANN's planned sample size was originally 384 domain names. With a simple random sample, this sample size would allow a percentage of valid records to be calculated with a standard error no greater than 2.5 percent (which allows a confidence interval to be the estimate +/- 5 percent). However, such a simple random sample would also result in a very costly survey with many countries in the world having only one or two selections.

We needed to select a larger sample size because of the more complicated sample design, which results in an effective sample size less than the total sample size. This is due to the geographic clustering. The ratio between the total sample size and the effective sample size is often referred to as the design effect (DEFF).

Rather than select domain names from all over the world, we selected a subset of clusters (countries) to be in the sample, and selected domain names from only these countries. This allows us control over how many countries would be in the sample. However, if some countries are more or less likely to have accurate registry records than others, the sample suffers a loss of power due to intraclass correlation (within country, the domain names are correlated, or more related to each other than to the rest of the world). This loss of power is called the design effect due to clustering

(DEFFc), and can be approximated by using the intraclass correlation (usually positive between 0 and 1) and the average cluster size (the average number of interviews obtained per cluster).

This decision (number of countries) impacts the design effect (the factor by which we need to increase the sample size from 384 to achieve an accuracy percentage to be calculated with a +/- 5 percent margin of error at the 95% confidence level), and therefore our recommended sample size.

Our preliminary sample size is 400 times our estimated design effect. We have rounded up 384 to 400 simply to be conservative. We compared many different choices for the number of non-certainty country selections. It should be noted that the certainty countries are completely defined by the number of non-certainty countries selected. As we increased the number of non-certainty country selections, the design effect (and therefore the necessary sample size of validations) decreased, but the costs (due to visiting more countries) increased. We chose the optimal number of non-certainty countries to be eleven, which then defined the five certainty countries (see below).

Selecting Domain Names from Selected Countries

Since the five certainty countries include almost 67 percent of the domain names in the frame, the certainty countries receive almost 67 percent (1,186) of the 1,571 sample selections. The other eleven countries all receive (up to) 35 domains each. In selecting the domain names within country, we sorted by gTLD so that every country's sample is a proportional sample from that country's domain names.

We initially hypothesize that 90 percent of the sample will be eligible (will be in the WHOIS directory when we begin data collection), and that we can achieve a response rate (resolving the accuracy of the WHOIS record) for 70 percent of the eligible domain names. Under these assumptions, our sample would result in 1,571 * .9 * .7 = 990 interviews. For a sample with this many countries, we have conservatively estimated a design effect of 2.47 (based on an intraclass correlation of 0.07), resulting in an effective sample size of at least 990/2.47 = 400, which allows a percentage to be calculated with a +/- 5 percent margin of error at the 95% confidence level.
It is important to note that for selected countries with less than 35 domains in the microcosm, all domains are selected. This does reduce the number of domains selected to 1,419. We expect the response rates and design effects above to be conservative, and that an effective sample size of at least 384, if not 400, will still be reached.

Sample for the WHOIS Accuracy Study

The sample is stored in an Excel spreadsheet (SAMPLE11_1419.XLS) and a comma-delimited file (SAMPLE11_1419.CSV) with 1,419 domain names selected. This sample size was determined from our choice to have 16 countries:

5 CERTAINTY countries:
United States, Canada, United Kingdom, Germany, and China

11 NON-CERTAINTY countries:

6 LARGE (> 16 domains in microcosm of 2400, > 31 percent selection probability):
Australia, Japan, Turkey, France, Spain, and Netherlands.

3 MEDIUM (> 5 domains in microcosm of 2400, > 10 percent selection probability):
Malaysia, Russia, and Sweden.

2 SMALL (< 6 domains in microcosm of 2400, < 10 percent selection probability):
Singapore and Ireland.

Tables 2 and 3 list the sample frequency by country and by top-level domain. Table 4 lists the variables in the sample file.

Table 2. Sample frequency by country.

Country_code	Country_name	Selected
US	United States	928
CA	Canada	77
GB	United Kingdom	71
DE	Germany	61
CN	China	49
AU	Australia	35
JP	Japan	35
TR	Turkey	23*
FR	France	35
ES	Spain	31*
NL	Netherlands	35
MY	Malaysia	8*
RU	Russia	10*
SE	Sweden	13*
SG	Singapore	5*
IL	Israel	3*
TOTAL		1,419*

*Many non-certainty countries have fewer than 35 domains among the frame of 2,400 domains. All domains in such countries are selected.

Table 3. Sample frequency by top-level domain.

gTLD	Selected	Percentage
com	1,066	75.12
net	162	11.42
org	102	7.19
info	64	4.51
biz	25	1.76
TOTAL	1,419	100.00

Table 4. Variables included in sample file SAMPLE8_1231.XLS/SAMPLE8_1231.CSV

Variable	**Description**
Domain_name	
Country_code	Two-character code for country of registrant from WHOIS directory
Country_name	Full name of country of registrant from WHOIS directory
gTLD	Top-level domain (i.e., com, net, org, info, or biz)
Country_strata	Certainty, Large, Medium, or Small

Weighting and estimation

A weight has been developed for the 1,419 selected domain names which corrects for the clustering by country, and enables expansion to the full universe of the five top gTLDs. Any analyses on all 1,419 domain names should use this weight, and all the tables in this report were calculated using this weight.

The standard errors in this report have been calculated with the SUDAAN software. This is because the ICANN Whois sample is not a simple random sample of all domain names on Earth. The ICANN Whois sample is a stratified (by gTLD) cluster (by country) sample. Generally speaking, stratification reduces standard errors while clustering increases standard errors.

The ICANN Whois involves heavy stratification (the certainty countries contain 83.6 percent of all domain names) and a large amount of clustering by country (to reduce the costs of attempting cases in many different countries), the standard errors sometimes are larger and sometimes are smaller than they would be for a simple random sample. The design effect, which is the ratio of the design-corrected standard error and the simple random sample standard error, differs greatly for the different tables in these reports. The design effects (and therefore the standard errors and margin of errors) are larger for the address delivery tables, indicating that there are strong differences by strata (and country) in the deliverability of addresses in the ICANN Whois database. However, the design effects are close to or smaller than 1 for the very first table on accuracy, which indicates that accuracy is less different by strata or country.

As can be seen from the tables, there is only one margin of error (out of forty-one) that exceeds 5 percent (and it is 5.1 percent), indicating that our sampling strategy did accomplish our goal of keeping the margin of error to plus or minus five percent.

Appendix 2: Registrant contact script

INTRO 1 (PERSONS NAME INCLUDED IN REGISTRANT DETAILS)

Hello, may I please speak with (name)?
My name is ____________, and I'm calling on behalf of ICANN, the Internet Corporation for Assigned Names and Numbers. We're doing a very brief survey about the internet. I have a few quick questions about the registrant information for [DOMAIN NAME].

INTRO 2 (BUSINESS OR NO NAME INCLUDED)

Hello, may I please speak with the person there who registered the domain name [FILL DOMAIN NAME] ? IF NEEDED: Who is responsible for your website?
My name is ____________, and I'm calling on behalf of ICANN, the Internet Corporation for Assigned Names and Numbers. We're doing a very brief survey about the internet. I have a few quick questions about the registrant information for [DOMAIN NAME].

May I ask who I am speaking with? (your first name will do, just in case we are disconnected and I have to ring back)

Section B1 When the registrant appears to be a business, organization, club, group, association etc (ORG)

BCONF Can you confirm that you did register the domain name [FILL DOMAIN NAME].?

ALTERNATIVE WORDING IF NEEDED: **We have [FILL NAME OF REGISTRANT] listed as the registrant of [FILL DOMAIN NAME]. Is this correct?**

IF THEY HAVE NO CLUE WHAT YOU ARE TALKING ABOUT, CHECK IF YOU ARE TALKING TO THE RIGHT PERSON

1.	Yes, immediate recognition and confirmation, no issues	GO TO SECTION B2
2.	Yes, but it took them some time to confirm DESCRIBE SITUATION	
3.	Yes, but as the interviewer you detected some issues DESCRIBE SITUATION	
4.	Unable to say LAST RESORT CODE – DESCRIBE SITUATION.	GO TO SECTION X
5.	No, they did not register site or authorize their name to be used to register the site DESCRIBE HOW THEY THINK IT HAS COME TO BE REGISTERED IN THEIR NAME	

Section B2 – ORG registrant, registration confirmed

BTYPE how would you describe [FILL NAME OF REGISTRANT]; is this a: READ OUT MOST LIKELY CATEGORIES

1. Registered or incorporated business, partnership or organization with employees?
2. Registered or incorporated business, partnership or organization with no employees?
3. An informal club or group
4. A potential business
5. Other (specify)

BREL In what capacity are you associated with [FILL NAME OF REGISTRANT]; ? READ OUT IF NEEDED

1. You are an employee of ORG
2. You own ORG
3. ORG is your client, to whom you provide web related services
4. ORG is a club/association/group you are involved in
5. Other (specify)

BACONF We have the registrant address recorded as [FILL ADDRESS OF REGISTRANT].

IF ADDRESS APPEARS DELIVERABLE: **If we posted an envelope to that address today, would it reach you?**

IF ADDRESS APPEARS UNDELIVERABLE**: This address appears to be (undeliverable/missing). Was that intentional?**

INTERVIEWER: PROBE TO SELECT BEST FIT CATEGORY

1. Address is correct as given – a letter posted would reach them GO TO BATYPE
2. Address is out of date – they have since moved GO TO BATYPE
3. Address is wrong, incomplete or missing – intentional
4. Address is wrong, incomplete or missing – but they don't know why, they thought they gave the full and correct details in registration GO TO BATYPE
5. Other (specify) GO TO BATYPE

BINTENT What were your concerns that led you to not provide a (full/accurate) address?

RECORD VERBATIM. IF THEY JUST SAY 'PRIVACY CONCERNS' ASK 'CAN YOU PLEASE TELL ME A BIT MORE ABOUT THAT?

BATYPE   And is (was) this address: READ OUT MOST LIKELY CATEGORIES

1. Your home (street) address
2. Your personal PO box
3. The home address of someone else associated with ORG
4. The PO box of someone else associated with ORG
5. The street/physical address of the head office of ORG
6. Some other physical address of ORG
7. A postal address (as in PO box) of ORG
8. other   DESCRIBE

Section P1 When the registrant appears to be a person, or the registrant name is missing completely

PCONF   Can you confirm that you did register the domain name [FILL DOMAIN NAME].?

1. Yes, immediate recognition and confirmation, no issues
2. Yes, but it took them some time to confirm DESCRIBE SITUATION
3. Yes, but as the interviewer you detected some issues DESCRIBE SITUATION

GO TO SECTION P2

4. Unable to say LAST RESORT CODE – DESCRIBE SITUATION.
5. No, they did not register site or authorize their name to be used to register the site DESCRIBE HOW THEY THINK IT HAS COME TO BE REGISTERED IN THEIR NAME

GO TO SECTION 4

Section P2 – PERSON registrant, registration confirmed

PNCONF   We have your name recorded as [FILL NAME] Is this your…

PROBE WITH MOST LIKELY CATEGORIES.

IF NAME COMPLETELY MISSING, CODE 7 AND MOVE TO NEXT QUESTION

a. Full name as appears on your license or passport?
b. Commonly known first and last name
c. Last name only
d. First name only
e. Initials or alias
f. Fake or nonsensical name DO NOT READ OUT
g. Name is completely missing DO NOT READ OUT
h. Some other name – specify

PACONF We have your address recorded as [FILL ADDRESS OF REGISTRANT].

IF ADDRESS APPEARS DELIVERABLE: **If we posted an envelope to that address today, would it reach you?**

IF ADDRESS APPEARS UNDELIVERABLE**: This address appears to be (undeliverable/missing). Was that intentional?**

INTERVIEWER: PROBE TO SELECT BEST FIT CATEGORY

1. Address is correct as given – a letter posted would reach them GO TO PADDTYPE
2. Address is out of date – they have since moved GO TO PADDTYPE
3. Address is wrong, incomplete or missing – intentional
4. Address is wrong, incomplete or missing – but they don't know why, they thought they gave the full and correct details in registration GO TO PADDTYPE
5. Other (specify) GO TO PADDTYPE

PINTENT **What were your concerns that led you to not provide a (full/accurate) address?**

RECORD VERBATIM. IF THEY JUST SAY 'PRIVACY CONCERNS' ASK 'CAN YOU PLEASE TELL ME A BIT MORE ABOUT THAT?

PATYPE   **And is (was) this address:** READ OUT MOST LIKELY CATEGORIES; IF ADDRESS MISSING CODE 7 AND MOVE TO NEXT QUESTION

1. Your home (street) address
2. Your personal PO box
3. The home address of a friend/relative
4. The PO box of a friend/relative
5. Your work address
6. Your school address
7. MISSING OR NONSENSICAL ADDRESS
8. other   DESCRIBE

Section W – WHOIS knowledge and use

WFREQ **Have you ever used the WHOIS service on the internet to look up who registered a particular domain name?** IF THEY SAY NO OR NEVER PROBE**: Have you heard of WHOIS?**

1. Never heard of WHOIS >> GO TO END
2. Know of WHOIS but never used it   >>GO TO END
3. Once or twice
4. Three to 20 times
5. 21 times or more
6. Refused >> GO TO END
7. Don't know (don't code here – probe to get nearest category above)

WUSE   **For what sort of purpose did you look it up? (multiple)**

a. To see what would appear of their own information if they purchased a domain name
b. To see who registered a particular domain name, but no intention to contact
c. To see if it was a legitimate site (e.g. before purchasing or giving information)
d. To see who registered, with the intention of contacting (specify why –
    i. Had a tech issue with the site
    ii. Was suspicious of the site – e.g. wanted to check whether it was a phishing site
    iii. Had a commercial interest in the domain name or related names (e.g. wanted that domain name, wanted to sell the owner similar domain names)
e. Other (specify)

GO TO END

-

Section X (ownership denied or uncertain). All registrant types.

ADDRECOG **The address we have recorded for the registrant of this site is [FILL ADDRESS OF REGISTRANT Do you recognize it?**

1. Yes - home physical address
2. Yes - own PO box
3. Yes – friend/relatives address
4. Yes – other (specify)
5. No recognition

DISPLAY REGISTRANT NAME: **FILL REGISTRANT NAME**
DISPLAY ADMIN CONTACT NAME: **FILL ADMIN NAME**

IF ADMIN CONTACT NAME NOT MISSING, AND NOT THE SAME AS THE REGISTRANT NAME, ASK:

ADMIN FILL ADMIN NAME is given as the administrative contact for this site. Do you know that person?

1. Yes
2. No GO TO THOUGHTS

ADMINREL **How is that person connected to you**?

f. Relative/friend
g. Employee/colleague
h. Internet service
i. Other (specify)

THOUGHTS **Do you have any thoughts as to why or how this site has been registered (in your name / in ORG name)**

SPECIFY

END

That was the last question. Thank you for your time. If you have any questions about this survey, please ring us on <phone number>

Interviewer: review the information which led to the contact details you were using. Do you think you were speaking with the right person?

1. Yes – fairly sure the person named as registrant is the person you were speaking with
2. Not sure – perhaps there is someone else of the same name we should try to locate SPECIFY WHY
3. Very likely not - you are fairly certain there has been a locating error SPECIFY WHY

Appendix 3: Privacy and proxy service identification and confirmation

Among the 1419 sampled cases, 364 cases were identified as being *potentially* privacy or proxy service providers. The net was deliberately drawn wide for the cases, because the original intention was for ICANN to contact the relevant service providers as part of a separate study on the prevalence of privacy and proxy services, and since those excluded from the subset would automatically be classified as *Neither privacy nor proxy*, borderline cases were included in this set. It was also originally planned that as a byproduct of ICANNs prevalence study, the accuracy of the associated WHOIS records would be established.

The initial classification for potentially privacy or proxy was done as follows:

1. where the registrant name contained the following words or phrase 'privacy', 'proxy', 'registration', 'registration service', 'identity', 'shield', 'guard', 'private', 'buy', 'rare', 'names', 'WHOIS', 'value', 'domain', and 'secure'
    - Examples: *Registration Private*, *Domains by Proxy Inc, Moniker Privacy Services*, and *WHOIS privacy services.*

2. where multiple domains contained the same registrant name, registrant organization, or registrant address.

3. where the registrant name or registrant organization may correspond to the name of a privacy service, proxy service, or multiple domain name holder.
    - Registrant Name Examples: *DowntownWebsites.com* and *DNSADMIN*
    - Registrant Organization Examples: *ServAlliance* and *United Online Pty Ltd*.

Thirteen cases were sufficiently borderline however to be retained in the main "individual registrant" study.

It became apparent however that the accuracy of the WHOIS records for the 351 cases remaining exclusively in the "potentially privacy/proxy" group was not going to be determined in sufficient time for the current accuracy study, and so a separate stream of work was established to assess the accuracy of these cases.

Assessing accuracy

In this first step, we were less concerned with identifying whether or not a service was privacy/proxy as whether the WHOIS information was accurate. For the registrants associated with just one domain, the process followed was that used by the individual cases, that is:

1. a search of business and individual listings to see if we could link the name and address and find a phone number through that linkage,
2. where no linkage was found in those searches, searches via Google to see if an association could be established elsewhere, and address based searches to obtain phone contact details

3. contact where possible with the registrant to ask if they acknowledged registration of the domain name.

For cases with multiple listings, we would look for association and contact details in similar ways, but plan only one contact to cover all names associated with the registrant. At this point we were sometimes directed to email. In others, they acknowledged having a service, but directed us to an online tool to check whether or not each particular domain name was covered by the service. The hypothesized main source of inaccuracy for those with multiple domain names would be “free riders” – that is, imposters who register a domain name but use information found in WHOIS for such services in place of their own.

Methods to Verify Domain not a free-rider:

1) Direct contact: Establish contact with the service to request confirmation that they were providing a service for the domain in question. For the larger cases, contact was first attempted through information obtained in the domain name’s WHOIS record as long as we had verified that those contact details were not inconsistent with other contact details provided by the service. If contact was established, we would ask if they were providing a privacy/proxy or other service to the domain in question. If they denied providing service to the domain name we would check the domain's current WHOIS record and if it still matched the original WHOIS record then this was coded as **9_Free_Rider**. It would otherwise be classified as “partial confirm”.

2) Privacy or Proxy services online contact tool :Some Privacy and Proxy services have an online contact tool which, when supplied with a domain name, will allow anyone to contact the registrant without revealing the registrants identity. It is possible to use this tool as a verification method, because if the domain name is not in the Privacy or Proxy service’s database it will return an error or domain not found message.[4] Conducting a search and not receiving an error message resulted in a final Name Type of **Privacy/Proxy Service** and deeming the record accurate. If the result returned an error and the domain name’s current WHOIS records matched the WHOIS records in the sample then this was coded as **Free_Rider**. If the WHOIS records did not match then Name Type was coded as “partial confirm” as we could no longer can verify that the registrant used the privacy service in the past.

3) Underlying registry match: Many registration services will provide underlying registry WHOIS information. These registration services will add in their services name as a 'registered through' field to the underlying registry data which is added to the domain's WHOIS information (see figures 5 and 6 below). Given this information, a final Name Type of **4_Privacy/Proxy Service** was assigned if the following criteria were met:
    a. If the registrants name or organization matched the registered through field
    b. If the registered through field matched the registrars name and the registrar only uses one privacy/proxy service when registering domains and this privacy/proxy service matched the registrants name/organization and address

[4] Not all Privacy and Proxy contact tools returned an error message for domains which were not in their database. To ensure that domains were not incorrectly verified we first conducted a test of the contact tool. The test consisted of supplying a valid domain and an invalid domain. If an error message was not returned on the invalid domain search then the contact tool was not used as a verification method.

Confirmation as a privacy/proxy service

The above process often clarified whether a service was indeed a privacy/proxy one or not. If however the status of the service was still in question, several additional tests were considered:

1. Is there any mention of the provision of a privacy/proxy service among any of the organizations literature;
2. Does a search of business records or does their online presence indicate they are primarily a different type of business (for example, many attorneys and web developers provide proxy or privacy type services to clients, but only as a byproduct of their main service to the client, and there is no evidence they provide privacy or proxy services independent of their other services).

Unless we were fairly certain we had a privacy/proxy service, we coded it as a multiple domain name holder, organization, or person, as most appropriate.

Appendix 4: GAO study replication

As a preliminary step in this study, we replicated part of the GAO study *Prevalence of False Contact Information for Registered Domain Names*, as described in the GAO report dated August 30, 2005.

The first objective of the GAO study was to determine the prevalence of "patently false" or incomplete contact data in the WHOIS service for the three "legacy" top level domains: .org, .net, and .com.

To accomplish this, 300 domain names from each of these gTLDs were randomly selected, and reviewed to identify data that are incomplete or patently false – data that appeared obviously and intentionally false without verification against any reference data, such as "(999) 999-9999" for a telephone number, "asdasdasd" for a street address, or "XXXX" for a postal code.

The following findings in respect of the registrant information were reported in the GAO study:

2005. **At least one field in the Registrant contact information was:**	**.com**	**.org**	**.net**	**overall**
Patently false	3.3%	3.0%	0.9%	3.0%
Incomplete	0.8%	2.1%	3.0%	1.1%
Unable to access WHOIS data	3.3%	1.3%	1.9%	3.0%
None of the above	**92.7%**	**93.7%**	**94.3%**	**92.9%**
Total	100.0%	100.0%	100.0%	100.0%

We repeated this exercise for all 2400 domain names selected in the first stage of sampling (see appendix 1 for detailed sampled design), and found the situation had not changed substantially:

2009 **At least one field in the Registrant contact information was:**	**.com**	**.org**	**.net**	**.biz/info**	**overall**
Patently false	1.6%	0.6%	2.2%	0.0%	1.5%
Incomplete	4.3%	1.8%	3.7%	4.4%	4.0%
Unable to access WHOIS data	2.4%	0.0%	1.5%	0.0%	2.0%
None of the above	**91.7%**	**97.6%**	**92.7%**	**95.6%**	**92.5%**
Total (Percent)	100.0%	100.0%	100.0%	100.0%	100.0%
Patently false	28	1	6	0	35
Incomplete	77	3	10	7	97
Unable to access WHOIS data	44	0	4	0	48
None of the above	**1652**	**163**	**253**	**152**	**2220**
Total (record counts)	1801	167	273	159	2400