NORC

Proposed Design for a Study of the Accuracy of Whois Registrant Contact Information

Developed by NORC at the University of Chicago for ICANN

NORC Project Reference: 6558, 6636

3 June 2009

Contents

Study overview

ICANN seeks to determine the accuracy of Whois information for the total population of domain names registered in generic Top Level Domains ("gTLDs").

Given the size and flux of the universe of domain names, a sample-based study is proposed which will enable unbiased estimates to be generated while still enabling sufficient investigation of selected domain names.

This document describes the sample design, the sample selected, and the proposed methodology.

Because of the processes involved, work towards two other studies can be included at minimum additional cost: a partial replication of the GAO study; and sample preparation for a study on the prevalence of privacy and proxy services. The work involved in these two studies is described in Appendix 1.

Proposed Sample Design

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 was designed to minimize cost, but did 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 studies covering large geographic areas where to have sample selected in every area would be cost-prohibitive. Prominent national area-probability studies (area-probability is the industry term for multi-stage cluster samples) done by NORC are the General Social Survey (conducted every 2 years), the Survey of Consumer Finances (every 3 years), and the National Longitudinal Survey of Youth (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.

ICANN has previously selected simple random samples from the top five gTLDs, which cover 98.4 percent of the domain names in the 15 gTLDs. ICANN has agreed to exclude the remaining 10 gTLDs from this study, but future studies could include them.

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%	
7	.name	287,442	0.3%	99.6%	
8	.travel	201,047	0.2%	99.8%	
9	.asia	159,682	0.2%	99.9%	
10	.cat	29,230	0.0%	100.0%	
11	.jobs	13,279	0.0%	100.0%	
12	.pro	7,994	0.0%	100.0%	
13	.coop	5,861	0.0%	100.0%	
14	.aero	5,414	0.0%	100.0%	
15	.museum	528	0.0%	100.0%	
	TOTAL	102,840,406			

Frame Used for NORC Sampling

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

For the rest of this document we will use the more descriptive word *Microcosm* to refer to this frame of 2,400.

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 the microcosm, 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 the 2,400 microcosm records. The table in Appendix 3 shows the distribution of countries.

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 selection probability (based on the number of domain names in our microcosm) 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 percent, 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 microcosm, which corresponded to having at least a 31 percent chance of being selected. This group was sorted by 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 microcosm, 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 have allowed a percentage of valid records to be calculated with a standard error no greater than 2.5 percent (which allows a confidence interval of +/- 5 percent). However, such a simple random sample also would have resulted 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 resulted 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 allowed 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 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 a 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 reduces 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.

Proposed Sample for the Whois Accuracy Study

We will deliver an Excel spreadsheet (SAMPLE11_1419.XLS) and a comma-delimited file (SAMPLE11_1419.CSV) with our sample of 1,419 domain names. 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 2,400, > 31 percent selection probability):
Australia, Japan, Turkey, France, Spain, and Netherlands.

3 MEDIUM (> 5 domains in microcosm of 2,400, > 10 percent selection probability):
Malaysia, Russia, and Sweden.

2 SMALL (< 6 domains in microcosm of 2,400, < 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

Proposed Name and Address Verification Methodology

Accuracy definition

The methodology for this phase is dependent on how accuracy is defined. ICANN determined the following definition, 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.

To further clarify:

- 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)

  Note that there is no requirement that the name be a legally held one, as would be required, for example, when opening a bank account or applying for a passport.

- 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.

Translating the definition into measurable factors

To develop a methodology, the definition above has been re-framed into quantifiable test conditions. We are proposing that all three of the following test conditions must hold for a WHOIS entry to be considered completely accurate:

1. The address given is a valid postal address, as specified in the above definition

2. The entity named as the registrant is independently associated with the address given; that is, there is some evidence *other than* the WHOIS entry that an entity of that name can be contacted at the address given, and

3. The registrant, once contacted using independently obtained contact information, acknowledges that they are the registrant of the domain name, and (if needed given the similarity between many domain names) recognizes the description of the web page associated with the domain name.

Why not simply go straight to test condition #3 using the contact information in WHOIS? A short questionnaire could be posted to the registrant at the given address, or we could just telephone or email the administrative contact. While this would be a very cheap way to proceed, it does not address the requirements given in the definition. For example, someone using a false name and a PO Box which is otherwise untraceable to them could still complete and return the questionnaire acknowledging ownership of the domain, and nothing in this process would alert us to the inherent inaccuracy of the Whois information for that domain.

Each of these test conditions can be treated as a binary condition (met vs. not met) but, because there are nuances as to whether the condition is fully met vs. partially met, the data collected will be more powerful if these shades of grey can be captured. That will enable the data to be reanalyzed for alternative definitions of accuracy.

The processes which will be followed to answer the above test conditions are described in the next section.

Step 1: Checking the mailing address

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

Registrant Address Type (single code, categorical)

1. Address completely missing
2. Address patently false (e.g., "99999" entered as street name)
3. Insufficient detail below city or state to classify as either physical or postal service
4. Street/physical address.
5. Postal service address (e.g., PO Box, RM Box)

Where the case is coded as either a street/physical or postal service, it will then be checked against postal data for accuracy, then coded on the following deliverability scale:

Registrant address deliverability (single code, ordinal)

1. Deliverable as given
2. Deliverable with minor automated changes (for example, postcode added for a valid street/city/state combination)
3. Possibly deliverable (for example, the street number is missing, but the street, city, state and postcode are all valid, or where a slightly different spelling of a street name would form a fully deliverable address)
4. Definitely undeliverable (for example, the post office mentioned does not exist, or the street or city named do not exist)

The outcomes of this stage (including any alternative or corrected addresses which may be found) will be added to the file for use where appropriate in subsequent steps.

It is important to note that the purpose of this test is to assess the quality of the registrant address information provided in WHOIS; it is not to find an actual address for the selected registrants. If in the course of our investigations we get updated address information for a registrant, we may use that information to obtain a current telephone number, and keep information on file regarding what we found the best address to be, but it will not change the outcome of this stage, which is an assessment of registrant address data in WHOIS at the time of the information extraction.

It is also important to note that even if we find this is a deliverable address, it does not tell us if it is the deliverable address of the registrant named. The subsequent steps are needed for that.

Step 2: Classifying the type of registrant

Because the ways in which we can proceed in later steps depends on the nature of the entity, the next step is to classify the registrant by type. The following classification, based on the information provided for registrant name, is proposed. The word "apparent" is included in the title to emphasize that at this stage we can code only according to the information provided in the name, and that in the course of investigation we may need to reclassify the entity.

Apparent Registrant Type (single code, categorical)

1. Name completely missing
2. Name patently false (e.g., "99999")
3. Natural person
4. Registered business, no person named
5. Registered business, person named
6. Other organization, person named
7. Other organization, no person named
8. Privacy/proxy service

Regarding this classification:

1. The reason we distinguish between "person named" and "no person named" when dealing with businesses or other organizations is because different databases often exist for people as opposed to businesses, and where a person is named we will have more search opportunities.
2. Registered business is a general categorization intended to distinguish legal entities as opposed to informal organizations, since the incorporation or registration process implies inclusion in a greater range of databases. Such organizations are generally recognized by the use of a suffix such as ltd, inc, plc or their international equivalents.
3. Privacy and proxy services will in many cases be registered businesses, but they are categorized separately because they are of particular interest (see Appendix 1).

Step 3: Finding an independent name/address association

Any case identified as being privacy/proxy will be passed to ICANN for direct confirmation with the service involved.

Cases where both name and address are either missing or patently false will be classified as inaccurate by definition and put aside, since no association is possible without both a name and an address.

For the remaining cases, the ideal association is that the entity named can be found listed with the given address in a reliable publically available database, such as in the residential white pages (when dealing with an individual), or the yellow pages or a business database such as Dunn and Bradstreet (when dealing with an organization as the registrant). Such a match will give the additional benefit of a telephone number by which to contact the registrant.

So, the first check will be against public telephone records, using a scoring system ranging from 9 (address match on street number, street, and suburb town with an unambiguous organization name match) down to 1 (no telephone number found for name or address). Cases with scores between 2 and 8 collectively will be called a partial match (or more simply, a telephone number worth trying), with the scores assigned on a partially subjective system. For example, any match with an unusual surname would score higher than the same match with a common surname.

Phone listing outcome (single code, ordinal)

1. No match

2-8 Partial match found (a phone number worth trying, actual score reflecting degree of match)

9. Strong match found (address at street level plus unambiguous name match)

Regarding this classification:

- For businesses and organizations where a person is named, searches will be performed on both person name and organization name, and a strong match on either will be coded as a strong match.
- The lack of such an association, however, does not immediately imply inaccuracy, but simply the need for further investigation. For example, the telephone book will yield a match only if the individual uses a single name, and has given his home address in WHOIS, and if he has a telephone account in his name.
- Conversely, any telephone number obtained in a strong match does not imply that further searches will not be needed, since the number may later be found to be disconnected.

Should further searches be needed, other publically available internet based search methods will be used. For example, we may Google the person and find an entry in a social networking site such as Facebook. However, given that such sites require no proof of identity, the case would score relatively low. On the other had, should that site lead us to discover that the address is the person's work address, for example, and we can confirm via the employer that the person does work there, then a higher score will be assigned. Because of the large range of possible searches, it is not feasible to list them all in advance; however we will track those used and record the source of any matches made.

Where available, we also will use subscription databases which are not generally available to the public, or available only for a fee. Many such databases are constructed from credit applications where the risk of being denied credit if the information is not found to be correct increases the trustworthiness of the information contained. Such databases are particularly useful in that often they will provide an address history for a person (or an occupant history for an address), and so may assist in resolving what might otherwise appear to be a non-match when in fact the registrant information is simply out of date.

Any cases which require further searches will have the following variable updated based on the "best outcome" among all further searches:

Other source match outcome (single code, ordinal) (The source of the match also will be recorded)

1. No match
2-8 Partial match found
9. Strong match found

International considerations

The extent to which independent sources of information exist and are accessible will differ between and among countries. Part of the setup process for each country will involve a review of the available databases and information sources, and documentation of the extent of coverage for each source. This will not only inform the extent of adjustment to the basic procedures needed for a country, but also how the data are analyzed.

For example, countries with highly organized infrastructures, low mobility rates, and low privacy concerns will have a larger proportion of entities associated with addresses than those from other countries. The question arises as to whether this means there should be a normalization process applied so these countries don't automatically end up with generally higher rates of association. This may or may not be appropriate – we will address this question once we have sufficient information on each country to see if these differences might lead to a biased analysis, and if so, what adjustments need to be made.

Step 4: Contacting the registrant

It is possible that a registrant trying to hide their identity gives a name and address other than their own. If they use the name and postal address of a real person, we will find a valid postal address and a valid association, so it is only upon contacting the named individual that we will find that the person does not know of the domain name. Given that most correspondence relating to a domain name goes to the administrative contact, many people could have their identities used by the real registrant in this way without ever being aware of it.

Telephone contact is recommended as the main procedure. Telephone surveys avoid the often prohibitively high costs of personal visits, and the very low response rates of mail and other self-completion surveys. They allow flexibility in the questions asked, and are able to be completed in shorter timeframes than most other survey methods.

Telephone numbers identified in the course of finding the association will be used to contact the registrant. Care will be taken to identify the source of any contact information obtained, as ultimately the source of information must be considered when assessing the likelihood that the person contacted is indeed the named registrant. For example, in some cases we will not be able to find any telephone number based on the registrant name or address alone. However, we might still be able to contact someone claiming to be the registrant through telephone numbers listed on the domain website, or through information provided about the administrative contact.

So, how do we distinguish such cases from those where the registrant protects their privacy by not having any listed telephone numbers and avoiding any activity which is likely to result in their information being included in a database that might be used by skip tracers (for example, by never applying for credit)? We cannot demand proof of identity, and there is little point in asking the typical "identity" questions (if they are trying to use another's identity, they probably can find out what their target's mother's maiden name was). We can use only hints such as the lack of independence of the contact information, slipups or inconsistencies the "registrant" may make when we are speaking with them (another reason for telephone contact as opposed to mail or email), and possibly the nature of any website (if any) the domain name leads to. This underlines the importance of not having a binary classification of accuracy, but rather something that captures all these "shades of grey."

Finding an appropriate contact within organizations

Where the registrant is not a person but a business or organization, if public listings of that entity exist, there will in most cases be a "front desk" phone number available. We will need to "navigate" through such organizations, asking to speak to someone with a technical or accounting function who is most likely to know about domain names registered for that organization. Although this method could result in being transferred among several people within an organization, it is preferable to going directly to the administrative contact named because the named person may not be authorized to register sites on the entity's behalf.

If this process of navigation proves unsuccessful, we will resort to asking to speak with the named administrative contact, and note that this happened.

Questions to be asked of the registrant

Appendix 2 contains a draft script for the telephone interview with individuals named as registrants. Because at this stage we are not certain of the full range of responses likely, we will conduct the first interviews (up to 100) using hardcopy, where the interviewer is not constrained in question wording or answer options. We will refine the script and answer options over these initial interviews, so that the majority of interviews can be completed within a CATI (computer assisted telephone interviewing) system which provides for more cost effective data collection and better control over the progress of the sample. Broadly, the script contains four sections:

1. Introduction and basic confirmation that the right person (or at least someone of the correct name) is on the phone;
2. A direct question about ownership of the domain name, including if needed a description of the site (if any) the name leads to. Provision is made here for the interviewer to record any concerns raised by themselves or the respondent;

3. If ownership is acknowledged, confirmation/correction of address and, if appropriate, administrative contact information. Additional questions on the respondent's relationship to the address and the administrative contact are proposed to provide contextual data for analysis and recommendations;
4. If ownership is not acknowledged, any details of recognition and concern of identity theft are collected.

If the respondent acknowledges ownership (Q2), confirms that the address is correct (Q3), and gives the interviewer no cause for concern that they are not the actual owner, then the entry will be deemed accurate.

Respondent relations

We are keeping the questionnaire as short as possible to minimize the impact on the registrants. However, like any survey and in accordance with industry standards, we need to have ways to reassure the respondents that no adverse impact will arise from their cooperation, and that the interviewer calling is indeed working on this survey and is not part of some scam. We will establish a 1-800 number for respondents to call should they have any questions about the process, and we will work with ICANN to prepare a short letter of endorsement that can be faxed, emailed, or posted to any registrant as needed. This letter will include the name and direct telephone number of a manager at NORC, as well as that of someone at ICANN. We will also work with ICANN to prepare a suitable package of respondent information to post on both ICANN and NORC's websites. Examples of information and letters provided to respondents of other NORC surveys can be found by going to the NORC website (www.norc.org), and clicking on the link "For Survey Participants" which appears in the center of the lower third of the home page.

Dealing with non response

As this is essentially a survey, there will be some level of non-response where we are either unable to find any appropriate contact information, or where we have good contact information but the registrant refuses to speak with us. Mail or email contact may be attempted as a last resort, bearing in mind the deficiencies mentioned previously in respect to self-completion surveys. However, unlike most surveys where non-response can represent a substantial problem, by the time we are attempting contact with a registrant we already have considerable information about the likely accuracy of its WHOIS information, including the outcomes of the previous two tests, the presence (or absence) of risk factors regarding accuracy (such as the nature and content of the website, if any), as well as correlating information with other domain names where we have been able to complete this final stage. There are well established statistical survey estimation methods which use such "auxiliary information" to make adjustments for non responding units and to assess potential bias. Any such methods used will be fully documented and explained in the final report.

Appendix 1: Partial replication of the GAO study, and initial classification of potential privacy and proxy services

As part of the sampling process, a random sample of 2,400 domain names was selected, and their Whois information extracted and provided to NORC. The final sample for the verification process will be drawn from this file, after coding for stratification elements such as country.

Considerable effort has already been expended by both ICANN and NORC in preparing and cleaning this file. At the moment, there is no common standard by which registrars collect or store the required WHOIS information, consequently addresses can be stored in as few as three fields for some registrars, and as many as six for others. Some require differentiation of first and last name from organization, some do not. Some force country to be included as part of an address, others do not. As a result, getting all WHOIS information for a sample taken from many registrars into a consistent data structure requires considerable work. Having a consistent data structure is important for several reasons:

1. to ensure that determinations made are based on the actual content of the information, and not simply a result of an unusual data structure. For example, registrant information may appear to be missing, but on closer inspection might be found included in an adjacent field; and

2. to reduce costs of the subsequent phases. For example, batch matching processes – where a list of names and addresses, for example, are run against a telephone database and are returned with an additional field populated with telephone number where available – are much cheaper than looking up every name individually. However, such batch runs require fields to be precisely defined and consistent.

In the process of cleaning this file and preparing it for the verification tasks, work towards two additional studies can be easily accommodated.

Partial replication of GAO study

This refers to 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:

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%
Total domain names (millions)	35.8	3.5	5.7	44.9

In the process of classifying the type of registrant for the verification study, we will need to classify cases where details are missing or patently false. Thus we will have partially replicated the GAO study, which will:

1. provide a point of comparison to the earlier study; and
2. assess whether the prevalence of patently false information about registrants has changed over the last four years.

There are however a few significant differences from the GAO study which need to be noted:

1. The GAO study produced separate estimates for the three gTLDs covered. Because the sample design for the current study has proportional representation (to provide better overall estimates, but not gTLD-specific ones), only the .com estimates are likely to have sufficient precision for comparison.
2. The GAO study examined registrant, administrative contact, and technical contact separately. This study is focused on the registrant information only.
3. At this stage it is not planned to replicate other components of the GAO study – e.g., submitting error reports for the patently false or incomplete entries, and determining whether they are corrected within 30 days.

Initial classification for prevalence of proxy and privacy service use

This study is being carried out primarily by ICANN; however, since the resolution of these cases is needed in order to complete the accuracy study, there are cost efficiencies achieved by having NORC and ICANN work together on this study.

Defining, and Identifying Proxy and Privacy Services

Proxy and Privacy are often used interchangeably, yet they are distinct services. However, there are several views on what constitutes Proxy and Privacy services, and pending constituent agreement on a definition for each. In the meantime we are approaching this issue by using two distinct phases of classification:

1. *The initial classification, Potentially Privacy or Proxy.* These cases will be identified by NORC in a coding exercise from which registrant details appear to be 3rd party arrangements (for example, by using proxy, hosting, or privacy in the registrant name). It is fully recognized that this alone is not sufficient for a final classification, since what appears as a proxy or privacy service might simply be 'faked' by the registrant by entering a name which appears to be such a service.

2. *The final classification* will be made by ICANN, and will depend on the outcome of the register query process. To be classified as such, the service will need to conform with the final agreed-upon definition, and the service will need to acknowledge during the query process that they are such a service for the domain name given. Estimates of prevalence will be made using only this final classification.

Implications for the Accuracy Study

Cases found in the final classification to be confirmed as privacy or proxy services will be classified as accurate as long as the address details were accurate for the service provider. However, since this is a qualitatively different type of accuracy from one where the registrant name is not obscured by a 3rd party, they will be reported as a special subtype of accuracy.

For cases where no legitimate service has confirmed they were acting for that domain name, further attempts to contact the registrar will be undertaken before the cases are classified with respect to accuracy.

Appendix 2: Draft script

This script is a starting point. It will be refined as we progress through the start of the survey.

1. **Good morning/afternoon, I am <…> from NORC at the University of Chicago. May I speak with <name>?**

   a. <name> is available, comes to phone or you are speaking with them GO TO 2
   b. <name> is at that number, but not currently available MAKE APPOINTMENT
   c. <name> no longer there – other contact details available COLLECT DETAILS
   d. <name> no longer there – no alternative contact details available END AND RETURN CASE TO LOCATING
   e. Never heard of <name> END AND RETURN CASE TO LOCATING

   PROCEED ONLY ONCE YOU ARE SPEAKING WITH <NAME>.

2. **I am calling about the domain name <domain name> which is registered in your name**. (IF APPROPRIATE): **It leads to a website which <describe>. Can you confirm that you did register that site?** (IF NEEDED): **You registered it with <registrar name>**

a. Yes, immediate recognition and confirmation, no issues b. Yes, but it took them some time to confirm DESCRIBE SITUATION c. Yes, but as the interviewer you detected some issues DESCRIBE SITUATION	GO TO SECTION 2
d. Unable to say LAST RESORT CODE – DESCRIBE SITUATION. e. 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 3

Section 2 (ownership confirmed)

3. **We have your name recorded as <name given for registrant>. Is that correct?**

   a. Yes
   b. Minor correction needed SPECIFY
   c. Major change needed COLLECT NEW NAME, AND PROBE FOR REASON IT HAS CHANGED AND NOT BEEN UPDATED

4. **We have your address recorded as <address given for registrant>. Is that still correct?**

   a. Yes
   b. Minor corrections needed SPECIFY
   c. Whole new address needed COLLECT NEW ADDRESS, AND PROBE FOR REASON IT HAS CHANGED

5. **And is that (this new) address your……**

   a. home physical address
   b. own PO box
   c. employer address
   d. school address
   e. other DESCRIBE

6. IS THE REGISTRANT NAME REPEATED AS THE ADMIN CONTACT?

   a. Yes GO TO END
   b. No

7. **<Name> is given as the administrative contact for this site. Is that still correct**?

   a. Yes
   b. No GO TO Q9

8. **How is that person connected to you**?

   a. Relative/friend
   b. Employee/colleague
   c. Internet service
   d. Other (specify)

9. THANK AND TERM. **Interviewer: Do you think you were speaking with the right person?**

   a. Yes – fairly sure the person named as registrant is the person you were speaking with and that they are known by the name given

   b. No/Not sure SPECIFY YOUR CONCERNS

Section 3 (ownership denied or uncertain)

10. Have you ever been a victim of identity theft?

   a. Yes PROBE – DO THEY THINK THIS SITE COULD BE RELATED?
   b. No/Not sure

11. The address we have recorded for the owner of this site is <registrant address>. Do you recognize it?

   a. Yes - home physical address
   b. Yes - own PO box
   c. Yes - employer address
   d. Yes - school address
   e. Yes - other DESCRIBE
   f. No recognition

12. IS THE REGISTRANT NAME REPEATED AS THE ADMIN CONTACT?

   a. Yes GO TO Q15
   b. No

13. **<Name> is given as the administrative contact for this site. Do you know that person**?

   a. Yes
   b. No GO TO Q15

14. **How is that person connected to you**?

   a. Relative/friend
   b. Employee/colleague
   c. Internet service
   d. Other (specify)

15. **Do you have any thoughts as to why or how this site has been registered in your name**?

THANK AND TERM.

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

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

Appendix 3: Country distribution

Rank	Country	Country Code	Region	Number of domain names in microcosm	% of group	Cumul-ative %	Sampling stratum	Whether sampled
Domains where country known				**2346**	**97.75%**			
1	United States	US	ARIN	1387	59.12%	59.12%	Certainty	1
2	Canada	CA	ARIN	115	4.90%	64.02%	Certainty	1
3	United Kingdom	GB	RIPE	106	4.52%	68.54%	Certainty	1
4	Germany	DE	RIPE	90	3.84%	72.38%	Certainty	1
5	China	CN	APNIC	73	3.11%	75.49%	Certainty	1
6	France	FR	RIPE	52	2.22%	77.71%	Large	1
7	Australia	AU	APNIC	50	2.13%	79.84%	Large	1
8	Netherlands	NL	RIPE	44	1.88%	81.71%	Large	1
9	Japan	JP	APNIC	37	1.58%	83.29%	Large	1
10	Spain	ES	RIPE	31	1.32%	84.61%	Large	1
11	Turkey	TR	RIPE	23	0.98%	85.59%	Large	1
12	Italy	IT	RIPE	22	0.94%	86.53%	Large	0
13	Korea, Rep.of (South)	KR	APNIC	20	0.85%	87.38%	Large	0
14	India	IN	APNIC	19	0.81%	88.19%	Large	0
15	Portugal	PT	RIPE	17	0.72%	88.92%	Large	0
16	Brazil	BR	LACNIC	16	0.68%	89.60%	Medium	0
17	Switzerland	CH	RIPE	14	0.60%	90.20%	Medium	0
18	Sweden	SE	RIPE	13	0.55%	90.75%	Medium	1
19	Cayman Islands	KY	ARIN	10	0.43%	91.18%	Medium	0
20	Russian Federation	RU	RIPE	10	0.43%	91.60%	Medium	1
21	Hong Kong	HK	APNIC	8	0.34%	91.94%	Medium	0
22	Malaysia	MY	APNIC	8	0.34%	92.28%	Medium	1
23	Poland	PL	RIPE	8	0.34%	92.63%	Medium	0
24	Saudi Arabia	SA	RIPE	8	0.34%	92.97%	Medium	0
25	Austria	AT	RIPE	7	0.30%	93.27%	Medium	0
26	Belgium	BE	RIPE	7	0.30%	93.56%	Medium	0
27	Panama	PA	LACNIC	7	0.30%	93.86%	Medium	0
28	Thailand	TH	APNIC	7	0.30%	94.16%	Medium	0
29	Georgia	GE	RIPE	6	0.26%	94.42%	Medium	0
30	Mexico	MX	LACNIC	6	0.26%	94.67%	Medium	0
31	New Zealand	NZ	APNIC	6	0.26%	94.93%	Medium	0
32	Norway	NO	RIPE	6	0.26%	95.18%	Medium	0
33	Ukraine	UA	RIPE	6	0.26%	95.44%	Medium	0
34	Argentina	AR	LACNIC	5	0.21%	95.65%	Small	0
35	Egypt	EG	AFRINIC	5	0.21%	95.87%	Small	0
36	Ireland	IE	RIPE	5	0.21%	96.08%	Small	1
37	Singapore	SG	APNIC	5	0.21%	96.29%	Small	1
38	Costa Rica	CR	LACNIC	4	0.17%	96.46%	Small	0
39	Czech Republic	CZ	RIPE	4	0.17%	96.63%	Small	0
40	Philippines	PH	APNIC	4	0.17%	96.80%	Small	0
41	Bulgaria	BG	RIPE	3	0.13%	96.93%	Small	0
42	Denmark	DK	RIPE	3	0.13%	97.06%	Small	0
43	Finland	FI	RIPE	3	0.13%	97.19%	Small	0
44	Iran, Islamic Rep.of	IR	RIPE	3	0.13%	97.31%	Small	0
45	Israel	IL	RIPE	3	0.13%	97.44%	Small	0

46	Nigeria	NG	AFRINIC	3	0.13%	97.57%	Small	0
47	South Africa	ZA	AFRINIC	3	0.13%	97.70%	Small	0
48	United Arab Emirates	AE	RIPE	3	0.13%	97.83%	Small	0
49	Viet Nam	VN	APNIC	3	0.13%	97.95%	Small	0
50	Virgin Islands, British	VG	ARIN	3	0.13%	98.08%	Small	0
51	Zimbabwe	ZW	AFRINIC	3	0.13%	98.21%	Small	0
52	Albania	AL	RIPE	2	0.09%	98.29%	Small	0
53	Antigua and Barbuda	AG	ARIN	2	0.09%	98.38%	Small	0
54	Belize	BZ	LACNIC	2	0.09%	98.47%	Small	0
55	Indonesia	ID	APNIC	2	0.09%	98.55%	Small	0
56	Romania	RO	RIPE	2	0.09%	98.64%	Small	0
57	Syrian Arab Republic	SY	RIPE	2	0.09%	98.72%	Small	0
58	Andorra	AD	RIPE	1	0.04%	98.76%	Small	0
59	Belarus	BY	RIPE	1	0.04%	98.81%	Small	0
60	Bermuda	BM	ARIN	1	0.04%	98.85%	Small	0
61	Colombia	CO	LACNIC	1	0.04%	98.89%	Small	0
62	Cyprus	CY	RIPE	1	0.04%	98.93%	Small	0
63	Dominican Republic	DO	LACNIC	1	0.04%	98.98%	Small	0
64	Gibraltar	GI	RIPE	1	0.04%	99.02%	Small	0
65	Grenada	GD	ARIN	1	0.04%	99.06%	Small	0
66	Jordan	JO	RIPE	1	0.04%	99.10%	Small	0
67	Kenya	KE	AFRINIC	1	0.04%	99.15%	Small	0
68	Lao People’s Dem.Rep.	LA	APNIC	1	0.04%	99.19%	Small	0
69	Latvia	LV	RIPE	1	0.04%	99.23%	Small	0
70	Lithuania	LT	RIPE	1	0.04%	99.28%	Small	0
71	Malta	MT	RIPE	1	0.04%	99.32%	Small	0
72	Pakistan	PK	APNIC	1	0.04%	99.36%	Small	0
73	Paraguay	PY	LACNIC	1	0.04%	99.40%	Small	0
74	Peru	PE	LACNIC	1	0.04%	99.45%	Small	0
75	Puerto Rico	PR	ARIN	1	0.04%	99.49%	Small	0
76	Saint Kitts and Nevis	KN	ARIN	1	0.04%	99.53%	Small	0
77	Samoa	WS	APNIC	1	0.04%	99.57%	Small	0
78	Slovakia	SK	RIPE	1	0.04%	99.62%	Small	0
79	Slovenia	SI	RIPE	1	0.04%	99.66%	Small	0
80	Tokelau	TK	APNIC	1	0.04%	99.70%	Small	0
81	Turks and Caicos Islds.	TC	ARIN	1	0.04%	99.74%	Small	0
82	Tuvalu	TV	APNIC	1	0.04%	99.79%	Small	0
83	Uruguay	UY	LACNIC	1	0.04%	99.83%	Small	0
84	Vanuatu	VU	APNIC	1	0.04%	99.87%	Small	0
85	Venezuela	VE	LACNIC	1	0.04%	99.91%	Small	0
86	Virgin Islands, U.S.	VI	ARIN	1	0.04%	99.96%	Small	0
87	Yugoslavia (split now)	YU	RIPE	1	0.04%	100.00%	Small	0

Domains where country not known	**54**	**2.25%**	
Repeated timeout error when tried to access	36	66.67%	66.67%
Domain not found in Whois	7	12.96%	79.63%
No country listed in Whois entry	4	7.41%	87.04%
Out of registry at time of country extraction	3	5.56%	92.59%
Other error	4	7.41%	100.00%