

Pilot Study for the WHOIS Accuracy Reporting System: Preliminary Findings

About the Pilot Study

On 8 November 2012, in response to the [Recommendations of the WHOIS Review Team](#) convened under the [Affirmation of Commitments \(AoC\)](#), the [ICANN Board approved](#) a series of improvements on the manner in which ICANN carries out its responsibilities regarding WHOIS. As part of these improvements, [ICANN committed](#) to proactively identifying potentially inaccurate gTLD WHOIS contact data and forwarding potentially inaccurate records to gTLD registrars for investigation and follow-up. To accomplish these tasks ICANN initiated the development of the WHOIS Accuracy Reporting System (ARS) – a framework for conducting repeatable assessments of WHOIS accuracy over time and publicly reporting the findings.

The Pilot Study is a precursor to the ARS, designed with the intention of isolating and testing key components of the system. Results from the Pilot Study will be used to inform the design of the ARS.

This Paper describes the preliminary findings from the Pilot Study. It is published to illustrate the application of the proposed methodology and approach using actual samples of WHOIS data. These preliminary findings will be explored at the Los Angeles ICANN meeting at the [All Things WHOIS Session](#) on 13 October 2014. A full study report, with updated findings and statistics, will be published following the [ICANN Los Angeles Meeting](#).

Methods and Approach

The Pilot Study was an international collaborative effort led by ICANN, with support from National Opinion Research Center (NORC) at the University of Chicago, DigiCert, Inc., Strikelron, Inc., The Universal Postal Union (UPU), and Whibse, Inc. Initial planning for the Pilot Study began on 27 August 2014, and data collection began on 4 September 2014. Accuracy validation of the WHOIS records selected for review was completed on a compressed schedule between 12 September 2014 and 19 September 2014. Analysis of the validation results concluded on 8 October 8 2014.

In the Pilot Study, from a sample of WHOIS records, these organizations validated the telephone, email, and postal address fields respectively. The validation was completed by examining the sample WHOIS records for various factors of syntactic and operational accuracy, which are explained below.

Syntactic perspective: The syntactic perspective examines WHOIS data with the intent of determining which data exists or is missing from within WHOIS data fields and assesses whether data meets specified standards of formatting and structure.

Operational perspective: The operational perspective examines WHOIS data with the intent of determining if contact information can be practically used to achieve the intended function of the data type and involves gathering and utilizing data external to the WHOIS record to inform an accuracy assessment.

A third perspective, identity, has been defined for possible inclusion in the ARS. Due to cost and time constraints, the Pilot Study did not undertake the identity perspective. However, discussions with the community, in addition to the findings of the Pilot Study, will inform the possibility of including the identity perspective in the ARS. The identity perspective is described below.

Identity perspective: The identity perspective seeks to confirm that WHOIS data corresponds to the person or entity being represented as the registrant by attempting to contact the registrant or entity using the information listed in the WHOIS record.

An accuracy assessment of WHOIS data, guided by any of the three perspectives should result in the data being categorized into a rating scale based on the degree to which it can be considered accurate. The rating scale categories established for the WHOIS accuracy studies describe accuracy within an inclusive range between “No Failure,” the highest degree of accuracy, and “Full Failure,” which represents the lowest degree of accuracy. For the Pilot Study, records were grouped into the superset categories of “Accurate” or “Inaccurate” which encompass the ARS accuracy rating scale categories as follows:

Accurate Categorization

- **No Failure:** Indicates that data are missing no significant information, and that the data provided are accurate using all three accuracy perspectives;
- **Minimal Failure:** Indicates that data are missing no significant information. The record may need clarification or might benefit from additional information, but the data provided are accurate; and
- **Limited Failure:** Indicates that the data are missing at least some significant information, and that the data present are to some degree considered inaccurate.

Inaccurate Categorization

- **Substantial Failure:** Indicates that the data are missing significant information, and that the data provided are mostly inaccurate.
- **Full Failure:** Indicates that significant data elements are missing, or that data provided are overwhelmingly inaccurate.

In the Pilot Study, ICANN contractors applied predetermined validation criteria to the email, telephone, and postal data of WHOIS records to assess accuracy from the syntactic and operational perspectives. The validation results were compiled into datasets, separated by data type and by perspective, and these datasets were then analyzed to produce accuracy statistics (for New and Prior gTLDs) stratified by gTLD, Geographic Region, and registrar.

Email Address Validation

Syntactic validation for email addresses could be described as examining the information contained within the syntax of the address. If an email address was determined to be present in the parsed WHOIS data, the syntax of the email was checked for conformance with RFC specifications.

Operational validation of email addresses was performed by checking the domain within the address for validity and responsiveness and attempting communication with mail servers to determine whether an email address was accurate and functional. The communication attempts returned information on whether the domain, mail server, and email address could be reached, or

provided a reason why they could not be reached, and this data was used to code email addresses into rating scale categories.

Telephone Number Validation

Syntactic validation of telephone numbers was performed by examining WHOIS records to determine if a telephone number was provided and, if so, evaluating the format of the number provided for compliance with expected telephone number requirements. Proper format was based on the country in which the number is located and, as a result of the assessment, telephone numbers were coded into a rating scale category.

Operational validation of telephone numbers was performed by evaluating each phone number to determine whether a connection could be established and whether additional information was needed to make a connection possible. Phone numbers were coded into rating scale categories based on whether the phone number as it was provided in the WHOIS record could connect successfully, whether the number required additional information in order to connect or whether a connection was not possible.

Postal Address Validation

Syntactic validation of postal address information was performed by examining WHOIS records to determine if a postal address was provided and, if so, whether the format of the address was consistent with expected norms and syntax requirements for the country in which it is located. When components of an address were missing or did not match syntax requirements, additional checks were made to see if the address could be considered serviceable. The expected norms and syntax requirements for these checks were based on guidance from the UPU July 2013 Postal Addressing Systems report¹. Postal addresses were coded to a rating scale category based on the outcome of the assessment.

Operational validation of postal addresses was performed by determining whether the address provided in a record was an existing address according to UPU's database of worldwide postal address information. Addresses that were syntactically accurate, whether they were accurate as provided or otherwise serviceable during the syntactic assessment, were included in the assessment and each record was subsequently coded to a rating scale category.

Sample Design

The Pilot Study began with the selection of a sample of 100,000 records from gTLD Zone Files. Because New gTLDs are a primary focus of the study, a high percentage of these domains were sampled, accounting for 25 percent of the domains represented in the sample (at the time of sampling, only 1.4 percent of domains were within the New gTLDs). In order to provide some information on all gTLDs, smaller gTLDs were oversampled. Specifically, all New gTLDs had at least 10 records selected (or all if fewer than 10 records were available) and all other ("Prior") gTLDs had at least 30 records selected (or all if fewer than 30 records were available). Once these records had been selected, the remaining records were selected proportional to the number of records in each gTLD (separately for New and Prior gTLDs), with the exception that the prevalence of the top four New gTLDs and the Prior gTLD ".com" be allowed selection of less than the full proportion of these records to allow slight increases in the proportion of smaller gTLDs. Within each gTLD, a systematic sample was specified (with a starting record

¹ Postal Addressing Systems. July, 2013. Berne: Universal Postal Union. ISBN: 978-92-95025-58-5

and skip interval) to implicitly stratify by any partial or complete sorting within the gTLD Zone File (newer domains at the top or bottom, for example). All of the data provided in this document is unweighted, and while smaller gTLDs and all New gTLDs are overrepresented, the numbers provided here are roughly representative. Weights will be calculated as the study progresses as part of the full ARS operation.

Once the records were selected for the sample, they were then queried via WHOIS lookups to produce data for the Pilot Study. A critical next step was to parse the raw WHOIS data so that the information in the records could be consistently interpreted during the validation phase of the study. After parsing, all email records moved forward to the syntactical and operational validation phase. The number of successful queries (and parsed records) was 98,821. A subsample of 10,000 records was selected for syntactical validation of postal addresses and telephone numbers. In order to provide estimates at the Geographic Region level, NORC at the University of Chicago determined the country (and Geographic Region) for as many of the domain names as possible, and oversampled the two smallest Regions: Africa and Asia/Australia/Pacific.

Within each Geographic Region, the file was serpentine sorted for systematic sampling to guarantee coverage of subgroups. The sorting variables were:

- 1) New versus Prior gTLDs;
- 2) RAA 2009 registrars versus RAA 2013 registrars;
- 3) gTLD;
- 4) Registrar;
- 5) Country; and
- 6) Place in Zone File.

Serpentine sorting makes neighbor records more similar than regular sorting (within new gTLDs, for example, RAA 2009 registrars are last, but within Prior gTLDs, RAA 2009 registrars are first to keep all RAA 2009 registrars together in the sorted list), and results in implicit stratification by the sort variables. A smaller subsample of 1,000 records was drawn (again, Africa and Asia/Australia/Pacific were oversampled) for operational validation of postal addresses and telephone numbers. The same set of sort variables were used to sort the 10,000 records in a serpentine way before selecting a systematic sample of 1,000 records. Table 1, below, shows the distribution by ICANN Geographic Region for each of the three sample sizes.

Table 1: Sample Sizes by Geographic Region, with Subsampling Rates

Geographic Region	Full Sample	10,000 Subsample	Selection Probability (%)	1,000 Subsample	Selection Probability (%)
Africa (AF)	617	617	100.0	100	16.2
Latin America/Caribbean (LAC)	3,380	1,000	29.6	150	15.0
Asia/Australia/Pacific (AP)	20,176	1,784	8.9	160	8.9
Europe (EUR)	23,417	2,071	8.9	186	8.9
North America (NA)	50,768	4,490	8.9	401	8.9
Missing (.)	463	38	8.9	3	8.9
TOTAL/OVERALL	98,821	10,000	10.1	1,000	10.0

Table 1b provides the sample sizes for the then largest gTLDs:

Table 1b: Sample Sizes for the Top Ten gTLDs

gTLD	Full Sample	10,000 Subsample	Selection Probability (%)	1,000 Subsample	Selection Probability (%)
.com	44,399	4,518	10.2	451	10.0
.net	12,539	1,261	10.1	129	10.2
.org	8,726	883	10.1	89	10.1
.info	4,725	456	9.7	44	9.6
.xyz (New)	3,282	293	8.9	26	8.9
.biz	2,215	223	10.1	24	10.8
.berlin (New)	1,504	134	8.9	12	9.0
.club (New)	1,162	141	12.1	17	12.1
.guru (New)	805	82	10.2	10	12.2
.mobi	781	84	10.8	9	10.7

The full sample was used for email validation. Syntactical verification for telephone numbers and postal addresses used the 10,000 subsample, while the Operational verification for telephone numbers and postal addresses used the 1,000 subsample. The analysis reports on subgroups with at least 20 records, and only the top seven gTLDs met this criterion in the 1,000 subsample.

It is important to note that the different sample sizes result in a very different uncertainty in the estimates below. Specifically, for the email verification, small differences are statistically significant while for the operational validation of telephone numbers and postal addresses, differences of several

percentage points might not be statistically significant. The full report will make it possible for readers to determine which differences are statistically significant. Table 1c gives standard errors and 95 percent confidence intervals for our overall estimates of accuracy:

Table 1c: Impact of Sample Size on Standard Errors

	Accuracy Rate: All gTLDs	Sample Size	Standard Error	95% Confidence Interval
Syntactical:				
Email	89.2	98,821	0.10	(89.0, 89.4)
Telephone	88.7	10,000	0.32	(88.1, 89.3)
Postal	81.2	10,000	0.39	(80.4, 82.0)
Operational:				
Email	89.2	98,821	0.10	(89.0, 89.4)
Telephone	72.4	1,000	1.41	(69.6, 75.2)
Postal	82.4	1,000	1.20	(80.0, 84.8)

Table 1c shows that the standard errors for Email accuracy are very small, while the standard errors for Operational accuracy for telephone and postal are much larger than the standard errors for Syntactical accuracy. The standard errors for subgroups within these three sample sizes are larger.

Findings

The ARS Pilot Study analysis results are summarized below to show top-level accuracy statistics by New and Prior gTLDs, by 2009 and 2013 RAAs, and by Geographic Region. These results are further broken down to show accuracy rates from syntactic and operational perspectives for email, postal address, and telephone number information. Overall accuracy rates for the 10 gTLDs and registrars that had the largest sample sizes in the study are also reported towards the end of this summary of findings.

As seen in Table 2, below, the accuracy of email, postal address, and telephone information was assessed from the syntactic and operational perspectives, and group statistics were computed for all gTLDs, 2009 and 2013 RAA, and for New and Prior gTLDs. Table 2 shows the rates of records which were assessed as accurate, which includes all records categorized as No Failure, Minimal Failure, and Limited Failure.

Table 2: Accuracy Percentages by Perspective, across RAAs and New vs. Prior gTLDs

	All gTLDs	2009 RAA	2013 RAA	Prior gTLDs	New gTLDs
Syntactical:					
Email	99.9	99.9	99.9	99.9	99.9
Telephone	88.7	90.1	89.9	88.0	90.9
Postal	81.2	85.9	81.0	81.8	79.1
Operational:					
Email	89.2	81.7	89.8	88.9	90.2
Telephone	72.4	79.6	73.5	74.3	66.5
Postal	82.4	85.7	82.4	82.4	82.4

Table 2 shows that email and telephone syntactical accuracy is not statistically significant for Prior and New gTLDs while syntactical accuracy is better for Prior gTLDs on postal addresses. The operational numbers for email addresses indicate that New gTLDs have slightly better accuracy than Prior gTLDs. Prior gTLDs do have more operational accuracy on telephone numbers, but the two groups are equal on operational postal address accuracy.

Table 2 also shows that the differences in syntactical accuracy between registrars under the 2013 RAA and registrars under the 2009 RAA for email, telephone numbers and postal addresses are not statistically significant. Operationally, registrars under the 2013 RAA have more accuracy for email addresses than registrars under the 2009 RAA. Registrars under the 2009 RAA have more operational accuracy in telephone numbers while the two groups are similar on postal addresses.

Overall, it is not surprising that lower syntactical accuracy occurs for postal addresses because more than one field is involved. Operationally, the accuracy of telephone is lowest while the accuracy of email addresses is highest.

Similar to Table 2, Table 3 reports on accuracy percentages of email, telephone, and postal address information by perspective, but groups the results by ICANN Geographic Region.

Table 3: Accuracy Percentages by Perspective, across ICANN Geographic Regions

	All Geographic Regions	Africa	Asia Australia Pacific	Europe	Latin America Caribbean	North America
Syntactical:						
Email	99.9	99.5	99.9	99.9	99.9	99.9
Telephone	88.7	67.4	82.6	85.9	88.9	96.1
Postal	81.2	67.4	74.9	91.2	92.0	79.2
Operational:						
Email	89.2	90.9	83.2	90.0	89.8	91.1
Telephone	72.4	52.0	57.5	65.6	76.7	85.5
Postal	82.4	69.0	71.9	93.6	92.0	81.8

Table 3 shows that domains in Africa have the lowest syntactical accuracy in postal addresses and telephone numbers, but syntactical accuracy in email addresses is similar across the Geographic Regions. Operationally, domains in Africa and Asia/Australia/Pacific have the lowest accuracy for both postal addresses and telephone numbers, but Africa and North America have the highest accuracy for email.

In contrast to the tables above, Tables 4 and 5 report on email validation for the gTLDs with the ten largest sample sizes in the Pilot Study. The names of these gTLDs are not reported because this is a pilot study, and the accuracy percentages for telephone numbers and postal addresses are not reported due to sample size concerns. Table 4 reports the Syntactical accuracy percentages for email addresses.

Table 4: Syntactical Email Accuracy Percentages for gTLDs with Largest Sample Sizes

TLD	Accuracy
gTLD A	100.00
gTLD B	100.00
gTLD C	100.00
gTLD D	100.00
gTLD E	99.99
gTLD F	99.95
gTLD G	99.94
gTLD H	99.92
gTLD I	99.89
gTLD J	99.88

Table 5 reports the Operational accuracy percentages for email addresses for the gTLDs with the largest sample sizes. Note that the ordering of these gTLDs is different from than that based on Syntactical accuracy.

Table 5: Email Accuracy Percentages for gTLDs with Largest Sample Sizes

TLD	Accuracy
gTLD A	91.62
gTLD D	96.53
gTLD B	93.63
gTLD J	92.54
gTLD F	91.46
gTLD I	91.19
gTLD E	90.88
gTLD C	90.65
gTLD H	88.93
gTLD G	88.02

Tables 6 and 7 report on the accuracy percentages for email addresses by registrar. Table 6 reports the Syntactical accuracy percentages for email addresses for the registrars with the largest sample sizes.

Table 6: Syntactical Email Accuracy Percentages for Registrars with Largest Sample Sizes

Registrar	Email Accuracy
Registrar A	100.00
Registrar B	100.00
Registrar C	100.00
Registrar D	100.00
Registrar E	100.00
Registrar F	99.97
Registrar G	99.95
Registrar H	99.95
Registrar I	99.93
Registrar J	99.92

Table 7 reports the Operational accuracy percentages for email addresses for the registrars with the largest sample sizes. As with the gTLDs with the largest sample sizes, the Operational accuracy order for registrars changes compared to the Syntactical accuracy order.

Table 7: Operational Email Accuracy Percentages for Registrars with Largest Sample Sizes

Registrar	Email Accuracy
Registrar F	94.6
Registrar D	94.4
Registrar I	94.0
Registrar A	94.0
Registrar E	93.9
Registrar C	92.6
Registrar G	92.6
Registrar H	92.5
Registrar B	92.3
Registrar J	89.6

Ongoing Work for the Pilot Study

At present, the tasks of data collection, validation, and analysis for the Pilot Study have been completed. Currently, ICANN is focusing on refining elements of the study that will be critical for establishing a replicable framework for implementing the ARS. The rules for determining accurate and inaccurate WHOIS records need to be considered further. The vendors that performed the validation steps provided reasonable assessments of postal address, email address and telephone number accuracy. However, more thought needs to be given as to whether the assessments fit the definitions of the accuracy categories (No Failures, Minimal Failure, Limited Failure, Substantial Failure, and Complete Failure). An evaluation of the methodologies will be conducted to determine ways to make phone, email and postal address accuracy evaluations comparable.

There is also a desire to combine email, postal and telephone evaluations into an overall accuracy score both within and across perspectives. Once the evaluation methods for each of the fields are made more comparable, methods for combining the information will be considered. Thus, there will be an overall syntactical assessment for each record, which can be aggregated to provide syntactical assessments for Geographic Regions, registrars and gTLDs. Similarly, an overall operational assessment can be done. Finally, a global assessment that combines syntactical and operational methods will be developed.

After this Pilot Study is complete, a report will be released to the ICANN Community for comment. The feedback provided will help to design a WHOIS Accuracy Reporting System that the community finds informative.

Implications of the Pilot Study

The Pilot Study was a highly informative exercise, and provided valuable lessons and caveats for the methods and procedures of the ARS. Some of the key lessons learned from the pilot are:

- **Slow Process:** Selecting samples of WHOIS records, turning the raw WHOIS information into parsed fields suitable for evaluation, and evaluating postal addresses, email addresses and telephone numbers takes time, and should not be rushed. The timetable for the Pilot Study was greatly condensed and should be lengthened when the WHOIS Accuracy Reporting System is launched.
- **Postal Address Validation:** Unlike email address and telephone number, a postal address is comprised of a number of fields. This makes it much more difficult to validate postal addresses and the outcomes may be quite different than those of email and telephone. The five category evaluation scale (No Failures, Minimal Failure, Limited Failure, Substantial Failure, and Complete Failure) may not necessarily be useful for postal address validation, while it may be useful for email and telephone validation.
- **Sample Size Issues:** It takes more time to validate postal addresses and telephone numbers than it does email addresses. Furthermore, it takes longer to perform operational validation than syntactical validation for both postal addresses and telephone numbers. Because of this, different sample sizes were used. This means that the accuracy measures based on smaller samples are less reliable than those the used on larger samples. In particular, operational validation for postal address and telephone number used a sample of 1,000 WHOIS records. This may be too small to provide good estimates for comparing Geographic Regions, registrars and gTLDs. As such, a larger sample should be considered in the future.