Statistical Journal of the IAOS 35 (2019) 201–209
DOI 10.3233/SJI-180453
IOS Press

Accuracy in contact information for website registrations

Steven Pedlow[a,*], John Lickfett[a], Ed Mulrow[a], Cyrus Jamnejad[b] and Jared Erwin[b]
[a]*NORC at the University of Chicago, 55 E. Monroe Street, Chicago, IL 60603, USA*
[b]*ICANN, 12025 Waterfront Drive, Los Angeles, CA 90094, USA*

Abstract. The Internet Corporation for Assigned Names and Numbers (ICANN) coordinates web address identifiers worldwide. Any individual, business, or organization that registers a domain name must provide names, addresses, emails, and phone numbers for the registration service called WHOIS. This data is managed by "registries", which are under contract with ICANN to operate top level domains like .com, .org, or new ones now in operation (.consult, for example).
Anyone can use WHOIS to search and identify the registered name holder of a domain name. The WHOIS Accuracy Reporting System (ARS) is a formal examination of the accuracy of contact information provided to registries and registrars. This project examines syntax (does the email contain an "@" symbol; does the phone number have the correct number of digits; does the mailing address have the required fields?) and operability (does the email bounce; does the phone number connect; is the postal address deliverable?) accuracy. This paper will provide statistics on Internet domains worldwide as well as the accuracy of domain contact information based on four studies over two years.

Keywords: ICANN, WHOIS

1. Introduction

The Internet Corporation for Assigned Names and Numbers (ICANN) was contracted by the U.S. government to coordinate website names and numbers during the early years of the Internet. Now, ICANN coordinates website names and numbers within generic top level domains (gTLDs). The most common gTLDs coordinated by ICANN are .com, .net, and .org, but ICANN started to register new top level domains in October, 2013. The largest of these is currently .xyz. ICANN also coordinates country-specific top-level domains like .uk. Certain top level domains such as .gov, .edu, and .mil (for military) are outside of ICANN coordination.

As you may know, any individual, business, or organization can register a domain name as a registrant. For example, www.norc.org is a non-profit organization website for which the ISO (Infrastructure, Security, and Operations) Administrator at NORC at the University of Chicago is the registrant. The domain name is the unit of analysis for this paper. Each gTLD is operated by a registry operator or sponsor, who is delegated the operation by ICANN through registry agreements. Registrars such as GoDaddy actually sign up registrants and collect website registration fees from the registrants. Registrars must be ICANN-Authorized through a Registrar Accreditation Agreement (RAA). Some registrars are still working under a 2009 RAA, but most are now working under a 2013 RAA, which has additional requirements. Domains may be grandfathered to the 2009 RAA if they were created or last revised their registration before the registrar switched to a 2013 RAA. In order to manage domains worldwide, ICANN has a need for data systems. WHOIS is one of ICANN's administrative systems that has been organized for this purpose.

Registrants are required to provide certain information to registrars. This includes the registrant name and contact information including email addresses, telephone numbers, and postal addresses. This information is often referred to as "WHOIS data". But the WHOIS

*Corresponding author: Steven Pedlow, NORC at the University of Chicago, 55 E. Monroe Street, Chicago, IL 60603, USA. E-mail: pedlow-steven@norc.org.

service is not a single, centrally-operated database. Instead, the data is managed by independent entities known as "registrars" and "registries". WHOIS isn't an acronym, though it may look like one. In fact, it is the system that asks the question, *who is* responsible for a domain name or an IP address? All of the information provided by registrants, as well as other information from the registrars is publicly available. One place you can look up the information for a particular domain name is at https://whois.icann.org/en/about-whois.

ICANN's Accuracy Reporting System (ARS) is a project to determine the accuracy of registration information for all domain names in gTLDs that are coordinated by ICANN. There were two motivations for the ARS. First, to proactively identify potentially inaccurate data and determine the rates of inaccuracy. Second, to forward potential inaccuracies to registrars for follow-up and correction. ICANN leads a team of vendors on the ARS project, including NORC. NORC is in charge of sampling, testing specifications, and analysis while Whibse parses the WHOIS data ICANN obtains from various databases, DigiCert carries out the testing for the email addresses and telephone numbers, and Universal Postal Union (UPU) carries out the testing for postal addresses.

The ARS examines nine contact fields. The three contact modes are email addresses, telephone numbers, and postal addresses. Separate contacts are required for three contact types: the registrant, a technical contact, and an administrative contact. However, these three contacts can be the same, and about 75 percent of the time, the information across these three contact types is the same. For NORC, our registrant is the Manager of the ISO department while the technical and administrative contacts are both the same Senior Engineer within ISO. The email address and telephone number for the registrant are not required under the 2009 RAA (but a registrant postal address is required), but they are required under the 2013 RAA. Even under the 2009 RAA, email addresses and telephone numbers are required for the technical and administrative contacts.

The ARS determines accuracy separately by syntax and operability. The accuracy testing criteria can be found at https://whois.icann.org/en/whoisars-validation. To be determined to be syntax accurate, the contact must satisfy all requirements for validity. For email addresses, all characters must be permissible, the "@" symbol is required, there must be characters before the "@" symbol (the "local" component), there must be a valid top-level domain at the end, and there must be a valid domain after the "@" symbol, but before a "." and the valid top-level domain. For telephone numbers, there must be only permissible numbers and formatting characters (dashes and parentheses), and a valid country code followed by the correct number of digits for that country (possibly including a valid extension number that is syntax accurate). For postal addresses, there needs to be an identifiable and syntax valid country (or country code), and the following fields also have to be filled in and syntax accurate, if they are required in that country: postal code, state or province, city, and street.

To be determined to be operably accurate, the contact has to be operable; in other words, the email cannot bounce, the telephone number has to connect without an error message such as that the number is invalid or disconnected, and the postal address must be mailable. The vendors for the ARS actually send emails to the email addresses and dial the telephone numbers. However, due to the necessary time lag and unavailability of receipt confirmation worldwide, the vendor (UPU) uses a tool to automatically determine the operability of the postal addresses.

The ARS started with a pilot "proof of concept" in 2014 and a report was published on December 23, 2014. We then ran a much-improved "phase 1" with only syntax accuracy in the first half of 2015, with a report published on August 24, 2015. Since then, we have completed five cycles in "phase 2" with both syntax and operability accuracy, with two cycles per year. Reports have been published on December 23, 2015 (Cycle 1), June 8, 2016 (Cycle 2), December 12, 2016 (Cycle 3), June 12, 2017 (Cycle 4), and December 19, 2017 (Cycle 5, the data of which is not included in this paper). Cycle 6 is currently underway and its report will be published in June, 2018. ARS information and reports are available online at https://whois.icann.org/en/whoisars.

In this paper, we focus on using data from the first four cycles of Phase 2 with both Syntax and Operability Accuracy. Beyond just the accuracy details, we can learn other interesting things about the Internet in Section 2, which shows some summary statistics about the domains in our study population, Section 3 will briefly discuss our sampling strategy, and Section 4 will focus on the accuracy of the contact information. In Section 5, we will summarize and provide a glimpse of the future for the ARS.

2. What does the Internet look like?

Prior to 2013, there were only 18 global top-level domains ("prior" domains). This excludes the country-

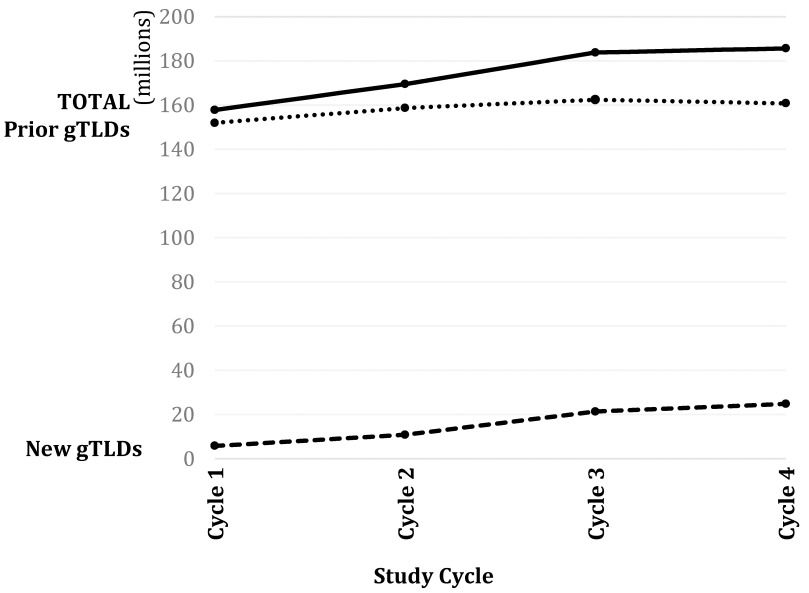

Fig. 1. Number of domains in prior and new gTLDs over time.

Table 1
Top 12 Generic Top-Level Domains, as of January 2017

gTLD	Domains	Percentage	gTLDtype
.com	125,792,045	67.8%	PRIOR
.net	15,120,252	8.1%	PRIOR
.org	10,458,297	5.6%	PRIOR
.xyz	6,023,285	3.2%	NEW
.info	5,378,675	2.9%	PRIOR
.top	4,311,646	2.3%	NEW
.biz	2,256,331	1.2%	PRIOR
.win	1,119,034	0.6%	NEW
.wang	862,057	0.5%	NEW
.loan	852,617	0.5%	NEW
.club	805,036	0.4%	NEW
.mobi	639,356	0.3%	PRIOR
Total	185,651,496	100.0%	n/a

code top-level domains as well as the top-level domains not supervised by ICANN (.gov, .edu, and .mil). In October, 2013, ICANN started accepting new top-level domains. Applicants who want to create and run a new top-level domain fill out an application that is reviewed at ICANN. Some new top-level domains are already quite popular, but .com still contains over two-thirds of all domains eligible for the ARS. Table 1 shows the 12 most popular generic top-level domains, by the number of registered domains.

The three most popular top-level domains are the prior gTLDs .com, .net, and .org, but the fourth most popular is the new .xyz gTLD. The idea for .xyz was that "it was created to merge generations x, y, and z." Google helped popularize .xyz by choosing the domain abc.xyz for its parent website, Alphabet Incorporated. Two other new gTLDs, .top and .wang, are both popular in China, and as we'll see later, domains are growing in Asia faster than anywhere else in the world. Figure 1 shows the growth in domains for Prior gTLDs and New gTLDs.

Figure 1 shows that the number of new gTLD domains doubled between Cycle 1 and Cycle 2 and doubled again between Cycle 2 and Cycle 3 (the scale makes these doublings difficult to see). This growth slowed between Cycles 3 and 4. Meanwhile, the growth in prior gTLD domains has slowed and actually decreased in Cycle 4. Overall, there was consistent growth of 12–14 million per cycle between Cycles 1 and 3, but a much smaller growth in Cycle 4.

As hinted above, the distribution by region of the world has been changing over time. This is shown in Table 2. Compared to the other (ICANN) regions, the growth in the Asia-Pacific region is much greater. In less than two years, the percentage of domains in Asia has increased from 22 percent to 33 percent. The region with the second largest growth is the Latin America/Caribbean region, but the percentage of domains in this region has only increased from 4 percent to 5 percent in the same time period. The number of domains in North America actually dropped from Cycle 3 to Cycle 4. The percentage of domains in North America has dropped below 50 percent of the total, from 53 percent to 43 percent in less than two years. Europe has continued to grow, but the percentage in Europe has still slightly decreased. The percentage of domains in

Table 2
Estimated number of domains over all four cycles (in millions)

Cycle	TOTAL (millions)	North America	Asia-Pacific	Europe	Latin Amer./ Caribbean	Africa	Unknown
Cycle 1	157.9	53.4%	22.0%	19.2%	4.0%	0.7%	0.7%
Cycle 2	170.0	50.3%	25.7%	18.4%	4.1%	0.7%	0.7%
Cycle 3	184.1	47.8%	28.7%	18.0%	4.5%	0.7%	0.3%
Cycle 4	185.7	42.7%	33.1%	18.3%	5.0%	0.6%	0.2%

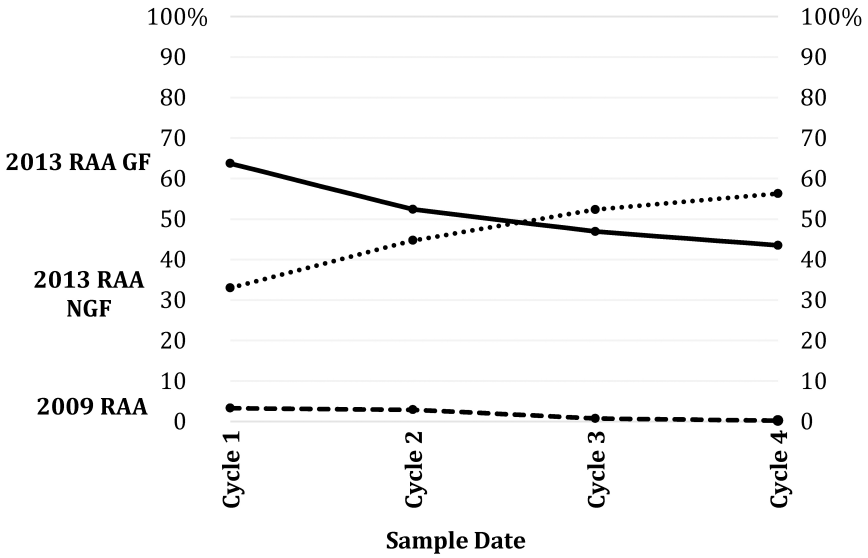

Fig. 2. Distribution of domains by RAA status over time.

Africa has remained consistent at less than 1 percent. In Cycle 3, we improved our ability to assign countries so that less than half a percent could not be identified rather than just under 1 percent.

As time proceeds the number of domains with registrars under the 2009 RAA is rapidly shrinking. The percentage of grandfathered 2013 RAA is also dropping. However, there are still enough grandfathered domains (that are only subject to the 2009 RAA requirements) that the ARS baseline accuracy tests use the 2009 RAA requirements. We still perform accuracy tests using the 2013 RAA requirements, but it is proper to only hold accountable those domains that have registrars under the 2013 RAA and are not grandfathered. In this report, all accuracy testing shown uses the 2009 RAA requirements. Figure 2 shows the distribution of RAA status over time.

By Cycle 1, the percentage of 2009 RAA domains had already dropped to under 3.5 percent, but the percentage in Cycle 4 is only 0.2 percent. In Cycle 1, more than 63 percent of the domains had registrars under the 2013 RAA, but needed to be grandfathered to the 2009 RAA requirements. This percentage dropped below 50 percent in Cycle 3, and is 43.5 percent in Cycle 4. The percentage of domains that can truly be held to 2013 RAA standards (2013 RAA non-grandfathered) has risen from 33.0 percent in Cycle 1 to 56.3 percent in Cycle 4. All statistics in this report use 2009 RAA standards because 43.7 percent of the domains still are not required to fulfill all of the 2013 RAA standards.

3. The ARS sampling strategy

Due to the lack of information available for all domains, a multi-stage sample is required. In fact, the only known information for our initial sampling is the number of domains in each generic top-level domain. An analysis sample of 12,000 is currently analyzed by vendors for accuracy. This sample size is large enough to obtain estimates in most cells of interest, but is small enough that the vendor analysis can be done in six weeks. To obtain sufficient records for some cells, we draw an initial sample of 200,000 records from gTLD "zone" files. The analysis subsample involves heavy oversampling of some cells.

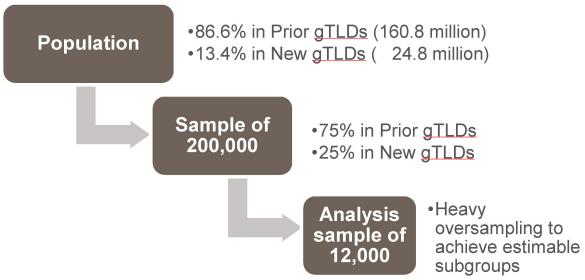

Fig. 3. Summary of the ARS sampling strategy.

At the time of the initial sample for Cycle 4, in January 2017, there were 185,651,496 domain names spread across 1,231 gTLDs. As shown in Fig. 1, approximately 87 percent of the domains were registered in one of the 18 prior gTLDs. Approximately 13 percent of domains in January 2017 were registered in new gTLDs. The overall number of new gTLDs has risen sharply, increasing from 678 in Cycle 1 to 1,213 in Cycle 4 eighteen months later.

Out of the 1,213 new gTLDs in Cycle 4, only 779 had at least one domain (434 new gTLDs did not yet have any domains). Of the 779, 61 had exactly one domain (these were excluded from our sample since it is typically an administrative domain for the gTLD) and the remaining 718 others had at least two domains. Adding together the 18 prior gTLDs and the 718 new gTLDs described above, the initial sample represented a total of 736 gTLDs. In order to have a sufficient number of new gTLD domains, the ARS project has always selected 25 percent of the initial sample from new gTLDs.

Of the initial sample of 200,000, WHOIS data are gathered and parsed successfully for almost all of the records (this percentage was 98.7 percent in Cycle 4). Some domains no longer exist while other records turn out to be registry-reserved domains and are excluded from the study. Our goal in sampling is to have 800 or more domains in each region for each RAA type. However, this is not possible for Africa or the 2009 RAA domains, so these initial sample domains are all selected for the analysis sample. Unknown region domains are selected at the North American rate (the smallest rate). Figure 3 summarizes the sampling strategy.

All of the estimates in this paper are weighted based on the Horvitz-Thompson weight [1], which is the inverse of the sampling probabilities at the initial and subsample stages. This includes population estimates by RAA Type and Region given above in Section 2.

4. Accuracy of the contact information

Our accuracy criteria are quite strict. For a domain to be determined to be completely accurate, all our tests must be passed for all nine contact fields: the email address, telephone number, and postal address for all of the registrant, administrative, and technical contacts. Below, we present accuracy estimates separately for the syntax requirements and actual operability. We do have a weaker standard that we call "immediately contactable", which requires only one of the six email addresses and telephone numbers to be operable. This simulates whether the domain name could actually be immediately contacted from the public information available. For Cycle 4, the immediately contactable rate was over 98 percent.

Figure 4 shows the syntax accuracy for the three contact modes. The domain is considered syntax accurate if and only if all three contact types (registrant, administrative, and technical) pass all the syntax tests. Almost all email addresses pass all of the syntax tests while the syntax accuracy for postal addresses and telephone numbers are between 80 and 90 percent. While the postal address syntax accuracy line is flat, the syntax accuracy for telephone numbers has been rising. The lowest line is the overall syntax accuracy (all nine contacts are syntax accurate, which has risen from 73.1 percent in Cycle 1 to 79.3 percent in Cycle 4. This rise is due to the rising syntax accuracy of telephone numbers.

Figure 5 shows the operability accuracy for the three contact modes. The domain is considered operably accurate if and only if all three contact types (registrant, administrative, and technical) pass the operability tests. Postal addresses show a higher operability rate, but postal addresses are only tested by a database tool; no actual mailing is attempted. Email address operability shows an increase in Cycle 4, but telephone number operability has dropped in Cycles 3 and 4 (while syntax accuracy is rising). We will be watching these trends in future cycles. Again, the lowest line is the overall syntax accuracy (all nine contacts), which has consistently been near 65 percent, except in Cycle 2 when it rose to 70 percent.

Now, we turn to the types of accuracy errors we see and how often they occur. For email addresses, there are very few syntax errors and the operability errors almost all fall into the category of "email bounced". Figure 6 shows the number of telephone number syntax errors in Cycle 4 for the administrative contact. We limit this graph to the administrative contact because the dis-

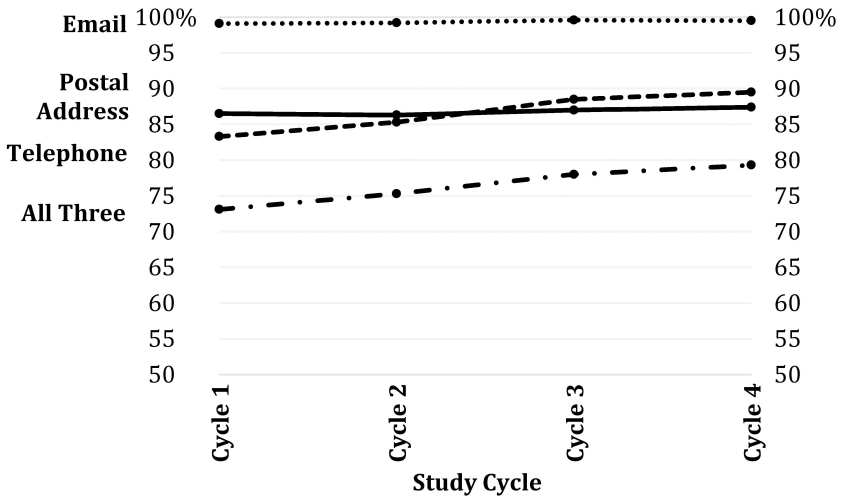

Fig. 4. Syntax accuracy for the three contact modes.

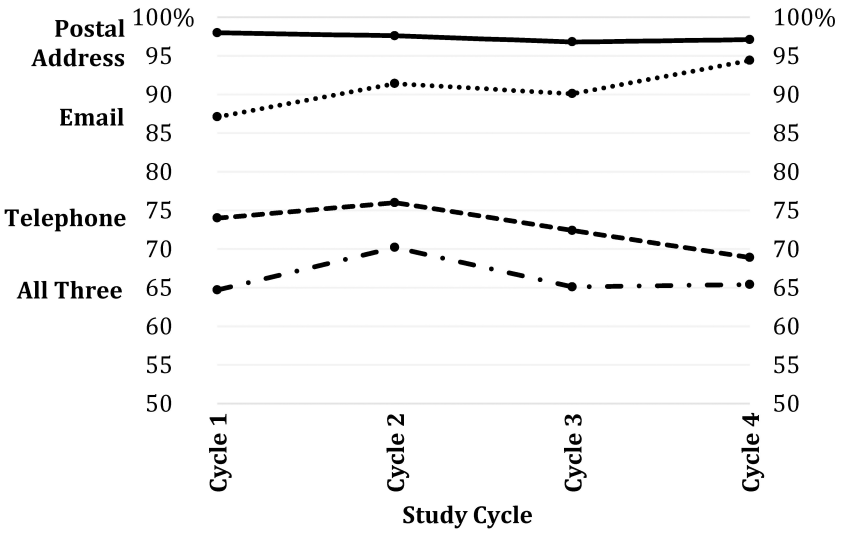

Fig. 5. Operability accuracy for the three contact modes.

tributions are the same for the registrant, administrative, and technical contacts, there is duplication among these telephone numbers, and using only the administrative contact means there are 12,000 possible errors (not 36,000).

Figure 6 shows that there are very few missing or unallowable telephone numbers. Since there are 12,000 telephone numbers, the unweighted error rate is 1,398/12,000 = 11.6 percent. Almost two-thirds of the telephone number syntax errors are due to the phone number having an incorrect number of digits, with most of the rest lacking a required country code.

Figure 7 shows the number of telephone number operability errors in Cycle 4 for the administrative contact. The unweighted error rate is higher for operability: 3,623/12,000 = 30.2 percent. Almost half of these errors are due to an invalid number being provided (an operator message is received) while about one-third of the errors are disconnected numbers (an operator message is received). "Other not connected" errors, which are errors for which no operator message is received (too few digits would be one possibility), only account for about one-sixth of the telephone number operability errors.

Figure 8 shows the number of postal address syntax errors in Cycle 4 for the administrative contact. The error rate (even unweighted) is not determinable from Fig. 8 for postal address syntax errors because

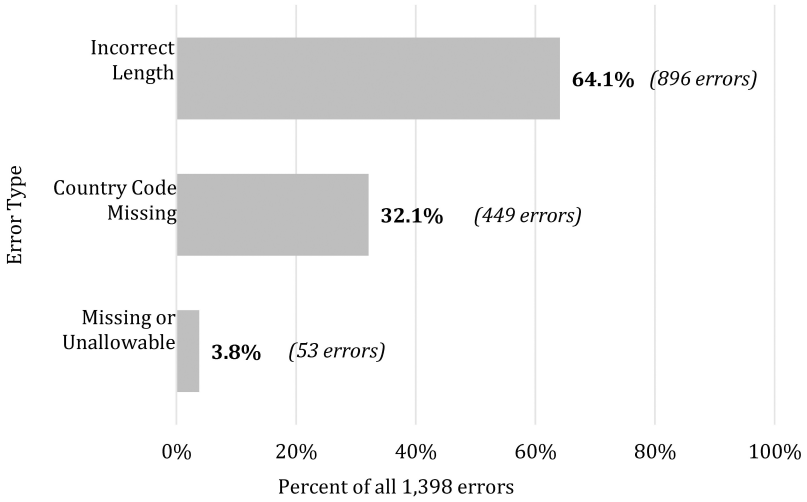

Fig. 6. Telephone number syntax errors for administrative contact.

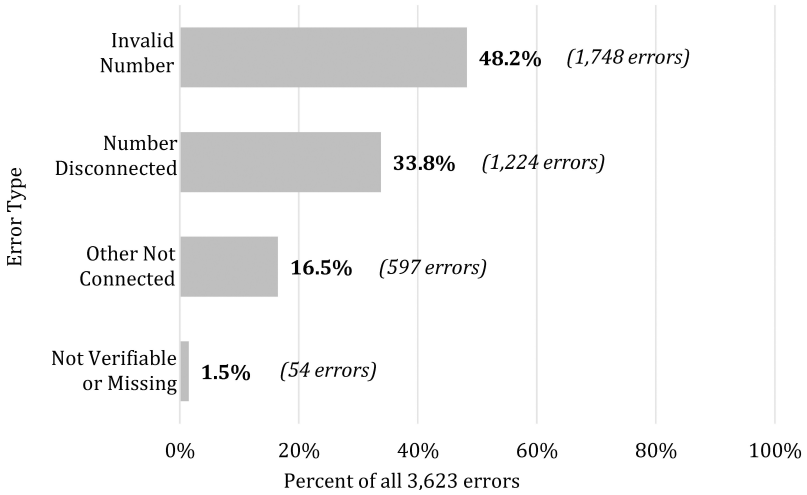

Fig. 7. Telephone number operability errors for administrative contact.

postal addresses can have multiple errors. Most syntax errors are required fields that are missing. Almost ninety percent of the syntax errors were for missing a required street, city, or postal code. Less than seven percent of the errors were for an entirely missing address or a missing or unidentifiable country code. We do not present any breakdown of the postal address operability errors since the operability accuracy determination for postal addresses is mostly a binary determination.

We can also compare accuracy across regions of the world, which we show in Table 3. The highest accuracy rates are in North America. Operability is lowest in Asia, which is at least partly explained by the difficulties of Chinese characters. Syntax accuracy is lowest in Africa, where the number of domains is also the lowest. It is important to remember that these numbers are for strict overall accuracy, which requires all nine contacts to be operable or to satisfy all the syntax requirements. Rates for immediate contactability are in the high nineties for all ICANN regions, which means that most registrants can be immediately contacted with publicly available information.

Finally, Fig. 9 shows operability rates across time by RAA Status. Domains whose registrars are under the 2013 RAA and who do not need to be grandfathered are generally newer domains than those that are grandfathered or have registrars under the 2009 RAA. Once again, Figure 9 contains overall operably accuracy rates (all nine contacts must be operable).

Table 3
Postal address syntax errors for administrative contact

	All	North America	Asia-Pacific	Europe	Latin Amer./ Caribbean	Africa
Syntax	79.3%	88.3%	68.8%	74.5%	78.1%	46.1%
Operability	65.4%	81.2%	42.1%	59.3%	74.2%	51.6%

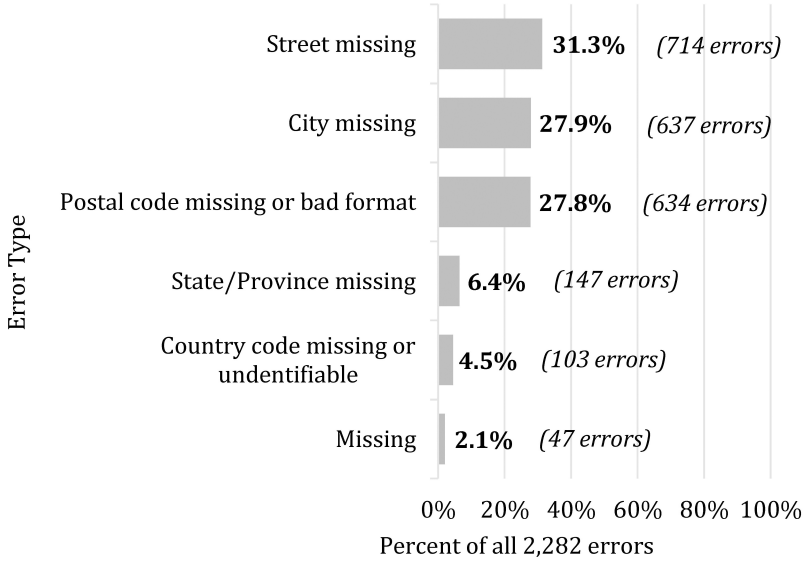

Fig. 8. Postal address syntax errors for administrative contact.

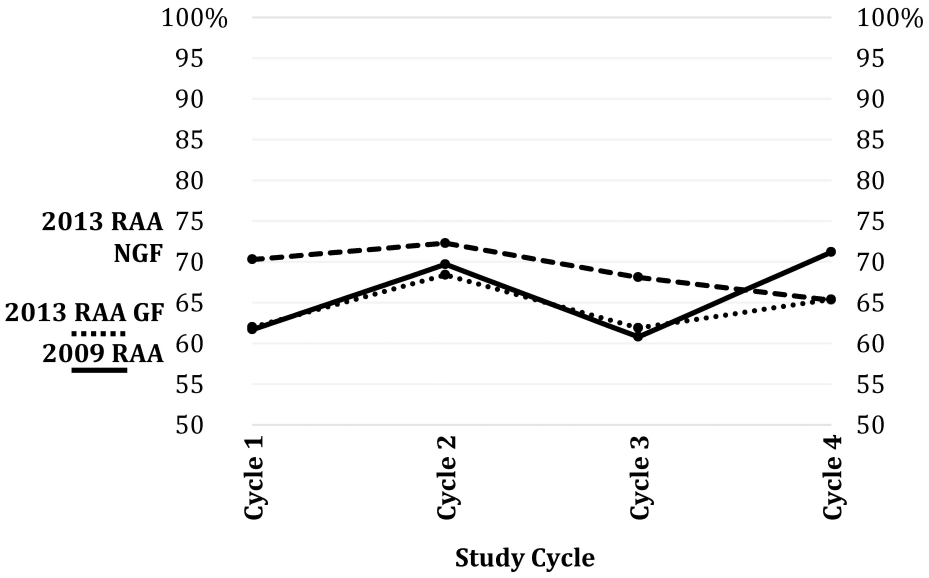

Fig. 9. Operability rates by RAA status over time.

It is not surprising that the 2009 RAA line is noisy since the number of records in this group is shrinking so fast. However, the other RAA groups also show some variability. In Cycle 4, the grandfathered and non-grandfather groups meet at 65 percent. However, the non-grandfathered subgroup has declined in the last two cycles while the grandfathered subgroup has varied around a 65 percent line. It seems four cycles are

not enough to detect a trend, so we will continue to monitor accuracy for the RAA Types.

5. Summary and the ARS future

The Accuracy Reporting System (ARS) examines both syntax and operability accuracy of WHOIS contact information over several dimensions, focusing on rates of conformance by contact mode (email, telephone or postal) to the requirements of RAAs (2009 RAA or 2013 RAA). For over 75 percent of domains, the contact information in the registrant, administrative, and technical contacts is identical for any one of the three contact modes, revealing why accuracy rates among the three contact types are all similar.

The ARS has grown into a system used for repeatable assessment, with a new cycle every six months. Over time, our measurement methods have improved, and the statistics in this paper reflect the improvements. Some of these improvements have occurred during the Compliance follow-up in which errors detected are forwarded on to registrars, who begin the process of getting errors fixed or disputing the errors. It is hoped that over time, the ARS will result in a reduced number of errors as registrars become better able to enforce the RAA requirements with registrants. On May 25, 2018, the European Union General Data Protection Regulation became enforceable. This will affect ICANN's monitoring work, so we eagerly anticipate the effect on Cycle 7 of the WHOIS ARS.

Acknowledgments

Many others from ICANN have contributed to this project, including Christopher Bare, Russ Weinstein, Grant Nakata, Owen Smigelski, Jonathan Denison, Harel Efraim, Matt Ashtiani, Margie Milam, and Steve Allison.

Additional NORC personnel include Zach Seeskin, Yongheng Lin, Amy Ihde, Becki Curtis, Fang Wang, Michael Zeddies, and Andrea Kaplan.

Reference

[1] Horvitz DG, Thompson DJ. A Generalization of Sampling Without Replacement From a Finite Universe. Journal of the American Statistical Association. 1952 Dec; **47**(260): 663-85.