The Harvard Kennedy School Misinformation Review[1]
August 2020, Volume 1, Special Issue on COVID-19 and Misinformation
Attribution 4.0 International (CC BY 4.0)
Reprints and permissions: misinforeview@hks.harvard.edu
DOI: https://doi.org/10.37016/mr-2020-036
Website: misinforeview.hks.harvard.edu

Research Article

Ambiguity in authenticity of top-level Coronavirus-related domains

During the novel coronavirus (Covid-19) crisis, citizens have been attempting to obtain critical information and directives from official government websites. These are usually hosted on top-level domains, such as coronavirus.mx. There is no reliable mechanism to verify these websites' authenticity, and the space is also shared by commercial entities selling related (or not) products and advertisements. This loophole is an urgent information security and misinformation problem that can be resolved by registering websites under restricted second-level domains or adopting existing methods of domain registrant identification.

Authors: Nathanael Tombs (1), Eleonore Fournier-Tombs (2)
Affiliations: (1) Solarisbank AG, Germany, (2) Faculty of Law, University of Ottawa, Canada
How to cite: Tombs, N.; Fournier-Tombs, E. (2020). Ambiguity in Authenticity of Top-level Coronavirus-Related Domains. *The Harvard Kennedy School (HKS) Misinformation Review*, Volume 1, Special Issue on COVID-19 and Misinformation
https://doi.org/10.37016/mr-2020-036
Received: April 16th, 2020. Accepted: August 1st, 2020. Published: August 31st, 2020.

Research questions

- To what extent can we determine the authenticity of top-level coronavirus-related websites that purport to be government websites?
- For what purpose would a non-governmental entity or company register a top-level coronavirus-related domain name?

Essay summary

- After a broad review of websites using coronavirus-related domain names, specifically coronavirus, covid19, covid-19, sars-cov-2 and sarscov2, we found that we could not verify the authenticity of over 80% of websites presented as government websites.
- Of the 303 websites surveyed, 90 (or nearly 30%) had unverified information and nearly half were squatting domains or 'under construction.'
- Government websites providing its citizens with life-critical coronavirus-related information should not be subject to this ambiguity, and should therefore not share the top-level domain name space with non-governmental individuals or entities. This finding will be critical in

establishing trusted communication channels between governments and their citizens during this crisis.

Implications

In this study, we present how misleading information was disseminated under top-level coronavirus-related domain names, based on data collected from 303 websites between April 5 and April 6, 2020. We find that the authenticity of a striking number of websites posing to be of government source cannot be verified, and of those that are overtly non-governmental, many are selling products, advertising, or domain names. We argue that this is due to the co-existence of governmental and non-governmental organisations in the same domain space, which undermines the authenticity and trustworthiness of the information being presented (Cherdantseva and Hilton, 2013).

In 1984, the naming convention for the first internet domains was developed by the Internet Engineering Task Force (IETF) (Internet Engineering Task Force, 2020). The document outlined the internet's first Top Level Domains (TLDs) including .gov, which it reserved for US Government domains. This ensured that US citizens could be confident that information and services provided by websites hosted on .gov domains would be reliably related to the authorities. It also laid out two letter country code TLDs (ccTLDs), based on the ISO-3166 standard. In 1994, a follow-up (Internet Engineering Task Force, 2020) was published, which outlined the use of these ccTLDs. The responsibility to determine how these domains were used, and whether government-specific Second-Level Domains (2LDs) were assigned, was delegated to country administrators. This led to two major discrepancies in the use of TLDs. The first was whether these domains could be registered by a global entity, as is the case for the Ionian Island's .io domains, or whether registration would be restricted to local entities, as is the case for China's .cn domains. In practice, however, major registrars offer trustee services. By offering local legal proxies to their registrants, the registrars allow international entities to bypass these regional restrictions and access otherwise restricted ccTLDs. This also serves to anonymize the real registrant. The second was the use of a reserved government 2LDs for which there was no convention. As an example, the Canadian government's primary website is canada.ca, whereas the Australian government's primary service website is australia.gov.au. The Canadian government uses their publicly available ccTLD, whereas Australian government leverages a reserved government 2LD.

As governments strive to give their citizens access to authoritative information and services relating to the coronavirus, they are using their country's ccTLD in order to register these domains. Given that anyone can register these domains, citizens have no authoritative way of differentiating between government websites, commercial websites, or malicious websites. Furthermore, domain profiteering is occurring, resulting in private individuals or companies registering corona-virus related domains in order to sell them at a considerable premium. The result is a network of domains where covid19.bz, .mx and .se live in parallel. The first is a seemingly legitimate government site asking for citizens' social security numbers and bank information, the second is a fake government website asking citizens to fill out an online questionnaire, and the third is a commercial website selling medical supplies.

Domain registration and content served by associated web servers can change rapidly without leaving an audit trail. Furthermore, web servers can be configured to serve different content to different audiences based on their devices, browsers or location. As a result, this study is conducted based on the samples taken from a data stream which is in perpetual movement.

In our study, we found that a striking 29.7% (90 out of 303 websites surveyed) of websites surveyed could contain misinformation, from aggregating misleading information and possibly impersonating government authorities, to selling unrelated products. A further 48.51% (147 out of 303 websites surveyed) were squatting the domains or not providing any information, either to become future websites

disseminating misinformation, or charging exorbitant sums to those attempting to provide authoritative information. This shows not only a striking opportunism when it comes to the pandemic, but also a security gap in the domain naming system that could easily be addressed.

As we discussed, a majority of websites appearing to be official government sites are registered by third parties. The fact that governments are sharing top-level domains with non-government individuals and entities makes it very difficult to confirm the authenticity of official websites. This loophole has allowed a number of malicious website developers to exploit users at a particularly vulnerable time.

While there has been some work on top-level domains and the authenticity of health information (Walther and Wang, 2004) research on website information credibility has largely been content-based, rather than domain-based (Sbaffi and Rowley, 2017). When it comes to misinformation related to the Covid-19 pandemic, studies have targeted social media, again with a focus on content (Rosenberg et al, 2020). This research therefore presents a rather unique perspective, with some applicable solutions, discussed below.

The first possibility would be for governments to limit themselves to restricted Second-Level Domains. This would allow users to validate the authenticity of the website by inspecting its domain. This could be considered for new websites, but would be impractical for wide roll-out since many government websites already use unrestricted ccTLDs. The second possibility would be for governments to adapt the emerging Registration Data Access Protocol (RDAP) to publish ownership information in a standard and consistent manner. This information is publicly searchable. While it is not approachable to most non-technical users, it could be leveraged by browsers or browser extensions to add a security indicator to websites, such as a security icon.

The relevance of this research is amplified by the fact that citizens around the world are turning to their government – and therefore government websites – for life-saving information. The ambiguity present in these top-level, seemingly authoritative domains could have critical implications, and should be addressed as soon as possible. Furthermore, the underlying reasons for which the authenticity of a government website cannot be properly established, apply to any scenario in which governments act quickly and proper protocol is not followed. A simple and more robust authentication method would be applicable in any crisis situation.

This research could be enhanced in the future in several respects. First, the study presents a portrait of the Covid-19 website landscape in early May 2020 – as the pandemic evolves, there will no doubt be more top-level domain websites registered. Second, it would be worthwhile to study evolving traffic to these websites, and to evaluate user response to the ambiguous information presented.

Findings

In total, we downloaded and analyzed screenshots from 303 websites. We obtained data samples from all 5 continents and 17 subregions. As can be seen in Table 1, we found a significant number of websites of unclear authenticity, including a significant gap between the unconfirmed and confirmed government websites. Of the 32 websites which prominently displayed government logos or copyright notices distributed over 29 countries, only 6 had verifiable ownership. Nearly 20% of websites sold commercial products or advertisements, nearly 10% had domains for sale, and 3 were verifiably malicious websites. We also noted that almost 40% of websites were in construction – websites that might fall in any of the other categories when they are completed.

Table 1. *Percentage of surveyed websites per category*

Category	Count	Percentage
Purports to be a government website	32	10.56%
Government website confirmed	6	1.98%
Domain for sale	30	9.9%
Website overtly malicious	3	0.99%
Domain a commercial initiative (advertisements)	41	13.53%
Domain a commercial initiative (products)	20	6.6%
Website in construction	117	38.62%

We propose that the following categories of websites were ambiguous enough to contain misinformation – government websites that could not be verified, overtly malicious websites and commercial initiatives. The other categories – domains for sale and websites in construction, while not misinformation per se, took advantage of the pressing need of governments to put up websites to provide accurate information. In a way, they benefited from the ambiguity in what constitutes misinformation that is accentuated by the top-level domain loophole. The table below shows the total breakdown of misinformation candidate websites.

Table 2. *Misinformation candidate websites in data surveyed*

Category	Count	Percentage
Possible misinformation	90	29.7%
Squatting domains and “in construction”	147	48.51%
Other websites	66	21.78%

Government domains: The challenge in establishing domain ownership

We were unable to definitively establish ownership for 81.25% - or 24 of 32 domains that appeared to be governmental. We found that establishing domain ownership, even when dealing with government domains, to be a difficult task, unless governments are using a restricted second-level domain. In order to establish ownership, we looked at the issuer of the site’s encryption certificates or performed a registry look-up for the domain. Both of these techniques presented their own challenges.

Certificates are issued by a third-party Certificate Authority in order to encrypt traffic between a user and a website using the HTTPS protocol. This third party can be a government entity, but is more often a non-government entity, which makes it an unreliable means to confirm domain ownership. Furthermore, web traffic from certain domains that appear to be governmental, such as Mauritius’ covid19.mu, is unencrypted over HTTP, meaning that no certificates will be available for verification. For example, the creator of the Firefox Browser, Mozilla, only recognizes five governments as Certificate Authorities, namely Hong Kong, Spain, Taiwan, Netherlands and Turkey (Mozilla Wiki, 2020; Mozilla CCADB, 2020).

Domain registry look-up, known as WHOIS data, is made available at the discretion of the country administrators and is not always available. As an example, Albania’s ccTLDs are administered by an organisation (AKEP) which does not offer these services (WIPO, 2020). If WHOIS data is available, it often proves to be unreliable. Government contractors often register domains under their personal names, or use anonymous proxies to obfuscate domain ownership (ICANN, 2020). Lastly, the collection and validation standards for WHOIS data are minimal (GoDaddy Domain, 2020), resulting in accuracy and reliability issues.

This problem is best illustrated by covid-19.tn, which appears to be a Tunisian government website; featuring a prominent government seal in the site's header and a government copyright notice in its footer. The encryption certificates are issued by a third party, the domain is registered by an individual, and the contact email is a Gmail address.

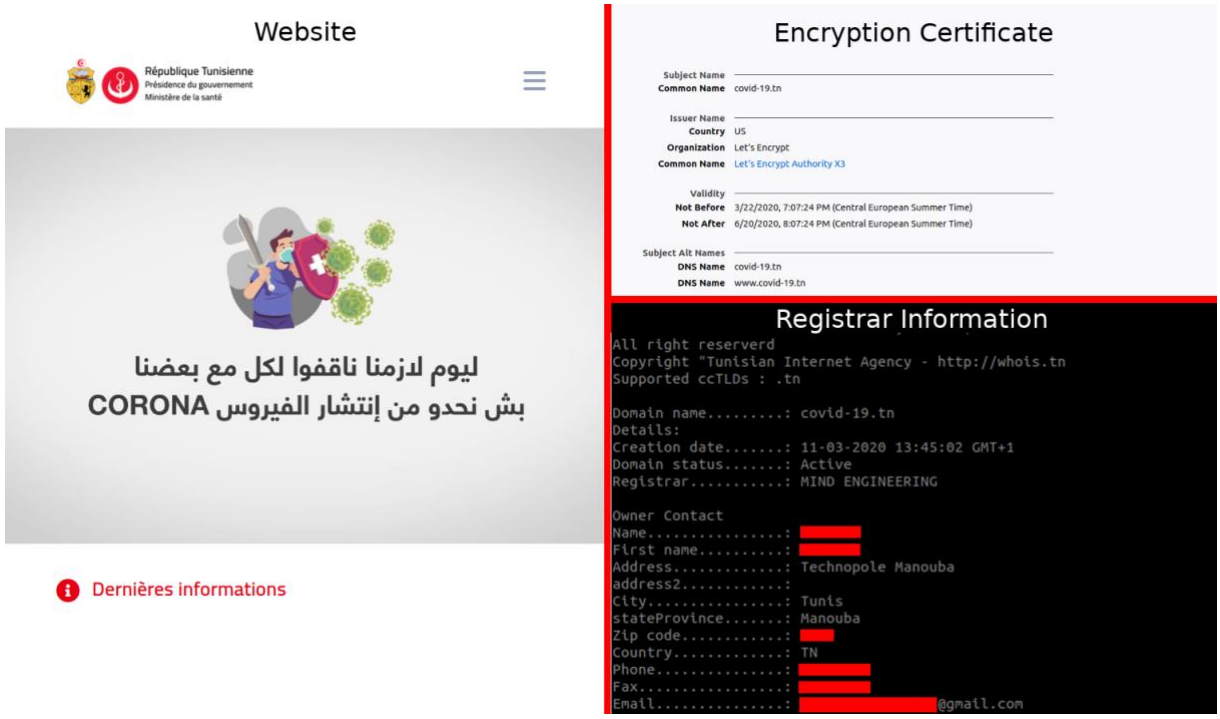

Figure 1. Unverified Tunisian government Covid-19 website

Another illustration of this issue would be the covid19.mc site which appears to be from the Government of Monaco. The registrant is not available for lookup, the servers are located in France, and the Certificate Authority (Let's Encrypt) is an American entity.

Domain squatting and commercial initiatives

In our examination of domain squatting, we chose to only look at those domains which were registered and explicitly serving a web page stating that they were available for resale. Of the 30 domains matching these criteria, eight had list prices, which averaged USD 10,889. These domain registrants benefitted from an arbitrage opportunity by rapidly registering emerging keywords. As a result, governments have to compete against domain squatters for access to their ccTLDs. One such example would be covid19.pl which was listed for 100,000 Polish Zolty (roughly USD 23,900). The first year cost of registering a .pl domain is USD 0.5 (GoDaddy Domain, 2020), meaning that this registrant is reselling at a 4,780,000% mark-up.

Commercial initiatives are those domains whose website either through advertising, direct product sales or affiliate links to marketplaces such as Amazon. These commercial websites often relied heavily on data and infographics distributed by reputable organizations. In our study, one such example was Moldova's covid-19.md, which was selling emergency supplies. Product links to alleged KN95 masks or hand sanitizer were intermixed with John Hopkins University data graphs, videos of the head of the WHO and official CDC infographics.

Figure 2. Covid-19 website using aggregated information to sell medical supplies

Overtly malicious websites

Upon review, three websites were included in this category. The first, covid19.mx, seemed to be a Ministry of Health website with an online coronavirus test, only to redirect users to paid surveys and advertising. The website was taken down shortly after data collection. The second, coronavirus.ws, points users to a pharma fraud website selling a number of counterfeit drugs, and belonging to the network of sites involved in the WordPress Pharma hacks which have been ongoing since 2015 (Kumar et al, 2018). The last, coronavirus.ma, redirects users to a lottery scam which encourages them to provide personal data to redeem a prize.

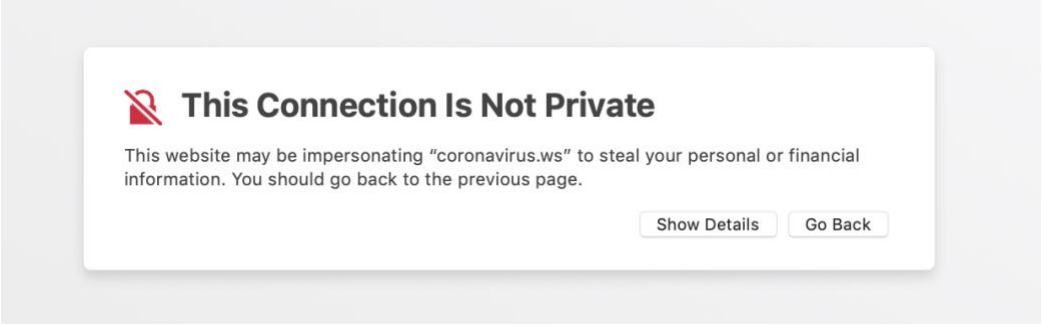

Figure 3. Browser warning when attempting to reach coronavirus.ws

Methods

In order to answer our research questions, we used a combination of automatic web scraping and manual coding of downloaded screenshots in order to answer these questions. We found that the web scraping

methodology allowed us to study as many websites as we could identify. We also found that the diversity of the websites required a manual approach when it came to analysis.

Establishing scope

The first step in conducting this study was to identify eligible domains. This was accomplished by combining variations of World Health Organization recognized names for either the COVID-19 virus or SARS-CoV-2 disease with eligible ccTLDs. These recognized names included coronavirus, covid19, covid-19, sarscov2, and sars-cov-2 (WHO, 2020).

Eligible ccTLDs were documented by the Internet Assigned Numbers Authority (Internet Assigned Numbers Authority, 2020). Geographic TLDs such as .barcelona or .africa were not included since they are not typically used by governments. Two ccTLDs were excluded; the first being **AQ** which stands for Antarctica and is a co-administered zone. The second is **US,** which has the dedicated **GOV** TLD for government domains.

This exercise yielded a total of 1165 potential domain names. Of the 1165 possible domains, 303 were registered at the time of the study, which were downloaded between May 5, 2020 09:03 EST and May 6, 2020 00:41 EST.

Obtaining data

Our primary analysis was performed visually using a coding manual. In order to ensure a standard, consistently rendered view of the websites, we developed a custom web scraper, which leveraged a web browser automation tool in order to emulate a user using the Firefox browser. It then rendered the website's home page, and took a full-page screenshot (SeleniumHQ, 2020; PyPi, 2020). We used the Yandex image translation API (Yandex, 2020) in order to identify the language of the website. If the screenshot was in a language other than English, French or Spanish, it was translated to English prior to analysis. As illustrated in Figure 5, we found a somewhat even breakdown of top-level domain websites using *coronavirus*, *covid19*, and *covid-19*, but no websites using *sarscov2* or *sars-cov-2*.

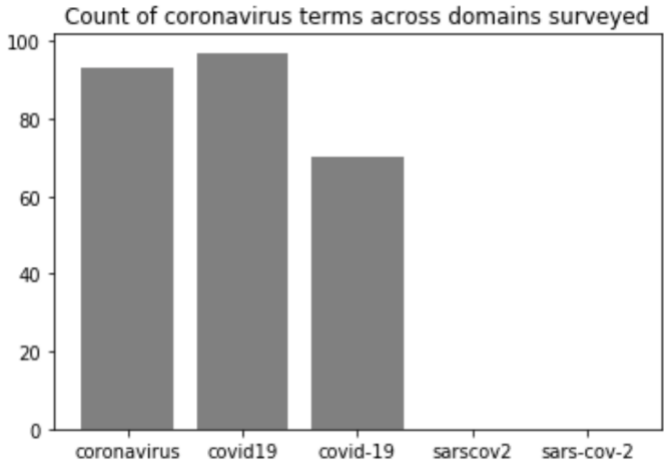

Figure 4. Distribution of recognizable terms across data sample

Coding manual

Once the data was downloaded, the individual screenshots were manually analysed using the coding protocol below. In each category, we screened manually for one of several possible flags. If one of the flags was identified, the resulting score was a 1. Otherwise, the category was recorded as a 0. The data was consensus-coded by both authors.

Table 3. *Coding manual*

Category	**Possible flags**
Seems to be a government website	• One or more government logos in the header. • Government copyright notice in the footer.
Government website confirmed	• The registry identifies the registrant as being a government entity. • The encryption certificate is issued by the government.
Domain for sale	• Explicitly for sale, either with a contact function or a list price.
Website overtly malicious	• Making users download malware. • Redirecting users to untrusted third parties. • Impersonating a trusted entity.
Domain a commercial initiative	• Selling products. • Serving advertising.
Domain a non-commercial initiative	• Appears run by non-profit organizations or universities.
Website in construction	• Serves a default web page stating that it is in construction. • Serves partial or incomplete content. • Serves an error, indicating that the web server is misconfigured. • Serves a blank page

For the sake of this exercise we chose to follow redirects. Therefore, if a domain redirected to a commercial website, as is the case with coronavirus.do redirecting to farmacard.com.do, then the domain itself was considered to be a commercial initiative.

Vocabulary and acronyms

In defining the elements of a URL, we will be using https://www.australia.gov.au/ as an example.

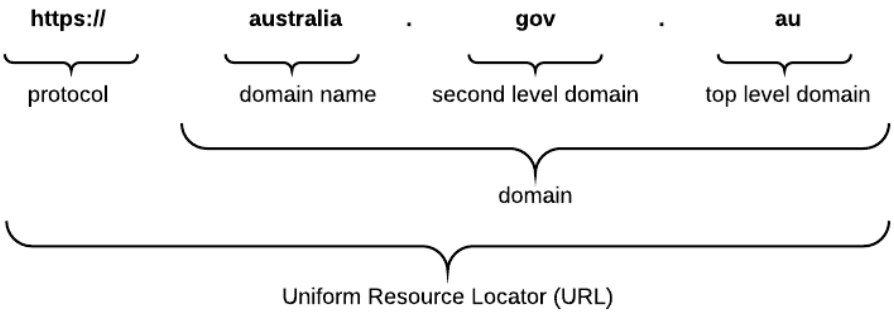

Figure 5. Elements of an URL

- Country Code Top-Level-Domains (ccTLD): a TLD based on 2 letter country codes, as laid out by the ISO-3166 standard. Our example TLD is also a ccTLD.
- Second-level Domain (2LD): A second-level domain is an optional domain directly before the TLD. The 2LD of our example would be GOV. The level of the domain is There is no depth limit to of Nth-level Domains. For instance the Government of Quebec registers its domains using a third-level domain; .GOUV.QC.CA.
- Domain Name: The identification string of a domain, preceded by the 2LD if relevant and the TLD. The Domain name for our example would be AUSTRALIA.
- Registry: The organization that manages TLDs.
- Registrant: The entity who registers a domain name.
- Registrar: The organization, accredited with the registry, that sells domain names to the public. An intermediary between the Registrant and the Registry.

Bibliography

Cherdantseva, Y., & Hilton, J. (2014). Information Security and Information Assurance: Discussion about the Meaning, Scope, and Goals. In Portela, I. M., & Almeida, F. (Ed.), *Organizational, Legal, and Technological Dimensions of Information System Administration*, IGI Global, 167-198. http://doi:10.4018/978-1-4666-4526-4.ch010

GoDaddy Domain. (2020). Name Search. *GoDaddy.* http://www.godaddy.com/domainsearch/find?checkAvail=1&tmskey=&domainToCheck=coronavirus-domain-example.pl

Kumar, P.R., Perianayagam H. R., and Perianayagam J. (2018). A Framework to Detect Compromised Websites Using Link Structure Anomalies. *International Conference on Computational Intelligence in Information System*. Springer, 72-84. https://doi.org/10.1007/978-3-030-03302-6

Internet Assigned Numbers Authority. (2020). Root Zone Database. *IANA*. http://www.iana.org/domains/root/db

Internet Corporation for Assigned Names and Numbers (ICANN). (2020). Accuracy. *ICANN.* whois.icann.org/en/accuracy

Internet Engineering Task Force. (2020). Domain Requirements. *IETF*. tools.ietf.org/html/rfc920

Internet Engineering Task Force. (2020). Domain Name System Structure and Delegation. *IETF.* tools.ietf.org/html/rfc1591

Mozilla. (2020). Mozilla Included CA Certificate List. *Mozilla Wiki.* https://wiki.mozilla.org/CA/Included_Certificates

Mozilla. (2020). Included CA Certificate List. Mozilla CCADB. https://ccadb-public.secure.force.com/mozilla/IncludedCACertificateReport

Rosenberg, H., Syed, S., and Rezaie, S. (2020). The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic. *Canadian Journal of Emergency Medicine*, 22(4), 418-421. https://dx.doi.org/10.1017%2Fcem.2020.361

Sbaffi, L., & Rowley, J. (2017). Trust and credibility in web-based health information: a review and agenda for future research. *Journal of Medical Internet Research,* (19)6: e218. https://doi.org/10.2196/jmir.7579

SeleniumHQ. (2020). SeleniumHQ/Docker-Selenium. *GitHub.* github.com/SeleniumHQ/docker-selenium

PyPi. (2020). Opencv-Python. *PyPi*. pypi.org/project/opencv-python/.

Walther, J. B., Wang, Z. and Loh, T. (2004). The effect of top-level domains and advertisements on health web site credibility. *Journal of Medical Internet Research,* 6(3): e24. https://doi.org/10.2196/jmir.6.3.e24

World Health Organization. (2020). Naming the Coronavirus Disease (COVID-19) and the Virus That Causes It. *WHO.* www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it

World Intellectual Property Organisation. (2020). Arbitration and Mediation CenterccTLD Database. *WIPO.* http://www.wipo.int/amc/en/domains/cctld_db/codes/al.html

Yandex. (2020). Yandex Image Translation API. *Yandex*. https://translate.yandex.com/developers

Funding
No funding was received for this research.

Competing interests
There are no competing interests reported by the authors of this paper.

Ethics
The authors of this paper only used publicly available data for the purposes of this paper. No human subjects were used for the completion of this research.

Copyright
This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are properly credited.

Data availability
All materials needed to replicate this study are available via the Harvard Dataverse:
https://doi.org/10.7910/DVN/3B0JXW