June 4, 2018

To: Goran Marby, CEO, ICANN
Cherine Chalaby, COB, ICANN
Rod Rasmussen, Chairman, ICANN SSAC

From: Dave Jevans, on behalf of the
AntiPhishing Working Group Members and Board of Directors

Dear Sirs,

The AntiPhishing Working Group (APWG) is an international coalition of private industry, government and law-enforcement actors and NGO communities who focus on financial fraud and related cybercrime identification and mitigation.

The APWG membership has followed ICANN community's efforts to define an interim plan to ensure that the existing public display (disclosure) of domain name registration data complies with the impending European Union General Data Protection Regulation (GDPR). The intent to redact information that identifies data subjects as defined by the GDPR for all domains regardless is over-prescriptive; in particular, redaction will hinder legitimate anti-spam, anti-phishing, anti-malware and brand protection activities, particularly efforts to identify related domains that are under unified (e.g., cyber attacker's) control.

We believe that data subject information can be published in Whois and can be protected using a commonly practiced cryptographic solution called secure hashing. Secure hashing will allow investigators and researchers to make use of non-reversible, encrypted data public Whois. we believe that secure hashes, combined with an accreditation plan to facilitate disclosure of protected data to authorized parties, satisfies the GDPR obligation while still providing investigators or researchers with the means to associate domain names with miscreant or criminal actors.

We respectfully request that ICANN organization, community and board consider the attached proposal, entitled Public Whois Attributes, Securely Hashed (WhASH) Hashing Point of Contact Details in Public Domain Name Whois. APWG considers the issue of redaction versus encryption a matter of DNS and domain name security, stability and resiliency. We therefore request that ICANN's board direct the Security Stability Advisory Committee to review and comment on the viability and utility of the proposal as a replacement for redaction.

Thank you in advance for your consideration,

Dave Jevans, Chairman, Anti-Phishing Working Group (apwg.org)

Public Whois Attributes, Securely Hashed (WhASH)

Hashing Point of Contact Details in Public Domain Name Whois

Joe St Sauver and David Piscitello, editors

Domain names allow users to access Internet communications or applications services using names rather than more difficult to remember Internet addresses. Domain names are public resources. They are available to natural persons or legal entities for an annual fee through a "lease" or registration service. During the domain name registration process, parties ("registrants") submit point of contact (POC) information. POC information serves various business or operational purposes. Domain name registrants are responsible for the accuracy of their registration data. They may provide legal entity or personal identifying data as points of contact, and while a registrant may identify an organization, in practice, little effort is made to distinguish legal from natural persons, nor is the registrant's citizenship status requested, even though that may be important when it comes to the applicability of some privacy protections.

Historically, point of contact information for domain names that are registered under generic Top-level domains (gTLDs) has been made publicly available through the Whois service. Certain country-code TLDs (ccTLDs) also publish point of contact information. The European Union's General Data Protection Regulation ("GDPR"), which will go into effect on May 25th, 2018, prohibits the public display or exposure of personal data of GDPR Data Subjects without their explicit and revokable consent. Other governments or jurisdictions may enact electronic privacy regulations as well. Such data protection legislations are intended to mitigate intrusive or abusive collection and misuse of personal identifying data and to penalize parties who do not comply to the regulation.

In anticipation of GDPR, certain operators of Whois services in the domain name community have elected to redact point of contact (POC) information from *all* domain registration data records, including records for legal persons (such as corporations, who are out of scope for the GDPR), and for non-European data subjects (who are also out of scope for the GDPR). This redaction was not done under the direction of a privacy authority but was decided upon by the individual registrars

Figure 1 juxtaposes the registrant's point of contact information for typical domain registration record pre-GDPR against a representative post-GDPR redacted record:

PRE GDPR	Post-GDPR redacted record
Registrant Name: John Michael Smith	Registrant Name: REDACTED FOR PRIVACY
Registrant Organization:	Registrant Organization:
Registrant Street: 123 ABC Lane	Registrant Street: REDACTED FOR PRIVACY
Registrant City: Pleasantville	Registrant City: REDACTED FOR PRIVACY
Registrant State/Province: South Carolina	Registrant State/Province: South Carolina
Registrant Postal Code: 98765	Registrant Postal Code: 98765
Registrant Country: US	Registrant Country: US
Registrant Phone: +18435555555	Registrant Phone: REDACTED FOR PRIVACY
Registrant Email: jmsmith@example.com	Registrant Email: contact-domain@domainregistrar.<tld> (or redacted)

Unintended consequences of redacting domain name registration data

Redacting *all* registration records which were formerly publicly available has unintended and undesirable consequences to the very citizens and residents that electronic privacy legislation intends to protect. Anyone can register a domain name, including cyber attackers. Domain names are critically important tools of cyber attackers' trades. Cyber attackers may register thousands of domain names for a ransomware, phishing, or malware infection attack. Some of these domains will be used to name command-control (C2) systems for

botnets, others for transmitting the hundreds of millions of spam email messages that contain harmful hyperlinks or attachments, and others still for hosting harmful or illicit content.

Cyber attackers have the same obligation to provide POC information when they register domain names, but they have no incentives to provide complete and accurate contact information for identifiers that they register; instead, they use any of several forms of deception, including:

- Providing patently false or incomplete information
- Submitting plausibly accurate data, for example, personal data stolen from obituaries,
- Impersonating identities, possibly identities associated with stolen credit card or other payment methods, e.g., PayPal or other digital payment accounts, or
- Compromising or gaining unauthorized access to domain name registration accounts ("hijacking").

Historically, legitimate domain name registrants have employed these same deceptions to protect their personal data or act anonymously even though they are aware that they could forfeit their registrations if inaccuracies are reported. Since the registrant is responsible for submitting accurate POC information, the accuracy of domain name registration data is not assured.

Certain domain registration data that merit data protection under GDPR are used for legitimate, non-commercial purposes. *The most important of these purposes is to prevent financial or actual harm.*

Today, investigators search for a particular registration data element (name, address, email address…) across large sets (often, millions) of registration data records to attempt to identify all of the domain names that are associated with a criminal attack such as a ransomware campaign.

The redacted data solution eliminates any arbitrary investigator's ability to search for a particular data across a database of registration data records.

GDPR may add a deception method for cyber attackers, i.e., they may provide POC information that identifies them as EU citizens or residents.

Today, for example,

1. An investigator associates a domain name with a ransomware campaign. The registration data of this single domain today provides search arguments – POC data, name server data, creation dates, etc. – that investigators use to search millions of domain name registrations with these data as search arguments; for example, search all registration records available using the value of Registrant Name as the search argument to find other records that have that same Registrant Name.
2. An investigator identifies a web page that hosts ransomware malware. The investigator examines the web site and finds that all of the hosted content is legitimate except for the single malware-infected file and concludes that the site was compromised and the domain registrant is a victim not a criminal actor. Today, the investigator can use registration data to contact the domain registrant victim.
3. A brand owner identifies one infringing domain name and searches Whois for additional infringing domain names registered by the same party. The brand owner may use the POC information to serve cease and desist orders to the party in violation of trademark or copyright law.
4. Law enforcement and private investigators identify the domain name registration behavior of a botnet that is composed of hundreds of thousands of infected computers. POC and other registration data are integral to the eventual identification and apprehension of the criminal conspirators. Victims – parties whose computers were infected – can also be identified and notified so that they can remediate the infection and begin to restore the integrity of their computers.

5. After determining that a domain registrant has malicious intent, a law enforcement investigator may look for other equivalent registrations by this same registrant to proactively disable other similar domain registrations before they can attract victims.
6. An investigator examines registration data and identifies inaccuracies and reports these through a registration data inaccuracy reporting process.

In these examples, if registration data POC details for all domain names are completely redacted, the process of identifying or attributing abuse or victimized domain names to *any* entity, legal or natural, will be hard or impossible. Consider a scenario where the redaction replaces POC information for all domain names with the string "Redacted for Privacy". This redaction hinders legitimate anti-spam, anti-phishing, anti-malware and brand protection activities, particularly efforts to identify related domains that are under unified (e.g., cyber attacker's) control**.** Searches are now pointless: a search on the argument "Redacted for Privacy" will return every domain name registration in the search results.

Urgency of registration data access: Timely intervention

Timely intervention is critical for many cyber-attacks. Users can fall victim to ransomware, malware or phishing attacks for as long as a URL points to a malicious executable, banking fraud (impersonation) web site, or other cyber-attack resource. Redacting data without offering an alternative to today's timely intervention gives criminal actors longer windows of opportunity, and makes it difficult to identify persons of interest in a criminal investigation

Redacted data also hinders research

Academic, government or private sector investigators or researchers can't study registration behaviors if the only data they have is a redaction string that is present in all domain registrations. In many cases, researchers seek aggregate results, e.g., statistics or metrics derived from the compilation of search matches against millions of records. In such cases, individual POC data is only important as a search argument. The searches are highly automated: researchers may never actually examine POC data during their research project and search results not meeting the criteria are typically not retained.

Domain name registration data are critical assets for investigators

Registration data often lead an investigator to further evidence of an online crime or threat (e.g., cyber bullying, scams, terrorism, online coercion or "blackmail"). In a very real sense, domain registration data often help investigators (re)construct a global crime scene. Some registration data identify assets that may be part of a criminal's infrastructure. For example, name server information may identify a single "authoritative" name server that hosts DNS information for hundreds of spam, counterfeit goods, or illegal pharmaceutical websites. Other domain registration data are what investigators call "pivot" data. For example, investigators can use one or more POC data elements as search arguments or to query other databases. In these search scenarios, investigators pattern match data to expand an investigation beyond a single domain, ensuring that takedowns avoid missing alternative domains created for backup and redundancy purposes by the cybercriminal. In many cases, such searching is done using automation and only the POC data that is relevant to investigators will be exposed.

Investigations into these listed harmful or criminal activities often benefit from the ability to associate one or many domain names that are under *unified control*. By unified control, we mean a set of domains, in some cases, possibly tens of thousands of domains, can be associated with a single POC or other registration data.

Employing cryptography to protect registration data from public disclosure

We recommend replacing plain text point of contact details with consistently **hashed values**, rather than redacting those POC details altogether. Consistently hashed values would allow an investigator or research to search registration data sets and to associate multiple domains that use the same POC details, while not disclosing the original POC data of a potential GDRP data subject. The hashing process, simply stated is as follows:

1. Identify the data elements that must be protected
2. The ICANN community chooses a secure hash algorithm, e.g., SHA-512.
3. Each domain name registry or registrar operator creates its own unique private key.
4. For each data element that must be protected, concatenate the data element with the private key.
5. Compute the hash for {data element, private key}.
6. Substitute the hash for the data element in displayed Whois or Port 43 responses.

For example, consider the case where a domain name registrant is named "JOHN MICHAEL SMITH" and the registry operator is dot TLD:

1. Dot TLD's hypothetical private key, known only to the registry operator, is 1DVQSBUN4F89y5gpWR83Gx8j5T6lZwSHuGA5prsqtFZ2mVlvb9pXPgeX6kRkg1qg
2. Dot TLD concatenates the registrant name with the private key.
3. The resulting hashed name value is 8296ec4e3921f7b890322ddce0c79d898fff379c554e077b35c2ee703123f020af44f368ba5dac4e4f9d2c9b818974503f2e4dad4a2aeb7843fa57b21539f569
4. This hash value is made public in displayed Whois or Port 43 responses.

An investigator can now use the hash 8296ec4e3921f7b890322ddce0c79d898fff379c554e077b35c2ee703123f020af44f368ba5dac4e4f9d2c9b818974503f2e4dad4a2aeb7843fa57b21539f569 to search for any registration data records that have that same Registrant Name. Searching is possible, but the name "JOHN MICHAEL SMITH" is not disclosed to arbitrary parties.

Figure 2 juxtaposes the registrant's point of contact information for typical domain registration record pre-GDPR against a one interpretation of a post-GDPR record with securely hashed values:

PRE GDPR	SHA256: Post-GDPR
Registrant Name: John Michael Smith	Registrant Name: **<hash>**
Registrant Organization:	Registrant Organization:
Registrant Street: 123 ABC Lane	Registrant Street: **<hash>**
Registrant City: Pleasantville	Registrant City: **<hash>**
Registrant State/Province: South Carolina	Registrant State/Province: South Carolina
Registrant Postal Code: 98765	Registrant Postal Code: 98765
Registrant Country: US	Registrant Country: US
Registrant Phone: +18435555555	Registrant Phone: **<hash>**
Registrant Email: jmsmith@gmail.com	Registrant Email**: <hash>**

This solution works for the future Registration Data Service (RDS) or Registration Data Access Protocol (RDAP) or existing Whois service. Since we are performing the crypto operation on data values, the solution should not require changes to the existing schema.

Encrypted data is more useful than redacted or null data

It may be possible to accredit certain parties to get un-redacted registrant data, but in the general case, redacted or null data has two shortcomings that a hash solution removes. Null data is not useful data: you cannot usefully search on null data, so accredited parties must ask for data before they can search. In contrast, a hash is a unique value that cannot be reversed to derive the original plaintext but the hash string is useful as an argument in a search *and* it can be used without disclosing data protected by GDPR or future privacy legislation.

Redaction constrains all analysis to parties that are eligible for authorized disclosure of protected data, presumably obtained through an accredication process. Using secure hashes also makes searchable registration data available to any arbitrary researcher. This is especially important for ad hoc or one-time research. Accreditation may not meet the diverse needs of academic or other research that can readily be conducted without exposing personal data. Consider the challenge of accrediting (thousands, perhaps tens of thousands of) university students whose research involves domain names. Accreditation must be completed in a timely and efficient manner so that the student can complete his project by the end of term or semester.

Employing hash strings as search arguments is satisfactory for many types of research where the plain text value is unimportant; for example, one might conduct research into bulk registration behavior, where the relevant registration data elements might be some set or subset of {creation date, registrant name, registrant email, registrant address, sponsoring registrar…}. It is sufficient to search the hash of the registration data that must be protected under GDPR; e.g., registrant name, email, and address. The researcher has no need for the plain text.

Generally, the hash solution lessens accreditation as an impediment to research or investigation and thus accommodates use cases where the researcher does not need to know the value associated with a hash to study registration behaviors. The hash solution can also narrow the scope of registration records that an accredited investigator may request: accredited investigator can use hash searches to pre-determine which records are relevant to their case before requesting access. Imagine an investigation of domains associated with a suspected cyber attacker. Investigators can do this with hashes just as they do this today with actual registrant names but they don't have to expose the name when sharing hashes among a diverse set of collaborators who may or may not be accredited. In some cases, investigators who do not have credentials may be the first responder. With the hash solution, they can play a role by sharing hash "intelligence". The hash solution thus increases the numbers of collaborators but does not expose protected data to parties not accredited.

"Thick" and "thin" registries

Each registry or registrar must create and keep secret its own private key. In the thin Whois scenario, if a registrant registers domain names in more than one registrar, the hashes of their protected data elements generated by registrar X will be different from those generated by registrar Y. If all registries go "thick", these issues become moot: only the registries will generate the hashes.

Compliance considerations

ICANN must ensure correct implementation of secure hashing. If a contracted party (registry or registrar) does not hash properly, then the solution is worthless. A compliance process must be able to identify if registry operator or registrar incorrectly hashes since no other party will be able to tell it was happening from the hash they obtain. ICANN compliance could audit sets of {data element, private key, algorithm} from registrars or registries, generate hashes as the contracted party would, and then compare to the hashes publicly disclosed. If errors in hash computation are discovered, the contracted party would need to redress.

Conclusion

Protecting personal identifying information from unnecessary, unauthorized disclosure or abuse has many benefits. However, the proposed implementation to satisfy data protection obligations - redacting *all* registration records which were formerly publicly available - has unintended and undesirable consequences to the very citizens and residents that electronic privacy legislation intends to protect. This proposal attempts to balance two public interests: the needs to provide data protection for natural persons *and* the needs to protect these same parties as well as legal entities all over the world against harm from misuse or abuse of domain names.

Appendix A. An alternative, stronger cryptographic method

This method provides assurance against disclosure of private keys. The registry or registrar generates a secure hash as described in the document. This hash serves as an intermediate value. ICANN organization or a trusted third party concatenates the intermediate value with its own private key and computes a "final hash".

Again, consider the case where a domain name registrant is named "JOHN MICHAEL SMITH" and the registry operator is dot TLD:

1. Dot TLD's hypothetical private key, known only to the registry operator, is 1DVQSBUN4F89y5gpWR83Gx8j5T6lZwSHuGA5prsqtFZ2mVlvb9pXPgeX6kRkg1qg
2. ICANN or the trusted third party's private key (to be concatenated with the first hashed value supplied by the Registry Operator) is 0fVnLld21qrby2KNBqoPmDOAPR4xkkK71EpF9hF0nfgQzND5OnON0ZyQ60bl5OFQ
3. Dot TLD concatenates the registrant name with its private key: **JOHN MICHAEL SMITH1DVQSBUN4F89y5gpWR83Gx8j5T6lZwSHuGA5prsqtFZ2mVlvb9pXPgeX6kRkg1qg**
4. The resulting intermediate hashed name value 8296ec4e3921f7b890322ddce0c79d898fff379c554e077b35c2ee703123f020af44f368ba5dac4e4f9d2c9b818974503f2e4dad4a2aeb7843fa57b21539f569 is supplied to ICANN or a trusted party by the registry operator for final hash computation
5. ICANN or the trusted party concatenates the intermediate hashed name value with its private key. The resulting final hash value is **1e637b377e87b841825bc7b574a5b73da036417848d245e995ab635864faeb4d0fed5e498680969f9ffa7cdd6e977ad7df6b679128e90a6a94424b42472e8ff1**
6. The final hash value is made public in displayed Whois or Port 43 responses.

This solution provides a resiliency from private key disclosure because two independent parties have hashed the values with different secret keys.

Frequently Asked Questions

Q: How did you compute the hash in your examples?

A. An example of producing such a hashed value on a Linux system with gsha512 sum (see www.gnu.org/s/coreutils/):

$ **echo**
"JOHN MICHAEL SMITH1DVQSBUN4F89y5gpWR83Gx8j5T6lZwSHuGA5prsqtFZ2mVlvb9pXPgeX6kRkg1qg" | gsha512sum | awk '{print $1}'
8296ec4e3921f7b890322ddce0c79d898fff379c554e077b35c2ee703123f020af44f368ba5dac4e4f9d2c9b818974503f2e4dad4a2aeb7843fa57b21539f569

$ **echo**
"8296ec4e3921f7b890322ddce0c79d898fff379c554e077b35c2ee703123f020af44f368ba5dac4e4f9d2c9b818974503f2e4dad4a2aeb7843fa57b21539f5690fVnLld21qrby2KNBqoPmDOAPR4xkkK71EpF9hF0nfgQzND5OnON0ZyQ60bl5OFQ" | gsha512sum | awk '{print $1}'
1e637b377e87b841825bc7b574a5b73da036417848d245e995ab635864faeb4d0fed5e498680969f9ffa7cdd6e977ad7df6b679128e90a6a94424b42472e8ff1

Q. "What registration data elements will you hash?"
A. All fields that are deemed to merit data protection under GDPR would be individually hashed. The exact set is under consideration.

Q. "Why are some parts of the previous illustration colored red, purple, green, brown or blue?"
A. Solely to help the reader see how the various pieces "flow together" (without use of color, long strings of random numbers can be hard to pick out of a composited command).

Q. "Why tack on a long private key rather than just hashing the raw field contents?"
A. Without the concatenation of the long private key, short names ("JIM WATT") might be subject to brute force attacks whereby every conceivable ASCII string of length <N> can be precomputed and then searched. Concatenation of a long private key makes such brute force attacks impractical.

Q. "Because each registry operator would have and use its own long private key, hashes won't be comparable across registry operator!"
A. This is true, and an intentional design choice. However, intra-registry operator, the hashes WOULD be consistent and comparable.

Q. "Why have ICANN 're-hash' already-hashed values?"
A. In the event that either the registry operator or ICANN fails to keep their long private key confidential, the underlying data will still remain protected.

Q. "Could the long private keys ever be rekeyed?"
A. Either the registry operator or ICANN could periodically change their private keys, perhaps bi-yearly. If/when this happens, any previously-cached data held by third parties would need to be "refreshed" (re-looked up) to ensure all domains are using the same "generation" "new" keys. (Old domains that are no longer active would cease to have comparable hashes post-rekeying, and would probably be discarded)

Q. "What about registrants using name variants, such as Joe Smith, Joey Smith, Joseph Smith, etc.?"
A. Obviously this would result in different hashed values, but many who abuse domain names are not willing to create a separate profile for each domain name they create.

Q. "Why did you pick SHA512 for your hash function? I think <insert hash function name> would be better!"

A. SHA512 was selected as an example of a strong hash function. If the community prefers an alternative hash function, that's fine, and will not affect the underlying process that has been proposed. Registries could even be given latitude to select different hashes, subject to minimum strength requirements.

Q. "Why did you pick a 64 character mixed case alpha numeric private key?"
A. Again, the hypothetical long private keys are just examples. A different private key character set or private key length, within reason, will not change the proposed process.

Q. "Do you *really* expect registries to use Unix command line tools to compute these values?"
A. No. The example computation shown in the footnote on the preceding pages is simply a way to demonstrate the sort of values that would result. In production/at scale, a dedicated program would produce the values.

Q. "Because the hashing is consistent, couldn't users get a plain text to hash mapping for individual values by registering a domain with a known value, and then seeing what hash results?"
A. Yes, on a case-by-case basis. That is, Samuel K. Anderson, having registered a domain in his name, could find the hash for his name simply by consulting the hashed value for that name in Whois. This is a known limitation, but not a serious one given the cost of purchasing each such value by registering a domain with unique POC data in each case.

Q. 'What about fields with a limited set of unique values, such as the "State/Province" field? If hashed individually, making just fifty queries, one for each state, would let you know the hashed value of each state within a given registry, wouldn't it?'
A. The proposal deals with the issue of fields with a limited set of unique values by treating address subfields (other than the country name) as a single concatenated field that get hashed as a single value.

Q. "What about bulk access to hashed values, e.g., for commercial Reverse Whois search providers?"
A. Bulk access to hashed values could be offered on a daily basis much in the way Zone File Access is currently provided.