From WHOIS to WHOWAS: A Large-Scale Measurement Study of Domain Registration Privacy Under the GDPR

Chaoyi Lu, Baojun Liu, Yiming Zhang, Zhou Li, Fenglu Zhang, Haixin Duan, Ying Liu, Joann Qiongna Chen, Jinjin Liang, Zaifeng Zhang, Shuang Hao and Min Yang
From WHOIS to WHOHAS: A Large-Scale Measurement Study of Domain Registration Privacy under the GDPR

"In this study, we report the first large-scale measurement study to answer these questions, in hopes of guiding the enforcement of the GDPR and identifying pitfalls during compliance. This study is made possible by analyzing a collection of 1.2 billion WHOIS records spanning two years. [...] Our findings of WHOIS GDPR compliance are multi-fold. To highlight a few, we discover that the GDPR has a profound impact on WHOIS, with over 85% surveyed large WHOIS providers redacting EEA records at scale. Surprisingly, over 60% large WHOIS data providers also redact non-EEA records. A variety of compliance flaws like incomplete redaction are also identified. The impact on security applications is prominent and redesign might be needed. We believe different communities (security, domain and legal) should work together to solve the issues for better WHOIS privacy and utility." READ MORE
When Systems Go Real-Name...

Defeats abusive acts effectively

Cellular networks

Transportation

Online activity
Domain Registration Goes Real-Name, Too

Supported by ICANN and government regulations

Registrant Name

Postal Address & Code

Phone

Email

ID card & Passport verification

ID card & Passport verification

ID card & Passport verification

Domain registration data required by AliYun)
Personal data of domain holders are collected

Names, addresses, phone numbers and emails

Stored by registrars and registries (WHOIS providers)
WHOIS: Real-Name for Domain Registration

Personal data of domain holders are published

Query-based access via WHOIS protocol
Web-based interface / WHOIS server (TCP port 43)

WHOIS query is open and free to everyone

<table>
<thead>
<tr>
<th>Domain Information</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name</strong>: ndss-symposium.org</td>
</tr>
<tr>
<td><strong>Registry Domain ID</strong>: D40220000003323312-LROR</td>
</tr>
<tr>
<td><strong>Nameservers</strong>:</td>
</tr>
<tr>
<td>aron.ns.cloudflare.com</td>
</tr>
<tr>
<td>yahir.ns.cloudflare.com</td>
</tr>
<tr>
<td><strong>Registry Expiration</strong>: 2021-08-15 17:22:32 UTC</td>
</tr>
<tr>
<td><strong>Updated</strong>: 2020-10-06 14:36:34 UTC</td>
</tr>
<tr>
<td><strong>Created</strong>: 2017-08-15 17:22:32 UTC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Contact Information</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Registrant</strong>:</td>
</tr>
<tr>
<td><strong>Organization</strong>: Internet Society</td>
</tr>
<tr>
<td><strong>Mailing Address</strong>: Virginia, United States</td>
</tr>
</tbody>
</table>

(Domain registration data of ndss-symposium.org acquired from lookup.icann.org on Jan 31, 2021)
Security Feeds on WHOIS, Heavily

Spam detection, domain takedown, vulnerability notification...

“Like other companies, Facebook uses Whois data in conjunction with our security technology and systems to help protect people from a range of abuse, spam, and other risks. For example, we have used Whois data and related DNS infrastructure to identify and take down tech support scams operated by spammers who make fraudulent use of domain names, phone numbers, and websites.”

“Microsoft includes Whois data with our security intelligence insights to provide additional context in investigations and threat detections. This context helps us more quickly triage security issues and implement protections for Microsoft and our customers.”

Sounds good, right?

Until...
General Data Protection Regulation

A high-level framework about protecting personal data

Personal data: information of identifying/identifiable natural person

Protects personal data *processing* (storage, disclosure, ...)
General Data Protection Regulation

A high-level framework about protecting personal data

Personal data: information of identifying/identifiable natural person

Protects personal data processing (storage, disclosure, ...)

Expanded territorial scope

Applies to processing of personal data of subjects in the EU

Regardless of where the processing takes place

Profound impact on Internet applications

Website cookies, online ads, privacy notices, ...
When WHOIS Meets GDPR

“WHOIS” becomes “WHOWAS”
Releasing personal data in WHOIS shall be consented

Guidelines published by ICANN on May 17, 2018
“Temporary Specification for gTLD Registration Data*” (TempSpec)
Applies to all gTLD registries and registrars

Collection of registration data
Is maintained.
Personal data is still collected at domain registration.

Access to registration data
Is restricted.
Tiered/layered access under legitimate purposes.

When WHOIS Meets GDPR

WHOIS publishing requirements of ICANN TempSpec

Replacing personal data with *redacted/anonymized* values

Providers decide the scope of data to be protected.

<table>
<thead>
<tr>
<th>Registration Data Fields</th>
<th>Data Subjects</th>
<th>Data Publishing Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name, Street, City, Postal Code, Phone, Fax</td>
<td>Registrant, Admin, Tech</td>
<td>1) Provide a <em>redacted value</em> (&quot;<em>substantially similar</em>&quot; to “redacted for privacy”), or</td>
</tr>
<tr>
<td>Organization, State/Province, Country</td>
<td>Admin, Tech</td>
<td>2) Provide an <em>empty value</em>, or do not provide the fields</td>
</tr>
<tr>
<td>Email Address</td>
<td>Registrant, Admin, Tech</td>
<td>Provide an <em>anonymized email address</em> or <em>web form</em> enabling communication with data subject</td>
</tr>
</tbody>
</table>

When WHOIS Meets GDPR

WHOIS publishing requirements of ICANN TempSpec

Replacing personal data with *redacted/anonymized* values

Providers decide the scope of data to be protected.
## Research Questions

### Data Publishing Changes of WHOIS Providers
- Are providers compliant to the TempSpec?
- How do they redact WHOIS data?
- Are there any compliance flaws?
- What is the scope of protected domains?

### Security Impact of WHOIS Data Loss
- How many security works rely on WHOIS?
- Do they use redacted WHOIS data?
- What are the security systems used for?
- How to remediate the loss of WHOIS?
Part I-A: Data Publishing Changes of WHOIS Providers (Methodology)
Methodology: Overview

Data-driven measurement study

*Latitudinal view:* covering a wide range of WHOIS providers

*Longitudinal view:* covering dates before/after GDPR went effective

A. WHOIS data collection

2-year parsed WHOIS data

B. Compliance Analysis (*GCChecker*)

Identify protected/redacted records and give compliance rankings

---

**WHOIS Data Collection**

- WHOIS server data
- WHOIS templates
- Parsed WHOIS dataset (2018-2019)

**GDPR Compliance Analysis (GCChecker)**

**Data Preprocessing**

- Grouping WHOIS records
- Preprocessing field values

**Compliance Analysis**

- Extracting TF-IDF features
- Clustering WHOIS strings
- NER annotator
- Classifying WHOIS providers
Methodology: WHOIS Data Collection

Challenge: WHOIS ecosystem is fragmented
    Hundreds of providers maintain WHOIS servers
    Format of WHOIS data is inconsistent

Solution: parsed historical WHOIS dataset from industrial partner
    Collects WHOIS of domains observed in its passive DNS
    Parsed by manually-generated templates
Methodology: WHOIS Data Collection

Overview of WHOIS dataset (Jan 2018 ~ Dec 2019)

12% EEA domains; 13% domains older than 10 years
Collected from port 43 of WHOIS servers (not from web WHOIS tools)

<table>
<thead>
<tr>
<th>Year</th>
<th>Count of</th>
<th>Creation Date</th>
<th>Registrant Region</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Record</td>
<td>Domain</td>
<td>Region</td>
</tr>
<tr>
<td>2018</td>
<td>659M</td>
<td>211M</td>
<td>218</td>
</tr>
<tr>
<td>2019</td>
<td>583M</td>
<td>215M</td>
<td>218</td>
</tr>
<tr>
<td>All</td>
<td>1.24B</td>
<td>267M</td>
<td>219</td>
</tr>
</tbody>
</table>
Methodology: Compliance Analysis

Challenge: different wording/language for redaction

TempSpec do not enforce the use of “redacted for privacy”
Methodology: Compliance Analysis

Challenge: different wording/language for redaction
TempSpec do not enforce the use of “redacted for privacy”

Solution: unsupervised clustering of WHOIS record groups
Replace records at scale → High textual similarity → Clusters → Few Outliers

Not compliant, %outlier is high

Compliant, %outlier is low
Methodology: Compliance Analysis

Design of GCChecker

Grouping WHOIS records: \((\text{provider}, \text{registrant\_region}, \text{data\_subject}, \text{week})\)
Methodology: Compliance Analysis

Design of GCChecker

**Grouping WHOIS records:** *(provider, registrant_region, data_subject, week)*

**Preprocessing:** normalize values, extract *TF-IDF features*
Methodology: Compliance Analysis

Design of GCChecker

**Grouping WHOIS records**: (provider, registrant_region, data_subject, week)

**Preprocessing**: normalize values, extract *TF-IDF features*

**Clustering**: DBSCAN finds *outliers*, NER refines clusters
Methodology: Compliance Analysis

Design of GCChecker

**Grouping WHOIS records:** *(provider, registrant_region, data_subject, week)*

**Preprocessing:** normalize values, extract *TF-IDF features*

**Clustering:** DBSCAN finds *outliers*, NER refines clusters

**Provider classification:** rank from on weekly outlier ratios

---

**GDPR Compliance Analysis (GCChecker)**

- **Data Preprocessing**
  - Grouping WHOIS records
  - Preprocessing field values

- **Compliance Analysis**
  - Extracting TF-IDF features
  - Clustering WHOIS strings
  - NER annotator
  - Outlier ratio
  - Classifying WHOIS providers
Part I-B:
Data Publishing Changes of WHOIS Providers
(Results of 143 large providers)
Scale of WHOIS Data Redaction

Over 85% large WHOIS providers are fully-compliant

Large: as of *EEA WHOIS records* collected

**Registrars: 73 / 89** (total domain share > 54%)

**Registries: 51 / 54**

Flawed implementations

- Missing protection of addresses
- Only protecting email addresses
- Others...

![WHOIS compliance of EEA records from registrars (corresponding with their domain share)]
Timeline of WHOIS Data Redaction

Over 80% fully-compliant providers completed in time
100 / 124 completed before May 25, 2018
Timeline of WHOIS Data Redaction

Over 80% fully-compliant providers completed in time
100 / 124 completed before May 25, 2018

Prominent efforts were taken *after* TempSpec (May 17)

Providers lack specific guidelines, thus chose to wait
Only *1 week* left for providers to take actions

Significant drop near GDPR effective date
(May 25, 2018)
Measures of WHOIS Data Redaction

Contact masking measures

TempSpec: Use redacted value / empty value / privacy protection services

<table>
<thead>
<tr>
<th>Category</th>
<th># Provider</th>
<th>Example provider and values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Redacted value</strong></td>
<td>58</td>
<td>ID-69 Tucows Domains Inc. (&quot;Redacted for privacy&quot;)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ID-2 Network Solutions, LLC (&quot;statutory masking enabled&quot;)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ID-625 Name.com, Inc. (&quot;non-public data&quot;)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ID-1505 Gransy, s.r.o. (&quot;not disclosed&quot;)</td>
</tr>
<tr>
<td><strong>Empty value</strong></td>
<td>63</td>
<td>ID-146 GoDaddy.com, LLC; Public Internet Registry (PIR)</td>
</tr>
<tr>
<td><strong>Privacy protection</strong></td>
<td>13</td>
<td>ID-1456 NetArt Registrar Sp. z o.o. (whoisdataprotection.com)</td>
</tr>
</tbody>
</table>
Measures of WHOIS Data Redaction

Email anonymization measures

TempSpec: Use web form / anonymized email that *facilitate communication*

Over 25% fully-compliant registrars do not offer such channel

<table>
<thead>
<tr>
<th>Facilitates Communication</th>
<th># Registrar</th>
<th>Interface</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>42 (72%)</td>
<td>Web form</td>
<td>(<a href="https://www.godaddy.com/whois/results.aspx">https://www.godaddy.com/whois/results.aspx</a>)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Email</td>
<td>(f***************<a href="mailto:7@proxyregistrar.email">7@proxyregistrar.email</a>)</td>
</tr>
<tr>
<td>No</td>
<td>21 (28%)</td>
<td>Web</td>
<td>(<a href="https://tieredaccess.com">https://tieredaccess.com</a>)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Email</td>
<td>(<a href="mailto:abuse@web.com">abuse@web.com</a>)</td>
</tr>
</tbody>
</table>
Scope of WHOIS Data Redaction

TempSpec lets providers decide what data to protect

Apply to EEA domains only / Apply in a global basis
Scope of WHOIS Data Redaction

TempSpec lets providers decide what data to protect

Apply to EEA domains only / Apply in a global basis

Most providers sanitize all WHOIS data → Bad news for researchers

At least 60% fully-compliant providers apply globally

Causing a global, escalated loss of WHOIS

Comparison of outlier ratio of EEA and non-EEA records
Scope of WHOIS Data Redaction

TempSpec lets providers decide what data to protect
   Apply to EEA domains only / Apply in a global basis

Most providers sanitize all WHOIS data → Bad news for researchers
   At least 60% fully-compliant providers apply globally
   Causing a global, escalated loss of WHOIS

Reasons?
   1 week time is short for complete plans
   Hard to determine what data is under scope
   Saves work to comply with future policies (e.g., CCPA)
Part II:

Security Impact of WHOIS Data Loss
WHOIS in Security Literature

Security papers published in 15 years of 5 conferences
Download all via custom crawler

https://www.ndss-symposium.org/ndss-program/2020-program/
WHOIS in Security Literature

69% works that use WHOIS rely on redacted data

31 papers covering a wide range of security topics

### Classified by security topics

<table>
<thead>
<tr>
<th>Domain Security</th>
<th>31.4%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobile &amp; Web</td>
<td>11.4%</td>
</tr>
<tr>
<td>HTTPS</td>
<td>8.6%</td>
</tr>
<tr>
<td>Privacy</td>
<td>8.6%</td>
</tr>
<tr>
<td>Cybercrime</td>
<td>11.4%</td>
</tr>
<tr>
<td>Others</td>
<td>14.3%</td>
</tr>
</tbody>
</table>

### WHOIS Usage

<table>
<thead>
<tr>
<th>WHOIS Usage</th>
<th>Paper examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infer domain ownership / measurement purposes</td>
<td>Halvorson15, Vissers15, Chen16, Liu17</td>
</tr>
<tr>
<td>Features for detection</td>
<td>Sivakorn19, Le Pochat20</td>
</tr>
<tr>
<td>Vulnerability notification</td>
<td>Stock16, Stock18, Roth20</td>
</tr>
<tr>
<td>Result validation</td>
<td>Paxson13, Van Ede20, Delignat-Lavaud14</td>
</tr>
</tbody>
</table>
WHOIS in Security Literature

69% works that use WHOIS rely on redacted data

31 papers covering a wide range of security topics

Registrant contact and email addresses are frequently used

Registrant contact: 29 papers (83%)

Admin/Tech contact: 15 papers (43%)

Email addresses: 26 papers (74%)

Classified by WHOIS fields
WHOIS in Security Literature

69% works that use WHOIS rely on redacted data

31 papers covering a wide range of security topics

Registrant contact and email addresses are frequently used

Other works not affected by WHOIS redaction

Use WHOIS fields that are not personal data

Creation date, Registrar info, Nameserver IP...
WHOIS in Security Literature

Hurdles for security researchers to access WHOIS

Over 70% WHOIS requests from security researchers are rejected*
Current tiered systems lack instructions

Remediation: a better format of tiered access / data redaction

Use RDAP protocol to control access
Use Fuzzy hashing to replace fixed values
Review and adjust current security systems

* https://docs.apwg.org/reports/ICANN_GDPR_WHOIS_Users_Survey_20181018.pdf
Part III:
Discussion & Summary
Discussion

GDPR’s impact on WHOIS is substantial

Most WHOIS providers *actively* and *extensively* redact personal data
A number of security works are affected due to WHOIS loss

Lessons learnt: Enforcing privacy policies is still a complex task

TempSpec leaves flexibility for providers, but not enough time
Checking tools are helpful to identify implementation flaws
The task requires more efficient collaboration across communities
## Recommendations to multiple stakeholders

<table>
<thead>
<tr>
<th>Party</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tech and legal authorities</td>
<td>Allow more lead time for more efficient discussions</td>
</tr>
<tr>
<td>Internet Supervisors (e.g. ICANN)</td>
<td>Develop more specific guidelines to avoid confusion</td>
</tr>
<tr>
<td>WHOIS providers</td>
<td>Review data protection implementations</td>
</tr>
<tr>
<td>Security researchers</td>
<td>Review and adjust security systems that rely on WHOIS</td>
</tr>
</tbody>
</table>
Search Engine for Security Papers

Search published security papers by keywords

Location: https://secpaper.cn/about


Trials and suggestions are welcome!

Credited to:
Fenglu Zhang @ Tsinghua
zfl20@mails.tsinghua.edu.cn
Summary

GDPR’s impact is profound on WHOIS

Large WHOIS providers *actively* and *extensively* redact WHOIS data
Implementation flaws need to be fixed
The *excessive data protection scope* causes global WHOIS loss

A wide range of security works need review or adjustment

Redacted WHOIS data is widely used by security literature

Lessons learnt

Multiple stakeholders need more efficient collaboration
Release compliance checking tool
From WHOIS to WHOWAS: A Large-Scale Measurement Study of Domain Registration Privacy Under the GDPR

Chaoyi Lu, Baojun Liu, Yiming Zhang, Zhou Li, Fenglu Zhang, Haixin Duan, Ying Liu, Joann Qiongna Chen, Jinjin Liang, Zaifeng Zhang, Shuang Hao and Min Yang