# WHOIS Right? An Analysis of WHOIS and RDAP Consistency

**Simon Fernandez**, Olivier Hureau, Andrzej Duda and Maciej Korczyński

ROW13 - June 4th, 2024

# Registration Information

# WHOIS and RDAP - Who?

When studying/blacklisting a domain, we may want to know:

- Who sold it?
- Who bought it? (Did they buy other domains?)
- When? (Did they buy many in bulk?)
- Who to contact in case of abuse? (To take it down)
- ...

# WHOIS and RDAP - Who?

When studying/blacklisting a domain, we may want to know:

- Who sold it?
- Who bought it? (Did they buy other domains?)
- When? (Did they buy many in bulk?)
- Who to contact in case of abuse? (To take it down)
- ...

We need *Registration Information*

# WHOIS

- Old protocol
- Insecure (unsigned & unencrypted)
- Widely spread
- Vague ”Human readable” format

# WHOIS - Example

```
Domain Name: GOOGLE.COM
Registrar WHOIS Server: whois.markmonitor.com
Updated Date: 2019-09-09T15:39:04Z
Creation Date: 1997-09-15T04:00:00Z
Registry Expiry Date: 2028-09-14T04:00:00Z
Registrar: MarkMonitor Inc.
Registrar IANA ID: 292
Registrar Abuse Contact Email: abusecomplaints@markmonitor.com
Name Server: NS1.GOOGLE.COM
Name Server: NS2.GOOGLE.COM
```

# WHOIS - Parsing Challenges

# WHOIS - Parsing Challenges

Language used:

# WHOIS - Parsing Challenges

Language used:

```
NOMBRE DE DOMINIO: epson.com.bo
CONTACTO TECNICO
Razón social: Markmonitor
Nombre Completo: Markmonitor Tech
Correo electrónico: ccops@markmonitor.com
País: Estados Unidos de America
Ciudad: Boise
Dirección: 391 N. Ancestor pl.
Teléfono: 12083895740

Fecha de activación: 2001-08-17
Fecha de corte: 2024-08-17
```

# WHOIS - Parsing Challenges

# WHOIS - Parsing Challenges

Date format:

# WHOIS - Parsing Challenges

Date format:

```
Creation Date: 01-02-03
```

# WHOIS - Parsing Challenges

Date format:

```
Creation Date: 01-02-03
```

- Febuary 3rd, 2001
- Febuary 1st, 2003
- March 2nd, 2002
- ...

# RDAP - Registration Data Access Protocol

In 2015, a new protocol is designed

- Using HTTP(S) for transport
- JSON data format
- Relatively well defined data types
- Not used by all TLDs

# RDAP - Example


```
"ldhName": "GOOGLE.COM",
"links": [{"value": "https://rdap.markmonitor.com/rdap/domain/GOOGLE.COM"}],
["registrar"], "publicIds": [{"type": "IANA Registrar ID","identifier": "292"}],
["abuse"], "vcardArray": ["email",{},"text","abusecomplaints@markmonitor.com"],
{"eventAction": "registration", "eventDate": "1997-09-15T04:00:00Z"},
{"eventAction": "expiration", "eventDate": "2028-09-14T04:00:00Z"},
{"eventAction": "last changed", "eventDate": "2019-09-09T15:39:04Z"},
{"objectClassName": "nameserver","ldhName": "NS1.GOOGLE.COM"},
{"objectClassName": "nameserver","ldhName": "NS2.GOOGLE.COM"},
```

# RDAP - Still not ideal

RDAP parsing difficulties:

- `"ns.ex.com"` or `["ns", "ex", "com"]`?

# RDAP - Still not ideal

RDAP parsing difficulties:

- `"ns.ex.com"` or `["ns", "ex", "com"]`?
- RFC 9083: directly references 17 other RFCs

# RDAP - Still not ideal

RDAP parsing difficulties:

- `"ns.ex.com"` or `["ns", "ex", "com"]`?
- RFC 9083: directly references 17 other RFCs
- ”The entity object class *can* contain the following members”

# RDAP - Still not ideal

RDAP parsing difficulties:

- `"ns.ex.com"` or `["ns", "ex", "com"]`?
- RFC 9083: directly references 17 other RFCs
- ”The entity object class *can* contain the following members”
- Chaotic `vCardArray` objects

# RDAP - Still not ideal

RDAP parsing difficulties:

- `"ns.ex.com"` or `["ns", "ex", "com"]`?
- RFC 9083: directly references 17 other RFCs
- ”The entity object class *can* contain the following members”
- Chaotic `vCardArray` objects
- …

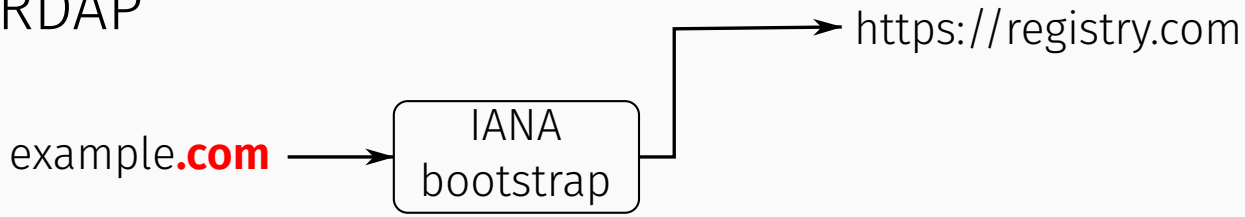
# WHOIS & RDAP - Servers & Records

## RDAP

example**.com**

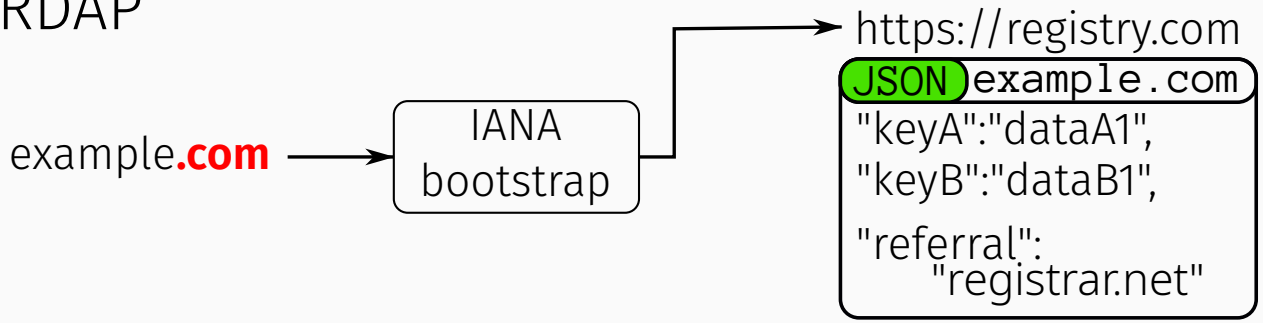
# WHOIS & RDAP - Servers & Records

## RDAP



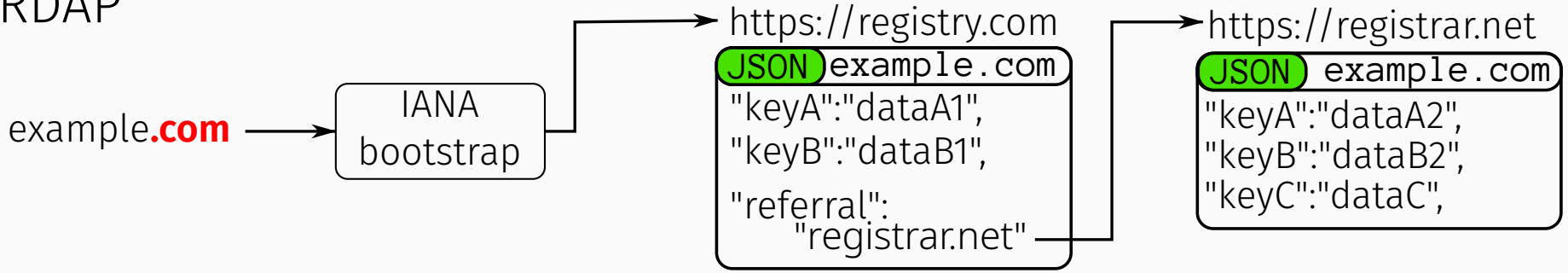
# WHOIS & RDAP - Servers & Records

## RDAP



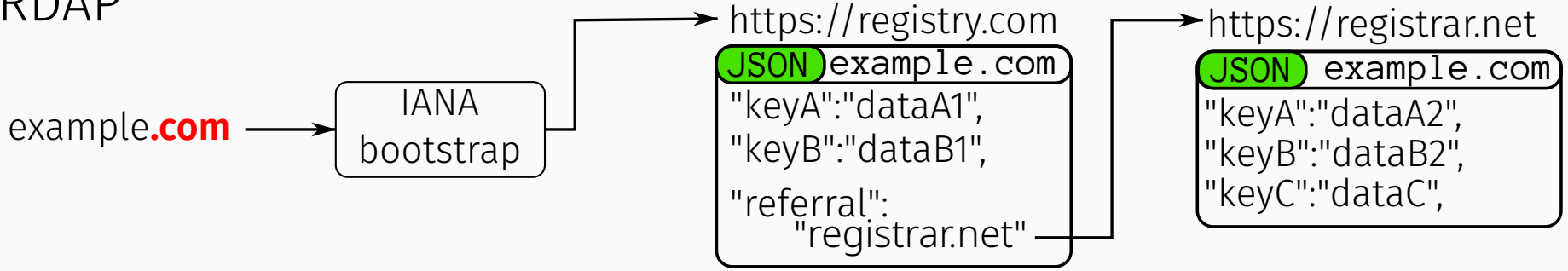
# WHOIS & RDAP - Servers & Records

## RDAP



# WHOIS & RDAP - Servers & Records

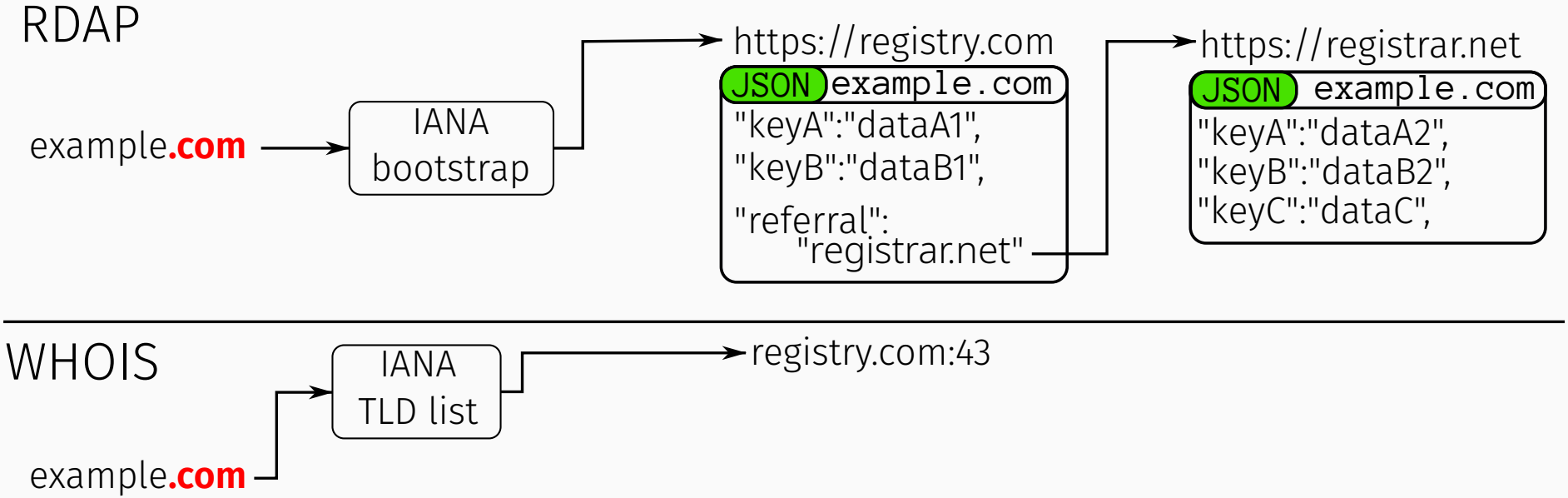
## RDAP




## WHOIS

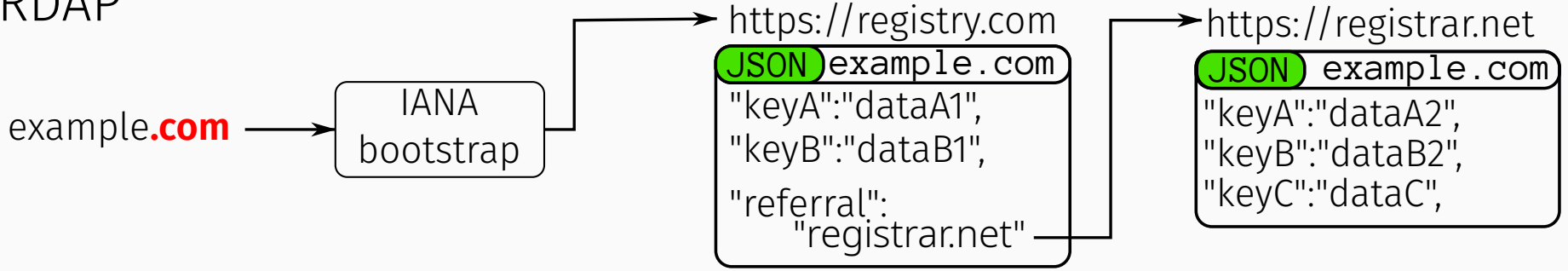
example.com

# WHOIS & RDAP - Servers & Records

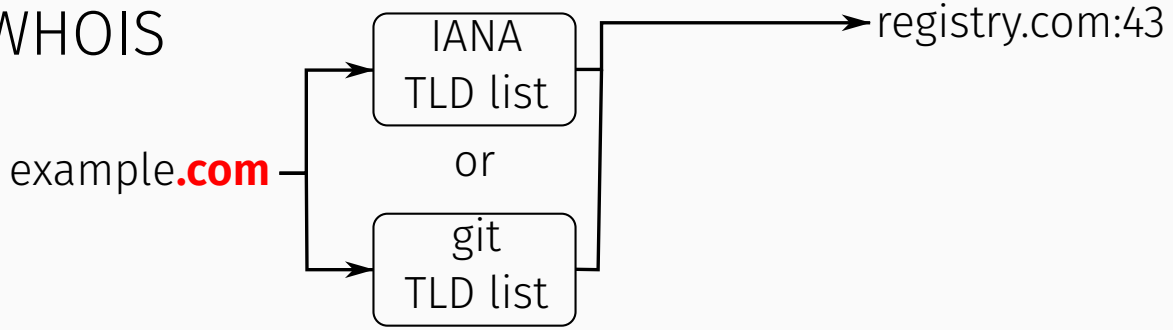


# WHOIS & RDAP - Servers & Records

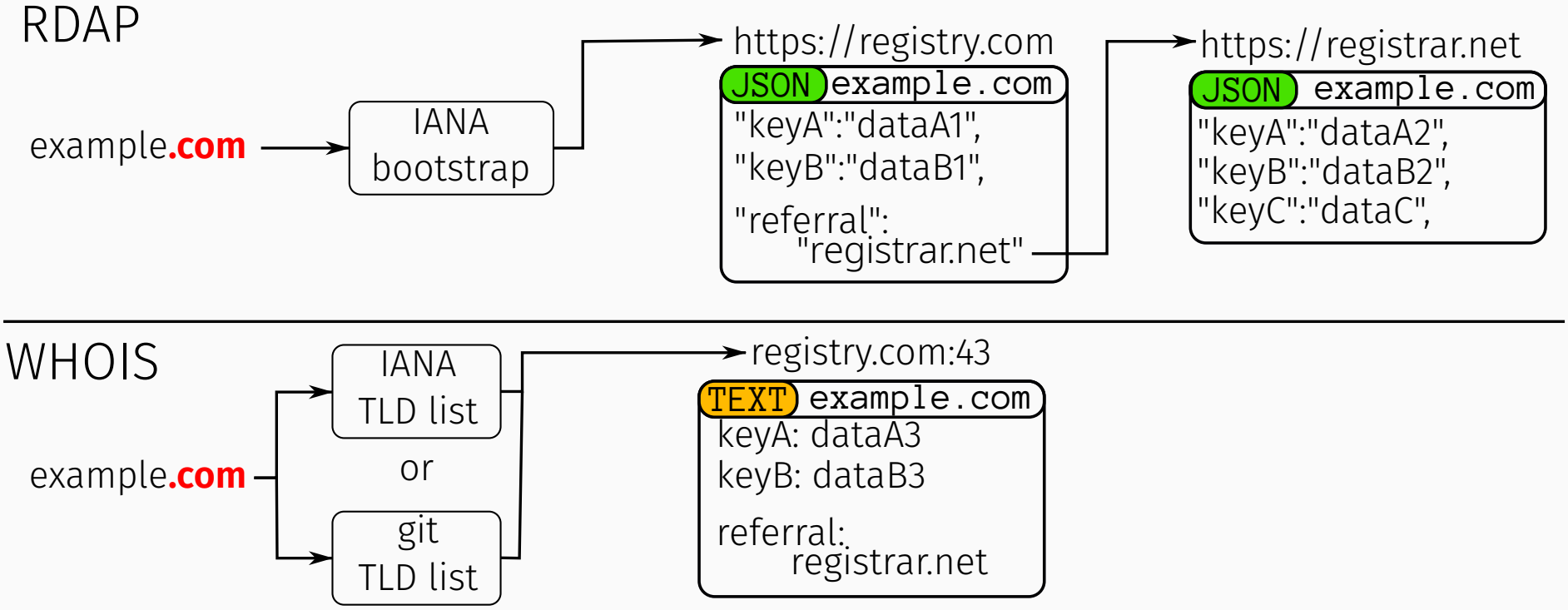
## RDAP



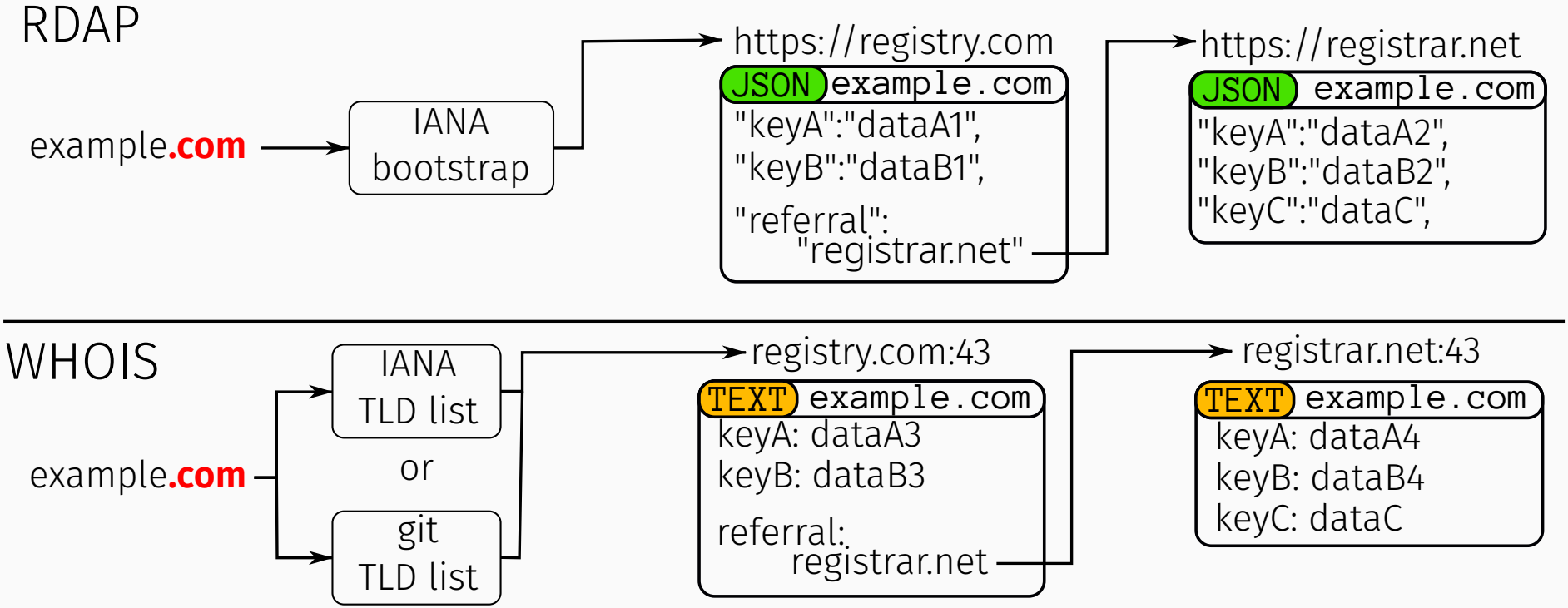

## WHOIS



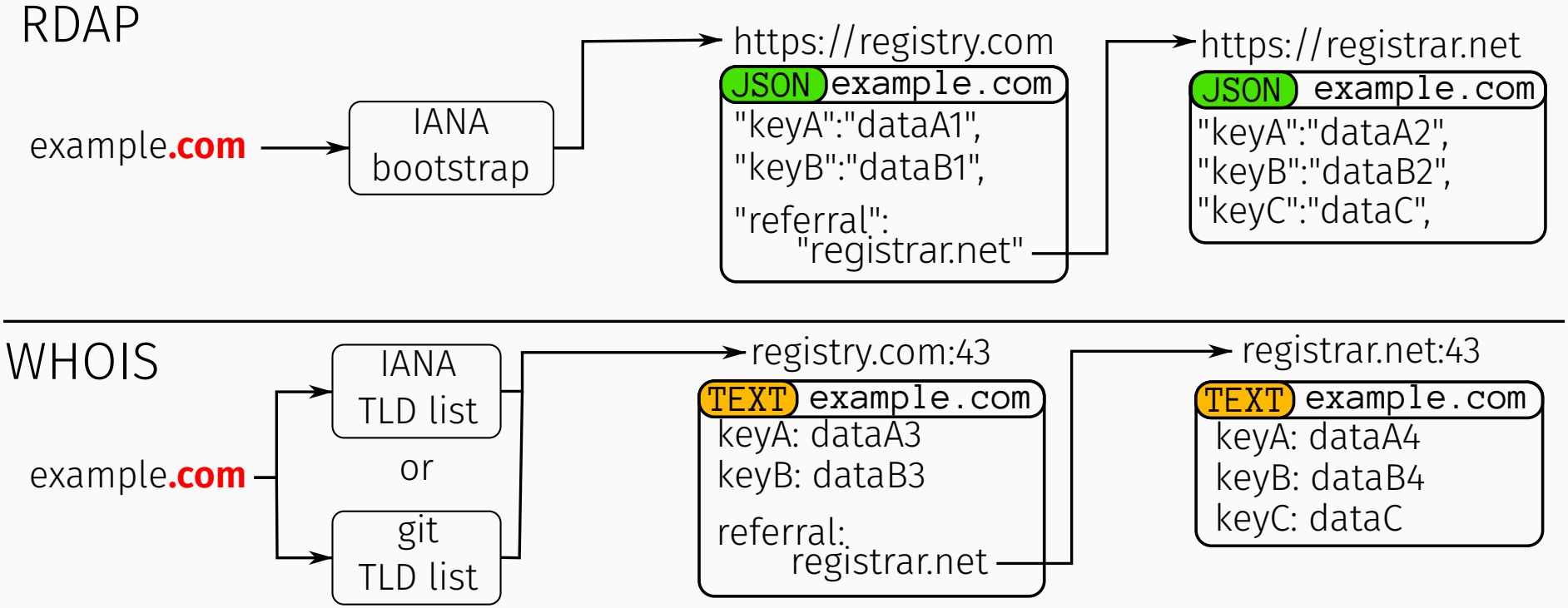
# WHOIS & RDAP - Servers & Records



# WHOIS & RDAP - Servers & Records



# WHOIS & RDAP - Servers & Records




*Research Question:*

Multiple servers and records. Are they coherent?

# Data Collection and Analysis

# Data Collection

- Start from a list of domains (CZDS, Passive DNS, Blacklists,...)

# Data Collection

- Start from a list of domains (CZDS, Passive DNS, Blacklists,...)
- Select 55M domains with both WHOIS & RDAP

# Data Collection

- Start from a list of domains (CZDS, Passive DNS, Blacklists,...)
- Select 55M domains with both WHOIS & RDAP
- Collect all their records $\rightarrow$164M records

# Data Collection

- Start from a list of domains (CZDS, Passive DNS, Blacklists,...)
- Select 55M domains with both WHOIS & RDAP
- Collect all their records $\rightarrow$164M records
- Parse the contents

# Data Collection

- Start from a list of domains (CZDS, Passive DNS, Blacklists,...)
- Select 55M domains with both WHOIS & RDAP
- Collect all their records $\rightarrow$164M records
- Parse the contents
- Check if the values are consistent

# Fields

Fields used by other research works & present in most records

- **Nameservers**: Authoritative servers for the domain
- **Creation & Expiration dates**: When the domain appeared and will expire
- **IANA ID**: Which registrar manages the domain
- **Emails**: Support and abuse mail addresses

# Results

# Inconsistencies

| Field | Data type | Missing rate | Domain inconsistency |
|---|---|---|---|
| `Nameservers` | List(Text) | 6.6% | 573,790 (1%) |
| `IANA ID` | Integer | 13.7% | 106,813 (0.2%) |
| `Creation date` | Date | 2.2% | 3,138,024 (5.7%) |
| `Expiration date` | Date | 2.7% | 2,424,951 (4.4%) |
| `Emails` | List(Email) | 14.8% | 18,958,821 (34.5%) |

# Inconsistencies

| Field | Data type | Missing rate | Domain inconsistency |
|---|---|---|---|
| Nameservers | List(Text) | 6.6% | 573,790 (1%) |
| IANA ID | Integer | 13.7% | 106,813 (0.2%) |
| Creation date | Date | 2.2% | 3,138,024 (5.7%) |
| Expiration date | Date | 2.7% | 2,424,951 (4.4%) |
| Emails | List(Email) | 14.8% | 18,958,821 (34.5%) |

# Nameservers

Multiple nameservers per record. Multiple types of mismatches.

- **Inclusion**: One set is a subset of the other
- **Intersection**: Both sets have a nameserver in common
- **Disjoint**: No common nameserver

# Nameservers

Multiple nameservers per record. Multiple types of mismatches.

- **Inclusion**: One set is a subset of the other
- **Intersection**: Both sets have a nameserver in common
- **Disjoint**: No common nameserver

| Case | Domains |
|---|---|
| All | 576,204 |
| Inclusion | 224,833 (39.1%) |
| Intersection | 23,934 (4.1%) |
| Disjoint | 343,994 (60.0%) |

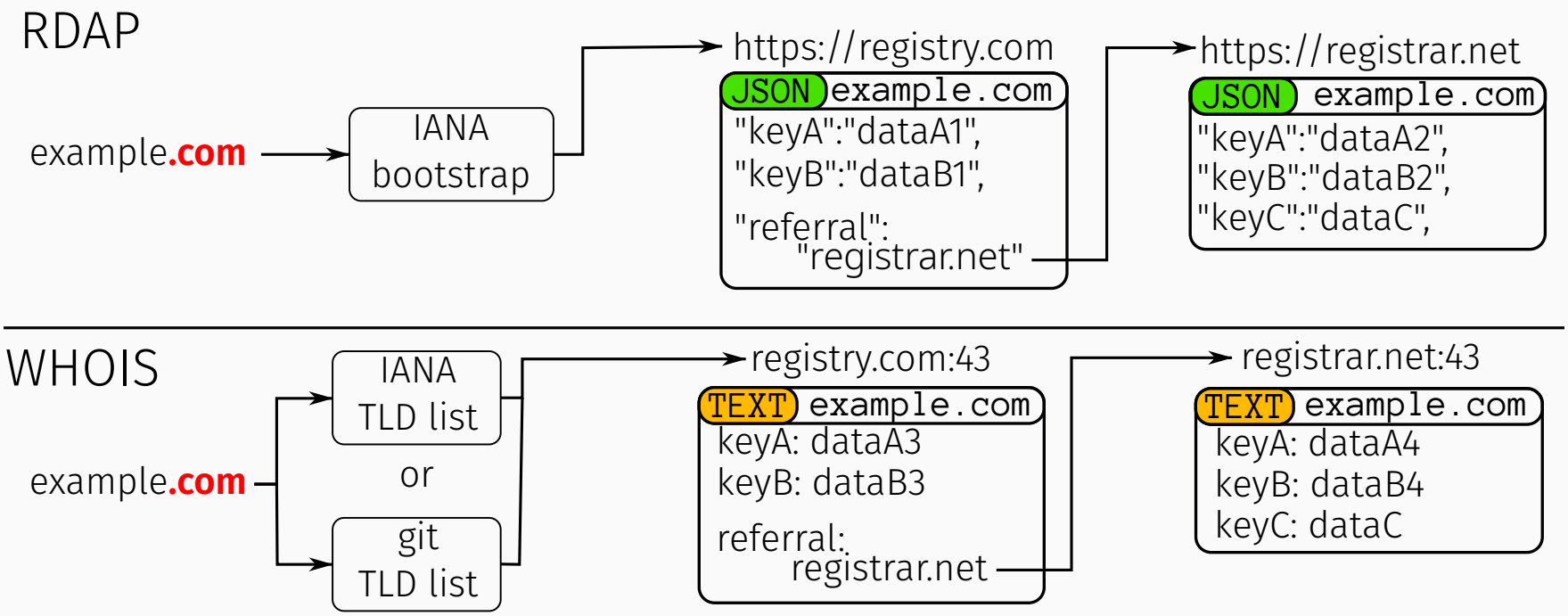
# Nameservers

Multiple nameservers per record. Multiple types of mismatches.

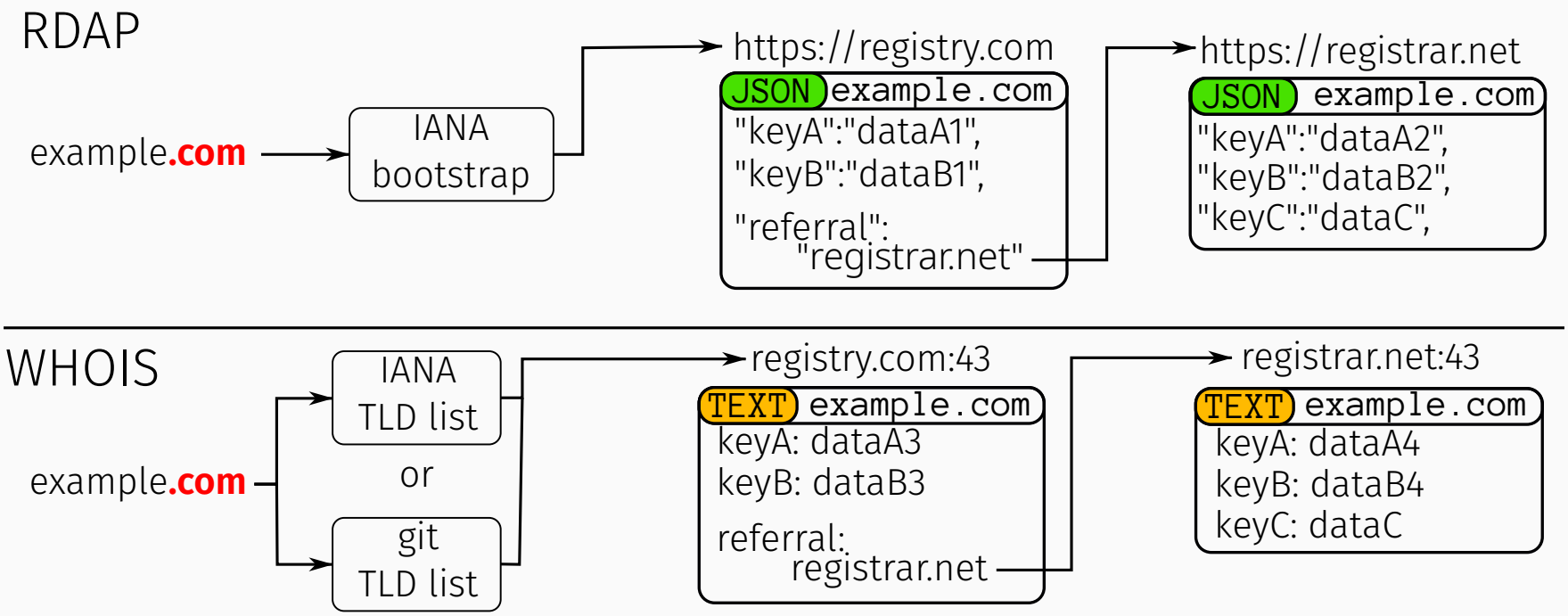
- **Inclusion**: One set is a subset of the other
- **Intersection**: Both sets have a nameserver in common
- **Disjoint**: No common nameserver

| Case | Domains |
|---|---|
| All | 576,204 |
| Inclusion | 224,833 (39.1%) |
| Intersection | 23,934 (4.1%) |
| Disjoint | 343,994 (60.0%) |

# Nameservers



# Nameservers




Inconsistencies can be within the same protocol (25.1%) or between protocols (74.9%).

# Nameservers

To check who is right, we need a ground truth. The DNS.

# Nameservers

To check who is right, we need a ground truth. The DNS.

The DNS has a way to find the authoritative nameservers.

We collected 300k **NS** records (at parent level).

When records are disjoint:

**WHOIS / RDAP**

# Nameservers

To check who is right, we need a ground truth. The DNS.

The DNS has a way to find the authoritative nameservers.

We collected 300k **NS** records (at parent level).

When records are disjoint:

| WHOIS / RDAP |
| --- |
| 21% / 78.5% |

# Other fields

For each field, new challenges and no source of truth:

# Other fields

For each field, new challenges and no source of truth:

- **Creation & Expiration dates**: is 1-day delta OK?
- **IANA ID**: wild ”internal usage”
- **Emails**: GDPR and proxies

# Conclusion

# WHOIS & RDAP - Conclusion

- Registration information: used by researchers & experts

# WHOIS & RDAP - Conclusion

- Registration information: used by researchers & experts
- Different sources of information (protocols, servers...)

# WHOIS & RDAP - Conclusion

- Registration information: used by researchers & experts
- Different sources of information (protocols, servers...)
- Parsing challenges: RDAP in the right direction, not there yet

# WHOIS & RDAP - Conclusion

- Registration information: used by researchers & experts
- Different sources of information (protocols, servers...)
- Parsing challenges: RDAP in the right direction, not there yet
- 164M records from 55M domains: ~5% are inconsistent

# WHOIS & RDAP - Conclusion

- Registration information: used by researchers & experts
- Different sources of information (protocols, servers...)
- Parsing challenges: RDAP in the right direction, not there yet
- 164M records from 55M domains: ~5% are inconsistent
- In most cases: no clear source of truth

# WHOIS & RDAP - Conclusion

- Registration information: used by researchers & experts
- Different sources of information (protocols, servers...)
- Parsing challenges: RDAP in the right direction, not there yet
- 164M records from 55M domains: ~5% are inconsistent
- In most cases: no clear source of truth
- Should be used with care

# Sharing Dataset & Analysis

**Dataset:** Parsed WHOIS and RDAP entries & DNS Records

https://doi.org/10.57745/RJX9XH

**Code:** Inconsistencies detection & Statistical analysis

https://github.com/drakkar-lig/whois-right-dataset

*WHOIS Right? An Analysis of WHOIS and RDAP Consistency*

# Sharing Dataset & Analysis

**Dataset:** Parsed WHOIS and RDAP entries & DNS Records

https://doi.org/10.57745/RJX9XH

**Code:** Inconsistencies detection & Statistical analysis

https://github.com/drakkar-lig/whois-right-dataset

*WHOIS Right? An Analysis of WHOIS and RDAP Consistency*

Thank you for your attention.

# Mail Inconsistencies

25% of mismatches are Disjoint

With GDPR:

- Removed: `REDACTED FOR PRIVACY`

# Mail Inconsistencies

25% of mismatches are Disjoint

With GDPR:

- Removed: `REDACTED FOR PRIVACY`
- Proxied: `3ceacab70b131276@privacy.com`

# Mail Inconsistencies

25% of mismatches are Disjoint

With GDPR:

- Removed: `REDACTED FOR PRIVACY`
- Proxied: `3ceacab70b131276@privacy.com`
- Specific: `whois@domain.com` & `rdap@domain.com`

# Mail Inconsistencies

25% of mismatches are Disjoint

With GDPR:

- Removed: `REDACTED FOR PRIVACY`
- Proxied: `3ceacab70b131276@privacy.com`
- Specific: `whois@domain.com` & `rdap@domain.com`

# Mail Inconsistencies

25% of mismatches are Disjoint

With GDPR:

- Removed: `REDACTED FOR PRIVACY`
- Proxied: `3ceacab70b131276@privacy.com`
- Specific: `whois@domain.com` & `rdap@domain.com`

`local@domain.com`

# Mail Inconsistencies

25% of mismatches are Disjoint

With GDPR:

- Removed: `REDACTED FOR PRIVACY`
- Proxied: `3ceacab70b131276@privacy.com`
- Specific: `whois@domain.com` & `rdap@domain.com`

`local@domain.com`

Disjoint down to ~10%. Resolves mismatches for ~20% of domains.