~~Defining a DNS Statistical Core~~
The DNS Core Census

Edward Lewis

DNS-OARC Workshop 32b
11 August 2020

ICANN

Never fall in love with a project name

⊙ The change in title reflects a change in the internal name for the work

The point of this talk

⊙ Describe a project meant to be a resource for researchers

⊙ Gauge interest level of the audience (which may be difficult remotely)

⊙ Seek a small group (off-line) of volunteers to review

Environment

- ⊙ This work sits at the intersection of the DNS and Domain Name Registries
  - ○ Domain Name Registries – the administrators of zones
    - • gTLD – "generic" or "sponsored" names
    - • ccTLD – "jurisdictional-based" names
    - • reverse map – IP address representations (IPv4 and IPv6)
  - ○ and a more non-celebrity elements

- ⊙ Just about any DNS measurement and analysis on the global public Internet will touch this environment in some way

Motivating use cases

⊙ *Is "the TLD" a class-of-2012 gTLD?*

⊙ *Is "the IDN" a ccTLD?*

⊙ *Are there ROA certificates for name server routes?*

⊙ *Is "this name" under an expected TLD?*

My desire in python code:

```python
import json,requests
zonej = requests.get('<path>/allzones.json').text
zones = json.loads(zonej)['CoreZones']

print (zones['example.']['category'])
>>> "IETFSpecialUse"

print (zones['arpa.']['IANA-registration-date'])
>>> "1985-01-01"
```

The fuzzy idea

⊙ Start at the Root Zone and map the "top-level" zones

⊙ Expand to include the sub-TLDs

⊙ Expand to include the Reverse Map zones

⊙ Include non-DNS domain names as well (to help recognize them when they "slip in")
  ○ Special Use or retired

What information is included?

- ⊙ A useful collection of zones and elements
  - ○ The nameservers used by the zones
  - ○ The addresses used by the nameservers

- ⊙ Meta/ancillary information
  - ○ Dates, status, category, routing attributes

- ⊙ Essentially, a **census** of the "conceptual top" **region** of the global, public operational DNS

What is a census?

- ⊙ Wikipedia:
  - ○ *A **census** is the procedure of systematically acquiring and recording information about the members of a given population. This term is used mostly in connection with national population and housing censuses; other common censuses include traditional culture, business, supplies, agricultural, and traffic censuses.*

- ⊙ How many TLDs sign with DNS Security Algorithm "5" ?
  - ○ If I have a census of DNS elements, I can measure this

Part of the "arpa." record
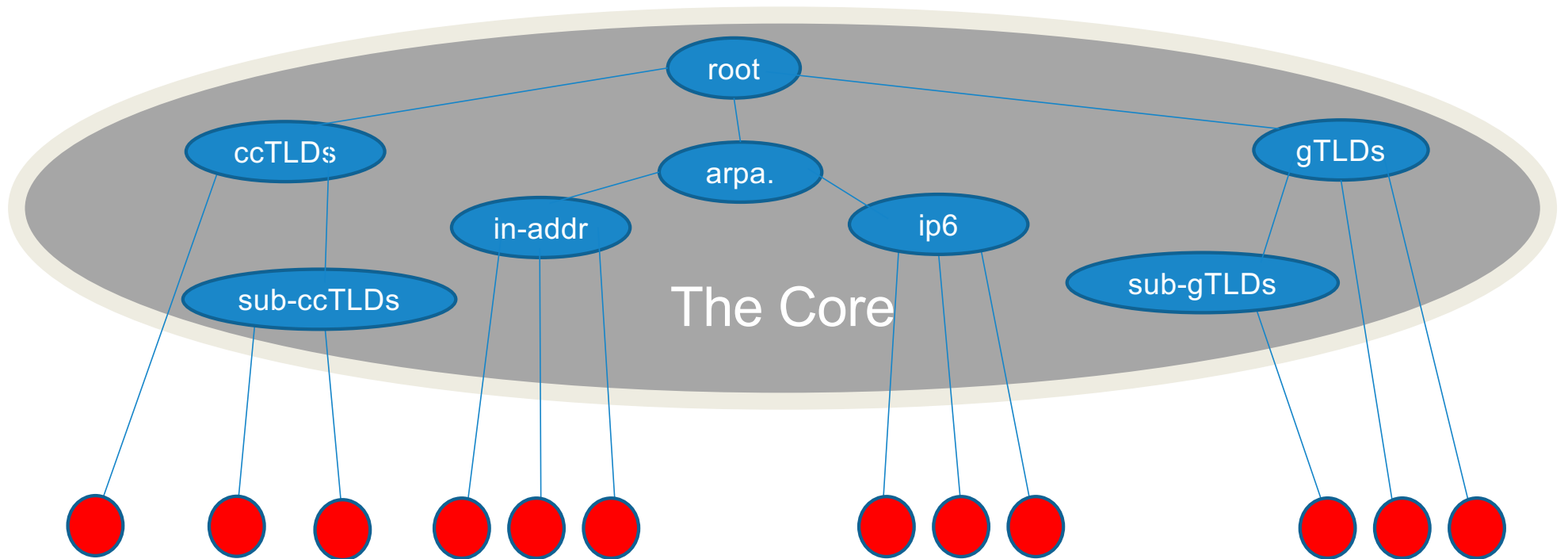
```
"arpa." : {
    "alabel" : "arpa.",
    "category" : "arpa",
    "status" : "ACTIVE",
    "jurisdiction" : "XA",
    "RNAME-field" : "nstld.verisign-grs.com.",
    "SOArrSIG" : [...
    "authnameservers" : [...
    "zonecuttrustanchors" : [...
```

What is the DNS core?

- A "census-designated place", again, Wikipedia:
  - *A **census-designated place** (**CDP**) is a concentration of population defined by the United States Census Bureau for statistical purposes only. CDPs ...are... the counterparts of incorporated places, such as self-governing cities, towns, and villages, for the purposes of gathering and correlating statistical data.*

- The significance is that the designation is solely for the purposes of "counting things"
  - No obligations are incurred by members
  - Does not establish any meaning

Cartoon of the DNS core

Creating the census ("discovery")

- ⊙ For the global public DNS root registry
  - ○ Root zone database, ICANN contractual management reports, etc.
  - ○ IETF's Special Use Domain Name registry

- ⊙ For many top-level zones
  - ○ AXFR – root zone, arpa, some RIR zones
  - ○ "AXFR over FTP" – some RIR zones

- ⊙ Sub-TLD hints - Mozilla Public Suffix List (as a start)

- ⊙ Addressing/Routing - Team Cymru's IP to AS number mapping

- ⊙ And lots of DNS queries

Principles

- ⦿ Representative
  - ○ Appropriately-selected range ("unbiased" is too strong of a word)

- ⦿ Stable
  - ○ Based on authorized sources
  - ○ Members of the core shouldn't "flap" in and out

- ⦿ Lightweight "discovery procedure and maintenance"
  - ○ Limit the amount of DNS queries, where needed
  - ○ Use public, bulk data sources, like zone transfers, as much as possible

Commercial registration boundary

⊙ *This is a conceptual, single slide, de-tour*

⊙ In the "registry-registrar-registrant" model, where are registrant delegations made?
  ○ For the vast majority of the gTLDs: "2LD" or second level
  ○ In some cases, deeper than "the second level"
  ○ In the IPv6 reverse map, it is very "deep" (RIR to LIR/ISP)

⊙ Determining this has proven to be helpful, to a degree
  ○ E.g., given a nameserver's domain name, who is the registrant?

The Achilles Heel of the discovery process

- ⊙ Detecting the sub-TLD elements
  - ○ Mozilla PSL lists "candidates" (which don't always exist anymore)
  - ○ Non-scalable, informal "scrape" from Wikipedia listings for TLDs

- ⊙ Candidates are tested, asking an appropriate-TLD's server
  - ○ Sometimes the answer is "I have that" – a subTLD
  - ○ Sometimes the answer is "My sibling has that" - a subTLD
  - ○ Sometimes the answer is "Someone else has that" – not a subTLD
  - ○ Sometimes the answer is "It doesn't exist" – not anything

- ⊙ And a few times the answer proves to be "incorrect"
  - ○ So there are some fluctuations in membership

Fixing that...

⊙ My "dream" – having TLDs indicate their subTLDs

⊙ My "it would work but is overkill" – having the TLD zone file
  ○ Applies to TLDs for whom I don't already have it

⊙ My "another crazy idea" – maybe use a passive DNS collection

What (else) needs work

- The names of (*struct Python dict**) keys
    - As an example these are dates:
        - `'gTLD-contract-signature-date'`
        - `'IANA-registration-date'`
        - `'gTLD-zone-delegation-date'`
        - `'gTLD-removal-date'`
        - `'gTLD-contract-terminated-date'`
        - `'IANA-entryintoservice'`
    - Is this nomenclature (naming scheme) good/bad?

- Other adjustments to what is stored, etc.
    - Some fields might be less useful, omitted data might be interesting

State of the project

⊙ Working but currently in a semi-secret location
  ○ Because it's not fully mature

⊙ The census is embodied in three files
  ○ `allzones.json` (latest)
    • `allzones-%Y-%m-%d-%H%M%S.json`
  ○ `allnameservers.json`
    • `allnameservers-%Y-%m-%d-%H%M%S.json`
  ○ `alladdresses.json`
    • `alladdresses-%Y-%m-%d-%H%M%S.json`

Use of the census so far

- RPKI ROA adoption by "top" DNS zones (TLDs and reverse map)
  - Uses the census data to measure ROA adoption across categories
  - The census data is used to create categories as well
  - This has "survived" some updates to the census files

- TLD DNSSEC Scorecard
  - Early in the process of increasing dependence on the Census

- To date, (only?) a few use cases which help shape the census files

What would help

- ⊙ Someone willing to write scripts in the name of measurement and analysis that can cope with changes to names of fields in structures
  - ○ Not an easy task; must be patient and flexible

- ⊙ I'm looking for a small group, not many
  - ○ I'm sure there will be changes, perhaps expansions

Wrap up (Questions & Commentary)

- ⊙ My email:
  - ○ edward.lewis@icann.org

- ⊙ ...or catch me in the hallway (well, not my hallway)...

Engage with ICANN

Thank You and Questions

Visit us at **icann.org**
Email: edward.lewisl@icann.org

@icann	linkedin/company/icann
facebook.com/icannorg	slideshare/icannpresentations
youtube.com/icannnews	soundcloud/icann
flickr.com/icann	instagram.com/icannorg