Beta Availability of two TLD Data Products

TLD Apex History and DNS Core Census

Edward Lewis

DNS-OARC 37
17 February 2022

ICANN

Purpose and Agenda

- Two data products are being made available for use, in “beta”
  - TLD Apex History
  - DNS Core Census (v010)
- The goal
  - The data sets are updated daily, not static – i.e., continually “fresh” data
  - The goal is to make it easy for scripts to access, digest and use the data
- What is the TLD Apex History?
  - The data
  - The “beta” questions
- What is the DNS Core Census?
  - The data
  - The “beta” questions

Setting the Scene

- There is a lot of information out there on the Internet
  - Much of it in web pages aimed at people’s eyes
  - But a computer “sees” differently than a human
  - An informational web table might be very useful to a human, but confusing to a script
- There are a lot of DNS information sources
  - Meta-data about the system, as opposed to the data managed/published by the system
  - Meta-data helps drive analysis, subdividing data sets according to criteria
  - Aggregating the data is not as simple as it should be
- Short-term projects are great for presentations and dissertations
  - Long-lived measurements are more useful for large, sustained investments/projects
  - Periodic repeating of a foundational study avoids “solving yesterday’s problems”
- The aspiration is to aggregate data in structured form to support the “long-haul”
  - What will it take?

Driver

- When studying the TLDs and other “top-level” names, having meta-data available is helpful when “slicing and dicing” the data
  - gTLDs versus ccTLDs and IDN gTLDs vs. IDN ccTLDs
  - Signed domains vs. un-signed domains
  - Geo-Regional groupings
  - Inner-workings of the industry (multi-tenant platforms)
- The data sets are not judgements, gradings or rankings, but collections of reference data
- This data is obtained from various ICANN sources, augmented by other sources

Availability

- The website is crude, no visually stunning HTML (yet), first goal is to make scripts happy as opposed to humans

- A summary list of URLs:

- https://observatory.research.icann.org/tld-apex-history/
  - https://observatory.research.icann.org/tld-apex-history/table/
  - https://observatory.research.icann.org/tld-apex-history/doc/
  - https://observatory.research.icann.org/tld-apex-history/code/

- https://observatory.research.icann.org/dns-core-census/
  - https://observatory.research.icann.org/dns-core-census/v010/
    - https://observatory.research.icann.org/dns-core-census/v010/table/
    - https://observatory.research.icann.org/dns-core-census/v010/doc/
    - https://observatory.research.icann.org/dns-core-census/v010/code/

Licensing

- The data in use comes from sources that do not have explicit licenses over their data. A few sources granted access but never formalized the bounds and limitations of use. In general, there is very little formality when it comes to licenses covering data use across the field
- ICANN will post a license for the data, which, in summary, will likely fit these rules
  - The data may be reused for any purpose (public or commercial) with attribution
  - The data may cease to be updated at anytime, ICANN is not liable for any consequences
    - Perhaps there are technical glitches, perhaps a source of data disappears
    - As a beta program, there's not enough resources to make this full production
  - The data may be removed from the website at anytime, ICANN is not liable for consequences
- Formal terms will be published on the web site in due time

Layout of the TLD Apex History

- https://observatory.research.icann.org/tld-apex-history/
  - https://observatory.research.icann.org/tld-apex-history/table/

  - A json and a csv file containing the entire history as recorded
  - The two are isomorphic, perhaps the json is unnecessary as it is ”bigger”

  - https://observatory.research.icann.org/tld-apex-history/doc/

  - A data dictionary in HTML and PDF for human eyes

  - https://observatory.research.icann.org/tld-apex-history/code/

  - A sample python3 script meant to show how to pull and display the fields
  - Not the code for assembling the data

Layout of the DNS Core Census

- https://observatory.research.icann.org/dns-core-census/
  - https://observatory.research.icann.org/dns-core-census/v010/
  - There is v002 data that may be moved here, and there may be a need for a future version
    - https://observatory.research.icann.org/dns-core-census/v010/table/
    - A catalog of the tables available in CSV
    - Currently (may change) a rolling 35-day window into the daily census runs
    - Each daily table is in a Gzip'd CSV (.csv.gz)

    - https://observatory.research.icann.org/dns-core-census/v010/doc/
    - Data dictionary in HTML and PDF for human eyes

    - https://observatory.research.icann.org/dns-core-census/v010/code/
    - A sample client in python, pulling the tables, decompressing, and displaying fields
    - Not the census-taking code

TLD Apex History – Origins

- Responding to an operator question, “what parameter settings should be used?” including when to rotate keys, which cryptography, what length of NSEC3 salt?
  - Many recommendations were in early, un-tested (by operators) Request for Comments
  - Default settings in various tools were initially un-tested
  - One thought: survey the pioneers and see how they are faring
  - This began surveys of what is at a TLD Apex

- What is needed to reverse-engineer how an operator “does” DNSSEC?
  - The SOA resource record set and its signatures
  - The NS resource record set and its signatures (helps)
  - The DNSKEYCDNSKEY resource record set and its signatures
  - The DS/CDS resource record set and its signatures
  - The NSEC or NSEC3PARAM record set and its signatures

TLD Apex History – Contents

- The history is a single table, updated daily
  - A number of observation “runs” occur daily, if a record is seen once, it is seen that day
  - The time granularity is daily
    - The purpose is to determine design choices, not measure (availability) performance
  - Uniqueness record is defined by a subset of its fields
    - E.g., a unique RRSIG resource record id defined by the owner, the type-covered and the key used to generate it, but not by the inception and expiration times nor the signature itself
  - For each unique record, the first and last date seen (consecutively) is recorded
  - For this reason, the table growth over time is manageable/fairly small
- Example entry:
  - <owner> <type> <first_seen> <last_seen> <...type-specific-fields...>
- The data does not cover validation, i.e., whether there were “broken” signatures

TLD Apex History – Why Publish this Data

- This data has been accumulating for some time
  - Having a long-term view of how DNSSEC has been configured yields insights on how operations takes on a protocol modification
- Benefits
  - For operators – see their own track record and track records of others in similar situations
  - For protocol developers – see what features of DNSSEC have deployed and how

TLD Apex History – My favorite “toy”

- I’ve used the history to visualize key lifecycles
  - The reason this code is not in the examples is that it involves a plotting package choice

- Seeing how keys are used reveals a lot about operations
  - From the detail of how keys are changed
  - To overall trends due to processes begin run in an automated fashion
  - Identifying unusual events, events that operators handle but developers never considered

- Surprises to the protocol developers
  - Keys used, removed and then used again (it happens)
  - Once two keys, same DNSSEC Security Algorithm, different bit lengths, same key tag seen simultaneously
    - One wonders how that happened with the state of tooling

TLD Apex History – Beta Questions

- Is the JSON format necessary? Or is it “just a thing cool kids do?”
  - Why not: it is 8 times as large and is equivalent in content
  - For such a basic table, JSON only adds “readability” which is not useful for a computer
- Are the resource records and fields sufficient?
  - What other data ought to be included?
  - For example, the ZONEMD record?
- Ought the coverage expand from TLD apex to include, say, reverse map zones?
  - Would such expansion be useful scope creep or just bulky scope creep?

DNS Census Core – Origins

- Is xn—example a ccTLD IDN or a gTLD IDN?
  - With IDNs it is especially hard to tell if a TLD is a ccTLD or a gTLD
  - “Hard” meaning, there’s no simple lookup that can be executed by a script
- “For South Asia, how many TLDs are signed?” – including the many IDN ccTLDs for India
- To answer these questions, a lot of exploration is needed
  - IANA’s web pages and whois service help
  - ICANN data on gTLD contracts, terminations, IDN fast-track
  - Region information defined by ICANN and by the UN (M49)
- And other questions: ROA coverage, services aggregation, commercial registration boundary
  - RIPE’s retired RPKI validator, replaced by Routinator, and Team Cymru’s IP to ASN
  - Zone files (a’la CZDS plus RIR published zones)
  - Public Suffix List (yes, with “due caution”)

DNS Census Core – Contents

- The DNS Core is a set of zones at the top of the globally public DNS
  - TLDs
  - ARPA's delegations
  - Delegations below in-addr.arpa and ip6.arpa managed by RIRs
  - Sub-TLDs, zones delegated but still managed by a TLD
  - Ends at the mythical "Commercial Registration Boundary"
- The census is an aggregation of information from many sources
  - Assigns a category, jurisdiction to each zone
  - Mostly repeats the original data, including what *is considered to be* useful
  - Correlates information about zones from different sources
  - Examines NS resource record sets, associated address records and pulls Autonomous System information to map a TLD to the network
- One (daily) run consists of 9 different tables working together

DNS Census Core – Publication

- The DNS Core Census consists of 9 tables
  - For each run 9 compressed csv tables are published
  - Files for a run include the start time of the run in the file name
  - The intention is to have one run per day, but this allows for testing (in alpha)
  - https://observatory.research.icann.org/dns-core-census/v010
    - Note the v010 (to distinguish from an earlier version, and maybe a future version)
    - code – plan to upload some code as example use
    - doc - data dictionaries
    - table – a rolling 35-day window of census runs
- Only compressed CSV is available due to size constraints
  - Size of the census is a concern
  - Only 35-days of data is published on the web, there is data going back almost a year for v010 (and a little further back for v002)

DNS Census Core – Version History

- The first public data version is V002
  - I used to mention the URL, but don't anymore
  - Data published (somewhere) from 2020 July until 2022 January
  - There are gaps, and data after September 2021 may be incomplete
  - There's a thought to translate this data into a subset of V010 data
- There were two non-public improvements, V003 and V004 as the idea expanded
  - These versions expanded the sources included in the census
- The second public data version is V010
  - There's unpublished history since May 2021 with very few gaps
  - Only the last 35 days (rolling) is public, due to size (disk space) concerns

DNS Census Core – Why Publish this Data

- This data is from many sources and sources that are in transition
  - HTML tables, one type of source, are optimized for eye-balls, not scripts
  - Where there are internal databases, the mapping to the web is not straightforward
  - When a data source is discontinued, a new means must be found
  - Some data is available in limited ways, this data is only included in a summarized fashion

- In short, a lot of work is needed to aggregate this data, especially as a routine
  - The investment to do this may be beyond the reach of some research efforts

- The experimental arm of this work (scope-creep) are
  - Determination of the commercial registration boundary
  - The inclusion of whether a valid ROA covers routes to name servers

DNS Census Core – Beta Questions

- The DNS Census Core is immature
  - Only a few public appearances (V002), there has been some feedback built in
  - But the track record is short
- Contents
  - Of what is there, what is useful/needless?
  - Of what is missing, what else ought to be included?
  - Is the non-data properly noted (‘_not applicable_’ kind of values)
- Publication
  - Is there a good/better way to store this long term? History is valuable.
  - A lot of data does not change day-to-day, but each run is separate, hints at a better way to lessen copies of the information
  - Would CSV alone be sufficient?
    - JSON amplifies the needed disk space if only to make it more visually appealing

Benefits/Costs to this Audience

- Intended benefits
  - A consolidated reference for information related to the top (core) of the global public DNS
  - Historical recording of the state of the core
  - Data from an authoritative source or from direct observations

- Costs
  - During a beta period provide constructive criticism
  - Be prepared for product errors, missed data runs
  - Be prepared for changes to format, data coverage

- Feedback
  - If you have feedback email me, my email is on the next slide

Engage with ICANN

Thank You and Questions

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

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