An Update on the DNS Core Census v0.1.0

Edward Lewis

ICANN 72 ccNSO TechDay
25 October 2021

Agenda

- DNS Core Census
  - Motivation
  - DNS Core
  - Census
- Sources of Data
- Assembly of Data
- Availability

The Work

- The DNS Core Census is a small project that has evolved over time
  - Began as a module to help analysis of other data collections
- An early version, numbered 0.0.2, was presented at DNS-OARC in August 2020 and then LacTLD in September 2020
  - Version 0.0.2 exists on the web and is publicly accessible
  - But not well publicized (and will move soon)
- Comments led to some internal versions and ultimately in a complete rewrite labelled 0.1.0
  - Instead of a python dictionary, use pandas.DataFrames
  - Add regional labels from UN (M49), IDN table data
  - Had to make the process more rugged against data feed issues and changes
- Making this version public is an on-going effort

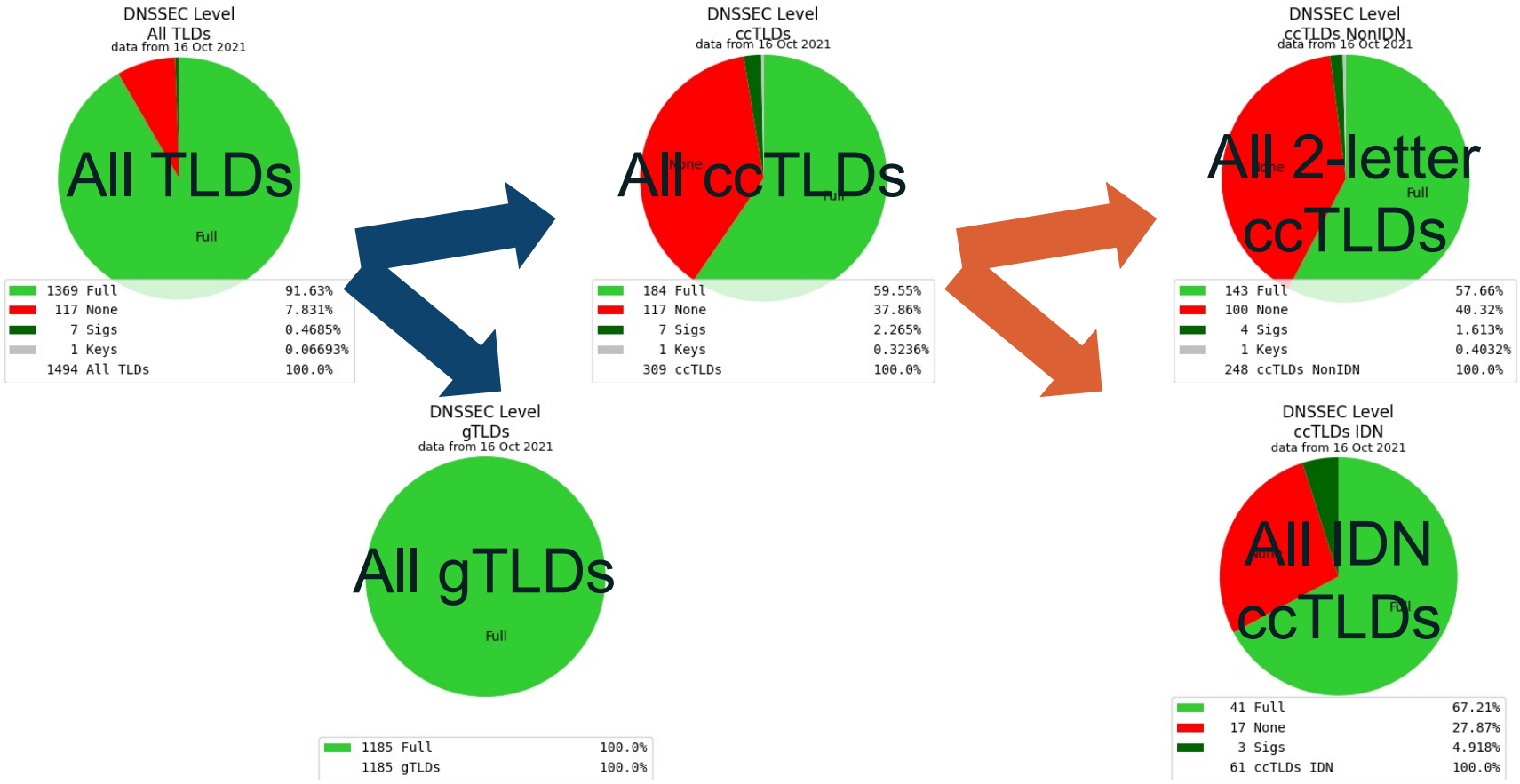
Origins of the Census

- “Is a TLD a ccTLD or a gTLD?”

- Why does it matter?
  - gTLDs and ccTLDs run under different rules, they behave differently

- It’s not just the gTLD vs. ccTLD division that is interesting
  - Regional
  - IDN or not
  - Many others

Divisions of TLDs (Level of DNSSEC Support)



Is an IDN TLD a gTLD or ccTLD?

- In the old days, it was simple to determine from a TLD name whether it was a ccTLD
  - ISO3166-2, alpha2 codes
  - But IDN ccTLDs changed that *xn--...example...*

- The subject question started the effort to build a census

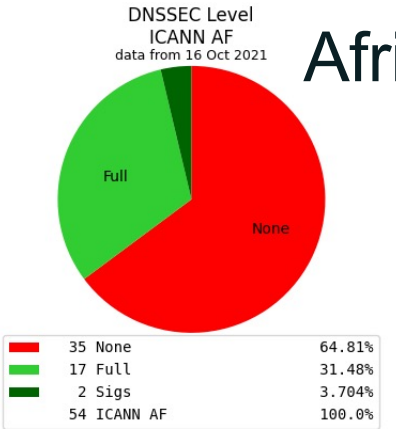
Scope Creep

- When analyzing observations about a TLD, looking for patterns in behaviors begs to know more meta-data
  - When did the TLD begin (and/or end) operation?
  - Which TLDs share the same DNS platform?
  - Which TLDs serve a particular geographical/geopolitical region?
  - What TLDs are large/medium/small and use NSEC3/elliptic curve keys/...?
  - Where are TLD name servers (addresses, routes, and autonomous system numbers)?
  - ...and more...

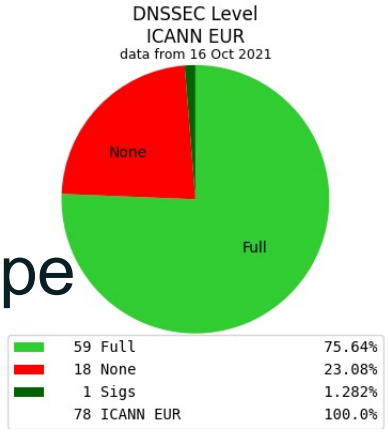
- For any given TLD, this information is available in many scattered sources, would be nice to simply collect it into one place
  - And maybe use it to create a history as well (i.e., do it daily and publish)

Geographic Divisions of (cc)TLDs (Level of DNSSEC Support)

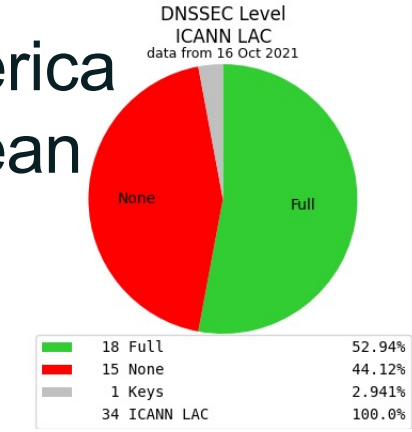
Africa



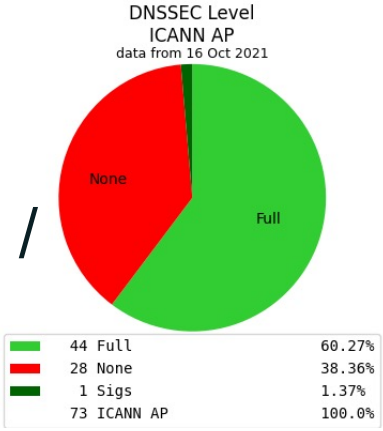
Europe



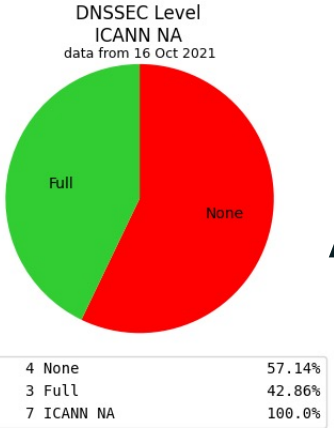
Latin America & Caribbean



Asia / Australia / Pacific



North America



Coverage Creep

- How much of the DNS ought to be included? (In other words: What is the DNS Core?)
  - Everything would be desirable, but “everything” is unmanageable
  - The root zone and its delegations is too small
  - Want something that is “the right size”, has a stable membership (definition) and is a “sensible” region of the DNS
- What is sensible?
  - There are many interesting regions of the DNS that will (probably) exhibit similar behavior
- Settled on:
  - ccTLDs, gTLDs
  - RIR run reverse map zones
  - And zones tying all of these together

Other Possible “Cores”

- Hi-volume/”popular” zones which get a lot of traffic
- Genres like “Social Media”, public sector management (governance), health, etc.
- Technically complex zones (using specialized DNS features like client subnet, etc.)
- There are many perspectives determining what is interesting

- All of these are worthy of measurement
- But membership is subjective, the lines are not clearly drawn (i.e., what’s technically complex?)

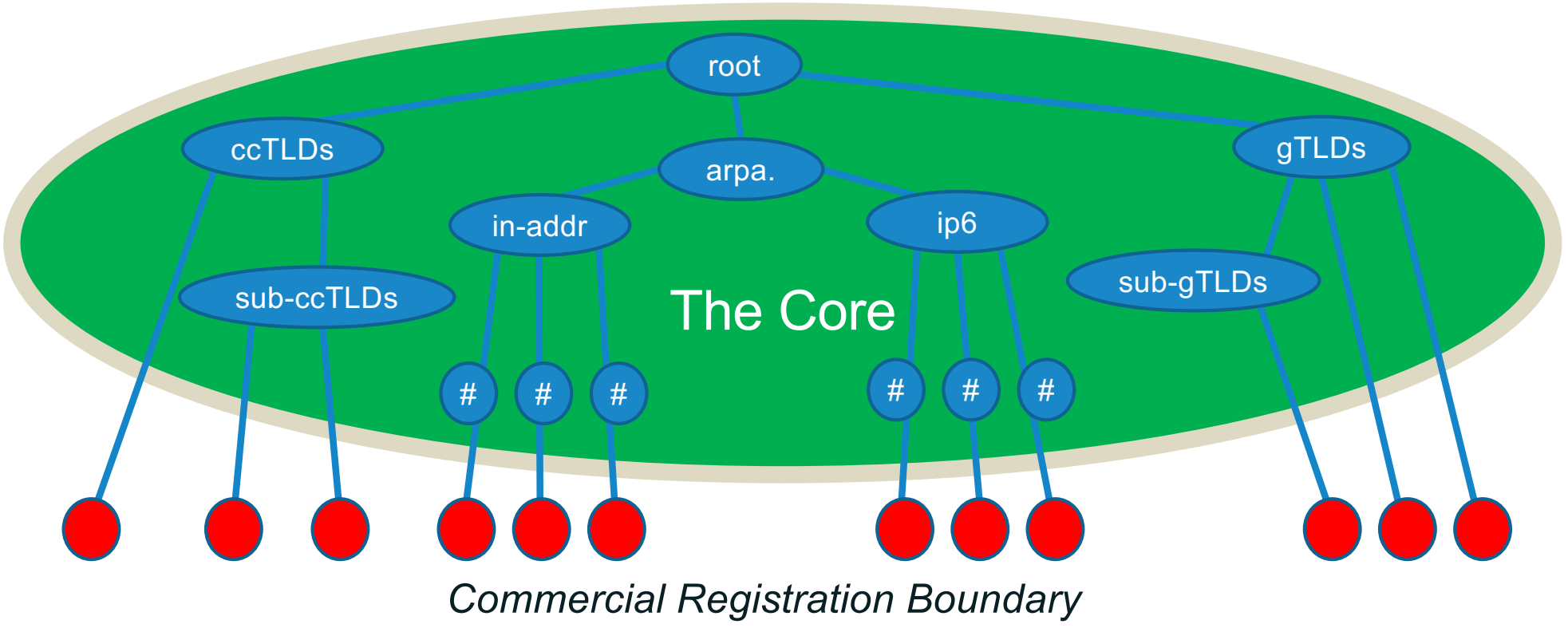
Defining a DNS Core

- The DNS Core (as defined here)
  - Vaguely: the elements of the DNS “close” to the root zone which primarily exist to delegate other zones
    - Top-level domains, including affiliated zones (sub-zones of a TLD)
    - Regional Internet Registries (IPv4 and IPv6 reverse map TLDs)
    - Other support zones or special names (”arpa.”, test and reserved names)
- These zones are generally run under guidelines set by a community
  - Admittedly, this may be a stretch to see
  - The operators of this portion of the DNS see the DNS itself as a primary service
  - Operations of these zones stick close to applicable standards of operations
  - These operations put a premium on stability, resiliency, as well as security

The DNS Core

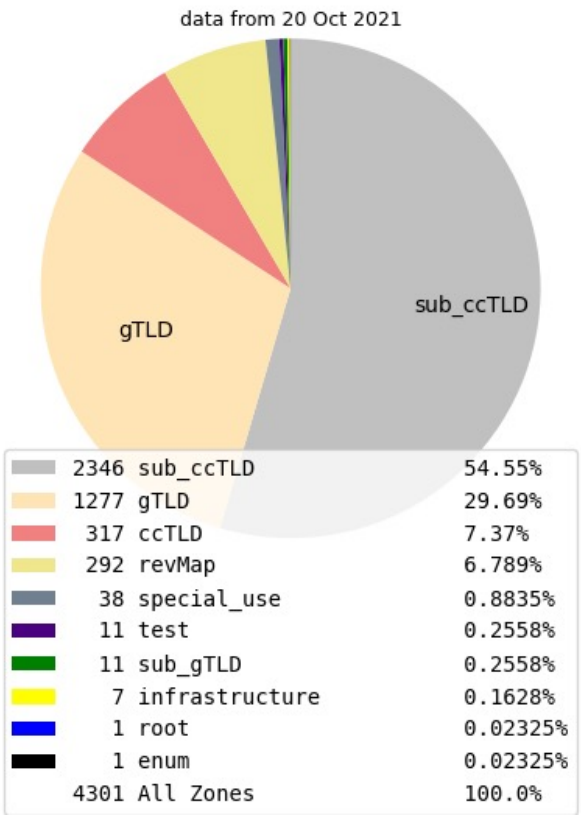
- Starts with the very top of the name space (“.”)
- The border is called ”The Commercial Registration Boundary”
  - Where “registrants” “pay” “for delegations”
  - Examples:
    - customer.gtld-example.
    - customer.category.gtld-example.
    - customer.city.province.cctld-example.
    - 358.455.258.in-addr.arpa. (Note: *invalid-on-purpose* example)
- The concept of the commercial registration boundary is still experimental
  - In some cases, the boundary is at the third label
  - In a few cases, the boundary is many labels deep (usually localities in a ccTLD)

Cartoon of the DNS core



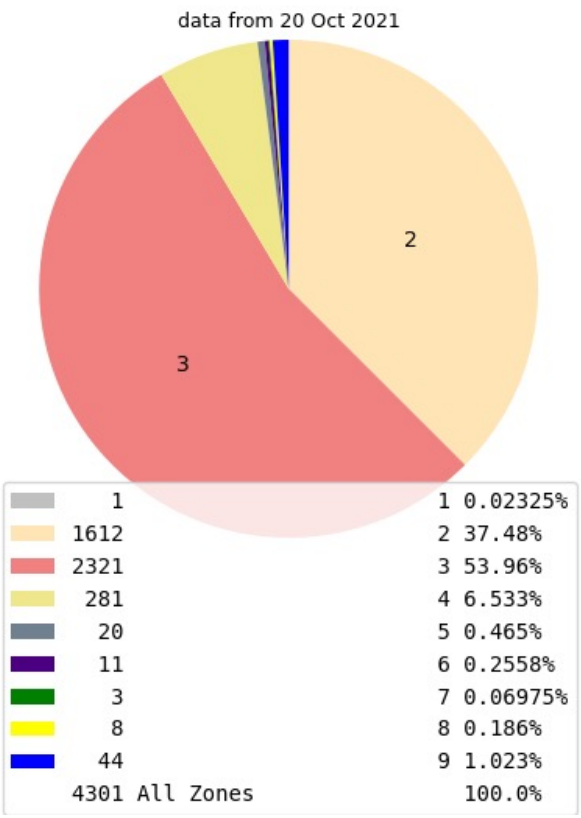
What does the DNS Core look like?

Zone Category
All Zones



2346	sub_ccTLD	54.55%
1277	gTLD	29.69%
317	ccTLD	7.37%
292	revMap	6.789%
38	special_use	0.8835%
11	test	0.2558%
11	sub_gTLD	0.2558%
7	infrastructure	0.1628%
1	root	0.02325%
1	enum	0.02325%
4301	All Zones	100.0%

Zone Depth
All Zones



1	1	0.02325%
1612	2	37.48%
2321	3	53.96%
281	4	6.533%
20	5	0.465%
11	6	0.2558%
3	7	0.06975%
8	8	0.186%
44	9	1.023%
4301	All Zones	100.0%

- Other scattered numbers
  - 4,301 Zones
  - 6,458 Name servers
  - 9,928 Addresses
  - 2,265 Route Origins
  - 524 AS Numbers
  - 8,038 DNSKEY records
  - 7,077 DS records
  - 20,639 RRSIG records
  - 14,144 IDN Tables

The Census

- Information regarding the core is
  - Complied daily
  - Stored as a set of 9 CSV files/9 JSON files, and as rows in 9 monthly database tables
  - The JSON files are “translations” from CSV, i.e., “simple” JSON

- Why not lead with JSON?
  - The “richness” of the interrelationships conflicted with the desire to make this data available in SQL-like database tables
  - ”Reverting” from JSON to CSV seems odd but necessary

Changing My “Goals” for the Census

- Originally, I wanted to build a “richly typed” data structure
  - Working in python – a python dictionary
  - Object-oriented in the sense that data about a zone would be organized together, with nameservers and addresses “hanging off of it”
  - V0.0.2 and some followers did this
  - Used a richly defined JSON structure (3, zones, name servers, addresses)
- But I came across a comment to make more use of pandas.DataFrames, tabular data
  - More compatible with open data designs, data analytics
  - This pivoted the work towards tables
  - Compatible with a SQL(like) database backend
  - First represented CSV (“backwards” from JSON) and later simulcast into JSON
  - 9 Tables are used
- I’m still thinking about this change

Sources of Data

- From IANA
  - Root zone database in XML: https://www.iana.org/exports/root-1.1.xml
  - Repository of IDN Practices : https://www.iana.org/domains/idn-tables
  - Special-Use Domain Names : https://www.iana.org/assignments/special-use-domain-names/special-use-domain-names.xhtml
- From ICANN
  - IDN ccTLD Fast Track String Evaluation Completion https://www.icann.org/resources/pages/string-evaluation-completion-2014-02-19-en
  - Registry Agreement Termination Information Page https://www.icann.org/resources/pages/gtld-registry-agreement-termination-2015-10-09-en
  - ICANN Geographic Regions https://meetings.icann.org/en/regions
  - gTLD contract status https://www.icann.org/resources/registries/gtlds/v2/gtlds.json
  - Various (non-ccTLD) zone files, see: https://www.dns.icann.org/services/axfr/
- More...

Sources of Data

- From the UN
  - Standard country or area codes for statistical use (M49) https://unstats.un.org/unsd/methodology/m49/overview/
- From the Regional Internet Registries
  - RIPE's RPKI validator https://rpki-validator.ripe.net/api/objects/validated
    - *Discontinued 16 September 2021*
    - Replaced by look-alike validated ROA list via a local Routinator instance
  - FTP or AXFR of zone files from the NRO plus the RIRs
    - Two sources required requests and whitelisting, others are open
- More...

Sources of Data

- From Team Cymru (via a DNS API)
  - IP to ASN Mapping Service https://team-cymru.com/community-services/ip-asn-mapping/
- The Public Suffix List https://publicsuffix.org/list/public_suffix_list.dat
  - This is used as a guide to identify, but not a definitive source of, sub-elements of TLDs
- (Other) DNS Queries
  - To fill in various records (SOA, NS, DNSKEY, DS, etc.)
  - To discover the commercial registration boundary

The Commercial Registration Boundary

- An “estimation” of the commercial registration boundary is performed
  - DNS queries are used
  - Targeting names suspected as being sub-elements of a TLD registry
  - Augmented by names in the Public Suffix List
- A name is considered to be part of a registry if
  - It is in the zone and not a cut-point (delegation)
  - It is a cut-point (delegation) and shares 1 or more name servers with the TLD
  - The test is recursive through zones, some fourth level zones have been identified
- This is still experimental
  - Other approaches have been tried
  - Unanswered: is finding this boundary worth all the work?

Assembly of Data

- Much (but not all) information from each source is passed through unaltered
  - Information that might be personally identifying is filtered
  - Keywords of values hint at the source
  - This is meant to keep the census as a utility for, not a product of, research
- The following are synthesized by the census:
  - CENSUS_START (start of run)
  - CENSUS_END (end of run)
  - CENSUS_SOURCES (where information has been obtained)
  - CENSUS_CATEGORY (divides into gTLD/ccTLD/revMap, etc.)
  - CENSUS_JURISDICTION (Two-letter code, with XA as global for non-jurisdictional items)

What Data is Covered? (Part 1)

- ZONES
  - A-label, U-label
  - Apex, zone cut, and registered technical information
  - Root database meta-data
  - gTLD contract meta-data
  - IDN ccTLD meta-data
  - Agreement Termination meta-data
  - IDN table references
  - UN regional names
  - ICANN regional name
  - Size data
  - non-delegated names (empty non-terminals, CNAME/DNAME owners)
- More...
- All ”where applicable or available”

What Data is Covered? (Part 2)

- NAMESERVERS
  - Name
  - Registered addresses (from zone file or registry database)
  - Cut-point (glue) addresses (DNS referral pointers) including the glue is from
  - Authoritative addresses (DNS response to an address query)
  - List of zones where the name server appears (cut-point or authoritative)

- Glue information and Authoritative information are supposed to be the same
  - But even in the DNS core, this is not always true

What Data is Covered? (Part 3)

- ADDRESSES
  - Address
  - Address family
  - Route Origins
  - Registered, glue, and authoritative use sets (what name servers claim the address)
- ROUTE ORIGINS
  - BGP (Autonomous System Number) information
  - Route Prefix
  - ROA status
- AUTONOMOUS_SYSTEM_NUMBERS
  - Registered name of the operator

What Data is Covered? (Part 4)

- DNSKEY_RECORDS, DS_RECORDS, RRSIG_RECORDS
  - Three separate tables
  - Individual fields within the named Resource Record (RR)
  - These tables are used to “flatten” the census into tables
- IDN PRACTICES
  - Topic – name of the language or script
  - Kind – whether the table covers a language or a written script
  - Version – registry supplied
  - Date – registry supplied
  - URL – location of the table

What Data is **not** Covered?

- I haven't included available ccTLD zone files to maintain consistency in the "coverage"
  - Most (numerically) ccTLD zone files are not readily available
  - I.e., I don't include sizes of ccTLDs
  - Don't have complete coverage of the "commercial registration boundary"

- This would help fill in the research data
  - But it is better to come from the source as coverage is uneven
  - I'm open to combining efforts, mindful of data-sharing concerns as well as the consistency mentioned above

Presentation of Data

- Data is stored in a SQL-like database that is not publicly accessible
  - Tabular, one table per nine topics (like ZONES, NAMESERVERS), divided by months
  - Convenient for use in python – Pandas DataFrames
  - This form of the data would be a model for any distribution platform that offers APIs for accessing tables in part or in whole

- Data is also published into a daily set of nine topics in CSV form and in JSON form
  - After engineering configuration freezes thaw, these files will be pushed to a publicly accessible server
  - (Not promising a timeframe!)

Next Steps

- Publish the data sets
  - Working on an organization, including documentation

- Feedback
  - Is the included data worth the effort?
  - Is there other data that would be helpful?
  - What is the best organization? Data representation?

- Feedback requires published data...that will come along...in the meantime, expressions of interest would be appreciated

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