ICANN's Open Data Initiative Pilot

Alain Durand, Filling in for Ed Lewis

ICANN 60
02 November 2017

ICANN

Agenda

- Session Overview
- Open Data Initiative Pilot project overview (slide 4)
- Developing a Census of Data Sets (slide 10)
- Developing Prototype Pipelines (slide 18)
- Next Steps (Slide 27)

In This Session

- ICANN's Path to Open Data, Where We Are (30 minutes)
    - Alain Durand (Standing in)
    - Suzanne Woolf
    - Marc Blanchet
- Community Panel: Uses of Open Data, Big vs. Open, Analytics vs. Visualization, etc.
    - Jay Daley, NZRS
    - Roland Laplante, Afilias
    - Christa Taylor, DotTBA
    - Jonathan Zuck, CCTRT

ICANN's Path to Open Data

- First Step, Office of CTO run Open Data Initiative Pilot Program
- Lessons learned, hand off to appropriate teams within ICANN organization
- Engineering and Information Technology is the home for technical activities
- Other activities have yet to have an agreed upon home

What is Open Data?

- Today ICANN has numerous web pages and published reports containing "stats"
- Why is that not sufficient?
  - (Not in) Machine readable formats
  - (Not guaranteed) Timely delivery
  - (Not an) Automated process
  - (In-)Completeness in scope and time
- Other Considerations
  - Open and free access

Current State of ICANN's Data Access

- "Spotty"
- A lot of data scattered across web pages
- A lot of data presented in reports (PDF, PPT)
- Some data downloadable
- Some projects pursuing open data developments
- Interesting data available upon request

What's Wrong with That?

- Researchers have to know where to look, what to ask for
- Data is often incomplete or is not delivered in a consistent timely manner
- Data must be manually scraped from display formats
- The process to produce is heavily manual
- Difficult to perform casual inspection/discovery of data

ODI Pilot Focus

- Automation of data delivery
- Downloadable, computer readable, formats; well catalogued
- Interface that allows for casual inspection
- Repeatability of the entire process (discovering data sets, etc.)
- Vendor selection

ODI Pilot

- Learning/discovery phase
- Build vs. Buy
- Assuming procurement of a tool, what, how, how much
- Need for a Data Set Census emerges
- Need to prototype automated delivery of data

The Data Set Census

- In order to manage data, the organization needs to know what it has
- Traditionally, this hasn't been tracked
- More than one effort to identify data running now
  - General Data Protection Regulation survey
  - Internal Engineering &Information Technology (E&IT) survey
  - And the Open Data Initiative
- A living list, long term life is probably Operations

Aren't there pots of data waiting to be made open?

- It appears that there are pots of data, so "*all we should need is to tap in, right?*"
- Well, sometimes the pots of data are not pots at all
- In some cases, there's a large pot of data that is encumbered
  - But reports drawn from it are public
  - Need to "build" the the data set behind the reports

For example

- Human Resources data
  - It is in one large database
- But the gender diversity of staff is reported
  - That "report" isn't a data set itself, it is extracted
  - Would need to create that as a separate data set for open data

What is a data set?

- What is seen in reports isn't always a single data set
- "Staged Data Sets"
  - Staged meaning "prepared"
- This is an important question for a few reasons
  - ICANN staff needs to know what to identify as a data set
  - This determines what is ready to present versus what must be staged
  - Vendor tool pricing is based on number of data sets

What data sets are to be managed?

- Some data sets are operationally significant
  - For example, reports to ICANN required under contracts
- Some data sets are in development
  - Data collected as part of research prior to realizing what it means
- Some data sets are temporary
  - Mail list of people interested in a topic

What data sets are candidates for open data?

- Some data distribution is encumbered by agreements
  - Purchased data is not available for redistribution
  - Some sets are temporarily held back from release (three months afterwards)
  - Some data is redacted
- Some falls under personally identifiable information protections
  - General Data Protection Regulations, as one
  - For some, salaries are pubic but not the bank transfer information for a staff member

ODI Data Release Process (An Overview)

- Identify all data sets for internal purposes, generate and maintain the census
  - In process
- Identify all data sets that can be made public
  - Once first census is finished, a "batch" operation
  - Ongoing, this will be a consideration for all new or newly identified data sets
- Prioritize data sets for open data access
  - Bottlenecks are mechanical, size of platform, development of distribution
- Formalize process by which data is opened

Current state of the Data Census

- Preparing/prepared a draft version
- Coordination with other surveys (GDPR, E&IT) is on-going
- Next stage is to complete coverage and fill in gaps identified
- Drive procurement for production phase
- Look towards a census that is maintained over the long haul

Data Pipelines

- Alongside the data census, document the process for managing data
- Will consider the legal framework for data release, and licensing use
- Will cover what data sets and staged data sets are to be considered for census
- This is less well-formed at this stage

Prototyping Activity

- Early pilots developed
  - In June 2017, opened access to four parallel implementations
  - Goal: get a feel for the tools, how they appear
- Currently building a second round of pilots
  - Goal: get a feel for internal "pipelines"
  - Pipeline – referring to the automated transfer of data from where it is maintained to the open data platform

Buy vs. Build

- E&IT standard operating procedure is to use off-the-shelf components, whether open source or commercial, instead of internal builds
- Commercial market for open data products has matured, especially in the government open data market which largely matches ICANN's needs
- Current assumption is that we will purchase a specific vendor's tool and build upon it
    - "Assumption" meaning, the pilot hasn't finished yet

Data Consumer's Buying Guide (I.e., the Community)

- Can data sets be downloaded "whole" or in "raw format"?
- Can data be interpreted (numbers into meaning)?
- Can data be drilled into?
- Can visualizations be created?
- Can work be "saved/downloaded"?

Data Producer's Buying Guide (I.e., the Organization)

- Can we present a narrative around the data?
- Can we deliver the data for access?
- Can we stage data sets (mirroring reports) for access?
- Can we leverage the platform for more reporting?
- Can we make use of the tool in different media outlets?

Early Pilots

- https://www.icann.org/resources/pages/odi-pilot-2017-06-27-en
  - Internally developed CKAN-based tool
  - Enigma vendor solution
  - OpenDataSoft vendor solution
  - Socrata vendor solution

First Data Sets

- One time delivery of data
  - I.e., delivered a CSV file to each of the efforts
- Chosen data sets
  - Registry Monthly reports, activity
  - Registry Monthly reports, transactions
- Some vendors added more, already public, data

Feedback

- Set up a mailing list:
    - https://mm.icann.org/mailman/listinfo/odi-pilot
- Predictably, little activity
    - No prior session held to announce the work, just blog posts
    - Not well publicized
    - Limited data to play with
- Merely a starting point

Second Round of Pilots

- Due by the end of December 2017
- Include more data sets
- Concentrate on the internals
  - Data "Pipelines", automated delivery
- Two vendor tools included this round
  - OpenDataSoft
  - Socrata

Progressing the Pilot

- The pilot (census and pipeline work) is a research activity
- To be useful the activity has to include the beneficiaries
  - E&IT, runs production services, will inherit the responsibility for making data open
  - Operations, responsible for making the organization function, will likely inherit the Census
- Internal coordination across ICANN lines of reporting is proceeding

Upcoming Milestones

- (Internal) Draft of the Data Census, "this week"
- Second round of pilots, December 2017
- (Internal) Final version of 2017 Census, March 2018
- (internal) Preliminary Open Data Process, March 2018
- Potential for more pilots, March 2018
- (Internal) Final Open Data Process, June 2018
- Handoff of Pilot into Production, FY2019