The ICANN GNSO "Business Constituency"

Comment on Open Data Initiative Datasets and Metadata

Status: FINAL

Version: 3

27-Jul-2018

Business Constituency Submission

GNSO//CSG//BC

Background

This document is the response of the ICANN Business Constituency (BC), from the perspective of business users and registrants, as defined in our Charter:

The mission of the Business Constituency is to ensure that ICANN policy positions are consistent with the development of an Internet that:

1. promotes end-user confidence because it is a safe place to conduct business
2. is competitive in the supply of registry and registrar and related services
3. is technically stable, secure and reliable.

Business Constituency Comment on Open Data Initiative (ODI) Datasets and Metadata

Introduction

The Business Constituency (BC) has long been interested in bringing the vast sources of anonymized infrastructure data that ICANN holds to public attention and allowing for the public processing and investigation of that data. As we've articulated in letters and meetings over the years, making such data open and available to anyone who wishes to examine it is an essential part of ICANN's responsibilities in regard to transparency and accountability, and underpins informed, responsible action on the part of ICANN's community, staff and Board. We welcome this long-awaited posting.

In these comments, the BC responds to the three questions explicitly asked in the original request for comments, making specific and actionable improvements for prioritization of the ODI datasets and the Metadata model used to support the publishing of that data. Most importantly, we emphasize that the ODI activity must move from a model where ICANN simply identifies data available to be published to a model where the emphasis is on the use and analysis of that data. We believe that a fundamental shift in the ODI program is required that focuses on utility of the data. We will discuss these points later in the document.

A. *What are your priorities for publication of datasets identified in the data asset inventory?*

There have been several occasions when the Business Constituency has identified specific datasets that should be made available immediately and these remain valid.

It should be noted that in our Panama meeting we met with the CTO Office (David Conrad and Matt Larson), who invited us to comment on the priority of ODI data items that should be provided first.

(Presumably, the need for priorities is driven by the fact that some of the datasets are not in a format that ICANN can easily move into the open data platform, an issue we address later in the document.)

In addition to this and many other meetings with ICANN Org on this topic, the BC has previously commented on ICANN Org's publication of data as follows:

- Jan-2018. Comment on Competition, Consumer Trust, and Consumer Choice Review Team (CCTRT) – New Sections to Draft Report of Recommendations
- Sep-2017. Comment on report of Statistical Analysis of DNS Abuse in gTLDs
- May-2017 Comment on Competition, Consumer Trust and Consumer Choice Review Team Draft Report of Recommendations for New gTLDs (see data/abuse part)

- January-2017 Letter from the CSG to Göran Marby, Steve Crocker and the ICANN Board and response - Letter from Göran Marby to the Commercial Stakeholder Group
- Sep-2015. Initial Report on Data & Metrics for Policy Making

Previous prioritization requests notwithstanding, the Business Constituency identifies seven broad areas of datasets that should be prioritized for the Open Data Initiative.

1 - Historical and ongoing zone file data

One of the most important tools to understanding the health, activity and market dynamics of the global DNS is the zone file system. The current Centralized Zone Data Service (CZDS), while well intentioned, suffers from both flaws of execution and scope. Subscribers regularly struggle with uncooperative registries – an ongoing compliance issue that merits action by ICANN Compliance. Our constituency has seen earlier RFPs that sought to examine the scope of DNS abuse and the overall health of the DNS. Those RFPs have, in the past, offered vendors access to zone file data collected by ICANN.

The Open Data Initiative should identify the datasets in the Data Asset Inventory that either are copies of zone file data or are artifacts of the processing of that data. Once identified, those datasets should be published in the Open Data Platform rather than forcing people and organizations to go through the registries. That data should include versioning and historical copies of zone files. The data should be published in a way that is completely open and transparent – in keeping with the goals of the ODI and not of the CZDS. Access to the data must be reliable and easily automated, with the ability to re-synchronize following a disruption and resumption in service access or delivery. The data formatting must be clearly specified and machine-readable. Given that many ODI data sets will be processed using automation, changes to data formatting and the rationale should be presented to the community well in advance of actual format change to provide data consumers ample time to accommodate for change.

Ideally ongoing current zone file data would be updated multiple times intra-day, but at a minimum no less often than once per day.

2 - Historical and ongoing WHOIS system performance and compliance datasets

Metrics that record historic and ongoing WHOIS system performance and compliance are critical for auditing the execution of this important function. We are aware that ICANN Org has access to a broad array of datasets in this regard. We request that you include disaggregated data that is attributed to individual registrars and registries. While we respect the need for anonymization of the raw data, the data should not be aggregated across the organizations responsible for collecting and publishing it.

We note that, later in the Data Asset Inventory, the WHOIS data accuracy dataset appears. We are, however, looking for datasets that go beyond the attempt to sample, examine and report on accuracy.

We also note in the Data Asset Inventory there are a large collection of datasets whose identifier begins with the string "cct." These data sets seem to be an ongoing collection of metrics on the topic of complaints. Such a group of datasets is an important contributor to any examination of DNS Abuse, but are separate from this request for broad and historic WHOIS system performance and compliance data.

3 - IANA Function Performance

A key priority for the BC in its advocacy for the IANA transition was to ensure that the IANA function is, and remains, fully accountable and transparent. While we see three identifiers related to IANA in the

inventory, we are surprised to not see the datasets collected to measure the function of IANA. Functional requirements are established in MoU's with IANA's stakeholders.

The IANA function operator regularly collects and reports on performance and related information. That data is reported upon in a variety of forums and is missing from the Data Asset Inventory. This is an oversight that could easily be fixed. The IANA function performance data is already in an easily digestible format and would be easy to add to an Open Data Platform. We believe it should not only be added to the Data Asset Inventory, but should also be a priority for immediate addition to the Open Data Platform.

4 - Compliance Data

We note that it is sometimes difficult to determine the precise content of datasets in the inventory based on their description. In particular, the datasets identified with "cct" in their string appear to often have compliance or complaint related information in them. For the purposes of prioritization, any data related to complaints and compliance should be an immediate candidate for inclusion in the Open Data Platform. This will be a key dataset for trend analysis, so historical information is a fundamental requirement of this category.

The datasets identified with "cct" in their string should be included in the Open Data Platform, but we note that there is no mention of historical data. In particular, we believe that the raw data of each complaint must be published with the complainant identity redacted. The subject of each complaint should not be redacted to enable attribution to the complainant.

5 - Pricing Data

We are aware that pricing data is being collected as part of the gTLD Marketplace work and may have been collected in previous years for other initiatives. However, a search of the Data Asset Inventory for the word "price" or the word "pricing" finds no results. We would like to see ICANN Org's current work on pricing included in the Data Asset Inventory since it is an essential element of understanding the health and abuse in the DNS marketplace. The BC understands that pricing is dynamic and that promotions affect pricing. What the BC seeks in obtaining pricing information is a means for ICANN to provide an objective, reproducible, data-determined answer to the question of whether pricing is an influencing factor to abuse.

The Business Constituency requests that regular, anonymized pricing data be collected and published through the Open Data Initiative. In addition, the BC seeks data on domain prices published for Sunrise registrations for those gTLDs that launch with a Sunrise period.

6 - DNSSEC Deployment and Implementation Data Beyond the Top-level

In general, DNSSEC deployment data is difficult to come by publicly. OCTO and the community have an interest in collecting and publishing usable data that helps the community understand both the current and historic level of deployment of DNSSEC – both at the root and at the second level. Publication of data about the root is relatively easy but appears not to be done in a consistent and public manner. On the other hand, publicly published data about the number of signed delegations, DS records, etc. would make it possible to see how effectively DNSSEC has been deployed and the progress that is being made. The Business Constituency requests that this data be included. The recent failure to be able to drive a decision about the KSK rollover based on data is an indication of how important consistent, published data is to the community.

7 - Fellowship Data

Coming from what amounts now to years of observations and research by different members of the BC, we find it pressing that we can understand the role that the Fellowship program plays in bringing business actors to ICANN. In order for us to correctly position ourselves in relation to the program and be able to interact with it in the most productive manner possible, a comprehensive dataset containing anonymized applicant information is necessary. As it stands, by only having data about selected Fellows, we are unable to understand if there are businesspeople applying and being rejected or if they are simply not being reached.

We see this data as an important companion to the self-funded research the constituency has been carrying out to increase our Global South membership, and would like ICANN to contribute in this effort by providing this dataset for our analysis.

B. Are there any errors or omissions in the data asset inventory?

To be more useful, the inventory needs to be categorized in a variety of ways—including potential use, source, complexity and format. The current inventory fails to provide enough information about the relationships between the datasets other than a linear survey of some available data sources. Finally, the metadata model fails to account for the fact that the source data may not be the same as the published data.

For each identifier in the Data Asset Inventory a description that is more specific than the column "Published as Data" is needed. Instead, when the Data Asset Inventory has "yes" in that column, it also should indicate the formatting of the source data. We recognize that there have been many, diverse formats for storing data at ICANN. However, in an inventory, it should be a relatively easy task to identify that format.

The Data Asset Inventory also does not provide any guidance on how raw data sources will be published. While we understand the need to import these datasets into a comprehensive and consistent platform, the time taken to do that is stretching out unacceptably. The BC needs ICANN Org to simply publish the raw data sources before importing them into the platform and applying appropriate metadata. We request that a set of data formats be chosen that are easily used by the community – and where datasets are available in those formats – and publish them immediately prior to the process of moving them to the open data platform. Once in the open data platform, these data sources should remain as a matter of record – a way for those using and investigating the data to audit and inspect the process of moving to the platform.

We also request that the Data Asset Inventory be categorized. In its current form, it appears as a simple list of data assets with some very basic metadata about each source. For instance, the CCT metrics or root zone metrics should be categorized as a group. Then, the establishment of priorities could proceed to meet community needs through identifying those groups that make the most sense for the highest priorities. As it stands, the Data Asset Inventory is a long and impressive list – but the nature of that list makes it difficult to provide specific suggestions for prioritization.

In addition, we believe that there is an important connection between the source data for ODI and the data-as-published. The inventory of available datasets mixes extremely simple datasets with highly complex databases with multiple data sources. It is essential to have a very rich set of metadata that represents the data—not just as-published—but also making the clear link between the source data and the data-as-published.

Finally, we suspect that there are several datasets missing from the Data Asset Inventory that have been produced through previous work – whether for OCTO, regional DNS Marketplace Evaluations, KSK Rollover work, or for MSSI as or support for CCT and ATRT2 Reviews. We discuss this further in the metadata vocabulary discussion below.

C. ***Does the proposed metadata vocabulary meet your needs?***

At an overview level, it is surprising that the Metadata model assumes that the source of the data is always ICANN. Besides being inflexible, we think that in practice this will not be true. At the Puerto Rico ICANN meeting we suggested that the community might be the source of new data for ODI – either through independent collection of data or as the production of artifacts and additional analytical work. ODI might also serve as a repository for data that is not produced directly by ICANN, but instead produced by third parties working on behalf of ICANN. We recognize that ICANN Org would need to create a review process that addresses things like data accountability in order to use non-ICANN generated data. While this is not an agreed-upon priority at this time, any proposed metadata vocabulary should be able to accommodate it.

Metadata should accurately reflect the actual source of the data. While it may be possible for the "publisher" to be confined to "ICANN," we do think metadata needs to include accurate source data – and, repeating, this is not always "ICANN."

The metadata includes the concept of "theme" which is broadly consistent with our discussion of "categories" when we discussed Data Asset Inventory above. The current Metadata approach includes the restriction "each one must in exist in the taxonomy being developed by ITI." It is difficult to talk about a piece of metadata when the content of that metadata is not presented. We would like to better understand why that taxonomy was not presented with the metadata, how flexible it will be to change it as circumstances change, and how well the taxonomy will reflect the needs of the users of the data.

Clearly, the metadata is one tool that users of the ODI will have to understand regarding the underlying datasets in the Data Asset Inventory. There is no metadata that describes the fields, contents and parameters of each record in the underlying dataset. We request that metadata describe not only the dataset as a whole, but also give the user of the dataset enough information to understand the format of the records, their content and potential use.

While we appreciate that the OCTO approach is a combination of prioritization and ease of import, what is not clear (at least from the Data Asset Inventory) is how the raw data sources are formatted.

It seems likely that – in some cases - ICANN will transform raw data into publishable data on the ODI platform. If so, the metadata must document both the metadata for the raw data and the transformed artifacts. Users of the ODI need to be able to understand both the underlying source data and the data as published by ICANN. The metadata model confuses the two. The metadata model should separate information about the source data from the data-as-published and then provide descriptors to document them both.

Finally, the Metadata Standard hints at future use of Media types as a descriptor for the format of the underlying data. For example, the Metadata Standard itself might be in the format "describedbyType" : "application/pdf." This is an interesting possibility, but we believe that there will be many data sources that are in a standard format, but not a media type published by the IETF. It is still valuable to have the "conformsTo" descriptor in these cases, but it would be essential to provide a mechanism where the underlying standard was not a Media Type.

D. ***Understanding the Context for ODI***

The ODI request for comments implies a priority inversion in the goals. To identify meaningful data and usable access, understanding the analyses and metrics is important. We should strive to know what we are looking for before we know if we are presenting it properly. While the community has made clear requests for specific data sets (and requested the addition of data sets that ICANN has/has access to but of which the community is not aware), the clear indication in all communications has been the need for results, metrics, analysis, etc. While the data is necessary for transparency and reproducibility, the need to productively use the data is what is driving this project. ICANN should illustrate the utility of how the ODI data will be beneficial before any declaration can be made of the success of its usefulness. We request that the ODI Initiative be updated to reflect that the community of data users will be asked to confirm whether the completeness, formatting and indexing of the data is sufficient for the desired analyses.

The varying forms of the raw data sets and the multitude of approaches to provide updates and deltas in the Data Asset Inventory indicates a distinct possibility that the ODI could result in open access to unusable data. The tabulation of meta-data seems incomplete as no illustration is provided as to how the data and meta-data will be useful in a meaningful analysis. This would seem to be a requirement before declaring the proper format.

The community has clearly asked for data that supports results (metrics, analysis, etc.). It is clear from numerous communications to the Board that ICANN has access to and/or responsibilities to facilitate access to vast amounts of data but must offer it in a productive way. Responses from the Board that opine on the difficulties of providing data, or the inability to immediately understand solutions to access or cull relevant data are insufficient. The entire OCTO should be considered a resource to investigate and facilitate data access initiatives.

A few notable citations of prior work published by the scientific community (and eScience communities) regarding data access and how these existing problems were addressed are included in the endnotes. We hope ICANN Org finds them useful.[1]

E. ***Selection of Open Data Platform***

The pace and substance of the original work of bringing four, competing, open data platforms to the community for comment was confusing. ICANN has indicated that "the RFP to choose an open data platform is almost complete and an announcement [would] be made by ICANN62 in Panama City". It is unclear however whether an open data platform has been selected.

This uncertainty about the choice of Open Platform Data Platform has increased our concern related to other aspects of the project, which will be built upon the platform. The urgency of our requests for raw data sets to be made available is not entirely due to our concerns about the completeness of the metadata vocabulary; it is also driven by our concern that selecting the data platform and formatting the data into it is likely to take longer than anticipated, preventing some types of analysis which could be performed by BC members on the raw data.

F. Conclusion

The BC looks forward to working with ICANN to support expeditious implementation of the Open Data Initiative to benefit the public and support informed, responsible action on the part of ICANN's community, staff and Board. Making this data open and available for public use is a key aspect of

ICANN's transparency and accountability responsibilities. We ask that ICANN give it the priority attention and support it deserves.

--

This comment was drafted by Denise Michel, Faisal Shah, Mark Datysgeld, and Mark Svancarek, with edits by Claudia Selli and Tim Chen.

It was approved in accord with the BC charter.

1

- Carly, Strasser, Kunze John, Abrams Stephen, and Cruse Patricia. "DataUp: A tool to help researchers describe and share tabular data." (2014). https://philpapers.org/rec/STRDAT-8
- Soyka, Heather, Amber Budden, Viv Hutchison, David Bloom, Jonah Duckles, Amy Hodge, Matthew Mayernik et al. "Using peer review to support development of community resources for research data management." (2017). https://darchive.mblwhoilibrary.org/handle/1912/9351
- Danielle Pollock. 2016. Understanding scientific data sharing outside of the academy. In *Proceedings of the 79th ASIS&T Annual Meeting: Creating Knowledge, Enhancing Lives through Information & Technology* (ASIST '16). American Society for Information Science, Silver Springs, MD, USA, Article 144, 5 pages. https://dl.acm.org/citation.cfm?id=3017447.3017591
- Kratz, John E., and Carly Strasser. "Making data count." *Scientific data* 2 (2015). https://www.nature.com/articles/sdata201539
- Tenopir, Carol, Elizabeth D. Dalton, Suzie Allard, Mike Frame, Ivanka Pjesivac, Ben Birch, Danielle Pollock, and Kristina Dorsett. "Changes in data sharing and data reuse practices and perceptions among scientists worldwide." *PloS one* 10, no. 8 (2015): e0134826. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0134826