Summary Report of Public Comment Preceding

Open Data Initiative Datasets and Metadata

Publication Date:	12 September 2018
Prepared By:	Matt Larson

Public Comment Proceeding	
Open Date:	11 June 2018
Close Date:	27 July 2018 (original) 10 August 2018 (extended)
Staff Report Due Date:	20 August 2018 (original) 31 August 2018 (extended)

Important Information Links
Announcement
Public Comment Proceeding
View Comments Submitted

Staff Contact:	Matt Larson	**Email:**	matt.larson@icann.org

Section I: General Overview and Next Steps

General Overview

On 11 June 2018, the Office of the Chief Technology Officer within ICANN (OCTO) posted Open Data Initiative Datasets and Metadata for public comment. The deadline to receive public comments was 27 July 2018. A request for extension was received and accepted, extending the deadline to receive public comments to 10 August 2018.

The Open Data Initiative is almost ready to start publishing datasets. A Data Asset Inventory has been created, which provides an initial list of the datasets that are potential candidates for publication, and associated metadata standards have been defined. The community input will be used to help determine the priorities for publishing datasets on the upcoming ICANN open data platform and amending any elements of the publication plan to address community input.

At the time this report was drafted, seven comments were submitted to the forum.

Next steps

OCTO will use the public comment as set out above. OCTO will forward on any public comments that require analysis and decision by another department within ICANN, such as requests to make data open that is currently only available under restricted terms or not at all.

Section II: Contributors

At the time this report was prepared, a total of seven (7) community submissions had been posted to the forum. The contributors, both individuals and organizations/groups, are listed below in chronological order by posting date with initials noted. To the extent that quotations are used in the foregoing narrative (Section III), such citations will reference the contributor's initials.

Organizations and Groups:

Name	Submitted by	Initials
Non-Commercial Stakeholder Group	Rafik Dammak	NCSG
Business Constituency	Steve DelBianco	BC
At-Large Advisory Committee	ICANN At-Large Staff	ALAC

Individuals:

Name	Affiliation (if provided)	Initials
Suriyaa Sundararuban	-	SS
Wisdom Donkor	Africa Open Data and Internet Research Foundation	WD
Mark W. Datysgeld	Business Constituency	MWD
Chokri Ben Romdhane	-	CBR

Section III: Summary of Comments

General Disclaimer: This section intends to summarize broadly and comprehensively the comments submitted to this public comment proceeding but does not address every specific position stated by each contributor. The preparer recommends that readers interested in specific aspects of any of the summarized comments, or the full context of others, refer directly to the specific contributions at the link referenced above (View Comments Submitted).

ICANN has received seven (7) comments from the community on the Open Data Initiative Datasets and Metadata. For ease of reference, comments submitted are included below organized by commenter. The text has been edited to remove any introduction of the author and to reformat it to enable paragraph numbering.

1. **Comments from Suriyaa Sundararuban (SS):**

1.1. I support the coming release of an ICANN Data Asset Inventory (DAI) system by ICANN.org and the Open Data Initiative. ICANN's intention that every dataset on this list that can be published as open data, makes the metadata, datasets and the work of ICANN more Open Source, transparent and easier accessible for everyone. I also think it's good that the Data Asset Inventory (DAI) supports standard formats such as CSV and PDF. This makes it easier for people to access the data files with different software and from different operating systems without causing any compatibility issues.

2. **Comments from Wisdom Donkor (WD):**

2.1. Asset and value potentials of data are widely recognized at all levels. Data collected or developed through public investments, when made publicly available and maintained over time, their potential value could be more fully realized. There has been an increasing demand by the community, that such updated data collected should be made more readily available to all, for enabling rational debate, increase transparency better decision making and use in meeting civil society and needs. Efficient sharing of data among data owners and inter-and-intra agencies along with data standards and interoperable systems is the need of the hour. Hence, there is the need to formulate a policy on ICANN data Sharing and Accessibility which could provide an enabling provision and platform for proactive and open access to the data generated through ICANN funds available with various Communities.

2.2. ICANN OPEN DATA POLICY should aim at providing an enabling provision and platform for proactive and open access to the data generated by various communities within ICANN. The objective of this policy is to facilitate access to ICANN owned shareable data (along with its usage information) in machine readable form globally in a periodically updatable manner, within the framework of various related policies, acts and rules of open data accepted policy, thereby permitting a wider accessibility and usage by the Public.

2.3. Different types of datasets generated both in geospatial and non-spatial form by communities within ICANN are supposed to be classified as shareable data and non-shareable data. Data management should encompass the systems and processes that ensure data integrity, data storage and security, including metadata, data security and access registers. The principles on which ICANN data sharing and accessibility should be based should include: Openness, Flexibility, Transparency, Quality, Security and Machine-readable.

2.4. *Identification of Resources (Datasets/Apps) and their organization under Catalogs*

2.4.1. As per open data policy, I expect ICANN communities to prepare it’s Negative List. The datasets which are confidential in nature and are in the interest of the global security is not opening to the public would fall into the negative list. However, all other datasets which do not fall under this negative list would be in the Open List. These datasets would need to be prioritized into high value datasets and non-high value datasets.

2.4.2. The data which are contributed to the ICANN OPEN DATA INITIATIVE Platform have to be in the specified open data format only. The data have to be internally processed to ensure that the quality standard is met i.e. accuracy, free from any sort of legal issues, privacy of an individual is maintained and does not compromise with any national authority. While prioritizing the release of datasets, one should try to publish as many high value datasets. Grouping of Related Resources (Datasets/Apps) should be planned and are to be organized under Catalogs. That way assessing becomes more easier.

I expect that, each communities within ICANN or ICANN data controllers should have its own criterion of high value and low value datasets, generally High value data are governed by following Principles

*Completeness *
2. Primary
3. Timeliness
4. Ease of Physical and Electronic Access
5. Machine readability 6. Non-discrimination
7. Use of Commonly Owned Standards
8. Licensing
9. Permanence
10. Usage Costs

2.5. *Data Formats*

2.5.1. I will recommend that data has to be published in open format. It should be machine readable. Though there are many formats suitable to different category of data. Based on current analysis of data formats prevalent in Government it is proposed that data should be published in any of the following formats:

- CSV (Comma separated Values)
- XLS (spread sheet- Excel)
- ODS (Open Document Formats for Spreadsheet)
- XML (Extensive Markup Language)
- RDF (Resources Description Framework)
- KML (Keyhole Markup Language used for Maps)
- GML (Geography Markup Language)
- RSS/ATOM (Fast changing data e.g. hourly/daily)

2.5.2. Rate of all data sets should meet the Tim Berners-Lee 5 star data classification. <https://www.google.com/search?q=tim+berners+lee+5+stars+open+data&source=lnms&tbm=isch&sa=X&ved=0ahUKEwibo6DI8d_bAhUJJcAKHVtBAPwQ_AUICigB&biw=1602&bih=796#imgrc=1nw1OIVHD4ZohM:>

2.6. *The ICANN open data initiative platform*

2.6.1. ICANN Open Data Initiative Platform should be setup with the primary purpose to collate access to ICANN Resources (datasets/apps) under Catalogs, published by different ICANN communities or entities in open format. It also provides a search & discovery mechanism for instant access to desired datasets. The Platform should also have a rich mechanism for public engagement. Besides enabling public to express their need for specific resource (datasets or apps) or API, it also should allow pubic rate the quality of datasets; seek clarification or information from respective Data Officer or data controller. The platform should have a strong backend data management system that can be used by Communities or entities within ICANN to publish their datasets through a predefined workflow. The platform should be integrated with visualization engine to allow the creation and viewing of visualization of the various datasets. The platform should have a dashboard to see the current status on datasets, visualizations, usage Metrics or analytics as well as feedback and queries from the public.

2.6.2. ICANN should encourage the integration of Communities component of the platform which that will help facilitates the forming of communities around published ICANN datasets. Example or could be app developers' community etc. This will give first hand input to development community for building new components, apps for the various ICANN communities for effective ICANN engagement. The key features to consider are listed below:

· Open Source Driven – Developed completely using Open Source Stack, facilitating cost saving in terms of software and licenses and also provisioning community participation in terms of further development of product with modules of data visualization, consumption, APIs to access datasets etc.

· Metadata – Resources (Datasets/Apps) should be published along with standard metadata along with controlled vocabularies on various communities, jurisdictions, dataset types, access mode etc. Besides facilitating easy access to datasets, this should be extremely useful in the future for integration of data catalogs.

· Social Media Connect–IT should support wider reach and dissemination of datasets, anyone can share the information about any dataset published on the platform with his/her social media pages on a press of a click.

· Public Engagement – The Platform should have a strong component of public

Engagement. Public can express their views as well as rate the datasets w.r.t three aspects (Quality, Accessibility and Usability) on the scale of 5. They can also embed the Resources (Datasets/Apps) in their blogs or web sites. Facility to contact the Data Officers should also be made available on the Platform.

· Community Collaboration – Public with specific interest can build communities and discuss online. ICANN open data policy and Platform should facilitate communities to open up online forums, blogs and discussions around various datasets, apps available on the platform. It also should provide a platform to express and discuss the kind of Datasets, APPs & APIs they would like to have. It should also give input to communities or entities as what kind of datasets is more useful and accordingly prioritize the release of those datasets.

2.7. *Metadata Elements for Catalogs/Resources and their Description*

1. *Catalog Title (Required): * The title of the dataset is very important aspect of the dataset

*Description (Required): * Provide a detailed description of the catalog e.g., an abstract determining the nature and purpose of the catalog.

*Keywords (Required): * It is a list of terms, separated by commas, describing and indicating at the content of the catalog. Example: rainfall, weather, monthly statistics.

Group Name: This should be an optional field to provide a Group Name to multiple catalogs in order to show that they may be presented as a group or a set.

Community name (Required): Choose the Communities/entities those most closely applies to your catalog.

Asset Jurisdiction (Required): This is a required field to identify the exact location or area to which the catalog and resources (dataset/apps) caters to viz. entire country, state/province, district, city, etc.

*2. ** Resources (Datasets/Apps)*

Category (Required): Choose from the drop down options. Is it a Dataset or an Application.

Title (Required): A unique name of the resource etc.

Access Method (Required): This could be “Upload a Dataset” or “Single Click Link to Dataset”.

Reference URLs: This could include description to the study design, instrumentation, implementation, limitations, and appropriate use of the dataset or tool. In the case of multiple documents or URLs, please delimit with commas or enter in separate lines.

* If Resource Category is Dataset

Frequency (Required): This should mentions the time interval over which the dataset is published on the ICANN Open Data Platform on a regular interval (one-time, annual, hourly, etc.).

Granularity of Data: This should mentions the time interval over which the data inside the dataset is collected/ updated on a regular basis (one-time, annual, hourly, etc.)

Access Type: This should mentions the type of access viz. Open, Priced, Registered Access or Restricted Access.

** If Resource Category is App *

App Type (Required): This should mention the type of App being contributed viz. Web App, Web Service, Mobile App, Web Map Service, RSS, APIs etc.

Datasets Used: Datasets used for making this app.

Language: Language used for app

Date Released: Should mention the release date of the Dataset/App.

Note: Should capture any information the contributor/controller wishes to provide to the data consumer or about the resource.

ICANN OPEN DATA Policy Compliance: This field should indicate if this dataset is in conformity with the with the ICANN Open Data Sharing and Access Policy.

2.8. *Capacity Building*

2.8.1. Finally, ICANN should make it possible to build the capacity of the data controllers with the ICANN communities. I will recommend two types of training modules both as offsite and onsite models should be envisaged. Each module would be for the duration of 2-3 days and should be within ICANN events. The modules would be:

· *Awareness and Sensitization Module* – for Data Officer or Controlers & other senior officers of the of the Communities

· *Data Contribution Module* – hands-on training for contributing datasets to the ICANN Open Data Initiative Platform, provide advisory on conversion of data to digital format to Data Contributors and Members of ICANN Communities.

2.9. I have work with the US government open data team, world bank open data team, Africa Open data collaborative, Open Data Institute, Open Data Canada, Global Open Data in Agriculture and Nutrition, India Open Data team, World Wide Web Foundation. I will be more than happy to contribute my time to this laudable initiative anytime any day.

3. Comments from Mark W. Datysgeld (MWD):

3.1. Having performed empirical research about ICANN and its workings, these are the datasets I see as priority and the reasoning behind my choices:

A) meeting-registrations: Studying participation in ICANN should rely much more on looking at the raw data and projecting long-term correlations than on the glimpses provided by the reports; I champion this as a priority because it is data that I am actually using for research, so I can attest from first-hand experience that it is useful but at the same time what we have right now is not good enough for serious inquiry.

B) meeting-session: In some ways related to the previous point, the continued evolution of ICANN depends on the community being able to understand what we are spending our time on and being able to paint at a broader picture of the activities being carried out.

C) fellowship: For proper evaluation of this program, a comprehensive dataset containing the applicant information is necessary, so that is can be crossed with the pool of selected fellows to look for inconsistences. The dataset can be liberally redacted, but what I see as a must is: stakeholder, country, age, and number of application attempts.

D) accountability family of indicators: While it is nice that these are available in an interactive dashboard, it is still much more desirable for research purposes that they be made available as raw data. it is hard to say which are more deserving of being prioritized, but I tend to think that subgroups 1, 4 and 5 are more aligned with the type of evaluation being done of ICANN at the moment.

E) gnso-list-statistics: As a member of the GNSO, I find it imperative to be able to better understand how the broader policy process works, and what patterns can surface from the analysis of this kind of data.

F) gac-members and gac-working groups: Having done extensive research on the GAC before, it is actually quite hard to get a historic notion of their workings, even though the data is available in a very scattered manner. This should be easier to study and I don't think it must be such a complicated dataset to work with from the ICANN side.

3.2. Also, below is a list of data that would be nice to have, but which I don't think has as much priority:

A) meeting-technical
B) financials
C) applications [ngTLD]
D) board-documents
E) ithi

3.3. I would like to thank all of those involved in moving this initiative forward in a way that attempts to listen to what the community actually wants in terms data instead of supposing what is needed or just going through the most convenient datasets. I will continue to support the development of the ODI in whatever capacity I can.

4. Comments from Chokri Ben Romdhane (CBR):

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

4.1.1. *Access to the Financial reports and SOs/ACs output should be the priorities of the open data initiative*

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

4.2.1. *GSE Working space activities and output (by regions and by workgroups) are not not invented

Example:

https://community.icann.org/display/MES/Middle+East+Working+Group <https://community.icann.org/display/MES/Middle+East+Working+Group>*

4.2.2. *Document * The public extract of the Data Asset Inventory is available <https://www.icann.org/en/system/files/files/odi-data-asset-inventory-11jun18-en.pdf>

4.2.3. *Landing Page column* *On click Target **URL ** truncated. * *Example: URL specified in **Landing Page column *http://stats.dns.icann.org/hedgehog/ *On click URL: *http://stats.dns.icann.org/h

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

4.3.1. *The Catalog object have (0,1) carnality and will include only Dataset, Looking to the ICANN structure and activities it will be useful to include the concepts sub-catalog(s) (or collection(s)) within a catalog.*

4.3.2. *Did the download URL Field content will be limited to a link to the dataset materials or It will be used as a resource to index the content of this dataset <https://project-open-data.cio.gov/v1.1/schema/#dataset> materials?* *Indexing dataset materials will be useful in order to implement powerful search engine on full-text and media, this will make Dataset materials more visible for the community. *

In another hand a few metadata vocabulary are really used within t*he data asset inventory, Example in the case of text and media files this metadata could be easily enriched by the properties of this files.*

4.3.3. *Access Field should depend on license tag.*

4.3.4. *Publisher Field cardinality is (1,1) : Dataset may be published by several Organisations not only one.*

5. **Comments from the Non-Commercial Stakeholder Group (NCSG):**

5.1. The Non-Commercial Stakeholder Group (NCSG) welcomes the opportunity to comment on the Open Data Initiative (ODI) data sets and the metadata that ICANN intends to publish along with each data set. This is a welcomed move forward for ICANN's bylaw-mandated principles of accountability and transparency, and we support the ODI's concept of bringing comprehensive access to raw data to the ICANN community in order to enable evidence-based policy development. While the publication of raw data in and of itself is a first step toward transparency, it is only when the right data is published in a means that can be meaningfully understood, that there is real value generated from this exercise. In pursuit of this objective, the NCSG has structured our comments to address (1) the proposed data sets, (2) the required data quality, and (3) the resources that the ICANN community requires in order to interrogate this data.

5.2. Proposed Data Sets

5.2.1. If properly executed, we believe the ODI will enhance the transparency and accountability of ICANN's activities. After thorough review, we believe that the data sets that should be published are those which are relevant to current and future topics in the community, practicable in use, and of good quality. They should allow the ICANN community to identify trends, to take note of pain points, and to monitor the effectiveness of activities and corrective actions that have been taken. However before we delving into the technicalities, we note our strong position on the need for ICANN to respect human rights, including the fundamental right to privacy, particularly as it relates to personal data which may be in ICANN's possession.

5.2.2. In technical terms, NCSG's priorities for publishing datasets on the open data platform have been set taking into consideration: (1) available data, and (2) NCSG data needs. We note that some of the data in the spreadsheet was missing, thus our prioritization comes based off of the limited data available to us. We have also flagged the data that we think is of least priority especially the ones that need to be only partially shared where personal information might be made public.

5.2.3. It is of critical importance to the NCSG that first amongst the published data sets be those that contain information related to, but not limited to, (1) accountability, (2) current policy development topics, (3) transparency of ICANN in terms of expenditure, including staff travel expenses, DIDP requests and Ombudsperson case dealings, as well as (4) certain metrics related to their effect on non-commercial and end-users, and (5) data related to ICANN's respect or lack thereof for human rights. The data sets which the NCSG has prioritized, based on the available information in the provided inventory, can be found in Appendix A *[Staff note: Appendix A was not included, see next point for details]*

5.2.4. *[Staff note: The link to Appendix A in the submitted comment directs to the following footnote:]* The Titles which NCSG prioritized - based on the available information in the provided inventory - can be found in this spreadsheet, graded as Necessary or High Priority. Also available on the spreadsheet, is the least prioritized Titles that we think should not be on top of the list as they are not as relevant Least relevant Slightly relevant but not of priority and Prioritized but not high on the scale. 5: Necessary - 4: High Priority - 3: Priority - 2: Slightly relevant - 1: Least relevant.

5.3. Data Quality

5.3.1. The NCSG is a strong advocate of evidence-informed policy making; in order to make our public interest-oriented contributions, we require data that is (1) timely and comprehensive, (2) accessible and usable, and (3) comparable, consistent, and interoperable.

5.3.1.1. Timely and Comprehensive

5.3.1.1.1. It is the position of the NCSG that in order for the ODI to be useful, ICANN must publish data in a timely manner according to a publicly published schedule. Data must be complete and a system of version control must be in place. If there are errors which are later corrected, these corrections must be made obvious.

5.3.1.2. Accessible and Usable

5.3.1.2.1. Data must be published in a format that allows for easy access and interrogation. It must be machine parsable (for instance, published in formats such as, but not limited to JSON and XML), provided via APIs, and it must be released under an open data license that permits free use, re-use, and distribution. In addition, it would be an asset for data to be visualised where useful.

5.3.1.3. Comparable, Consistent, and Interoperable

5.3.1.3.1. Data must be published in a consistent and comparable manner to allow for comparable units to be compared with one another with the passage of time.

5.4. Community Resource Requirements

5.4.1. Access to data only advantages those who have access to interrogate and analyse said data. Unlike many commercial stakeholders who participate in

ICANN processes in order to achieve business objectives, as a community of volunteers the NCSG pursuing public interest-oriented objectives, the NCSG does not have a long list of data scientists who we can call on in order to parse this information. We therefore would like ICANN to make independent data analysis support and training available to us. We believe this request to be consistent with recommendation 10.5 of the Accountability and Transparency Review 2 report, accepted by the ICANN Board in 2014, which called for ICANN to *"facilitate the equitable participation in applicable ICANN activities, of those ICANN stakeholders who lack the financial support of industry players."*

5.5. Errors in the Data Set Inventory Spreadsheet

5.5.1. In our review of the data set inventory spreadsheet, the NCSG observed a number of issues with the data that we ask be corrected in a future update of this resource. These include:

5.5.1.1. the "description field" for some identifiers is missing, for instance in sla-monitoring, transaction-reports, operator-reports, activity-reports, [srt]-expenses, when it is mentioned in the metadata vocabulary document that this field is always required

5.5.1.2. the landing page of the identifier [multiple] is missing, although it is said to be published, and published as a page (for reference: Column A, Row 18 - Appendix A).

5.5.2. There are also **omissions**. Data sets which the NCSG is interested in, but are not mentioned in the inventory, include:

5.5.2.1. Number of emails sent to the ICANN policy support staff off of a mailing list, whether a response was sent to the stakeholder or not, and whether or not ICANN staff took action following this email. This should be sorted by stakeholder affiliation;

5.5.2.2. Full details of any events sponsored by ICANN, including event name, location, date(s), whether the support was one-off or recurring, and the nature of the support (if financial, including figures in USD);

5.5.2.3. Any reporting of inability to participate in an ICANN meeting due to the location of the meeting and/or visas issues;

5.5.2.4. All meeting transcripts including the ones related to closed meetings;

5.5.2.5. Details on ICANN-funded face-to-face meetings, including the number of attendees in each meeting room (physical and online), use of scheduling apps by unique participants, and the date of registration and subsequent badge collection;

5.5.2.6. How many email messages have been sent to the Ombudsperson; how many informal complaints have been received; how many formal complaints have been received; and how many cases have been forward (if any) or rejected; and

5.5.2.7. Tracking effects of policy changes at ICANN.

5.6. In terms of the proposed metadata vocabulary, as far as our discussions have gone, we believe the title vocabulary is sufficient enough for a start, however a lot of information is missing under those titles.

5.6.1. The fields "keyword" and "theme" are declared as "always required" in the metadata description but keyword field is populated with only ten out of 231 lines, while the column "theme" was never really utilized at all. We would like to better understand the role of this field and the reason to have it if there will be no information crowding it.

5.6.2. There are also two titles that are flagged as restricted (tld-zones and L-Root), and we would appreciate a brief explicit reasoning as to why they are restricted.

5.7. As this initiative aims to increase the accountability and transparency of ICANN, we are looking forward to the successful roll-out of this initiative. However, there is a need for care and responsibility when personal information is being disclosed (which should not be used as an excuse not to publish information that is in the public interest, such as details on staff travel expenditure), and there is also a strong need for periodic reviews of the ODI, including of its data sets, in case there is necessary information not found in the inventory.

5.8. The NCSG is grateful for the opportunity to comment on this issue, and we trust that our comments will be taken into consideration. We are available to clarify our recommendations, if necessary.

6. Comments from the Business Constituency (BC):

6.1. Introduction

6.1.1. 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.

6.1.2. 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.

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

6.2.1. 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.)

6.2.2. 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: *[Staff note: The following may have been intended by the BC to be links to other documents but no links were in the submitted comment]*

6.2.2.1. Jan-2018. Comment on Competition, Consumer Trust, and Consumer Choice Review Team (CCTRT) – New Sections to Draft Report of Recommendations

6.2.2.2. Sep-2017. Comment on report of Statistical Analysis of DNS Abuse in gTLDs

6.2.2.3. May-2017 Comment on Competition, Consumer Trust and Consumer Choice Review Team Draft Report of Recommendations for New gTLDs (see data/abuse part)

6.2.2.4. 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

6.2.2.5. Sep-2015. Initial Report on Data & Metrics for Policy Making

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

6.2.4. 1 - Historical and ongoing zone file data

6.2.4.1. 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.

6.2.4.2. 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.

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

6.2.5. 2 - Historical and ongoing WHOIS system performance and compliance datasets

6.2.5.1. 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.

6.2.5.2. 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.

6.2.5.3. 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.

6.2.6. 3 - IANA Function Performance

6.2.6.1. 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.

6.2.6.2. 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.

6.2.7. 4 - Compliance Data

6.2.7.1. 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.

6.2.7.2. 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.

6.2.8. 5 - Pricing Data

6.2.8.1. 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.

6.2.8.2. 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.

6.2.9. 6 - DNSSEC Deployment and Implementation Data Beyond the Top-level

6.2.9.1. 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.

6.2.10. 7 - Fellowship Data

6.2.10.1. 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.

6.2.10.2. 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.

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

6.3.1. 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.

6.3.2. 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.

6.3.3. 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.

6.3.4. 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.

6.3.5. 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.

6.3.6. 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.

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

6.4.1. 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.

6.4.2. 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.”

6.4.3. 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.

6.4.4. 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.

6.4.5. 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.

6.4.6. 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.

6.4.7. 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.

6.5. D. Understanding the Context for ODI

6.5.1. 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.

6.5.2. 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.

6.5.3. 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.

6.5.4. 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.

6.5.4.1. 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

6.5.4.2. 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

6.5.4.3. 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

6.5.4.4. Kratz, John E., and Carly Strasser. "Making data count." *Scientific data* 2 (2015). https://www.nature.com/articles/sdata201539

6.5.4.5. 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

6.6. E. Selection of Open Data Platform

6.6.1. 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.

6.6.2. 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.

6.7. F. Conclusion

6.7.1. 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.

7. Comments from At-Large Advisory Committee (ALAC)

7.1. The ALAC appreciates the opportunity to comment on ICANN’s Open Data Initiative. The ALAC applauds this ICANN initiative to keep the ICANN Community informed of the data it collects and the resolve to publish collected data assets in as openly form as reasonably permissible.

7.2. Centralized, easy access to properly organized data repository

7.2.1. It is noted that the identified datasets are published at various locations. While the ALAC understands that different groups within the ICANN Community, and even within ICANN Org, have varying interest and use for different datasets, it is recommended that all the datasets to be **published at a single, centralized online location** which is **easily accessible** to all interested parties.

7.2.2. Descriptions for each dataset should be specific and unambiguous, and perhaps supported by a form of **simple keyword-based taxonomy** which allows each dataset to be tagged to provide supplemental user-guided context to otherwise general descriptions. This would make the datasets more understandable and searchable as well.

7.2.3. Of great interest to the ALAC are the online means made available to query the collected data. While we appreciate that it may be difficult for ICANN to develop and/or provide a common tool which would satisfy the data querying and analysis needs of every group within the ICANN Community, nevertheless, the ALAC proposes that ICANN engage in some effort to develop or license an **tool that would enable the ICANN Community to undertake basic querying of user-selected datasets.** Alternatively, the ALAC would appreciate if ICANN can suggests readily available, cost-effective online tools for querying and analysis the datasets. Education of the recommended tool(s) is also crucial. Paramount to both approaches, however, and for the overall success of this initiative, is the continued adoption of the three dimensions of data openness which the ALAC supports.

7.3. Types and value of data collected, lack of discernable information

7.3.1. While it has embarked on a laudable start with 231 named datasets, from the ALAC's perspective, it is **not only difficult for us to identify those of most interest to our group, but also those which possess discernable derivative value.**

7.3.2. Certainly, ICANN meeting demographics and the data specifically associated with At-Large participants/members rank high on our list, as do those related to competition, consumer trust and consumer choice. But of greater interest to the ALAC is data that is not readily identifiable or discernable from the datasets listed in https://www.icann.org/en/system/files/files/odi-data-asset-inventory-11jun18- en.pdf.

7.3.3. Most obvious is a lack of exhaustive data about contractual compliance and the actions it takes. This is arguably one of the most critical areas of ICANN's operations and other than some specific data sets compiled for the CCT Review, there appears to be nothing.

7.3.4. Another example that is of interest to At-Large is data associated with the Fellowship. The URL listed implies that the only information to be provided is a list of fellows along with the country and interest area. Absent however are the demographics about the Fellowship applicants (ie those who succeeded versus those who did not). Such critical data is needed to indicate to what extent information about the Fellowship Programme is reaching certain parts of the world, which would in turn facilitate fact-driven corrective action (if necessary) and for planning purposes.

7.3.5. Yet another example that is of interest to us is data associated with the membership of At-Large, in terms of participation rates.

7.3.6. Taking the above-mentioned examples further, there is a need to identify and capture (if not already present) metrics-based downstream data for datasets where there is a sequence of actions to be taken or for which some level of success or effectiveness needs to be measured for programme assessment and planning purposes. For our purposes, downstream data that can certainly inform on the effectiveness of various programmes include, but not limited to, the following:-

7.3.6.1. Contractual compliance: measurements of corresponding action taken, time taken to resolve, patterns of non-compliance, plausible trigger events/reasons for non-compliance

7.3.6.2. CCT-related complaints: types of complaints, time to resolve, patterns of domain name abuse etc, plausible trigger events

7.3.6.3. Fellowship programme: participation metrics of returning fellows versus first-time or one-time fellows, transition from fellows to active community membership

7.3.6.4. Membership, related to ALS and individual members:

7.3.6.4.1. diversity metrics of by country, region, gender, economy, disability status etc,

7.3.6.4.2. participation metrics in At-Large in policy development, education & outreach activities, direct & remote participation in meetings

7.3.6.4.3. travel-related metrics such as difficulties in obtaining travel support, visas, difficulties with Travel Constituency etc.

7.4. Uniformity of and responsibility for data

7.4.1. Understanding the methodology of how data which is of interest to us will be accumulated is also an important consideration. It should be noted that data which is or may be of interest to the ALAC currently resides in separate repositories -- eg those data collected and controlled exclusively by ICANN Org for ICANN operations versus those data collected by ICANN staff for the ALAC which reside, for all intents and purposes, behind the ALAC website and wiki ("the ALAC's repositories").

7.4.2. In this context, some preliminary questions arise:

7.4.2.1. For the data that already exists on the web, are there conceivably duplicates of data residing in separate repositories?

7.4.2.2. Will new data continue to be collected and stored in the existing manner? If yes, how will ICANN ensure that the two stay in sync with each other?

7.4.2.3. For the purposes of the open data platform, will ICANN Org be querying data in the ALAC's repositories?

7.5. Privacy rights

7.5.1. The ALAC supports the need to consider privacy rights and recognizes ICANN's legal obligations in processing and publishing data containing personal elements but cautions against withholding personal data to the point of rendering the data worthless. The approach of anonymizing data may be called for if even such data is NOT made publicly available and this should be applied in general.

7.5.2. In very specific cases where personal data is needed to be shared, and without which would render the data worthless to a user, then ICANN should consider placing confidentiality obligations on users who have been specifically identified and authorised to receive data containing personal elements, to do so on a limited license basis. As an example, limit sharing and use of Fellowship participant data to just the ALAC and not At-Large.

7.6. Conclusion

7.6.1. Thus, it would be useful if ICANN Org could assist in re-generating a list of datasets with suggestions on what downstream or upstream information can possibly be gleaned from each dataset. The ALAC believes such an exercise would assist both ICANN Org and the ICANN Community to better understand whether the range of data being collected is sufficiently complete and what related data is available to explain changes in the data, and if not, those that can and ought to be collected.

7.6.2. Once a revised list of datasets is established, it should be submitted for public comment. It is far easier to critique such a list than create it from scratch.

Section IV: Analysis of Comments

General Disclaimer: This section intends to provide an analysis and evaluation of the comments submitted along with explanations regarding the basis for any recommendations provided within the analysis.

OCTO appreciates all comments and suggestions added to the public forum for the Open Data Initiative Datasets and Metadata. All comments received are favorable to the general idea of open data.

Comments can be broadly categorised as follows:

1. Comments in support of the importance of the topic
2. Comments on how data should be published
3. Comments on the content of the data asset inventory
4. Comments on the metadata schema

Responses to specific comments are given below following the comment to which they apply. For readability, portions of the comment are elided with the notation [...], enabling the specific point addressed to be clearly identified.

1. Comments from Suriyaa Sundararuban (SS):

1.1. [...] I also think it's good that the Data Asset
Inventory (DAI) supports standard formats such as CSV and PDF. [...]

Response 1: The Open Data Platform (ODP) that will be used to publish the open data, will make that data available in a variety of machine readable formats, including CSV, ODATA and JSON.

2. Comments from Wisdom Donkor (WD):

2.1. [...] Hence, there is the need to formulate a policy on ICANN data Sharing and Accessibility which could provide an enabling provision and platform for proactive and open access to the data generated through ICANN funds available with various Communities.

Response 2: A policy covering data generated by various ICANN communities, as distinct from that generated or collected by ICANN Org would require a community based process to develop and as such is out of scope for this initiative.

2.2. ICANN OPEN DATA POLICY should aim at providing an enabling provision and platform for proactive and open access to the data generated by various communities within ICANN. [...]

See **Response 2**.

2.3. Different types of datasets generated both in geospatial and non-spatial form by communities within ICANN are supposed to be classified as shareable data and nonshareable data. Data management should encompass the systems and processes that ensure data integrity, data storage and security, including metadata, data security and access registers. The principles on which ICANN data sharing and accessibility should be based should include: Openness, Flexibility, Transparency, Quality, Security and Machinereadable.

See **Response 2**.

Response 3: ICANN is developing a data governance framework that defines how ICANN manages data. The principles will be those set out in the Open Data Charter with a minor amendment to replace the word "Citizen" with "Community".

2.4. *Identification of Resources (Datasets/Apps) and their organization under Catalogs*

2.4.1. As per open data policy, I expect ICANN communities to prepare it's Negative List. The datasets which are confidential in nature and are in the interest of the global security is not opening to the public would fall into the negative list. However, all other datasets which do not fall under this negative list would be in the Open List. These datasets would need to be prioritized into high value datasets and non-high value datasets.

Response 4: In the full Data Asset Inventory (DAI), which is used internally with ICANN for data governance purposes, each dataset has a classification that specifies whether or not that dataset can be published. When future versions of the DAI are published the classification field will be included.

2.4.2. The data which are contributed to the ICANN OPEN DATA INITIATIVE Platform have to be in the specified open data format only. The data have to be internally processed to ensure that the quality standard is met i.e. accuracy, free from any sort of legal issues, privacy of an individual is maintained and does not compromise with any national authority.

Response 5: The internal ICANN data governance framework covers internal requirements for the quality of the data and ensuring privacy regulations are complied with.

While prioritizing the release of datasets, one should try to publish as many high value datasets. Grouping of Related Resources (Datasets/Apps) should be planned and are to be organized under Catalogs. That way assessing becomes more easier.

Reponse 6: As noted in the request for public comment, the prioritization of datasets for publication will use a combination of community priorities and the difficulty of data publication.

I expect that, each communities within ICANN or ICANN data controllers should have its own criterion of high value and low value datasets, generally High value data are governed by following Principles

*Completeness *

2. Primary

3. Timeliness
4. Ease of Physical and Electronic Access
5. Machine readability 6. Non-discrimination
7. Use of Commonly Owned Standards
8. Licensing
9. Permanence
10. Usage Costs

Response 7: All data will be governed by the same set of principles as set out in **Response 3**. Please note the following specific response to the principles proposed here:

- All data will be identically licensed using the Open Database License.
- No charge will apply to use of the data.
- All data will be available in machine readable formats as set out in **Response 1**.

2.5. *Data Formats*

2.5.1. I will recommend that data has to be published in open format. It should be machine readable. Though there are many formats suitable to different category of data. Based on current analysis of data formats prevalent in Government it is proposed that data should be published in any of the following formats:

- CSV (Comma separated Values)
- XLS (spread sheet- Excel)
- ODS (Open Document Formats for Spreadsheet)
- XML (Extensive Markup Language)
- RDF (Resources Description Framework)
- KML (Keyhole Markup Language used for Maps)
- GML (Geography Markup Language)
- RSS/ATOM (Fast changing data e.g. hourly/daily)

See **Response 1**.

2.5.2. Rate of all data sets should meet the Tim Berners-Lee 5 star data classification. <https://www.google.com/search?q=tim+berners+lee+5+stars+open+data&source=lnms&tbm=isch&sa=X&ved=0ahUKEwibo6DI8d_bAhUJJcAKHVtBAPwQ_AUICigB&biw=1602&bih=796#imgrc=1nw1OIVHD4ZohM:>

Response 8: The intention of the Open Data Initiative is to publish all data as three star data initially and over time raise that to four star data.

2.6. *he ICANN open data initiative platform*

2.6.1. ICANN Open Data Initiative Platform should be setup with the primary purpose to collate access to ICANN Resources (datasets/apps) under Catalogs, published by different ICANN communities or entities in open format. It also provides a search & discovery mechanism for instant access to desired datasets. The Platform should also have a rich mechanism for public engagement. Besides enabling public to express their need for specific resource (datasets or apps) or API, it also should allow pubic rate the quality of datasets; seek clarification or information from respective Data Officer or data controller. The platform should have a strong backend data management system that can be used by Communities or entities within ICANN to publish their datasets through a predefined workflow. The platform should be integrated with visualization engine to allow the creation and viewing of visualization of the various datasets. The platform should have a dashboard to see the current status on datasets, visualizations, usage Metrics or analytics as well as feedback and queries from the public.

Response 9: The market in open data publication platforms is diverse with different platforms having a different subset of the full range of features listed. The first priority is for features associated with publication to skilled data consumers. The second priority is for features supporting the manipulation and exploration of that data. Features for public engagement are one of the lowest priorities. Until a final decision on an ODP vendor is chosen, ICANN cannot confirm what features will be supported by the ODP.

2.6.2. [...] The key features to consider are listed below:

· Open Source Driven – Developed completely using Open Source Stack, facilitating cost saving in terms of software and licenses and also provisioning community participation in terms of further development of product with modules of data visualization, consumption, APIs to access datasets etc.

Response 10: While open source products are being considered as part of the selection process for an ODP this is not a requirement of the system.

· Metadata – Resources (Datasets/Apps) shoulb be published along with standard metadata along with controlled vocabularies on various communities, jurisdictions, dataset types, access mode etc. Besides facilitating easy access to datasets, this should be extremely useful in the future for integration of data catalogs.

Response 11: A proposed metadata standard has been provided as part of this request for public comment.

· Social Media Connect–IT should support wider reach and dissemination of datasets, anyone can share the information about any dataset published on the platform with his/her social media pages on a press of a click.

· Public Engagement – The Platform should have a strong component of public Engagement. Public can express their views as well as rate the datasets w.r.t three aspects (Quality, Accessibility and Usability) on the scale of 5. They can also embed the Resources (Datasets/Apps) in their blogs or web sites. Facility to contact the Data Officers should also be made available on the Platform.

· Community Collaboration – Public with specific interest can build communities and discuss online. ICANN open data policy and Platform should facilitate communities to open up online forums, blogs and discussions around various datasets, apps available on the platform. It also should provides a platform to express and discuss the kind of Datasets, APPs & APIs they would like to have. It should also give input to communities or entities as what kind of datasets is more useful and accordingly prioritize the release of those datasets.

See **Response 9**.

2.7. *Metadata Elements for Catalogs/Resources and their Description*

1. *Catalog Title (Required):* The title of the dataset is very important aspect of the dataset

Description (Required): Provide a detailed description of the catalog e.g., an abstract determining the nature and purpose of the catalog.

Keywords (Required): It is a list of terms, separated by commas, describing and indicating at the content of the catalog. Example: rainfall, weather, monthly statistics.

Group Name: This should be an optional field to provide a Group Name to multiple catalogs in order to show that they may be presented as a group or a set.

Response 12: These fields are already present in the proposed metadata standard.

Community name (Required): Choose the Communities/entities those most closely applies to your catalog.

Response 13: A community name is unlikely to be relevant to many datasets and may unnecessarily limit the use and understanding of a dataset and so is not included.

Asset Jurisdiction (Required): This is a required field to identify the exact location or area to which the catalog and resources (dataset/apps) caters to viz. entire country, state/province, district, city, etc.

Response 14: It is not clear if this comment proposes a geotag for the jurisdiction of the data or the area it covers. If the former then this is not necessary as all data is published by ICANN under an open data. If the latter then there is already a ‘spatial’ field specified in the proposed metadata standard.

*2. ** Resources (Datasets/Apps)*

Category (Required): Choose from the drop down options. Is it a Dataset or an Application.

Response 15: It may not be clear that the metadata standards proposed are solely those for datasets that are to be published and not for the DAI itself. The proposed metadata standards therefore are only for data that is public via the ODP as datasets.

Response 16: All datasets to be made available through the ODP will be made available as datasets.

Title (Required): A unique name of the resource etc.

Response 17: This field is already present in the proposed metadata standard.

Access Method (Required): This could be “Upload a Dataset” or “Single Click Link to Dataset”.

Response 18: It is not clear what this comment proposes.

Reference URLs: This could include description to the study design, instrumentation, implementation, limitations, and appropriate use of the dataset or tool. In the case of multiple documents or URLs, please delimit with commas or enter in separate lines.

Response 19: This field is already present in the proposed metadata standard.

* If Resource Category is Dataset

Frequency (Required): This should mentions the time interval over which the dataset is published on the ICANN Open Data Platform on a regular interval (one-time, annual, hourly, etc.).

Granularity of Data: This should mentions the time interval over which the data inside the dataset is collected/ updated on a regular basis (one-time, annual, hourly, etc.)

Response 20: These fields are already present in the proposed metadata standard.

Access Type: This should mention the type of access viz. Open, Priced, Registered Access or Restricted Access.

See **Response 15**.

** If Resource Category is App *

App Type (Required): This should mention the type of App being contributed viz. Web App, Web Service, Mobile App, Web Map Service, RSS, APIs etc.

Datasets Used: Datasets used for making this app.

Language: Language used for app

Date Released: Should mention the release date of the Dataset/App.

Note: Should capture any information the contributor/controller wishes to provide to the data consumer or about the resource.

See **Response 16**.

ICANN OPEN DATA Policy Compliance: This field should indicate if this dataset is in conformity with the with the ICANN Open Data Sharing and Access Policy.

See **Response 15**.

2.8. *Capacity Building*

2.8.1. Finally ICANN should make it possible to build the capacity of the data controllers with the ICANN communities. I will recommend two types of training modules both as offsite and onsite models should be envisaged. Each module would be for the duration of 2-3 days and should be within ICANN events. The modules would be:

· *Awareness and Sensitization Module* – for Data Officer or Controlers & other senior officers of the of the Communities

· *Data Contribution Module* – hands-on training for contributing datasets to the ICANN Open Data Initiative Platform, provide advisory on conversion of data to digital format to Data Contributors and Members of ICANN Communities.

Response 21: This is not in the scope of the Open Data Initiative project nor would it appear to be within the limited remit of the ICANN organization.

2.9. [...]

3. Comments from Mark W. Datysgeld (MWD):

3.1. Having performed empirical research about ICANN and its workings, these are the datasets I see as priority and the reasoning behind my choices:

A) meeting-registrations:[...]

B) meeting-session:[...]

C) fellowship: [...]

D) accountability family of indicators: [...]

E) gnso-list-statistics: [...]

F) gac-members and gac-working groups: [...]

3.2. Also, below is a list of data that would be nice to have, but which I don't think has as much priority:

A) meeting-technical
B) financials
C) applications [ngTLD]
D) board-documents
E) ithi

Response 22: These priorities are all noted.

3.3. [...].

4. Comments from Chokri Ben Romdhane (CBR):

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

4.1.1. *Access to the Financial reports and SOs/ACs output should be the priorities of the open data initiative*

Response 23: This preference is noted and will be used when setting publication priorities.

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

4.2.1. *GSE Working space activities and output (by regions and by workgroups) are not not invented
Example:
https://community.icann.org/display/MES/Middle+East+Working+Group
<https://community.icann.org/display/MES/Middle+East+Working+Group>*

Response 24: This omission will be investigated further.

4.2.2. *Document * The public extract of the Data Asset Inventory is available <https://www.icann.org/en/system/files/files/odi-data-asset-inventory-11jun18-en.pdf>

Response 25: It is not clear what this comment refers to.

4.2.3. *Landing Page column* *On click Target **URL ** truncated. * *Example: URL specified in **Landing Page column *http://stats.dns.icann.org/hedgehog/ *On click URL: *http://stats.dns.icann.org/h

Response 26: This error is noted.

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

4.3.1. *The Catalog object have (0,1) carnality and will include only Dataset, Looking to the ICANN structure and activities it will be useful to include the concepts sub-catalog(s) (or collection(s)) within a catalog.*

Response 27: The proposed metadata standard includes the field 'isPartOf' which provides this functionality. The contents of this field are still to be added.

4.3.2. *Did the downloadURL Field content will be limited to a link to the dataset materials or It will be used as a resource to index the content of this dataset <https://project-open-data.cio.gov/v1.1/schema/#dataset> materials?* *Indexing dataset materials will be useful in order to implement powerful search engine on full-text and media, this will make Dataset materials more visible for the community. *

In another hand a few metadata vocabulary are really used within t*he data asset inventory, Example in the case of text and media files this metadata could be easily enriched by the properties of this files.*

Response 28: The download URL will be determined by the ODP and will already have been discovered before the metadata is retrieved and so that URL is not included within the metadata. This decision will be revisited following this comment.

4.3.3. *Access Field should depend on license tag.*

See **Response 7**.

4.3.4. *Publisher Field cardinality is (1,1) : Dataset may be published by several Organisations not only one.*

Response 29: There are many multiple complexities around data governance and quality that would come into play if ICANN were to publish data from third parties, such as the potential for accidental disclosure of personal information, the need to keep the data up to date, and the potential for a dataset license to change. Consequently, the ICANN ODP will only directly publish data produced or collected by ICANN Org. It is a standard approach within the open data community for each publishing organization to publish their own data and for the open data publication platforms to be federated, which is where the data published on one platform is indirectly available through another platform. It is intended that the ICANN ODP will support federation.

5. **Comments from the Non-Commercial Stakeholder Group (NCSG):**

5.1. [...]

5.2. Proposed Data Sets

5.2.1. If properly executed, we believe the ODI will enhance the transparency and accountability of ICANN's activities. After thorough review, we believe that the data sets that should be published are those which are relevant to current and future topics in the community, practicable in use, and of good quality. They should allow the ICANN community to identify trends, to take note of pain points, and to monitor the effectiveness of activities and corrective actions that have been taken. However before we delving into the technicalities, we note our strong position on the need for ICANN to respect human rights, including the fundamental right to privacy, particularly as it relates to personal data which may be in ICANN's possession.

Response 30: ICANN complies with all relevant laws in the processing and publication of data.

5.2.2. [...]

5.2.3. It is of critical importance to the NCSG that first amongst the published data sets be those that contain information related to, but not limited to, (1) accountability, (2) current policy development topics, (3) transparency of ICANN in terms of expenditure, including staff travel expenses, DIDP requests and Ombudsperson case dealings, as well as (4) certain metrics related to their effect on non-commercial and end-users, and (5) data related to ICANN's respect or lack thereof for human rights. The data sets which the NCSG has prioritized, based on the available information in the provided inventory, can be found in Appendix A *[Staff note: Appendix A was not included, see next point for details]*

5.2.4. *[Staff note: The link to Appendix A in the submitted comment directs to the following footnote:]* The Titles which NCSG prioritized - based on the available information in the provided inventory - can be found in this spreadsheet, graded as Necessary or High Priority. Also available on the spreadsheet, is the least prioritized Titles that we think should not be on top of the list as they are not as relevant Least relevant Slightly relevant but not of priority and Prioritized but not high on the scale. 5: Necessary - 4: High Priority - 3: Priority - 2: Slightly relevant - 1: Least relevant.

Response 31: These preferences are noted and will be used when setting publication priorities.

5.3. Data Quality

5.3.1. The NCSG is a strong advocate of evidence-informed policy making; in order to make our public interest-oriented contributions, we require data that is (1) timely and comprehensive, (2) accessible and usable, and (3) comparable, consistent, and interoperable.

5.3.1.1. Timely and Comprehensive

5.3.1.1.1. [...]

5.3.1.2. Accessible and Usable

5.3.1.2.1. [...].

5.3.1.3. Comparable, Consistent, and Interoperable

5.3.1.3.1. [...].

See **Response 3**.

5.4. Community Resource Requirements

5.4.1. Access to data only advantages those who have access to interrogate and analyse said data. Unlike many commercial stakeholders who participate in ICANN processes in order to achieve business objectives, as a community of volunteers the NCSG pursuing public interest-oriented objectives, the NCSG does not have a long list of data scientists who we can call on in order to parse this information. We therefore would like ICANN to make independent data analysis support and training available to us. We believe this request to be consistent with recommendation 10.5 of the Accountability and Transparency Review 2 report, accepted by the ICANN Board in 2014, which called for ICANN to *"facilitate the equitable participation in applicable ICANN activities, of those ICANN stakeholders who lack the financial support of industry players."*

See **Response 21**.

5.5. Errors in the Data Set Inventory Spreadsheet

5.5.1. In our review of the data set inventory spreadsheet, the NCSG observed a number of issues with the data that we ask be corrected in a future update of this resource. These include:

5.5.1.1. the "description field" for some identifiers is missing, for instance in sla-monitoring, transaction-reports, operator-reports, activity-reports, [srt]-expenses, when it is mentioned in the metadata vocabulary document that this field is always required

5.5.1.2. the landing page of the identifier [multiple] is missing, although it is said to be published, and published as a page (for reference: Column A, Row 18 - Appendix A).

Response 32: It was not made clear in the public comment that the copy of the data asset inventory provided for this public comment is still a work in progress and will continue to be completed.

5.5.2. There are also **omissions**. Data sets which the NCSG is interested in, but are not mentioned in the inventory, include:

5.5.2.1. Number of emails sent to the ICANN policy support staff off of a mailing list, whether a response was sent to the stakeholder or not, and whether or not ICANN staff took action following this email. This should be sorted by stakeholder affiliation;

Response 33: This dataset request is noted and will be passed to the relevant team for consideration.

5.5.2.2. Full details of any events sponsored by ICANN, including event name, location, date(s), whether the support was one-off or recurring, and the nature of the support (if financial, including figures in USD);

Response 34: This dataset request is noted and will be passed to the relevant team for consideration.

5.5.2.3. Any reporting of inability to participate in an ICANN meeting due to the location of the meeting and/or visas issues;

Response 34: This dataset request is noted and will be passed to the relevant team for consideration.

5.5.2.4. All meeting transcripts including the ones related to closed meetings;

Response 35: Multiple datasets listed in the data asset inventory include transcripts and it is unclear from this comment if there are any identified gaps other than closed meetings. The Open Data Initiative is not a mechanism to bypass community rules on confidentiality and so a request for transcripts for closed meetings cannot be resolved through this initiative.

5.5.2.5. Details on ICANN-funded face-to-face meetings, including the number of attendees in each meeting room (physical and online), use of scheduling apps by unique participants, and the date of registration and subsequent badge collection;

Response 36: This dataset is provided by the set of datasets with an identifier beginning with the prefix 'meeting-'.

5.5.2.6. How many email messages have been sent to the Ombudsperson; how many informal complaints have been received; how many formal complaints have been received; and how many cases have been forward (if any) or rejected; and

Response 37: This dataset request is noted and will be passed to the Ombudsperson for consideration.

5.5.2.7. Tracking effects of policy changes at ICANN.

Response 38: It is not clear what this requested dataset would contain.

5.6. In terms of the proposed metadata vocabulary, as far as our discussions have gone, we believe the title vocabulary is sufficient enough for a start, however a lot of information is missing under those titles.

5.6.1. The fields "keyword" and "theme" are declared as "always required" in the metadata description but keyword field is populated with only ten out of 231 lines, while the column "theme" was never really utilized at all. We would like to better understand the role of this field and the reason to have it if there will be no information crowding it.

See **Response 32**.

5.6.2. There are also two titles that are flagged as restricted (tld-zones and L-Root), and we would appreciate a brief explicit reasoning as to why they are restricted.

Response 39: This request will be passed to the relevant team for consideration.

5.7. [...] However, there is a need for care and responsibility when personal information is being disclosed (which should not be used as an excuse not to publish information that is in the public interest, such as details on staff travel expenditure), and there is also a strong need for periodic reviews of the ODI, including of its data sets, in case there is necessary information not found in the inventory.

Response 40: The ICANN policy as stated by the President and CEO is that "our default position must be that all of our data is open data unless there are good reasons to treat it otherwise". Also, see **Response 32**.

5.8. [...]

6. Comments from the Business Constituency (BC):

6.1. Introduction

6.1.1. [...].

6.1.2. [...].

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

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

Response 41: If those previous occasions and prioritization requests can be identified then they will be used when setting publication priorities.

6.2.2. [...]

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

6.2.4. 1 - Historical and ongoing zone file data

6.2.4.1. 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.

6.2.4.2. 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.

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

Response 42: The Open Data Initiative is not a mechanism to bypass community PDPs and consequently this request for zone data cannot be resolved through this initiative.

6.2.5. 2 - Historical and ongoing WHOIS system performance and compliance datasets

6.2.5.1. 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.

6.2.5.2. 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.

Response 43: It is not clear if this dataset exists but this request will be passed to the relevant team for consideration if it does.

6.2.5.3. 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.

Response 44: Dataset with an identifier beginning with ‘cct’ are those relating to “Competition, Consumer Trust” as developed by the CCT review team.

6.2.6. 3 - IANA Function Performance

6.2.6.1. 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.

6.2.6.2. 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.

Response 45: This data is covered by the dataset ‘pti-naming-performance’ and this request will be passed to the relevant team.

6.2.7. 4 - Compliance Data

6.2.7.1. 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.

6.2.7.2. 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.

Response 46: This request will be passed to the relevant team for consideration. Also see **Response 44**.

6.2.8. 5 - Pricing Data

6.2.8.1. 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.

6.2.8.2. 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.

Response 47: The purpose of the Open Data Initiative is to improve access to data that ICANN already holds by making that data open if possible. The omission of gTLD Marketplace data is because that work is still in progress and the datasets are not yet available. The collection and availability of pricing data is a topic for consideration as part of the gTLD Marketplace work.

6.2.9. 6 - DNSSEC Deployment and Implementation Data Beyond the Top-level

6.2.9.1. 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.

Response 48: It is not clear if this dataset exists but this request will be passed to the relevant team for consideration if it does.

6.2.10. 7 - Fellowship Data

6.2.10.1. 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.

6.2.10.2. 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.

Response 49: This data is covered by the dataset ‘fellowship’ and will be passed to the relevant team.

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

6.3.1. 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.

6.3.2. 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.

6.3.3. 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.

Response 50: It is not clear what this comment means with the references to ‘raw data sources’ and ‘applying metadata’. To clarify, data within ICANN is stored a multiplicity of systems and formats and a single dataset may contain fields that are confidential or contain confidential information. This comment seems to indicate there is an alternative to using an open data platform but that is not the case. For example, a significant amount of ICANN data is held in enterprise applications and it would neither appropriate nor secure for the community to be given direct access to those systems and so the data has to be extracted from that system in order to be published. The question then is how best to publish that data so that it meets the principles set out in **Response 3** and confirms to three-star open data.

The Open Data Initiative is building an infrastructure and process for this data to be piped from those source systems into a single open data platform, with an intermediate step to ensure privacy rules are met and to ensure that data conforms to the principles set out in **Response 3**. The data is then available from the open data platform in a variety of formats as explained in **Response 1**, which ensures that all data is equally accessible.

The metadata is published alongside the data and may optionally be retrieved by data consumers who wish to learn more about the dataset. The concept of 'applying metadata' to data does not make sense in this context.

A data governance framework is being put in place to ensure the data meets the principles as set out in **Response 3** and maintain community trust in that data.

6.3.4. 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.

See **Response 27**.

6.3.5. 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.

See **Response 50**.

6.3.6. [...]

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

6.4.1. 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.

6.4.2. 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."

See **Response 29**.

6.4.3. 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.

Response 51: The taxonomy being developed through the ITI was not available at the time this public comment was published. The theme field will be filled in when that taxonomy is available.

6.4.4. 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.

Response 52: It is planned to make that level of detail available through the open data platform though the exact mechanism for that is still to be decided. The ultimate goal is to publish a machine readable data dictionary for each dataset.

6.4.5. 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.

6.4.6. 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.

See **Response 50**.

6.4.7. 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.

Response 53: This response addresses future plans that are as yet uncertain and cannot be committed to. The ‘describedBy’ metadata field is a pointer to a data dictionary, which we hope to publish for each dataset as explained in **Response 52**. The data dictionary will be published in a machine readable format such as JSON or CSV (for the avoidance of doubt, PDF is not a machine-readable format) and the ‘describedByType’ indicates which machine-readable format.

6.5. D. Understanding the Context for ODI

6.5.1. 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.

Response 54: The philosophy behind open data in general and the open data initiative specifically is encapsulated in this quote from the ICANN President and CEO: "It is not our role to judge what is useful to publish, just as it is not our role to predetermine the usage that people will make of the data. When information is shared, it increases in utility and value as people analyze it, combine it, and enhance it in ways we cannot predict.". This is the start of a process that will develop over time, with the uses and value of data being discovered on the way. If during that process, problems with the data are discovered then those will need to be addressed.

6.5.2. 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.

Response 55: It is not clear what this comment is referring to with "multitude of approaches to provide updates and deltas". The use of metadata is a separate, well established field of data science and it is out of scope to explain that here.

6.5.3. [...]

6.5.4. [...]

6.6. E. Selection of Open Data Platform

6.6.1. 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.

Response 56: Unfortunately, the process for choosing an open data platform has taken a lot longer than expected. An announcement will be made when that decision is confirmed.

6.6.2. 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.

See **Response 50** and **Response 55**.

6.7. F. Conclusion

6.7.1. [...]

7. Comments from At-Large Advisory Committee (ALAC)

7.1. [...]

7.2. Centralized, easy access to properly organized data repository

7.2.1. It is noted that the identified datasets are published at various locations. While the ALAC understands that different groups within the ICANN Community, and even within ICANN Org, have varying interest and use for different datasets, it is recommended that all the datasets to be **published at a single, centralized online location** which is **easily accessible** to all interested parties.

Response 57: ICANN plans to publish all open data on a single open data platform.

7.2.2. Descriptions for each dataset should be specific and unambiguous, and perhaps supported by a form of **simple keyword-based taxonomy** which allows each dataset to be tagged to provide supplemental user-guided context to otherwise general descriptions. This would make the datasets more understandable and searchable as well.

Response 58: As specified in the proposed metadata standards, all datasets will be described with two metadata fields that are relevant to this comment: ‘keyword’ and ‘theme’, with the former being freeform and the latter restricted to categories from the standard taxonomy being developed by the ITI.

7.2.3. Of great interest to the ALAC are the online means made available to query the collected data. While we appreciate that it may be difficult for ICANN to develop and/or provide a common tool which would satisfy the data querying and analysis needs of every group within the ICANN Community, nevertheless, the ALAC proposes that ICANN engage in some effort to develop or license an **tool that would enable the ICANN Community to undertake basic querying of user-selected datasets.** Alternatively, the ALAC would appreciate if ICANN can suggests readily available, cost-effective online tools for querying and analysis the datasets. Education of the recommeneded tool(s) is also crucial. Paramount to both approaches, however, and for the overall success of this initiative, is the continued adoption of the three dimensions of data openness which the ALAC supports.

Response 59: An expected feature of the open data platform is limited data querying and visualisation. For anything beyond this see **Response 21**.

7.3. Types and value of data collected, lack of discernable information

7.3.1. [...]

7.3.2. [...]

7.3.3. Most obvious is a lack of exhaustive data about contractual compliance and the actions it takes. This is arguably one of the most critical areas of ICANN’s operations and other than some specific data sets compiled for the CCT Review, there appears to be nothing.

Response 60: This request will be passed to the relevant team for consideration.

7.3.4. Another example that is of interest to At-Large is data associated with the Fellowship. The URL listed implies that the only information to be provided is a list of fellows along with the country and interest area. Absent however are the demographics about the Fellowship applicants (ie those who succeeded versus those who did not). Such critical data is needed to indicate to what extent information about the Fellowship Programme is reaching certain parts of the world, which would in turn facilitate fact-driven corrective action (if necessary) and for planning purposes.

Response 61: The description provided for the dataset 'fellowship' states "Fellowship application data; selection of applications; collection of participant reports"

7.3.5. Yet another example that is of interest to us is data associated with the membership of At-Large, in terms of participation rates.

Response 62: It is not clear what specific data this comment is proposing but the request will be passed to the relevant team for consideration.

7.3.6. [...]

7.3.6.1. Contractual compliance: measurements of corresponding action taken, time taken to resolve, patterns of non-compliance, plausible trigger events/reasons for non-compliance

7.3.6.2. CCT-related complaints: types of complaints, time to resolve, patterns of domain name abuse etc, plausible trigger events

See **Response 60**.

7.3.6.3. Fellowship programme: participation metrics of returning fellows versus first-time or one-time fellows, transition from fellows to active community membership

Response 63: This request will be passed to the relevant team for consideration.

7.3.6.4. Membership, related to ALS and individual members:

7.3.6.4.1. diversity metrics of by country, region, gender, economy, disability status etc,

7.3.6.4.2. participation metrics in At-Large in policy development, education & outreach activities, direct & remote participation in meetings

Response 64: This request will be passed to the relevant team for consideration.

7.3.6.4.3. travel-related metrics such as difficulties in obtaining travel support, visas, difficulties with Travel Constituency etc.

Response 65: This request will be passed to the relevant team for consideration.

7.4. Uniformity of and responsibility for data

7.4.1. Understanding the methodology of how data which is of interest to us will be accumulated is also an important consideration. It should be noted that data which is or may be of interest to the ALAC currently resides in separate repositories -- eg those data collected and controlled exclusively by ICANN Org for ICANN operations versus those data collected by ICANN staff for the ALAC which reside, for all intents and purposes, behind the ALAC website and wiki ("the ALAC's repositories").

Response 66: The Open Data Initiative considers all data collected or created by ICANN as potentially in scope for publication, applying the existing community rules around confidentiality of that data. In that context, the data collected by ICANN staff on ALAC activities that are not considered confidential are regarded as candidates for publication just as any other ICANN collected data.

7.4.2. In this context, some preliminary questions arise:

7.4.2.1. For the data that already exists on the web, are there conceivably duplicates of data residing in separate repositories?

7.4.2.2. Will new data continue to be collected and stored in the existing manner? If yes, how will ICANN ensure that the two stay in sync with each other?

7.4.2.3. For the purposes of the open data platform, will ICANN Org be querying data in the ALAC's repositories?

Response 67: It is recognised that ALAC data is held on the web site as the current system of record and this provides some challenges when building a reliable and repeatable pipeline to push that data to the open data platform. These challenges are being discussed with the relevant teams within ICANN.

7.5. Privacy rights

7.5.1. The ALAC supports the need to consider privacy rights and recognizes ICANN's legal obligations in processing and publishing data containing personal elements but cautions against withholding personal data to the point of rendering the data worthless. The approach of anonymizing data may be called for if even such data is NOT made publicly available and this should be applied in general.

7.5.2. In very specific cases where personal data is needed to be shared, and without which would render the data worthless to a user, then ICANN should consider placing confidentiality obligations on users who have been specifically identified and authorised to receive data containing personal elements, to do so on a limited license basis. As an example, limit sharing and use of Fellowship participant data to just the ALAC and not At-Large.

Response 68: The scope of the open data initiative is restricted to data that can be published as open data in accordance with the principles set out in Response 3, and as such the proposals in this comment are out of scope.

7.6. Conclusion

7.6.1. Thus, it would be useful if ICANN Org could assist in re-generating a list of datasets with suggestions on what downstream or upstream information can possibly be gleaned from each dataset. The ALAC believes such an exercise would assist both ICANN Org and the ICANN Community to better understand whether the range of data being collected is sufficiently complete and what related data is available to explain changes in the data, and if not, those that can and ought to be collected.

7.6.2. Once a revised list of datasets is established, it should be submitted for public comment. It is far easier to critique such a list than create it from scratch.

See **Response 54**.

No further responses.

End of report