At-Large Workspace: Open Data Initiative Datasets and Metadata

Public Comment Close	Statement Name	Status	Assignee (s)	Call for Comments Open	Call for Comments Close	Vote Open	Vote Close	Date of Submission	Staff Contact and Email	Statement Number
06 August 2018	Open Data Initiative Datasets and Metadata	ADOPTED 13Y, 0N, 0A	Alan Greenberg Justine Chew	26 July 2018	03 August 2018	06 August 2018	09 August 2018	06 August 2018	Matt Larson matt.larson@icann.org	**AL-ALAC-ST-0818-01-01-EN**

Hide the information below, please click here >>

Original Close for Comments: 27 July 2018

Extension granted until 06 August 2018

Brief Overview

Purpose:

To seek community views on the list of datasets that ICANN has available to publish and the metadata that ICANN intends to publish along with each dataset. The comments will be used to determine the priorities for publication.

Current Status:

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 RFP to choose an open data platform is almost complete and an announcement will be made by ICANN62 in Panama City.

Next Steps:

The feedback 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 feedback.

Section I: Description and Explanation

This public comment is intended to help guide the next stage of the Open Data Initiative and to provide a detailed insight for potential data consumers and other interested community members.

Over the last few months a Data Asset Inventory has been created within ICANN org. This lists the preliminary set of datasets that ICANN holds along with associated attributes, such as the system of record and data format. It is the intention that every dataset on this list that can be published as open data, will be published over time. That will be a lengthy and complex process and could take some time to complete. Consequently, we are seeking feedback that will help us prioritize the publication of datasets. The Data Asset Inventory is available in both CSV and PDF formats as detailed in section III.

In addition, we have defined a metadata vocabulary for the metadata that will be published alongside the data and we are seeking feedback on this metadata vocabulary. This vocabulary is detailed in section IV below and is available in both CSV and PDF formats as details in section III.

The specific questions we seek feedback on are as follows:

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

   The next major stage in the Open Data Initiative is the lengthy process of publishing datasets on the upcoming open data platform. This stage could take some time, so it is important that community priorities are taken into account and the highest priority data are released first.

   The datasets within ICANN are stored in a variety of formats in a variety of systems of record. In many cases, custom code will need to be developed by ICANN staff to publish data, with any required redaction or aggregation applied. A process which can vary widely in both time and cost. Accordingly, we do not intend to translate community priorities directly into a prioritized list for publication and will instead use a prioritization model that combines ease of publication and community priority.
2. *Are there any errors or omissions in the data asset inventory?*

   Creating an inventory of datasets is complex process as this is a cutting-edge subject and staff of ICANN org, as with many other organizations, are still learning what makes a dataset. There is therefore the distinct possibility that there may be errors or omissions in the inventory and for that reason we seek feedback on the inventory as provided.

   Even if you do not know whether or not ICANN holds a specific dataset but you would like see that dataset published by ICANN then please let us know.
3. *Does the proposed metadata vocabulary meet your needs?*

   The metadata vocabulary is based on the Project Open Data Metadata Schema v1.1 with minor amendments. We have chosen this standard over other standards such as DCAT, due to its simplicity, greater applicability and ease of processing. This choice does not preclude us later adding additional metadata schemes to our published open data.

Section II: Background

The term "Open Data" has a very precise meaning. Data are open if anyone is free to use, re-use or redistribute them, subject at most to measures that preserve provenance and openness. There are three dimensions of data openness:

1. The data must be legally open, which means they must be placed in the public domain or under liberal terms of use with minimal restrictions.
2. Data must be technically open, which means they must be published in electronic formats that are machine readable and non-proprietary, so that anyone can access and use the data using common, freely available software tools.
3. Data must be practically open, which means they must be publicly available and accessible on a public server, without password or firewall restrictions.

Open data in this context means specifically tabular data – that is, data that would normally be stored in a spreadsheet, data file, or database. Data might also be stored in a more structured data format such as JSON or XML. To be clear, what will not go into the open data platform is information in a broader sense – e.g., policy documents, application forms, or email messages. Instead, these documents are part of the Information Transparency Initiative.

The Open Data Initiative is the program within ICANN to identify and publish all datasets as open data that do not have restrictions that require them to be confidential.

The Open Data Initiative is a personal goal of the President and CEO, who has written a blog post setting out his reasons for promoting open data, his expectations for how ICANNN will deliver this, and his vision of how open data will benefit ICANNN org and the ICANNN community.

There are a number of different initiatives underway related to the publication of data, which need disambiguating:

- The Open Data Initiative, which this public comment relates to and which is explained above.
- The Information Transparency Initiative, which relates to simplifying access to documents and other non-tabular information. This initiative is independent of and complementary to the Open Data Initiative.
- The Identifier Technology Health Identifier (ITHI), which is a project gathering specific data relating to the health of identifier technologies. The output of this project will be published part of the Open Data Initiative.
- Domain Name Marketplace Indicators, which is a project gathering specific data relating to the domain name marketplace. The output of this project will be published as part of the Open Data Initiative.

As noted above, the open data will be published on an upcoming Open Data Platform. ICANNhas been undertaking an RFP for several months seeking a Software-as-a-Service (SaaS) open data product from an established vendor. This RFP used a detailed list of requirements and a thorough process of assessment including presentations and demonstrations with multiple ICANN staff and contractors involved. This process is now in the final stage of deliberations and negotiations with a chosen vendor.

Section III: Relevant Resources

The public extract of the Data Asset Inventory is available in CSV format at https://www.icann.org/en/system/files/files/odi-data-asset-inventory-spreadsheet-11jun18-en.csvand in PDF format at https://www.icann.org/en/system/files/files/odi-data-asset-inventory-11jun18-en.pdf.

The metadata vocabulary as referred to in the Metadata Standard is available in CSV format at https://www.icann.org/en/system/files/files/odi-metadata-vocabulary-spreadsheet-11jun18-en.csvand in PDF format at https://www.icann.org/en/system/files/files/odi-metadata-vocabulary-11jun18-en.pdf.

Section IV: Additional Information

The ICANN metadata standard follows the Project Open Data Metadata Schema V1.1 and consists of a metadata vocabulary with the following customisations:

1. without the USG tagged fields
2. Without some fields that are not needed
    a. rights - all our data is public
    b. temporal - to avoid confusion between this update and the main dataset
    c. distribution - refers to URLs the open data platform should generate
    d. issued - too complex to determine when a dataset was first published
3. With values specified for some fields - publisher, accessLevel, license
4. With custom restrictions on some fields
5. With different requirements for which fields must be present

A link to the metadata vocabulary is given in section III.

Section V: Reports

FINAL VERSION SUBMITTED (IF RATIFIED)

The final version to be submitted, if the draft is ratified, will be placed here by upon completion of the vote.

FINAL DRAFT VERSION TO BE VOTED UPON BY THE ALAC

The final draft version to be voted upon by the ALAC will be placed here before the vote is to begin.

This draft posted by Alan Greenberg, 5 August 2018, 09:44 UTC-4. it is a minor modification of the comment drafted by Justine Chew which intern was based on an original draft by Alan Greenberg plus many further comments.

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.

Centralized, easy access to properly organized data repository

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.

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.

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.

Types and value of data collected, lack of discernable information

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.**

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.

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.

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.

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

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:-

- Contractual compliance: measurements of corresponding action taken, time taken to resolve, patterns of non-compliance, plausible trigger events/reasons for non-compliance
- CCT-related complaints: types of complaints, time to resolve, patterns of domain name abuse etc, plausible trigger events
- Fellowship programme: participation metrics of returning fellows versus first-time or one-time fellows, transition from fellows to active community membership
- Membership, related to ALS and individual members:
    - diversity metrics of by country, region, gender, economy, disability status etc,
    - participation metrics in At-Large in policy development, education & outreach activities, direct & remote participation in meetings
    - travel-related metrics such as difficulties in obtaining travel support, visas, difficulties with Travel Constituency etc.

Uniformity of and responsibility for data

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").

In this context, some preliminary questions arise:

- For the data that already exists on the web, are there conceivably duplicates of data residing in separate repositories?
- 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?
- For the purposes of the open data platform, will ICANN Org be querying data in the ALAC's repositories?

Privacy rights

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.

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.

Conclusion

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.

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.

DRAFT SUBMITTED FOR DISCUSSION

The first draft submitted will be placed here before the call for comments begins. The Draft should be preceded by the name of the person submitting the draft and the date/time. If, during the discussion, the draft is revised, the older version(S) should be left in place and the new version along with a header line identifying the drafter and date/time should be placed above the older version(s), separated by a Horizontal Rule (available + Insert More Content control).

Posted by Alan Greenberg, 26 July 2018\

The ALAC appreciates the opportunity to comment on ICANN's Open Data Initiative.

Although a number of the data assets are of interest to At-Large, with 231 entries in the list, it is difficult to identify those of most interest to our group. Certain easy access to ICANN meeting demographics and the data associated with At-Large are high on our list. The ALAC suggests that once initial priorities are established, this ordered list be submitted for public comment. It is far easier to critique such a list that create it from scratch.

Perhaps of more importance is the data that is not identified here. The most obvious gap (unless it is there under a cryptic name) is 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.

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 suceeded plus those who did not). This could provide critical data on to what extent information about the Fellowship is reaching certain parts of the world.

Understanding the methodology of how the data will be accumulated is of interest. For the data that already exists on the web, will:

- The new data be derived (scraped from the web); or
- The web data will be constructed from the data tables; or
- The two reside independently.

If the latter, how will ICANN ensure that the two stay in sync with each other?

Lastly, of great interest are the tools that will be made available for the ICANN community to use to extract and process the data. The utility of the entire project will greatly hinge on the availability and capabilities of such tools.