AT-LARGE ADVISORY COMMITTEE
ALAC Statement on Open Data Initiative Datasets and Metadata

Introduction

Alan Greenberg, At-Large Advisory Committee (ALAC) Chair, and Justine Chew, member of the Asian, Australasian and Pacific Islands Regional At-Large Organization (APRALO), developed an initial draft of the Statement on behalf of the ALAC.

On 26 July 2018, the first draft of the Statement was posted on its At-Large workspace. On the same day, ICANN Policy Staff in support of the At-Large Community sent a Call for Comments on the Statement to the At-Large Community via the ALAC Work mailing list.

On 05 August 2018, an updated draft of the Statement was posted to the same workspace. This draft posted by Alan Greenberg was a minor modification of the comment drafted by Justine Chew, which in turn was based on an original draft by Alan Greenberg plus many further comments from the At-Large community.

On 06 August 2018, the ALAC Chair submitted comment, and requested that Staff open an ALAC ratification vote.

In the interest of time, the ALAC Chair requested that the Statement be transmitted to the ICANN public comment process, copying the ICANN Staff member responsible for this topic, with a note that the Statement is pending ALAC ratification.

ALAC Statement on Open Data Initiative Datasets and Metadata

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.

<u>Centralized, easy access to properly organized data repository</u>

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.

<u>Types and value of data collected, lack of discernable information</u>

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.