IETF Data

We retain a good amount of (public) data about various aspects of the IETF, but we do not have a systematic or strategic approach to the data we collect, analyze, and publish about many of these aspects. The data that we do have is siloed in various places — the datatracker (including stats — https://datatracker.ietf.org/stats/), the mailing list archives (https://mailarchive.ietf.org/arch/), the meeting registration system (which powers https://www7.ietf.org/about/administration/reports/), the meeting survey results (https://www.ietf.org/how/meetings/admin/meeting-surveys/), the RFC database (https://www.rfc-editor.org/), and the financial statements (https://www.ietf.org/about/administration/financial-statements/), among others. This makes answering simple questions about the IETF difficult.

We have toyed with the idea on and off of defining some key metrics that it would be useful to be able to track over time, but we’ve been waiting to move forward with this until the ED is in place.

Just to give an example:

Perhaps the most pressing piece of foundational information that we are missing about the IETF is how many people participate. This is an important piece of information because it helps us size all of our efforts (meetings, RFC editing resources, investment in tools, etc.) and because evaluating this metric over time could inform (but of course not fully define) our perceptions about the health and relevance of the IETF.

There are potentially many questions embedded in this single question, for example:

1. Can we determine how many people participated in the IETF in a given year?
   a. How many people sent email to at least one IETF mailing list?
   b. How many people authored at least one draft?
   c. How many people published at least one RFC?
   d. How many people remotely attended at least one meeting?
   e. How many people attended at least one meeting?
   f. How many people attended at least one hackathon?

2. Can we characterize participation?
   a. How many people did at least one of 1(a-d) and at least one of 1(e-f)? [Written and in-person engagement]
   b. How many people did only 1(a) and not any of 1(b-f)? [Email engagement only]
   c. How many people did one or more of 1(a-d) but none of 1(e-f)? [Remote engagement only]
   c. How many people did only 1(e or f) and not any of 1(a-d)? [In-person engagement only]
The major hurdle is of course that an individual person is not necessarily consistently represented across all of the data sources: people use different email addresses and names for themselves within and across all of these data sources. It seems like a threshold question is whether a methodology for correlating identifiers to people could be developed to give sufficient confidence in the accuracy of the data produced.