Observations about IETF process measurements

Jari Arkko
Ericsson Research
jari.arkko@piuha.net

1. Introduction

This paper discusses the author's experiences in designing statistics collection tools for various IETF processes [1,2].

The paper focuses largely on making data available, rather than attempting to use it for some decision making, such as determining if there's a problem or improvement opportunity in some part of the process. There's certainly a lot of interesting work in that area but that is not the focus of this paper.

The uses are important, though, and they affect what data should be made available and in which way it should be visualized. The author has found that there's a fair amount of demand for various uses, e.g., authors are interested in how their publications are referenced or about getting an easy listing of what documents they have, companies are interested in how their efforts and topics compare to what other companies are doing, for diversity it is interesting to understand geographical split of various levels of activities, same for gender distribution, and so on. Many of the author's current tools have been built due to specific user requests, or similar interests that the author himself had. Without a grass-roots demand from those who want something, it is more difficult to develop tooling. Hence, an organizational, top-down fashion of development is not easy. Or this may be the case, if not sufficient attention is given to interacting with the user base. Unfortunately, such interaction may not always be easy, as the prospective users may not be part of IETF or LLC development processes, IAB workshops, or even respond to polls or questions when something new is proposed for development [4].

The rest of this paper is organized as follows. Section 2 discusses different architectural approaches to collecting statistics, and Section 3 discusses some challenges that the author has experienced.

2. Approaches

Measuring processes that are largely human processes may not be easy. In many cases, there is no formalism within the process, or if there is (such as in the IETF Data Tracker's database objects), it may not address all the information needs that a statistics collection might wish to have. For instance, an RFC may exist in the system, but whether its authors are properly registered in the database may depend on when the RFC was published.

And even when the authors are registered, the system may not know who they worked for at the time, where they live, or what gender they are. Not all information is synchronized and cross-referenceable. For instance, IETF mailing list discussions often employ different e-mail addresses than those listed in IETF RFCs or the Data Tracker. And of course, there should be no requirement for a rigid alignment and tracking.

Indeed, some information may be such that we don't want the system or the public to be aware of; some information may legitimately be private information.

And even when there is information, it may be incorrect or have mistakes – including some cases where the author's own name was misspelled in a specification. One can hope that it was perhaps a co-author that made this error.

Nevertheless, the author of this paper argues that that even with informal processes, there is a lot of opportunities for understanding what is going on in the process. The next section discusses some technical approaches to coming to that understanding.

2.1 Data improvement, heuristic, heuristic-assisted-with-human approaches

There are three fundamental approaches to dealing with such imprecise, partial, and unreliable information.

2.1.1 Manual improvement of information

The first approach involves careful manual improvement of the available information. In the case of the IETF, this typically involves adding information to the official IETF data bases, such as the Data Tracker, RFC Editor's databases, and so on. For instance, new data may be added to the official IETF data bases to understand some aspect of the IETF process, participants, or documents more fully.

Such additions are feasible in many cases, but not so easy in other ones. For instance, the number of documents that the IETF produces is so small that one could contract someone to go through all of them and look at some aspect, such as author names or contributors or what type of a document it is. The documents also have a lot of history associated with them, and people often the involved people could be asked about unclear cases. But while documents are a relatively constrained set of information, all email on IETF lists is not, for instance. Or comments spoken at meetings. Let alone hallway interactions.

It should also be observed that even when the careful improvement process is feasible, some uncertainties may remain. For instance, the author names or affiliations associated with a particular IETF document are not always recognized by the tools that the author of this paper implemented, because the data was not there, there was some ambiguity, or no easy information beyond the RFC was available. Any manual data improvement process may run into similar issues.

Typically, when a decision is made to collect additional information, it may also only start at given time and not cover earlier information. In many cases going back to past time may even be impossible, e.g., not all previous IETF meeting attendees might be reachable or willing to respond to queries.

2.1.2 Heuristic derivation of information

The second approach is a more heuristic one. In this approach the available hard information (e.g., which RFCs we have) is complemented with a heuristic process to try to make a best guess about something that only exists in informal manner or is expressed in non-machine-readable human text. It might perhaps be possible to use "machine intelligence" processes for something like this as well – this would be an exciting research project for future improvements in IETF process analysis.

2.1.3 Heuristics augmented with human assistance

A third approach is one where an otherwise heuristic process prompts the operator/designer of the tooling to provide some information that the heuristic process unable to determine. That is, a human act as a last resort information finder, perhaps periodically checking "errors" or "unrecognized objects" that the tools report. This is the choice that the author of this paper implemented for his tools.

For instance, while most RFCs have a section titled "Author's Addresses", sometimes its title is different, or it is formatted differently. Or, again, misspelled! The system typically knows when it is unable to find some information. In particular, it can be useful to tabulate and highlight frequently occurring issues that warrant some action. If a thousand RFCs are missing their author section, that's probably a problem in the heuristics rather than the RFCs. Additional rules can then be added to cover, for instance, "Author's Adresses" [8], or "Authors' Information" or "Editor's Address". Or even be able to correct the information in a single RFC that fails to be properly understood, e.g., due to mixing free-form text and semi-formal aspects such as author lists.

2.1.4 End results

These data improvement, heuristic, heuristic-augmented-with-human-assistance approaches all must live with imperfections in the result. This may, however, be fine for many or most use cases. All of them have an initial cost in building the solution, and all but the heuristic approach has a significant ongoing maintenance cost that scales with the amount of activity tracked. This cost may or may not be feasible in different cases.

2.2 Visualization

Data that exists may also be visualized. Traditionally, statistics are visualized as graphs, percentages, time series or similar.

But it is important to observe that the same data can be used for very different purposes. For instance, data about the progression of a specification through a standards organization can be used to show:

- The progress and status for a single work item. This can be useful to understand what happened with a particular process.
- The progress for a group or area within the organization, or even the whole organization. This can be useful to understand where time is spent, what parts of an organization need improvements, understand bias, etc.
- The status of different topics that a given person or company is working on. This can be useful to act as a "dashboard" to see what action a particular person can take next.

3. Challenges

Several challenges have made work with statistics and data collection harder.

3.1 Improved privacy

Keeping people's information private is very important. Heightened attention to privacy has, fortunately, been a trend in the last years. For instance:

- Legal frameworks have been introduced that set rules for processes involving collection of information, such as GDPR [5].
- There has been increased interest in understanding factors that may be more privacy sensitive than previously considered factors. One example of this is gender.
- Corporate policies that reflect the interest of the corporations to keep personal data safe.
- Privacy-sensitive measurement and statistical approaches.

This obviously complicates and limits information collection and presentation. Of course, as an open organization IETF necessarily has some visible information – for instance, email threads and meeting discussions are public, documents have author names and affiliations, and so on. The author's approach to tooling has been to present available public information in more readily available form – such as being able to process informal data into statistics – but not to provide any new information.

Nevertheless, statistics tools should avoid unnecessarily exposing information, even when that can be determined through analysis processes. For instance, the author's tools have evolved in the 2010's such that the location and affiliation data is no longer produced, except as an aggregate calculation. Of course, what constitutes necessary information, or a privacy concern is in some cases a matter of opinion. An organization may decide that collecting additional information from its participants (with permission) is desirable.

It is also necessary to ensure that there are mechanisms to correct information available through statistics tools, that the people who appear in the results know why their information is there, where information has been collected, and what they can do if they wish their information to be removed. For these purposes a privacy policy and contact mechanisms are necessary. In the case of the author's tools the privacy policy is listed in [3], and the has been a small number of cases where information was requested to be removed, which the author complied with.

The new policy also interacts with the heuristics-assisted-with-human approach. In the author's tools, for instance, a new policy is that any extra information entered that relates to a person needs to come from the persons themselves. This has proven to be a workable model.

3.1.1 Avoiding collecting new data

Sometimes it is possible to avoid collecting some information, even if it is interesting from a research or organizational perspective. For instance, the author has decided against using any gender-related data in his tools. Even the IETF itself only collects gender data in limited situations, for meeting attendance, and only from participants willing to provide that information.

Avoiding the collection or use of some information obviously makes it harder to provide some types of statistics, for instance. But it does not entirely block any analysis in all cases. For instance, aggregate gender data may still be provided in a pool of named persons, using heuristic methods, with a downside of not being able to determine the correct results in all cases. As an example, best popular tools for classifying names to gender have an approximate 15% failure rate (that includes non-classification) [6].

3.2 Evolving interfaces and formats

The author's tools have been in existence since late 1990's, when they were written as a temporary hack. They are still a hack, but in those 25 years the IETF's formats for documents, information in the Data Tracker, available APIs, and processes have evolved greatly. Constant maintenance is needed to enable the use of information sources. The maintenance burden is particularly hard on data that is "screen-scraped" and frequent changes due to formatting issues.

Anecdote: the most troublesome screen scraping part of the author's tools has been the use of population data from Wikipedia to calculate how many RFCs per capita are produced in different countries. The format of the Wikipedia page appears to be in constant flux. Indeed, that part of the tools is currently (again) broken [7]. The design decision to base the calculations on a public web page was a monumentally bad one, even if it can be argued that there would have been some changes to any suitable machine-readable database on the matter as well, in those 25 years.

3.3 Evolving data sources

Besides interfaces and formats, the actual data sources tend to evolve as well. For instance, at the time the author's tools were started, there were limited programmable APIs, so documents were a primary source of information. In the intervening years, the IETF Data Tracker had started to store much information that can be directly queried by programs. But information that only existed in IETF mailing lists has in part migrated to GitHub issue lists and updates, various instant messaging platforms, and so on. These changes will continue.

4. References

[1] IETF document statistics, J. Arkko. At https://www.arkko.com/tools/docstats.html.
[2] Authorstats, J. Arkko. Description at https://www.arkko.com/tools/authorstats.html.
[3] Privacy policy, J. Arkko. Available at https://www.arkko.com/tools/stats-privacy.html.
[4] Statement of Work for Extensions to the IETF Datatracker for Author Statistics. R. Housley. RFC 7760 . Available at https://datatracker.ietf.org/doc/html/rfc7760.
[5] DIRECTIVE 95/46/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. EC. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31995L0046&rid=5.
[6] Comparison and benchmark of name-togender inference services. L. Santamaria and H. Mihaljević. See https://peerj.com/articles/cs-156.pdf.
[7] Distribution of RFCs According to the Countries of their Authors, per Capita. J. Arkko. See https://www.arkko.com/tools/rfcstats/d-countryeudistrcap.html.
[8] Definitions of Managed Objects for Bridges with Traffic Classes, Multicast Filtering, and Virtual LAN Extensions. RFC 4364. Available at https://datatracker.ietf.org/doc/html/rfc4363.