

Internet Engineering Task Force (IETF)
Request for Comments: 7760
Category: Informational
ISSN: 2070-1721

R. Housley
Vigil Security
January 2016

Statement of Work for Extensions to the IETF Datatracker for Author Statistics

Abstract

This is the Statement of Work (SOW) for extensions to the IETF Datatracker to provide statistics about RFCs and Internet-Drafts and their authors.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at <http://www.rfc-editor.org/info/rfc7760>.

Copyright Notice

Copyright (c) 2016 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1.	Introduction	2
2.	Purpose	2
3.	Statistics	3
3.1.	Documents	3
3.2.	Authors	4
3.3.	Affiliation of Authors	4
3.4.	Countries of Authors	5
3.5.	Continents of Authors	5
4.	IETF Meeting Attendees	6
4.1.	Countries of IETF Meeting Attendees	6
4.2.	Continents of IETF Meeting Attendees	6
5.	Existing Code	6
6.	Deployment	7
7.	Thoughts on Future Enhancements	7
8.	Security Considerations	8
9.	Informative References	8
	Author's Address	8

1. Introduction

A prominent member of the IETF community has developed a set of tools to produce statistics about the authors of RFCs and Internet-Drafts. These tools analyze the documents themselves to produce statistics about the documents and their authors. The goal of the IETF Datatracker enhancements described in this document is to provide similar statistics and ensure that the software is maintained as part of the IETF information services. While some data may still need to be extracted from the documents themselves, as much data as possible should be maintained in the IETF Datatracker database [DATATRACKER].

Current statistics are available on the web at
<http://www.arkko.com/tools/docstats.html>.

The code that is used to produce these statistics is available at
<http://www.arkko.com/tools/authorstats.html>.

2. Purpose

Author statistics allow the community to understand where work is being done and by whom. The statistics make it visible which individuals, companies, and geographic regions are the most active contributors. The statistics also show how these are changing over the years.

Some of the statistics provide "nice to know" information; however, others are sometimes used to refer to a particular participant's contributions in the IETF or used to study trends within IETF work. For instance, the IETF has been trying to increase the diversity of participants, and the statistics are one way to see the impact of those efforts. Also, the most active individuals are potential candidates for various leadership positions.

3. Statistics

The enhancements to the IETF Datatracker shall provide statistics and graphs about documents, document authors, author affiliation, author country, and author continent.

The statistics should also include trends relating to IETF meeting attendees, which the current tools do not track.

For the purposes of these requirements, "recent Internet-Drafts" and "recent RFCs" cover documents that have been published in the last five years.

3.1. Documents

The statistics shall provide insight into the number of authors per document. The current web page presents the statistics and a bar chart. The current web page can be seen at <http://www.arkko.com/tools/rfcstats/authdistr.html>.

The statistics shall provide insight into the size of the documents. The current web page presents the statistics and a bar chart. The current web page can be seen at <http://www.arkko.com/tools/allstats/pagedistr.html>. With the planned change in document format, some other way to measure document size might be more appropriate, such as word count.

Additionally, statistics about the document format that was used by the authors should be provided, which is not provided by the current tools.

The statistics shall provide insight into the use of various specification techniques such as ABNF, ASN.1, C code, CBOR, JSON, and XML. The current web page does not include all of these techniques. The current web page can be seen at <http://www.arkko.com/tools/allstats/formatdistr.html>.

3.2. Authors

The statistics shall provide insight into the distribution of authors according to the number of documents they have authored for recent Internet-Drafts, recent RFCs, and all RFCs. The current web pages that provide similar information include the statistics and a bar chart. The web pages are available at <http://www.arkko.com/tools/stats/authactdistr.html>, <http://www.arkko.com/tools/recrfcstats/authactdistr.html>, and <http://www.arkko.com/tools/rfcstats/authactdistr.html>.

The statistics shall provide insight into the relative impact of authors by the number of their RFCs that are cited by other RFCs. The current web page can be seen at <http://www.arkko.com/tools/rfcstats/hindextop.html>.

3.3. Affiliation of Authors

The statistics shall provide insight into the affiliation of authors for recent Internet-Drafts, recent RFCs, and all RFCs. The current web pages that provide similar information include the statistics and a bar chart. The web pages are available at <http://www.arkko.com/tools/allstats/companies.html>, <http://www.arkko.com/tools/stats/companydistr.html>, <http://www.arkko.com/tools/recrfcstats/companydistr.html>, and <http://www.arkko.com/tools/rfcstats/companydistr.html>.

The statistics shall provide insight into the way that affiliations of RFC authors have changed over the years. The current web page can be seen at <http://www.arkko.com/tools/rfcstats/companydistrhist.html>.

3.4. Countries of Authors

The statistics shall provide insight into countries of authors for recent Internet-Drafts, recent RFCs, and all RFCs. It has been useful to provide country-based statistics, and it has also been useful to provide statistics showing the European Union (EU) as a single "country" for the sake of comparison with other large countries. The current web pages that provide similar information include the statistics and a bar chart. The web pages are available at <http://www.arkko.com/tools/rfcstats/countries.html>, <http://www.arkko.com/tools/stats/d-countrydistr.html>, <http://www.arkko.com/tools/stats/d-countryeudistr.html>, <http://www.arkko.com/tools/stats/countrydistr.html>, <http://www.arkko.com/tools/stats/countryeudistr.html>, <http://www.arkko.com/tools/recrfcstats/d-countrydistr.html>, <http://www.arkko.com/tools/recrfcstats/d-countryeudistr.html>, <http://www.arkko.com/tools/recrfcstats/countrydistr.html>, <http://www.arkko.com/tools/recrfcstats/countryeudistr.html>, <http://www.arkko.com/tools/rfcstats/d-countrydistr.html>, <http://www.arkko.com/tools/rfcstats/d-countryeudistr.html>, <http://www.arkko.com/tools/rfcstats/countrydistr.html>, and <http://www.arkko.com/tools/rfcstats/countryeudistr.html>.

The statistics shall illustrate the change in number of Internet-Draft and RFC authors per country over time. The current web page can be seen at <http://www.arkko.com/tools/rfcstats/countrydistrhist.html>.

3.5. Continents of Authors

The statistics shall provide insight into continents of authors for recent Internet-Drafts, recent RFCs, and all RFCs. The current web pages that provide similar information include the statistics and a bar chart. The web pages are available at <http://www.arkko.com/tools/stats/d-contdistr.html>, <http://www.arkko.com/tools/recrfcstats/d-contdistr.html>, and <http://www.arkko.com/tools/rfcstats/d-contdistr.html>.

The statistics shall illustrate the change in number of Internet-Draft and RFC authors per continent over time. The current pages can be seen at <http://www.arkko.com/tools/rfcstats/d-contdistrhist.html>.

4. IETF Meeting Attendees

The enhancements to the IETF Datatracker shall provide statistics and graphs about the country and continent of IETF meeting participants.

4.1. Countries of IETF Meeting Attendees

The statistics shall provide insight into countries represented by attendees at each IETF meeting. Country-based statistics have been presented in the plenary session for many years. For consistency with the author statistics discussed in Section 3 of this document, the statistics will include a way to show the EU as a single "country" for the sake of comparison with other large countries. The statistics for each meeting should be accompanied with a pie chart that shows the eight countries with the highest number of attendees and "other".

The statistics shall illustrate the change in number of IETF meeting attendees per country over time. Again, for consistency with the author statistics discussed in Section 3 of this document, the statistics will include a way to show the EU as a single "country".

4.2. Continents of IETF Meeting Attendees

The statistics shall provide insight into continents represented by attendees at each IETF meeting.

The statistics shall illustrate the change in number of IETF meeting attendees per continent over time.

5. Existing Code

Since the new code will be driven by the Datatracker database to the greatest extent possible, the existing code may be of limited value. The existing code was intended as a temporary solution and requires a rewrite. However, a set of heuristics used by the code may be useful. These heuristics are provided in a separate rule database and are used as a last resort when there is otherwise too little information. The heuristics include author aliases, some recognized authors and some recognized affiliations, domain name data for determining location and affiliation, and mappings for some ways that people represent their countries in a post address.

Authors are not consistent about the way their names appear in various documents. For example, one document may include their given name and another document may include a nickname. The Datatracker

database provides a way to capture aliases, but not all of the aliases in the documents have been added to the database.

The current Datatracker database does not have tables for heuristics other than author aliases that are used in the current tool. Appropriate tables to hold the additional heuristics from the current rule database should be added to the Datatracker database in a manner agreed by the group of people that maintain the Datatracker source code.

A workable web interface, possibly using Django Admin, to update the new heuristics tables shall be provided.

The current code can be found at <http://www.arkko.com/tools/authorstats.html> and is openly available but without any warranty.

The software is split in two parts, with the code itself being separate from the heuristics database. The two main components of the code are authorstats, which produces the statistics and generates the statistics web pages, and getauthors, which performs document analysis.

6. Deployment

The current tools analyze the documents themselves to produce statistics. Some of the data needed to produce the statistics is not currently in the Datatracker database. This development effort will include adding the capability to capture this data in the Datatracker database and populate it for all RFCs and Internet-Drafts posted over the last five years. It may be cost-effective to leverage the existing code to extract the information and then verify it one time.

The URLs for the current tools exist in many places in the Web. Once a suitable replacement tool is available, the author of the original tools has promised to provide a suitable form of redirection.

7. Thoughts on Future Enhancements

With the planned change in document format, some of the information obtained through heuristics might be more directly extracted from the XML file. Once this format is being used by a significant number of authors, a future effort might move away from heuristics toward extraction from the XML file. To support this approach, it would be desirable for the new XML schema to make the author's continent of residence available, even if it not used in the formatting of the human-readable document.

Currently, the Datatracker has information about document authors, but not other contributors. If information is added to the Datatracker in the future to cover contributors, then the statistics can be expanded to cover contributors as well as authors.

8. Security Considerations

This document contains the SOW for enhancements to the IETF Datatracker to provide author statistics. These enhancements do not affect the security of the Internet. The enhancements provide statistics about documents that are available to the public without prior authentication, and the statistics will also be available to the public without prior authentication.

9. Informative References

[DATATRACKER] Internet Engineering Task Force, "IETF Datatracker", <<https://datatracker.ietf.org/>>.

Author's Address

Russ Housley
Vigil Security, LLC
918 Spring Knoll Drive
Herndon, VA 20170
United States

Email: housley@vigilsec.com