The ietfdata Library

Stephen McQuistin
University of Glasgow
sm@smcquistin.uk

Colin Perkins
University of Glasgow
csp@csperkins.org

1 ACCESSING IETF DATA

We have developed the open-source `ietfdata` library [3], a set of Python 3 libraries that enable access to various IETF data resources, including the Datatracker [1], RFC index [4], and IETF IMAP e-mail archives [2]. The library provides a set of classes that model the various entities represented in the dataset, including people, documents, meetings, and e-mail messages. This enables programmatic, Pythonic access to the wide range of data that the IETF makes available.

Importantly, `ietfdata` appropriately regulates access and makes use of caching techniques to reduce the load on the IETF's infrastructure. The library can be connected to a MongoDB database instance, which it will use to cache requests and their results so that these can subsequently be served locally. The cache is also used to store a mirror of the IETF mail archive. Once the cache has been populated for the first time, subsequent updates will retrieve only the new e-mail messages.

Our library has proved useful in supporting a wide range of research projects, including finding new drafts to test processing tools, and in enabling bibliometric analyses. As part of the workshop, we hope to work with others in the community to help integrate the `ietfdata` library with their workflows, to help simplify data access and support their ongoing research activities, and to begin to form new collaborations with others interested in mining the available data.

2 NEXT STEPS

While the `ietfdata` provides coverage for much of the Datatracker's API, there are a few gaps that remain. An immediate next step is to expand coverage to include all of the available endpoints. Beyond that, the library could be expanded to enable appropriate post-processing across different datasets (e.g., linking Datatracker data to e-mail senders).

We would encourage API access to the Datatracker to continue to be as expansive as possible. As the IETF collects more data, where possible this should be made available via the Datatracker's API. Providing access to this data with appropriate versioning and timestamping enables caching, reducing the burden on the API's infrastructure.

Further, we would like to begin a conversation on how to share datasets based on IETF data, establish guidelines and recommendations on how data about the IETF should be shared more broadly with the research community, and identify the ethics, privacy, and legal implications of doing so.

3 SUMMARY

We have made extensive use of the `ietfdata` library in our recent paper, *Characterising the IETF Through the Lens of RFC Deployment* [6]. In this work, we use the library to examine the shifts and trends within the standards development process, showing how protocol complexity and time to produce standards has increased over time and develop statistical models of factors that lead to successful uptake and deployment of protocols.

This work is an output of the UK EPSRC-funded *Streamlining Social Decision Making for Improved Internet Standards* project [5]. This project aims to develop methods to improve distributed decision making in large organisations, using a combination of natural language processing and social network analysis. While the project aims to develop tools and techniques that are widely applicable, the openness and transparency of the IETF, in documenting its processes via public e-mail archives, meeting minutes, and other metadata, makes it a good organisation upon which to develop them.

This highlights the benefits to the IETF as an organisation in being open, transparent, and expansive about the data that it makes available. Projects, such as ours, can use the available data to develop tools and techniques that are beneficial to the IETF.

4 ACKNOWLEDGEMENTS

This work is supported by the UK Engineering and Physical Sciences Research Council, under grants EP/S033564/1 and EP/S036075/1.

REFERENCES

[1] IETF Datatracker. `https://datatracker.ietf.org`.
[2] IETF Mail Archive. `https://mailarchive.ietf.org`.
[3] `ietfdata` library. `https://github.com/glasgow-ipl/ietfdata`.
[4] RFC Index. `https://www.rfc-editor.org/rfc-index.html`.
[5] Streamlining Social Decision Making for Improved Internet Standards. `https://sodestream.github.io`.
[6] Stephen McQuistin, Mladen Karan, Prashant Khare, Colin Perkins, Gareth Tyson, Matthew Purver, Patrick Healey, Waleed Iqbal, Junaid Qadir, and Ignacio Castro. Characterising the IETF Through the Lens of RFC Deployment. In *Proceedings of IMC 2021*, 2021.