The Challenge of Delivering Open OSS Data for Research

Alex S. Moura[1], Artur Ziviani[2], Leobino N. Sampaio[3], Rafael L. Gomes[4]
[1]Rede Nacional de Ensino e Pesquisa, RNP, Brazil
[2]National Laboratory of Scientific Computing, LNCC, Brazil
[3]Federal University of Bahia, UFBA, Brazil
[4]State University of Ceará, UECE, Brazil
alex@rnp.br, ziviani@lncc.br, leobino@ufba.br, rafa.lopes@uece.br

For many years, the Brazilian research community has been requesting access for more and diverse datasets from the Brazilian Research and Education Network (*Rede Nacional de Ensino e Pesquisa - RNP*), which introduced the internet in Brazil in 1989. Since 2018, the Monitoring Technical Committee (*Comitê Técnico de Monitoramento*, CT-Mon), coordinated by RNP, developed a poll[1] to query researchers to answer the following question: Where should resources be invested that can generate the greatest benefits for network research? The resulting study produced a rich feedback with contributions from another RNP R&D working group named IpeAnalytics which mapped different kinds of data, tools and the restrictions, barriers and challenges at RNP to gather Operations Support Systems (OSS) network measurements data from RNP's network. The positive outcome granted CT-Mon to receive funds to publish, in 2020, an open call for proposals, which selected the MicroMon Working Group to develop a platform with automations to gather OSS network measurements from diverse sources and tools, organize, anonymize whenever needed, and deliver useful datasets for researchers. We have seen other initiatives for sharing measurement data, proposed by the Internet2 Community Metrics and Telemetry Project[2], with the intent to improve interdomain support for the core missions of the Universities, Regionals, Backbones, and research laboratories.

Some other initiatives in the context of the RNP's efforts are worth mentioning herein. The Global Network Advancement Group (GNA-G) is a global partnership with participants from R&E networks worldwide, with the goal to better align their collective resources and make the country-to-country interconnections more efficient for global science collaborations and transnational education. The Research Data Alliance (RDA) is an international organization focused on the development of infrastructure and community activities that reduce barriers for data sharing and exchange, and the acceleration of data driven innovation worldwide. RDA includes data science professionals from multiple disciplines that are building the social and technical bridges that enable open sharing and reuse of data, following and promoting principles like those of the F.A.I.R. (Findable, Accessible, Interoperability and Reusability) model. Ideally, joining forces among organizations and initiatives like NSF, CAIDA, GNA-G, IRTF, IETF, RDA, RIPE and others could leverage knowledge and resources to develop standards for data collection based on body of knowledge for best practices, operational procedures and platforms capable of storing, organizing, and sharing data sets. The results would become an invaluable contribution to the global research community providing global OSS network measurements at a large scale and production-grade level, which would leverage novel research to allow network operators to evolve their networks. Despite desirable, this approach would be a too large effort, and very slow to converge. Therefore, more attainable approaches are needed. Hence, it would be essential to discuss what would allow secure, fast storing and sharing of large data sets for research in areas like performance and security. There are both technical and non-technical points worth addressing. Some of the challenges are very high speed network and high precision measurement data acquisition, long term and accessible storage for large volume network flows, how to define useful data filtering strategies, how to ensure data sharing in compliance with local laws and regulations like GDPR, and how to ensure compliance and anonymization that will not compromise data usefulness. Regarding the potential value of data, how to define an interaction model so that the sharers have an interest (and incentive) to share? Given that it is somehow possible to establish value for the data, maybe NSF can model some "cashback style" approach that may incentive data sharing? Recognizing the importance of those who share data is crucial and establishing incentive mechanisms to allow the recognition in an organized way is important. Examples of such are the recent possibility of data papers that allow attribution of a DOI for a document describing a companion dataset that in turn allows an organized and accountable citation of the dataset by those who use it for further research, thus enabling the recognition of the data sharer by the number of citations it gets.

The National Academies of Engineering Science and Medicine (NASEM) published a report[3] about the benefits of Reproducibility and Replicability in Science and main global research funding institutions like Wellcome Trust, Bill and Melinda Gates Foundation and National Institutes of Health (NIH) started to demand the elaboration of a Data Management Plan when submitting research projects, in addition to requiring the publication, in open access, of the data set underlying the articles resulting from the funded research. NSF could start a discussion on how to promote open data sharing of some part of network OSS aggregated data that would not compromise operations security to leverage network research. At RNP, the top five data of interest for researchers were (1) active network performance measurements, (2) NetFlow traces with raw data truncated at first 200 bytes, (3) infrastructure monitoring – CPU, RAM, Cache, I/O, IRQ – of servers, routers and services, (4) detailed complete physical network topology with nodes and edges and (5) equipment configurations. Other network information of interest were circuit utilization (volume of bits and packets) and routing tables from each router. There are no open data standards, neither standard procedures as guidelines for sharing some kinds of network information like those, and this could be another area where NSF could help make improvements.

Researchers can develop and validate new R&D faster with access to the quality datasets in a timely fashion. This could be a game changer for future research. RNP's CT-Mon is interested in the concepts that will be explored at this workshop because we support Latin America's largest national R&E network, multiple research testbeds and a broad variety of scientific communities and initiatives, supporting public policies, and providing support to network science and empirical, applied network research, enabling support for national and international networking research. The topics outlined for this workshop constitute key areas that will assist in determining next generation data infrastructures and services and evolve internet metrics analysis that will support its evolution.

[1] Open Datasets for Research in Academic - http://bit.ly/2Xm8bL2
[2] Internet2 Community Metrics and Telemetry Project: http://bit.ly/38oburK
[3] Reproducibility and Replicability in Science: https://www.nap.edu/read/25303/chapter/1