Making OSS Network Data Available to Network Researchers

alex@rnp.br

Nov 10th 2020 - Géant Telemetry and Big Data Workshop

The Problem

Why researchers need data from production networks?

To validate and develop new research in computer network field, scientists need to build networks and perform experiments and analysis. This can be done:

- using **laboratories and/or *testbeds*** with physical network hardware;
- using **simulators** - Ex.: ns3 or mininet etc;
- using **emulators** - Ex.: Cisco Packet Tracer, Cisco VIRL, Juniper vLabs, NRELabs, GNS3, EVE-NG etc.

- None of the options above can reproduce with fidelity what happens in a network operating "in production"
- Some kinds of research in areas like performance, high availability, security and traffic engineering require real data captured for analysis, hypothesis validations and production of new insights and knowledge
- Networks are rich sources of data for scientific research
- Some network data can be opened without compromising its own security or violating current laws or individual rights, when handled adequately

Motivation

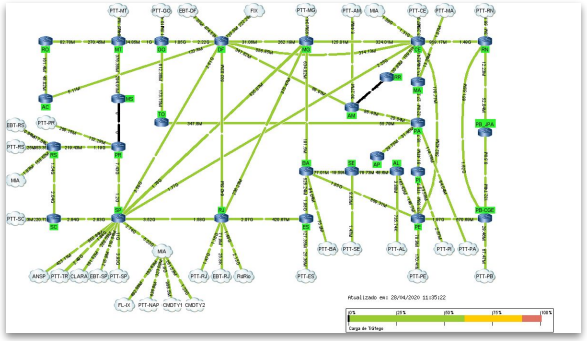
- The **RNP Evaluation Commission (CAA)** made a request, which was backed by the researcher Edmundo Souza e Silva, to have measurements data from the network available to serve as input for new scientific research projects in computer networking field
- Since 2018 there were significant advances towards meeting this demand at RNP:
  - The Monitoring Technical Commitee (CT-Mon) conducted a survey among some of the community members to identify datasets of most interest for them and for all the community
  - Network performance measurement data from the backbone were posted in the first RNP open data repository as a result of the efforts of the Research Data Network Repository Work Group (GT-RDP)*
- However, the CAA recommendations have been hampered by the lack of internal processes and an adequate infrastructure for the collection of data in an agile way. Today, if a researcher requests data about the RNP network, this requires manual data collection and long deadlines to answer each request received.
- To solve this problem, considering the recommendations of both CAA and CT-Mon, we proposed a new project and opened a public call in 2020, aiming to develop a solution that can, in an automated workflow, ingest OSS data, organize, anonymize (when needed), and publish datasets for researchers following FAIR principles and open standards. The selected project was the MicroMon Work Group

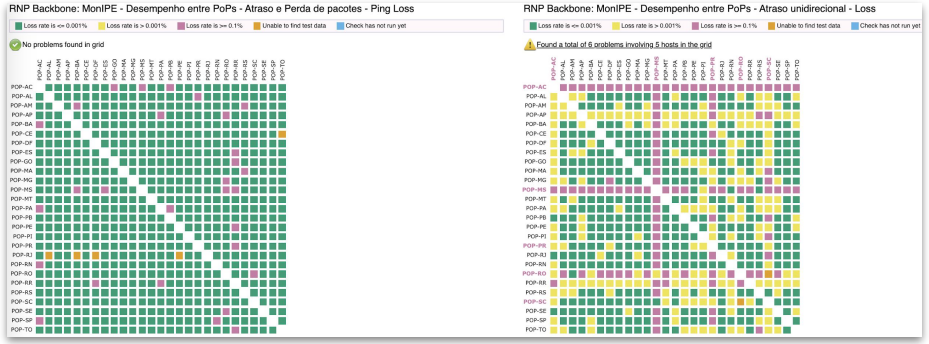
* https://dadosabertos.rnp.br/dataverse/redeipe

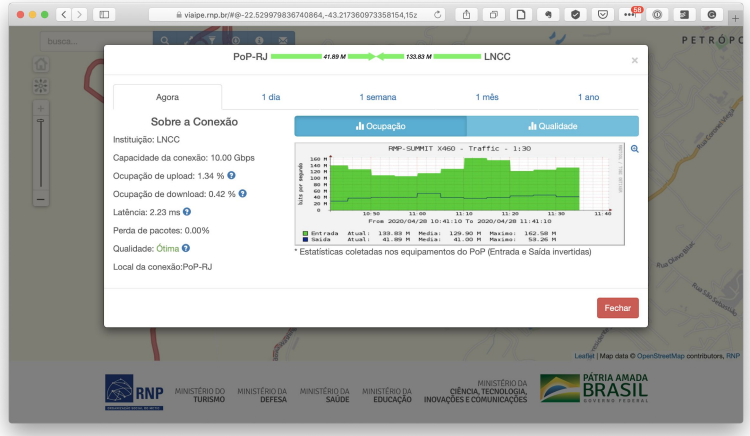
Network data

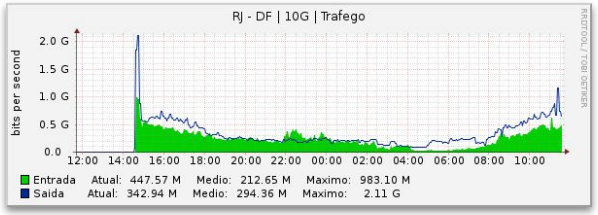
CT-Mon has verified that RNP does have several measurement and visualization tools but lacks internal processes, resources and integrations to collect, organize and share data sources of interest for researchers

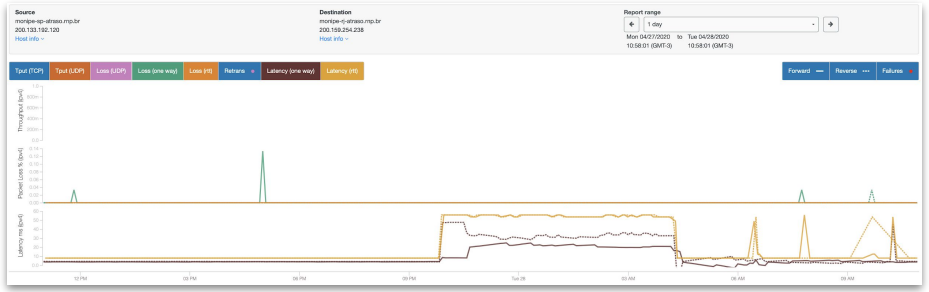
- Public visualization tools of network utilization, performance
- Data sources not available

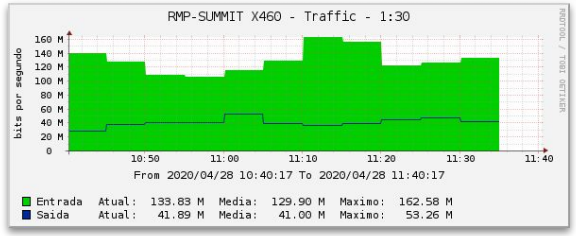












Data requested by researchers

Surveys carried out between 2018 and 2020

- The potential research interests are themselves broad
- Some may lie with classic OSS data, e.g., applying machine learning to predict component failures based on observed data, some may wish to explore specific details of routing protocols or routing policies, and others perhaps data related to network flows and network performance

NETWORK DATA OF INTEREST FOR RESEARCH COMMUNITY (2018)

- Active performance measurements
- NetFlow traces with raw data (truncated at 200 bytes)
- Monitoring of infrastructure (CPU, RAM, buffers, cache, I/O, IRQ) of servers, routers and services
- Details of the complete physical topology (nodes and edges)
- Router configurations
- Experimental service logs
- Utilization of each circuit and PoP
- Routing tables for each router
- Service tickets

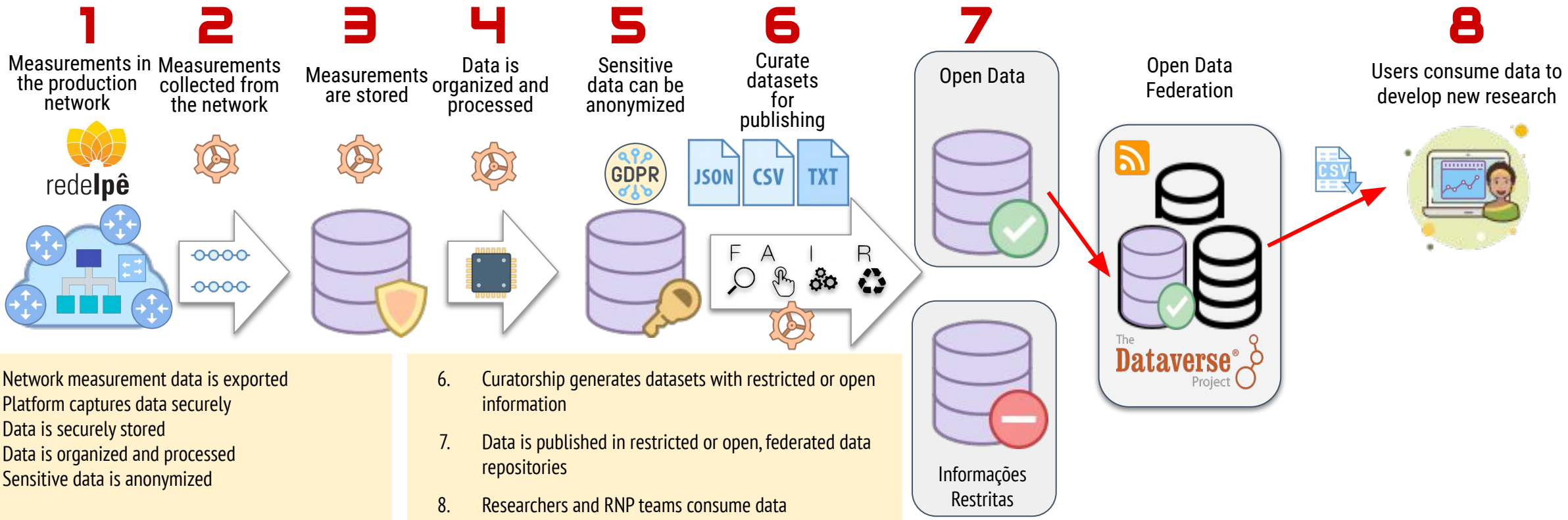
MEASUREMENTS FOR ANALYSIS OF COVID-19 IMPACTS (2020)

- Internal traffic level on the network
- Traffic level at the points of exchange of traffic with peers like commercial networks
- Application and services measurements (e.g. web conferencing, cloud services)
- Traffic and performance matrices from the network and peerings
- VPN traffic from neighbor ASNs to NREN client institutions
- Traffic from specific customers (e.g.: University Hospitals, HPC Centers)
- Dump of BGP updates received by NREN edge routers
- Periodic dump of routing tables from edge routers (before BGP best-path selection). It allows checking the occurrence of changes in routes concomitant with changes in traffic patterns.
- Periodic dump of IGP routing protocols messages

Project Overview

Project Objectives: Automated Network Data for Research

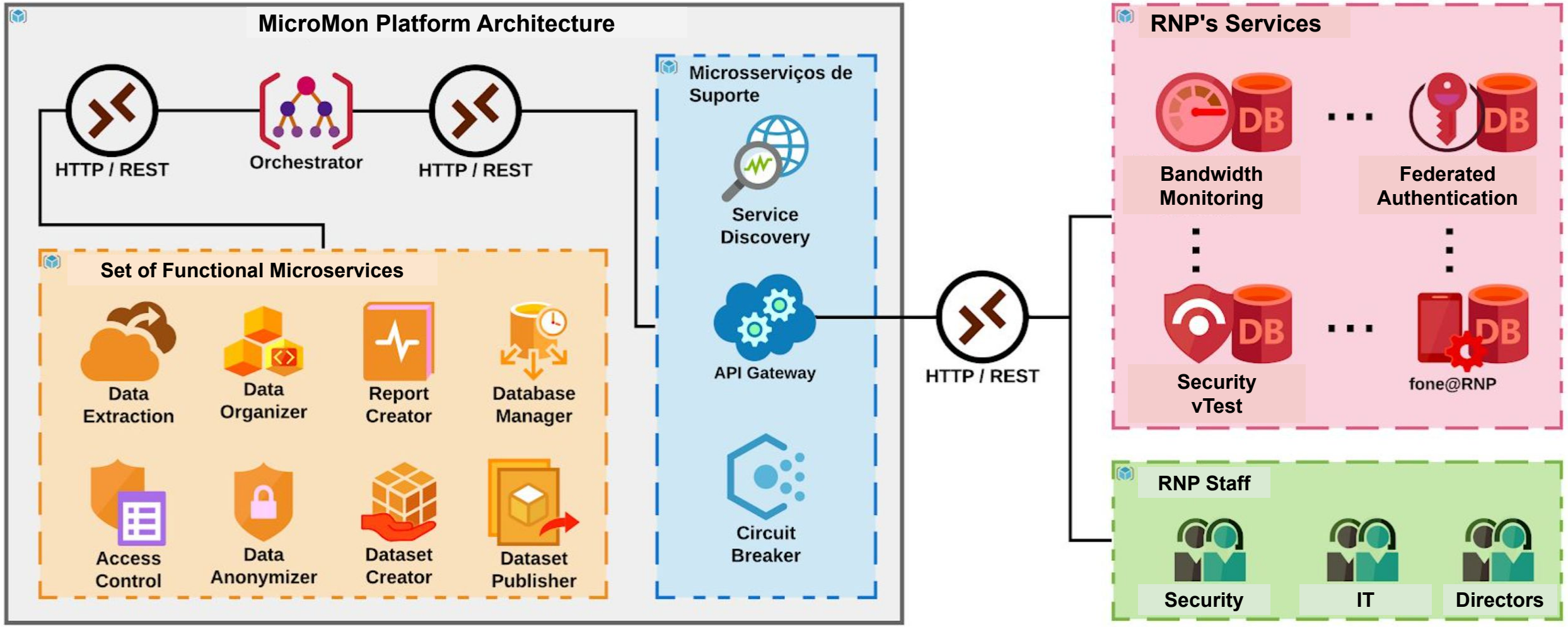
- Develop and deploy a new internal process and solution capable of collect, store, organize, anonymize and share - in an automated fashion, OSS data with both the research community and RNP's teams



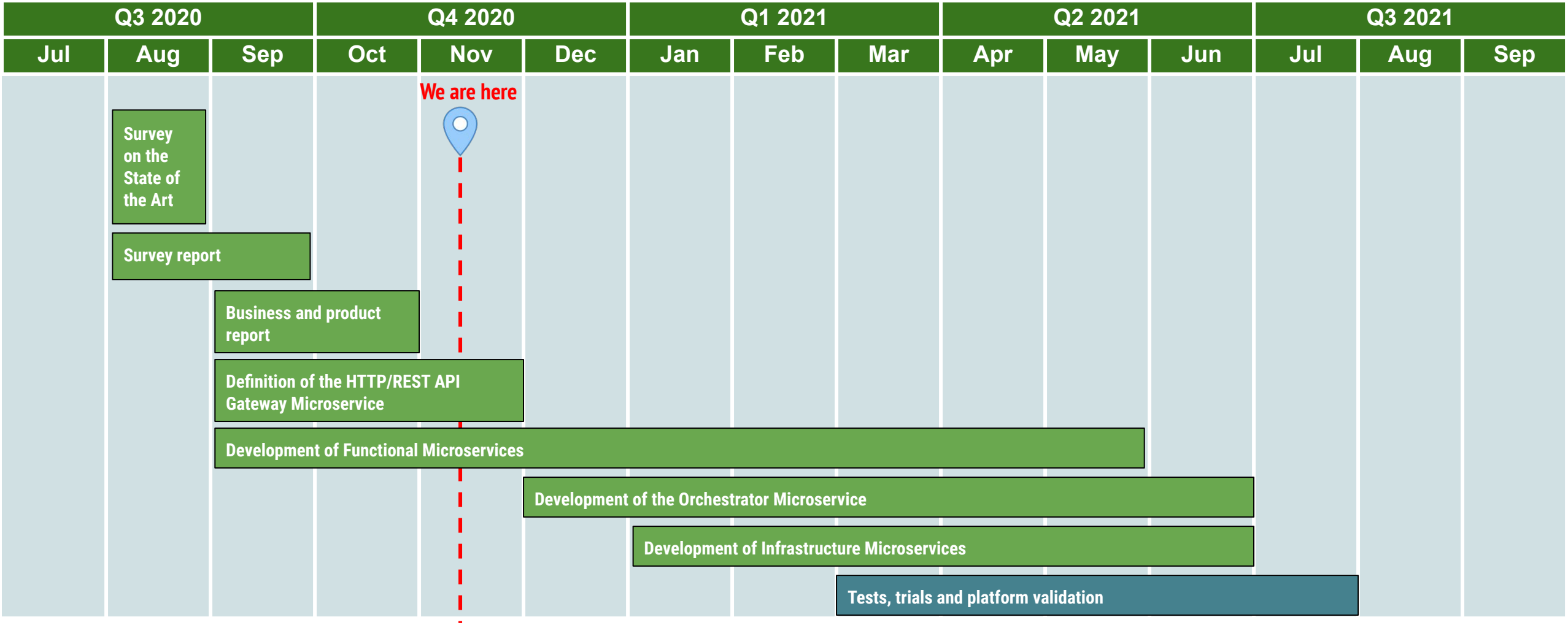
1. Network measurement data is exported
2. Platform captures data securely
3. Data is securely stored
4. Data is organized and processed
5. Sensitive data is anonymized

6. Curatorship generates datasets with restricted or open information
7. Data is published in restricted or open, federated data repositories
8. Researchers and RNP teams consume data

Architecture



Roadmap


Q3 2020
Q4 2020
Q1 2021
Q2 2021
Q3 2021
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
We are here
Survey on the State of the Art
Survey report
Business and product report
Definition of the HTTP/REST API Gateway Microservice
Development of Functional Microservices
Development of the Orchestrator Microservice
Development of Infrastructure Microservices
Tests, trials and platform validation

Challenges (1/3)

R&E networks share OSS data for research? How?

- Increasingly, operators, including NRENs and campus teams, are being asked for access to OSS data by network researchers interested in evaluating their research work against real network data
- How Research and Education Networks (RENs) and other network operators within the community can support requests from third-party researchers within the academic community for collecting and making available production OSS and related data from their networks for research purposes?
- What kind of network data can be open and, at the same time, still be valuable for networking researchers without compromising operations security or user privacy (e.g.: GDPR regulation)?

Challenges (2/3)

R&E networks share OSS data for research? How?

- Considering NRENs as instruments that can be used for research purposes by the academic communities, how R&E Network operators are handling requests?
- What data do we already collect in our OSS? What additional data might researchers want access to?
- What are the issues around sharing this data with third parties, especially network researchers, but also other NREN operators for cases of international multi-domain network (performance) troubleshooting?
- Considering advanced NRENs networks as instruments for CS scientific domain – how does your NREN support your researchers today?
- How network traffic, topology datasets, etc is collected and made available for researchers in the computer networking field? Are there existing best practices to be followed?

Challenges (3/3)

R&E networks share OSS data for research? How?

- Should (N)RENs publish their datasets in open data repositories? Are there any open standard for this purpose?
- Are there RENs sharing OSS data following F.A.I.R. principles?
- How can network-related data be shared (and to what level of detail) with networking researchers without compromising operations security or user privacy, considering the GDPR regulations, even if NDAs are being used?
- Which kind of anonymisation can or should be applied to sensitive data without compromising research objectives? What can be learnt from experts in the privacy field?

Thank You!

Alex Moura

alex@rnp.br