# Project Status Report

**Reporting period: 04/01/2022 - 07/30/2022**
**Project title:**
**Mid-Scale RI-1 Design Project (M1:DP):**
**Designing a Global Measurement Infrastructure to Improve Internet Security (GMI3S)**
OAC-2131987

**Principal Investigator:** kc Claffy, Bradley Huffaker (UCSD), David Clark (MIT)
**Project Manager:** Elena Yulaeva
**Lead Institution:** CAIDA, UCSD
**Other Institutions:** NSRC (U Oregon)
**Cognizant PO:**

## Table of Contents



# 1. Summary of project status

A brief summary of project's overall status on technical progress, cost and schedule performance.

| | | |
|---|---|---|
| Award Duration | Start date: 10/01/2021 | Planned close out: 09/30/2024 |
| Project Finish Date | Planned Early Finish: | Estimated Early Finish: |

| | | |
|---|---|---|
| Project Cost | Total project cost: 7,865,527 | Estimate-to-Completion: 6,490,382 |
| Cost Contingency | Budgeted contingency: 375,000 | Remaining contingency: 375,000 |
| Project %-complete 16% | | |

# 2. Near-Term Milestones

Include milestones with the scheduled dates or actual/forecast dates that are in current and the next reporting period, and milestones (with past scheduled dates) that are delayed to future reporting period. (**Completed deliverables have bold font dates**.)

| WBS | Subsystem | Milestone | Scheduled Date | Actual date (A) /Forecast Date (F) |
|---|---|---|---|---|
| 1.1 | **1.1.2.2** | Complete first draft of data needs report (based on 1.1.1) | 02/28/2023[1] | 02/28/2023 (F) |
| | **1.1.2.3** | **Preliminary Data catalog created** | **06/30/2022** | **06/30/2022(A)** |
| | 1.1.2.4 | Database of Peers (RV) created | 10/31/2022 | 10/31/2022(F) |
| | **1.1.3.1** | **Create inventory of all CAIDA machines** | **08/31/2022** | **6/30/2022 (A)** |
| | **1.1.3.2** | **Create inventory of all NSRC machines** | **08/31/2022** | **6/30/2022 (A)** |
| | 1.1.3.6 | Active measurements needs compiled | 03/31/2023 | 03/31/2023(F) |
| | 1.1.3.7 | DNS monitoring needs compiled | 04/30/2023 | 04/30/2023(F) |
| | 1.1.4.1 | Monitor specification report draft posted for internal feedback | 03/31/2023 | 03/31/2023(F) |
| | 1.1.5.1 | Telescope data monitor software prototyped | 03/31/2023 | 03/31/2023(F) |
| | 1.1.5.2 | Two-way traffic data monitor software prototyped | 01/31/2023 | 01/31/2023(F) |
| | 1.1.5.3 | BGP data monitoring software prototyped | 2/28/2023 | 2/28/2023(F) |
| | 1.1.5.4 | Active data monitoring software prototyped | 2/28/2023 | 2/28/2023(F) |
| | 1.1.6.3 | Two-way traffic monitor deployment | 1/31/2023 | 1/31/2023(F) |
| | **1.1.6.4** | **BGP data monitoring deployment** | **07/31/2022** | **07/31/2022(A)** |
| | 1.1.6.5 | Active probing measurements data monitoring deployment | 2/28/2023 | 2/28/2023(F) |
| | 1.1.6.7 | 10 nodes of at least one measurement deployed | 3/31/2023 | 3/31/2023(F) |
| | 1.1.7.1 | Software to support active probing measurements deployed | 9/30/2022 | 9/30/2022(F) |

[1] Originally the due date was Oct 31, 2022, but we decided to merge this with the Data Needs report prepared in Task 1.1.9. Community feedback to 1.1.9 will yield additional content for our data needs report. We will complete this task as part of subtask 1.1.9.8: Publish Final Report. Therefore, the due date will be changed to 2/28/2023.

| | | | | |
|---|---|---|---|---|
| | 1.1.8.1 | Preferred Network Function Virtualization framework selected | 9/30/2022 | 9/30/2022(F) |
| | 1.1.8.2 | Study and experiment with Network Function Virtualization frameworks for collection nodes and analysis machines started | 10/31/2023 | 10/31/2023(F) |
| | **1.1.8.4** | **Put all RV collectors on VMs (in conjunction with ILANDS)** | **08/31/2022** | **08/31/2022(A)** |
| | 1.1.9 | Integrated report "Internet infrastructure security vulnerabilities" and "Data Needs" published | 2/28/2023 | 2/28/2023(F) |
| | **1.1.9.1** | **Combine "Internet infrastructure security vulnerabilities" and Data Needs reports into a single report for increasingly wide review<br>Second draft posted for internal feedback** | **4/30/2022** | **4/30/2022(A)** |
| | **1.1.9.2** | **Internal feedback processed, draft (2) posted for internal feedback** | **6/30/2022** | **6/30/2022(A)** |
| 1.2 | 1.2.1 | Data storage hardware requirement documented – draft posted for stakeholders' review | 3/31/2023 | 3/31/2023(F) |
| | 1.2.2.2 | Data storage systems specifications documented– draft posted for stakeholders' review | 2/28/2023 | 2/28/2023(F) |
| | **1.2.3.1.1** | **Create metadata template for ASrank** | **6/30/2022** | **6/30/2020(A)** |
| | 1.2.3.1.2 | Create metadata templates for other ongoing data sets | 3/31/2023 | 3/31/2023(F) |
| | 1.2.3.1.3 | Create metadata templates for the datasets completed in the last 5 years | 10/31/2022 | 10/31/2022(F) |
| | 1.2.3.2 | Research the state-of-the-art metadata approaches | 10/31/2022 | 10/31/2022(F) |
| | 1.2.3.4 | YR1 Data and metadata standards specifications published | 10/31/2022 | 10/31/2022(F) |
| | **1.2.4.1** | **Report on existing state-of-the-art anonymization tools starting with Cryptopan and ONTA** | **09/30/2022** | **6/30/2022(A)** |
| | 1.2.4.3 | Specification of tools for data curation and documentation, YR1 report created | 3/31/2023 | 3/31/2023(F) |
| | 1.2.5.1 | Unified web interface to download heterogeneous datasets designed | 3/31/2023 | 3/31/2023(F) |
| | 1.2.5.3. | Improve existing API | 2/28/2023 | 2/28/2023(F) |
| | 1.2.5.4 | Document data and metadata APIs | 3/31/2023 | 3/31/2023(F) |
| | 1.2.8.1 | Report on the latest big data storage/management technologies | 3/31/2023 | 3/31/2023(F) |
| 1.3 | 1.3.1.3 | Prototype GMI3S Data Catalog | 10/31/2022 | 10/31/2022(F) |
| | 1.3.1.2 | Documentations of the existing tools and datasets improved | 10/31/2022 | 10/31/2022(F) |
| | 1.3.1.6 | Metadata databases to increase data accessibility created | 1/31/2023 | 1/31/2023(F) |

| | | | | |
|---|---|---|---|---|
| | 1.3.2.1 | Report on the gaps between privacy-preservation techniques and network and security research needs | 12/31/2022 | 12/31/2022(F) |
| | 1.3.2.3 | Gaps that privacy techniques can support identified, report created and shared | 1/31/2023 | 1/31/2023(F) |
| | 1.3.2.4 | Create taxonomy of data | 12/31/2022 | 12/31/2022(F) |
| | 1.3.2.6 | Design and prototype authentication/authorization solution | 12/31/2022 | 12/31/2022(F) |
| | 1.3.3.1 | A list of existing data sharing policies and best practices compiled and shared with community | 3/31/2023 | 3/31/2023(F) |
| | **1.3.3.2** | **Biannual meetings with at least two NSF-funded projects** | **07/31/2022** | **7/31/2022(A)** |
| | 1.3.3.3 | List of the DoD needs and requirements compiled | 3/31/2023 | 3/31/2023(F) |
| | 1.3.3.5 | Lessons learned documented | 3/31/2023 | 3/31/2023(F) |
| | 1.3.3.6 | Report on other countries' approaches | 3/31/2023 | 3/31/2023(F) |
| | **1.3.4.2** | **Quarterly meetings conducted** | **7/31/2022** | **7/31/2022(A)** |
| | 1.3.4.4 | Case study comparing the darknet dataset use by three entities documented and shared | 3/31/2023 | 3/31/2023(F) |
| 1.4 | **1.4.1** | **GMI quarterly workshop conducted** | **06/30/2022** | **6/30/2022(A)** |
| | 1.4.2.2 | Virtual collaboration environment evaluated | 09/30/2022 | 09/30/2022(F) |
| | 1.4.2.2 | Virtual collaboration environment evaluated and improved | 9/30/2022 | 9/30/2022(F) |
| | 1.4.3.3 | Video tutorials on nodes deployment and management created | 3/31/2023 | 3/31/2023(F) |
| | **1.4.3.4** | **Quarterly calls conducted, minutes shared** | **07/31/2022** | **07/31/2022(A)** |
| | **1.4.3.5** | **Project Presentations** | **09/30/2022** | **07/31/2022(A)** |

# 3. Technical progress highlights

This section summarizes highlights of progress of the current period, by near-term tasks.

## 1.1 Design infrastructure for Data Acquisition.

### 1.1.1 *Report on Internet infrastructure security vulnerabilities*.

As stated in our previous biannual report, we had completed a first draft[2] which we are now circulating for comments. We discussed the report at the GMI-Traffic, GMI-DNS and GMI-BGP work group meetings and are incorporating their suggestions into the document. We are planning another round or two of working meetings with GMI sub-teams to refine and expand the document.

[2] Username: `reviewer; Password: MSR1-view-m1-m36`

***Other relevant reporting on Internet infrastructure security vulnerabilities:***

***(a) Published Response to FCC NOI on BGP Routing Vulnerabilities.***

David Clark, Testart (MIT), and Claffy (UCSD) submitted a response to the FCC's Notice of inquiry regarding Internet routing security. We published this on our web site: Comments before FCC in the matter of Secure Internet Routing. Many of the other comments (mostly from industry and its representatives) called on the role(s) the government(s) should play with respect to funding measurement, data gathering, and research. SDSC Communications is currently working on a news story that will be published at the SDSC site.

***(b) Submited final version of "Challenges in Measuring the Internet for the Public Internet" to* Journal on Information Policy.** Expected publication July 2022. This paper serves as a framing document for our vision of GMI, and we will send it out to our Strategic Advisory Council (See Task 1.5 toward the end of this report.)

***(c) New study explaining benefit of additional BGP data collection to support BGP security.***

Clark and Claffy started to work on a paper based on Clark's recent MANRS+ presentation to NIST, Comcast, and British Telecom, and the Internet Society. (This presentation is now on version 5.0 after feedback from these four stakeholders and the GMI BGP internal sub-team. We are not publishing it yet since we will write a paper for blind peer review on this topic. We have shared it with PM Kevin Thompson.) We plan to submit a draft to Usenix in October to get feedback on the data architecture we believe is necessary to advance routing security operations and research.

**1.1.2. *Complete Report on Data Needs*.**

We completed a first draft report on data needs (based on the vulnerabilities identified in our report for Task 1.1.1 above). (Note that for Task 1.1.9 we integrate these two reports, mapping vulnerabilities to data needed to study and mitigate them. We've done this already since we found it easier to write both documents that way.) We sent an initial draft to the team in April and discussed it at the workgroup meetings (GMI-traffic, GMI-BGP and GMI-DNS) in May and June. We have incorporated most of their feedback and will do another two rounds of meetings, focused on specific sections of this document that need deeper thought. This document includes our preliminary data catalog as a set of tables listing relevant data sets for each set of vulnerabilities (Task 1.1.2.3).

We participated in research efforts (co-funded by other sources) only to the extent necessary to understand strengths and weaknesses of existing data sets, and gaps to consider for a GMI Implementation project. In this context we had several exchanges with commercial stakeholders (Spamhaus, Level3, Akamai, Amazon) to better understand the ability of existing data sets (provided by both industry, NGOs, academics) to provide insights to the study of security vulnerabilities of Internet infrastructure.

**1.1.3. *Assessment of monitor hardware and software needs.***

We decided we needed to restructure the Project Execution Plan to provide more granularity to the WBS, starting with this task. We now have six (7) sub-tasks under this task: two (2) that map to create current inventories of hardware and software prototypes at UCSD and UO, and five (5) that map to different categories of data we need to support. We then have sub-sub-tasks to separately track progress on hardware and software needs. Here is the refined structure for this Task 1.1.3, reflected in the updated WBS and milestones tables (Appendix A). We only include near-term due dates in the milestones table.

**Task 1.1.3 Based on results from 1.1.1 and 1.1.2, and community meetings, develop a monitor hardware and software design specifications**

| |
|---|
| 1.1.3.1 Create inventory of all CAIDA machines |
| 1.1.3.2 Create inventory of RouteViews-related hardware |

1.1.3.3 Compile Telescope data monitoring needs: for the initial test suite
1.1.3.3.1 Telescope monitor hardware requirements
1.1.3.3.2 Telescope monitor software requirements
1.1.3.4 Compile two-way traffic monitoring requirements (10 monitors)
1.1.3.4.1 Two-way traffic monitor hardware requirements
1.1.3.4.2 Two-way traffic monitor software requirements
1.1.3.5 Compile BGP data monitoring needs
1.1.3.5.1 BGP monitor hardware requirements
1.1.3.5.2 BGP monitor software requirements
1.1.3.6 Compile Active measurement needs
1.1.3.6.1 Active measurement hardware requirements
1.1.3.6.2 Active measurement software requirements
1.1.3.7 Compile DNS data monitoring needs
1.1.3.7.1 DNS data monitor hardware requirements
1.1.3.7.2 DNS data monitor software requirements

We did extensive work on this task, which we use the above numbering to report progress on.

***1.1.3.1 Create inventory of all CAIDA machines*.** Dan Andersen manually created an initial wiki page of all CAIDA hardware on the machine room floor. For a more sustainable solution, Bradley Huffaker, Dan Andersen, and Leo Pascual created a first draft of semi-automatic inventory system tracking all CAIDA virtual and hardware machines and software, and how they map to services and data shared with the research community. We are designing this to be as automated as possible, e.g., to use configuration and crontab files to create and update elements and links between them. We are also designing and developing mechanisms to map these software and hardware systems to datasets and software objects in the catalog, aiming to provide a unified view of the resources required to collect, process, curate, maintain and service, and share individual datasets (e.g. telescope, AS rank) and software services (e.g. BGP Stream).

***1.1.3.2 Create inventory of all RouteViews-related hardware.*** Ryan Leonard completed this inventory, which UO will publish as a project (internal) webpage. Ryan is working on adding additional RouteViews-related UO Data Center physical assets to this inventory. Next quarter we will discuss how to make our two inventory systems more consistent, or even to merge them.

***1.1.3.3 Compile Telescope data monitoring needs: for the initial test suite*.** We discussed the telescope data monitoring (hardware and software) needs during our May 2, 2022 GMI-traffic meeting. Based on feedback from these scientists. we decided to prioritize putting together an infrastructure that allows data analysis of near-real-time traffic streams. Ricky and Dan are now designing a Spark cluster, leveraging leftover hardware (Our hardware order was 9 months delayed due to supply chain issues; it did not arrive till June, so we adopted some donated Gordon nodes from SDSC's decommissioning of that HPC system, which we are using that to test and prototype designs.). We had to deploy a 10 GB link between Gordon's 1 GB nodes and CAIDA's racks. We are planning to finish this task by the end of September. Meanwhile we also helped researchers get access to our XSEDE HPC allocation at Expanse, which allowed them to access the telescope data from its nodes. CAIDA collaborators (U of Twente and MERIT) successfully ran their tools in this environment. They will provide their recommendations to the Telescope monitors needs report.

***1.1.3.4 Compile Two-way traffic monitoring needs*.** (formerly 1.1.4) We continued our work on putting together requirements for two-way traffic data monitoring hardware. Senior Personnel Dan Andersen made progress in a development of a prototype 100G capture machine (Related to this

project but funded by CCRI project CNS-2120399.) to capture 2-way traffic. He is revamping the 100G capture server design to use only one server (rather than one server for each direction of the link). Dan installed new SSDs into our first 100G capture machine and configured and tested the RAID array to make sure the speeds were sufficient for capturing two-way traffic. He ordered all the hardware needed to test Mellanox cards in these systems capturing packets across 100G links. We hope to be testing actual live capture by August.

***1.1.3.5 Compile BGP data monitoring needs***. This topic is multi-faceted since there are so many aspects of BGP relevant to security research. To begin, we are prototyping some tools that performs processing of RouteViews and RIPE BGP routing tables into a database to facilitate efficient analysis of per-RV-peer routing tables. We would like to get a better sense of how much redundancy is in the routing table that we could leverage to improve storage efficiency store in a monitoring and data collection system. The question also has operational security implications as some of the redundancy may be unintended and could facilitate malicious disruptions. Matthew Luckie (visiting scholar from University of Waikato) is leading this effort. He used all publicly available views of BGP data provided by RouteViews and RIPE RIS to construct the best publicly available view of the Internet's routing system as of 1 February 2022. In total, we observed 1,017,341 unique IPv4 prefixes announced by 73,530 ASes across 958 sessions with RouteViews and RIPE RIS collectors. We examined the first class of redundant prefix -- more-specific prefixes that are also covered by less-specific prefixes, where the origin and the single upstream are the same for all observed routes. In total, our dataset contained 245,714 prefixes with these properties. This might be an opportunity for efficiently encoding paths into an AS paths database.

We also spent some effort designing mechanisms to tie BGP data to other AS-level data sources, e.g., Spoofer. Combining data sources will be a key focus of Year 2 of the project.

***1.1.3.6 Compile Active measurement and monitoring needs***. Nothing to report. (See **Task 1.1.5.4.**)

***1.1.3.7 Compile DNS measurement and monitoring needs***. The DNS ecosystem is far more complex than any other aspect of the Internet architecture, and we will consider a variety of data needs. We are working with three other groups in the community doing different types of DNS measurement, and our primary work in this area this quarter as been in gathering feedback from those three groups (and other DNS researchers via the GMI-DNS WG meeting) to compile a list of DNS vulnerabilities, data that could help study them, and who has access (or could have access) to such data. (See also **Task 1.1.5.5**.)

**1.1.4. *Monitor Specifications Report (Y2 deliverable).*** Nothing to report. This report comes later.

**1.1.5. Develop *Monitor Software Prototype.*** Like **Task 1.1.3,** we restructured the Project Execution Plan to provide more granularity to this task. We now have five (5) sub-tasks under this task, one for category of data we currently plan to support. (We may expand this later based on research community feedback.). We report progress on design efforts for these 5 sub-tasks.

**1.1.5.1 *Telescope Monitor Software Prototype.*** We made significant progress in prototyping software for postprocessing raw **Telescope pcap traffic files**. Dr Mok supervised a UCSD Computer Science Masters (Max Gao) to investigate the use of NSF-funded HPC infrastructure at SDSC to speed up the analysis of data collected by the UCSD network telescope. We evaluated the performance of several libraries (libavro in C, goavro in Golang, pyavro in Python) for parsing telescope data in FlowTuple format, an Apache Avro-based data format that the STARDUST project used for aggregated telescope data. Researchers are now complaining that this format is inefficient to parse, i.e., it takes over a week to process a week's worth of FlowTuple data. We deployed Apache Spark on SDSC Expanse to parallelize analysis on the data. We achieved a 1500X performance improvement

over using our original library run in a single virtual machine. The techniques that we developed made analysis on long-term telescope data feasible.

***Programmable Hardware***. To handle the increasing amount of darknet traffic, we investigated the feasibility of using P4 switches and/or SmartNICs to accelerate the processing of packets and improve the reliability of the telescope infrastructure. Dr. Mok studied recent advances in network functionality built with P4, such as Heavyhitter detection, network telemetry, and sketches/data structures designed for P4 switches with limited resources. He explored the feasibility of using SmartNICs to offload some telescope packet processing functions from CPUs. He plans to conduct pilot tests on CloudLab, a NSF-funded testbed which recently deployed 15 SmartNICs in its platform, to examine in-line packet processing performance and estimate the benefits of deploying SmartNICs in network telescope or traffic monitors.

***1.1.5.2 Two-way traffic monitoring software.*** Nothing to report. See 1.1.3.2 on hardware progress.

***1.1.5.3 BGP data monitoring software.*** We explored a variety of steps toward automation and security as we consider our BGP monitoring design.

- Software to monitor infrastructure availability: NSRC developed a proof-of-concept GitHub Actions Workflow that will on-demand attempt to 'Ping All' of Route Views' in-service inventory. The Route Views team has investigated the many features of GitHub Enterprise that will facilitate sustainable management of the infrastructure. The University of Oregon has not yet rolled out GitHub Campus Program, so in the meantime, NSRC plans to pay out of pocket to enable GitHub.com Enterprise features to enable full features on Route Views (private) Infrastructure as Code git repositories. Image and link (see screenshot at https://www.caida.org/funding/msri-gmi3s/reports/github.com_routeviews_inventory_run.png)
- Software to enable automation : Route Views staff created an open source *routeviews* Python package that contains some key tools for managing Route Views infrastructure. There is a tool that enables automatically mailing a Route Views collector's peers in the case of an outage. There is also a tool that automatically generates peering requests for Route Views. This package leverages PeeringDB APIs and uses the Registration Data Access Protocol (RDAP) as a fallback.
- Delivered the above Python package (described above) to PyPI.org. Established Continuous Integration and Delivery (CICD) for this package. Delivering to PyPI.org enables Route Views Maintainers to easily download and use these tools. Additionally, we can leverage it using GitHub Actions (now running on GitHub Runners at hosted by UO).
  https://github.com/routeviews/Python-Package-Template
  https://pypi.org/project/routeviews/
- Prototyped BMP stack across private and public Kafka cluster, accessible from stream.routeviews.org functionalities (see diagram https://www.caida.org/funding/msri-gmi3s/reports/RouteViews-Infrastructure-Diagrams-2204.pdf) . This work will inform our design of high available architecture for streamed routing data.
- Working on capturing Route Views BMP infrastructure as Ansible code for the two (public and private) Kafka clusters.

***1.1.5.4 Active measurement software***. Our work on active measurement software included prototyping functionality in *scamper* (most well-known open-source active measurement software) to support its deployment on the RouteViews platform, including TLS support. Matthew Luckie added functionality to scamper to perform TLS authentication as a client to a remote controller. This feature obviates the need for the remote controller to have an IP-address-based firewall. Instead, the remote controller requires TLS authentication and verifies the scamper client presents a certificate

that it trusts -- i.e., is signed by a specific internal CA that we control. We still have to do the RouteViews side design and development of this TLS transaction. Understanding how effectively we can use the RouteViews platform to also conduct active measurements will inform our specification in **Task 1.1.3.6** above.

***1.1.5.5 DNS measurement software***. We spent some time understanding existing software systems that have been useful to the research community, but have no permanent home, i.e,. sustainable funding model. One system that CAIDA was asked to take over 2 years ago is called DNZ Zone Database (originally designed and developed by Ian Foster as http://dns.coffee), which is a repository (of flat files) and PostgreSQL database of Top Level Zone files gathered through ICANN's Centralized Zone Data Repository (and other sources) for the last ten years. Ian has been working with us to transition it to CAIDA cyberinfrastructure and use it as the basis for designing a more comprehensive system of DNS data and metadata and linking such data to other measurements.

Ryan Koga designed and developed a new log storage system that will assure data quality control, which has been a serious and persistent problem with the previous project. Elena Yulaeva contacted various zone files registries to gain access to missing zone files, another persistent issue for anyone trying to use the CZDS. We now download more than 1300 zone files daily. Ryan worked on transitioning of the software system to https://dzdb.caida.org , and began to investigate and experiment with the API. He added features in response to researcher requests, to assess the extensibility of the software platform and thus its feasibility as part of a GMI Implementation.

***1.1.7 Data Acquisition Component Evaluation***.

As with other tasks, we need to consider different data sources separately: Telescope, two-way traffic, BGP, Active Measurement, and DNS. One emphasis of this phase of the project is ascertaining the extent that we can leverage industry partners and their commercial infrastructure to support scientific research on the Internet. This quarter we had several meetings with engineering staff of ExpressVPN to discuss the possibility of using some of their 4,000 servers deployed at 146 locations around the world to perform sophisticated and demanding active measurements. In exchange, they are interested in partnership with a research group that can yield open-source tools and methods to reveal insight into how to optimize their peering and transit (what node is best connected to where) and traffic flow. They are also interested in improving the transparency and accountability of the VPN industry. They have agreed to provide us with access to several of their nodes for the purposes of experimentation and exploration of a mutually beneficial collaboration.

***1.1.8 Prototyping virtualization capabilities.*** Paul Biglete explored the possibility of Kubernetes Infrastructure deployment at CAIDA to support orchestration and virtualization needs. He succeeded in prototyping a Kubernetes cluster, and in Kubernetes containerization of various CAIDA APIs. After a thorough evaluation of the Kubernetes infrastructure security issues, we have concluded that it is going to add more complexity than benefits for our needs (and our IT staff), and have turned instead to the more recent, flexible, and lightweight *Nomad* software package. Our investigation is ongoing.

***1.1.9 Combine "Internet infrastructure security vulnerabilities" (1.1.1) and Data Needs (1.1.2) reports into a single report for increasingly wide review.*** We completed **Task 1.1.9.2**: Process internal feedback and **Task 1.1.9.3**: Draft (2) post for internal feedback. While we solicit further internal input, we have initiated **Task 1.1.9.4**: Draft (3) publish for community feedback. The target due date for the first round of community feedback is October 31, 2022. We have also advanced the integrated report by expanding the section on DDoS attacks and the data relevant to analysis of those attacks. There is still much to do on this document, but we are making steady progress.

## 1.2 Design Infrastructure for Data Management

***1.2.1.1 Data Storage hardware requirements***. The inventory we are designing (**Task 1.1.3.1**) will allow us to map datasets and software to hardware requirement. The metadata for each dataset and database curated by CAIDA contains an up-to-date size of the dataset, which is updated daily. We conduct periodic reviews of the data growth and plan future purchases based on these projections.

***1.2.2 Data storage systems specification***. Our Data Management Infrastructure group is meeting weekly to discuss the data management infrastructure hardware and software needs.

***1.2.3 Design Data and Metadata standards***. Our goals are to (1) provide a clean high-level representation of the data in each dataset; (2) provide a method for searching datasets by the type of data they contain; and (3) provide data provenance. In pursuit of this goal, Bradley Huffaker designed a representation of each Dataset as a collection of *DataTables* each with its own ordered set of Columns. Each Column will have a name, optional *DataType*, optional description, and optional example. Ken Keys proposed a JSON schema-based alternative. We are currently comparing these two options and will decide by September. By October we will create metadata templates for CAIDA datasets completed in the last 5 years.

***1.2.3.2 Research the state-of-the-art metadata generation approaches.*** We reviewed best practices of creating metadata for Internet measurements. We studied Joel Sommers' NSF-funded work on automatic metadata generation for active measurements. In the process of trying to classify publications in the catalog, we also designed some techniques for automatic generation of classification and other meta-data, e.g., dataset used.

**Use of cloud for storage and querying**. We have begun our investigation of putting RouteViews BGP data into the Google Cloud Platform as a public data set (like M-Lab) for use with the BigQuery system. We have begun engagement with Google on technical details. We will dedicate time during a GMI-BGP working group meeting next quarter to discuss existing and potential schemas for putting BGP RouteViews data into the BigQuery columnar storage.

***1.2.4 Tools for data curation and documentation.***

***1.2.4.1 Explore existing state-of-the-art anonymization tools***. David Clark reviewed existing literature, identified the most important schemes, and provided a summary (https://www.caida.org/funding/msri-gmi3s/reports/Task_1.2.4.1.pdf ) . Clark's list of the most promising schemes reveals a pattern widely observed across academic research, and especially network measurement. That is, many projects with creative ideas are implemented, but not sustained. The space is littered with expired projects. He compared and contrasted features of the most important active projects: CryptoPan, Tcpreplay, TraceWrangler, ONTAS, and PINQ. We will continue to update the survey if and as more tools emerge, but his exploration established the current state of the art.

***1.2.5 Data and Metadata APIs.*** We began to consider future designs for existing CAIDA APIs.

***1.2.6 Software Development Libraries.*** We focused on design and development of software systems and libraries for automating collection, curation, processing and searching of Internet Topology Data Kits. We are researching the possibility of rolling out a simplified JSON format for the ITDK and will prototype it next year. Our current focus has been on curating RTT data that provides

geolocation constraints, including automatically inferring when an RTT sample is unreliable because a system local to a vantage point is forging responses as if it were the destination.

***1.2.7 Tools for additional data sources integration.*** Nothing to report.

***1.2.8 Tools for dissemination.*** Nothing to report.

## 1.3 Design Infrastructure for Broad Usability

***1.3.1 Data discovery tools development.*** We invested efforts into further development of CAIDA's Internet science resource catalog which we are using to design and prototype a GMI catalog. We refined search functionalities by adding ontology-based annotations. We started adding external datasets to the catalog (based on the data needs report). Currently the catalog indexes 92 CAIDA and 13 external datasets. We are developing new functionality to map catalog objects and their resources to funding sources and create points of contact for each object. Leo Pascual and Bradley automated the process of programmatically detecting and updating the start and end dates for datasets stored in the catalog as well as their sizes on disk, Swift object storage, or database.

***Create DOIs for DataSets (Task 1.3.1.3.4).*** To enable consistent and correct citations of CAIDA datasets, Elena Yulaeva started minting DOIs for publicly available dataset. CAIDA owns an EZID account and can generate DOIs with **10.21986/CAIDA 10.21986/S6. CAIDA** shoulders.[3] We decided to describe dataset objects following the *DataCite* Metadata schema that is more appropriate for dataset citation and discovery purposes than other EZID schema options. We started with datasets that are requested/downloaded by the largest number of users. We will create DOIs for all CAIDA datasets indexed in the catalog by the end of Y2Q1.

**Create metadata databases to explore their value and feasibility**. Thomas Krenc designed and prototyped a BGP metadata database (Co-funded by CCRI project CNS-2120399 ). The two global BGP data collector projects (RouteViews and RIPE RIS) add millions of MRT files containing routing table dumps and updates to their archives every decade, to support debugging and research. These archives expand as the number of collectors and peers grows, but also because of routing dynamics in an increasingly complex global network. The BGP metadata database – *BGPmeta* -- allows the user to select only those MRT files that contain a specific resource, thus avoiding unnecessary download and processing of data. It maps unique numeric identifiers (prefix, ASN, and community) to MRT files they occur in. It provides mappings for all archived and new MRT files. The user queries the database with a numeric identifier (prefix, ASN, or community) and the database returns a list of related MRT file information (the database does not return the actual MRT files). Researchers can use the returned MRT file information to reconstruct a URL via which others can download the file. The database can serve as a broker by any tool that utilizes archived MRT files, to effectively improve performance.

***1.3.2 Understand landscape of software for disclosure controls.*** David Clark investigated gaps between privacy-preservation techniques and network and security research needs. There are several classes of disclosure controls, including differential privacy (DP), secure multi-party computation (SMPC), homomorphic encryption and synthetic datasets. Our initial focus has been differential privacy, which offers great promise but also raises great challenges. David Clark illustrated DP behavior by implementing one of the commonly used DP algorithms for adding noise (Laplace noise distribution) to a query result. His example showed how one can map noise to a probability of privacy loss that could make sense to a policymaker, but the example also illustrates

[3] https://ezid.cdlib.org/learn/id_basics#IdentifierBasics-shoulders

the problem with estimation of potential harm caused by privacy loss for different sorts of queries. Clark outlined several important insights in the DP approach and identified three promising open-source software packages that implement the basic DP primitives. The complete report can be found at https://www.caida.org/funding/msri-gmi3s/reports/Task_1.3.2.pdf

**1.3.2.*2 Conduct workshops twice a year***. We did not organize a workshop to bring together diverse experts to find common grounds and develop solutions that DHS was sponsoring a similar workshop in June: Workshop on Privacy Enhancing Technologies | Center for Accelerating Operational Efficiency.[4] None of us could attend that workshop but we did reach out to the chair and ask if there would be a report and could we talk with him afterwards. We will follow up with him a month after the workshop to give him a chance to finish the report (hopefully). We are planning to have our own online meeting on this topic in Y2Q1 of this project (Oct-Dec 2022).

***1.3.3.1 Work with industry partners to compile a list of existing data sharing policies and best practices.*** We got sidetracked on this project when Amazon asked us to create a data sharing license for them with one of our BGP-related (RouteViews-derived) data sets, and we thought it would make an excellent test case. worked with industry to create data sharing policies and agreements. After multiple calls with Amazon and meeting with the UCSD legal department and office of innovation and commercialization we sent Amazon a draft of the license agreement. Unfortunately, Amazon's response included additional requests and conditions that prevented UCSD from moving ahead with licensing. The areas of disagreement were: (1) Amazon wanted a multi-year agreement that automatically renewed, but we do not have sustained funding for that data set and are not prepared to provide guarantees for such data availability for many years in the future without sustained funding (2) Amazon wanted to shift commercial-scale liabilities to UCSD, and our pricing did not include this; (3) Amazon did not want to acknowledge credit for these data to CAIDA or the funding agencies for the data; (4) Amazon wanted any affiliate within the Amazon family of companies be able to create a commercial product with the data. We need to know (in order to report to funding agencies) who we are doing business with and where and how those parties use the data.

***1.3.3.2 Work with academic researchers (esp. with other NSF-funded projects) to leverage labor and experience starting with NSF-funded projects RSOC (Research Security Operations Center) and CCRI CLASSNET (USC-ISI).*** We had a phone call with Von Welch shortly before he departed IU to explore the best way to leverage RSOC resources for this project. He pointed us at the Stingar folks at Duke as a more fruitful avenue for this year.

***1.3.3.3 Work with DoD to compile their needs and requirements.*** We had unfortunate news from DARPA who said they could no longer fund the telescope out of their Searchlight program and that we needed to create some recharge mechanism for DARPA PIs to be able to pay UCSD directly for the data they use. (This news led to accelerated momentum on Task 1.3.3.4 below.)

***1.3.3.4 Work with UCSD OCGA and Rev-Up to develop agreements to share sensitive data.*** We continued our work with SDSC's *RevUp* program to develop agreements to share sensitive data for commercial use. RevUp reviewed the draft of the MOU that we want to sign with Domain Tools who will produce new products from and share CAIDA telescope data.

Elena Yulaeva participated in SDSC's RevUp program's May 3-6 On-Ramp workshop[5] which educated academic researchers about revenue generating approaches to sustain academic projects.

[4] https://caoe.asu.edu/workshop-privacy-enhancing-technologies
[5] https://www.sdsc.edu/services/rev-up.html

***1.3.3.5 Identify lessons learned from previous data sharing efforts.*** In progress, will report next quarter.

***1.3.3.6 Investigate how Europe and other parts of the world are approaching this same issue. .*** In progress, will report next quarter.

***1.3.3.7 Pursue agreements with commercial data providers (Kentik, Farsight).*** In progress, see Task 1.3.3.4 above.

***1.3.3.8 Design new data sharing agreements based on lessons learned above and share with community.*** We worked on pursuing agreements with commercial data providers. Dr. Claffy followed up with Verisign about access to rapid zone updates for .com and .net. She also tried to move forward with the Telescope SIE data (See Task 1.3.3.4).

## 1.4 Design Infrastructure for outreach

***1.4.1 Conduct quarterly workshops for each GMI workgroup.*** We conducted quarterly GMI-Traffic, GMI-BGP and GMI-DNS workgroup meetings, to ensure our Infrastructure Design me

***1.4.1.1 GMI-traffic meetings***. The first workshop on April 8 was an introductory workshop where Telescope data users presented their work (GMI-traffic Presentations 2022-04-08) . The second workshop on May 2 was dedicated to the following three issues: (1) How to sustain the UCSD telescope instrumentation, (2) What is the most effective way for community to benefit from the value of the UCSD Network Telescope, (3) Next steps to implement recommendations. Participants discussed the value of CAIDA telescope data, why it is important, and opportunities to improve the ease of access. Attendees also discussed their darknet data needs and offered sustainability suggestions, The summary of the workshop is at https://www.caida.org/funding/msri-gmi3s/reports/Summary_GMI-Traffic_meeting_2022-05-02 .pdf.

After this workshop kc held several smaller meetings with users currently dependent on the Telescope data for their research, to ascertain potential sources of sustainable support. One of our largest users is a research group in Germany, and they highlighted the difficulty of arranging a joint funding situation between Germany and U.S. universities.

The third GMI-Traffic workshop took place on July 15. It was structured with fewer talks and more opportunities for longer conversations. Raphael Hiesgen from Germany (in the research group mentioned above) reviewed his uses of telescope data and research results and drilled down on his recent study of scanners seeking remote-code-execution (RCE) of Log4Shell severe vulnerabilities. We then discussed related uses of telescope data for studying vulnerabilities and attacks, current and future arrangements for accessing telescope data, and starting task force on macroscopic trend analysis of DDoS attack ecosystem.

***1.4.1.2 GMI-BGP meetings***. We had our first GMI-BGP workshop on May 6. The goal of the meeting was to discuss BGP measurement efforts and collected metadata. This series of meetings is supported by two new CAIDA projects that seek to advance the state of BGP measurement: (1) CCRI ILANDS – creating new instrumentation to enable research combining traffic+BGP data (5-year NSF-funded project), and lower barriers to scientific research using both types of data; and [this project:] (2) NSF MSRI GMI3S – what data, tools and capabilities are needed to accelerate scientific research of Internet infrastructure security problems. Ryan Leonard presented the status of the

RoutViews project. Ben Cox presented his *bgp.tools* platform – user friendly tool to look at the current BGP state of the network. Ben presented an update on this topic at RIPE meeting in May 2022 https://ripe84.ripe.net/presentations/110-The-unending-misery-of-bgp.tools-RIPE84-1.pdf. Ethan B. Katz led discussion of "How to run safe and ethical BGP experiments" – we discussed codes of conduct in the community and how researchers could collectively approach execution of research methods with operational implications. Our summary of the meeting is at https://www.caida.org/funding/msri-gmi3s/reports/Summary_GMI-BGP_meeting_2022-05-06.pdf

***1.4.1.3 GMI-DNS meetings***. We had a kick-off GMI-DNS group meeting on May 24. The goal of the meeting was to solicit comments from DNS experts on the Domain Name System (DNS) section of the "Data Needs for Internet Security" document, a key deliverable of the MSRI GMI3S project. Participants suggestions and comments included:

- Classify/taxonomize barriers to various types of data collection and identifying common barriers (consider adding new "status" column to table, or fleshing out "Limitations")
- Evolve data use policies and best practices (what you can and cannot do with the data)
- Start with what the problems are, then map to data.
- Analyze how DNS data has greater privacy challenges than other data, e.g., BGP.
- Show high value of DNS transparency (zone updates) over currently available data
- Include "Commercial relevance": dimension (facilitates sustainability of data collection)
- Need (sometimes extensive) metadata for each DNS dataset that provides an accurate view of data collection structure, allowing metrics calculations
- Need methodologies to inform vantage point selection, to avoid bias in inferences
- Discussion of DNS analogue of MANRS, how to consider (cf. ICANN KIND-DNS list)[6]
- DNS abuse trends and how to approach the lack of data needed for research. Now have contradicting reports on the state of DNS abuse (from ICANN, DNS Abuse Institute, Interisle), and how researchers cannot get to the root of differences in abuse levels and trends without improved data access.
- Creating Zones of Trust by combining technological tools and best practices. Participants will add comments into the Vulnerabilities report, and will discuss the updated document further at the next meeting. Our summary of the meeting is at https://www.caida.org/funding/msri-gmi3s/reports/Summary_GMI-DNS_meeting_2022-05-24.pdf

We note that we still have seven meeting topics (Topology, Active Measurement, DDOS, Data Policy, Internet Economics, Security, Data Ops) to workshop, so to be efficient we will be consolidating some of them. We will merge GMI-Topology and GMI-Active measurement meetings and begin them next quarter. We will fold GMI-DDOS meetings into GMI-Traffic, and the first DDOS-focused GMI-Traffic meeting will be next quarter. The remaining four topics (Data use policies, Economics, Security, and DataOps) we are keeping separate for now, there is little overlap among them.

***1.4.2.2 Virtual Collaboration Environment***. We are extensively using MatterMost for our daily communications. We created channels for various topics and workgroups and added linked to documents, git repositories, etc to the channel headers. We also use github/gitlab and Google workspace collaboration tools. In August Elena Yulaeva will administer a short survey to evaluate effectiveness of existing virtual collaboration environment and to identify any additional needs.

***1.4.3 Modules to scale STEM workforce development.*** In June 2022 we had our first meeting on an online *Network Infrastructure Data Science* course design (which UCSD's CSE department is

[6] https://mm.icann.org/mailman/private/kindns-discuss/

extremely interested in), and discussed candidate assignments and capstone projects for both undergraduate and graduate level courses

***1.4.4 Quarterly calls with stakeholders.*** We continued CAIDA's weekly SALON calls (Studies in Architecture and Legislation Of Networks) which is a small tech policy forum we started as part of our previous DIBBS project. We use this forum as one early mechanism of stakeholder engagements since we exchange ideas with policy stakeholders and economics researchers on political and economic questions whose study is stalled by lack of empirical data on the Internet infrastructure. Topics this year have included: (a) the FCC's NOI on routing security, which highlighted many of the problems this project is trying to address (and we submitted a comment in response to this NOI). (b) our Measurement of the Internet in the Public Interest paper as it we revised it for the *Journal of Information Policy*. (c) the FCC's struggle to measure broadband deployment, and approaches to addressing this gap. (d) empirical data needs for a proposed Digital Platform Agency or Bureau of Cyber Statistics. (e) Measuring the Fragility of Internet Access Networks (our Usenix submission for another project). (f) two participants' submissions to the FCC Broadband Nutrition Label request for comments. (g) economic and policy implication of efforts to use network architecture and policy against Russia in the Ukraine conflict. (h) the biggest threats facing the DNS today. (j) review of and feedback on our Data Needs document that is a primary deliverable for this year

***1.4.5 Present project outcomes at various meetings***. We presented project vision, status, and outcomes at meetings. PI Claffy attended the MERIF workshop on June 1-3 in Madison, WI MERIF Workshop 2022. She made a presentation about CAIDA GMI3S project https://www.caida.org/catalog/media/2022_gmi3s_merif/gmi3s_merif.pdf and was one of the panelists for a panel on Data Sharing, Curation and Management. While there we met with two NSF PMs for this project (Deep Medhi and Kevin Thompson) to give them a status report.

Working with users of data we anticipate being a part of the GMI Implementation project, we assisted with 7 papers that were submitted to AMC's Internet Measurement Conference in May 2022. We will provide a list of papers in our next report when we find out which of them were accepted in August 2022.

## 1.5 Project management

***1.5.1.2 Periodically review and update PEP.*** In April we reviewed our PEP WBS dictionary and rearranged some Jira tasks and added issues to align our workplan to the team structure and team members functions. We propagated these changes back to our PEP WBS Dictionary (Table 5 of Project Execution Plan).

***1.5.2 Project Controls.*** We attended to several tasks in this area.

**1.5.2.1** We continued our biweekly CAIDA/MIT/UO meetings
**1.5.2.2**. We continued RV/CAIDA weekly technical meetings
**1.5.2.3** We documented all deliverables and assigned responsibilities in Jira. We propagated this information into PEP (see updated WBS and milestones tables in Appendix A)
**1.5.2.4** We submitted our first biannual report in April
**1.5.3.1** We compiled a list of strategic advisory council members and sent email invitations. The initial list of invitees (everyone accepted our invitation to serve) includes:

Jennifer Rexford, Princeton
Nick Buraglio, Esnet
Suzanne Wolff, PIR

Roland van Rijswijk-Deij, U. Twente/SIDN Labs
Romain Fontugne, Internet Initiative Japan (IIJ)
Avi Freedman, Kentik
Joe St Sauver, Domain Tools
In future we plan to invite representatives of AT&T, Comcast, Google, and Fastly.

# 4. Issues and major risks

Following up on previously mentioned issues: We finally received our first hardware order, but we are 9 months behind where we expected to be regarding developing and deploying prototypes. We have compensated by focusing on other parts of the project, and re-using existing hardware.

**Human Resource (HR) Retention**. As reported in the last quarterly, NSRC had planned to hire a new FTE that will be 50% RouteViews supported by this project, to replace loss of RouteViews Technical Lead David Teach to industry. NSRC is now prioritizing post this job card in August or September, so we hope to have someone to train up by January. In addition, RouteViews Technical Lead John Kemp offered to help address the technical backlog and train up new staff.

Unfortunately, NSRC had another HR hit as Hervey Allen, serving as PI and Project Manager for the UO/NSRC side of this project, has been out on medical leave since early May. NSRC veteran Hans Kuhn came in to fill the management gap with this project.

On the UCSD side, we lost a primary IT (dev-ops) staff person (Paul Biglete) to industry, again for higher salary and longer-term career opportunity. He only stayed with us two years, again illustrating the difficulty of talent retention in this IT climate. It took significant cycles from our primary Systems Administrator Dan Andersen (and others in the staff, including PI Claffy) to offboard him, which required understanding absolutely everything he was doing and try to get it documented. PI Claffy also talked to SDSC about paying for 50% of a young SDSC systems administrator (Victoria Nguyen) to take on some of the easier tasking while Dan retains the most complex responsibilities. Dan spent considerable time training up Victoria and an REU to take on some sysadmin tasks. We have posted a new job card to try to hire a full-time Internet systems administrator focused on traffic monitors for this project.

**Inflation and salary increases**. UCSD is providing a 4.5% salary increase across the board to all staff. This will impact our labor budgets. With inflation above 10%, this will not help sufficiently with competitiveness with industry salaries. We will likely need to use the contingency budget to hire new staff at competitive salaries, and then provide salary equity increases to existing staff.

# 5. Cost and performance summary data

| Subsystems (L2 or L3 WBS) | Budgeted Cost | Cumulative Actual cost | Invoiced but not paid subcontracts | Equipment committed | Work % completed |
|---|---|---|---|---|---|
| **1.1** | **2,731 k** | **229K** | | **21K** | **8%** |
| **1.2** | **1,946 k** | **441K** | | | **22%** |

| | | | | | |
|---|---|---|---|---|---|
| **1.3** | **1,390 k** | **405K** | | | **22%** |
| **1.4** | **932 k** | **185K** | | | **20%** |
| **1.5** | **491K** | **119K** | | | **20%** |
| **Project total** | **7,865K** | **1,379K** | **0K** | **21K** | **18%** |

# 6. Approved changes to the project baseline (if any)

We have adjusted the PEP to provide more granularity and coherence but not changed the project baseline.

# 7. APPENDIX A

**Updated Work Breakdown Structure (WBS)**
**Dictionary (PEP Table 5)**

| Code | Element | Definition |
|---|---|---|
| 1.1 | Design Infrastructure for Data Acquisition | Design measurement platform that can perform a range of data collection tasks, as well as supporting new vetted experiments by the research community |
| 1.1.1 | Report on Internet Security Vulnerabilities | Conduct meetings and workshops to review security risks and map to data needed (see Task 4). Draft preliminary report on Internet infrastructure security vulnerabilities that the GMI project will gather/manage/share data to address. |
| 1.1.2 | Data Needs Report | Identify target data sets and collection protocols.<br>1.1.2.1 Create a slide deck with detailed information about existing CAIDA datasets used for security research (see preliminary list)<br>1.1.2.2 Complete Data Needs draft report (based on 1.1.1)<br>1.1.2.3 Create preliminary data catalog<br>1.1.2.4 Create database of all peers (RouteViews) |
| 1.1.3 | Document Monitors Requirements | Based on results from 1.1.1 and 1.1.2, and community meetings, develop a monitor hardware and software design specifications.<br>1.1.3.1 Create inventory of all CAIDA machines<br>1.1.3.2 Create inventory of RouteViews-related hardware<br>1.1.3.3 Compile Telescope data monitoring needs: for the initial test suite<br>1.1.3.3.1 Telescope monitor hardware requirements<br>1.1.3.3.2 Telescope monitor software requirements<br>1.1.3.4 Two-way traffic monitoring requirements (10 monitors)<br>1.1.3.4.1 Two-way traffic monitor hardware requirements<br>1.1.3.4.2 Two-way traffic monitor software requirements<br>1.1.3.5 BGP data monitoring<br>1.1.3.5.1 BGP monitor hardware requirements<br>1.1.3.5.2 BGP monitor software requirements<br>1.1.3.5.3 Analysis of traffic engineering artifacts in BGP that create redundancy in data collection<br>1.1.3.5.4 Improve algorithm, implementation, data curation and API to access AS relationship data and correlate it with other data sources, e.g., Spoofer<br>1.1.3.6 Active measurement needs<br>1.1.3.6.1 Active measurement hardware requirements<br>1.1.3.6.2 Active measurement software requirements<br>1.1.3.7 DNS data monitoring needs<br>1.1.3.7.1 DNS data monitor hardware requirements<br>1.1.3.7.2 DNS data monitor software requirements |

| | | |
|---|---|---|
| 1.1.4 | Monitors Specification Report | 1.1.4.1 Integrate results of 1.1.3.(3-7) into the first draft of monitor specifications report<br>1.1.4.2 Conduct internal review<br>1.1.4.3 Process feedback from internal review<br>1.1.4.4 Send draft for community feedback<br>1.1.4.5 Process feedback from community review<br>1.1.4.6 Post final document |
| 1.1.5 | Develop Monitors Soft-Ware prototypes | Based on results from 1.1.1, collect and implement necessary software for specified suite of tests. Use quarterly community meetings to discuss modular software architecture/design<br>***1.1.5.1 Telescope data monitoring software prototype***<br>1.1.5.1.1 Explore use of P4 technology for telescope<br>***1.1.5.2 Two-way traffic data monitoring software prototype***<br>1.1.5.2.1 Explore Princeton's P4-based anonymization tools for "anonymized live traffic capture"<br>***1.1.5.3 BGP data monitoring software***<br>1.1.5.3.1 Write and present RV data processing pipeline<br>1.1.5.3.2 Document RV maintenance<br>1.1.5.3.3 Document & Diagram "Route Views Orchestration Strategy<br>1.1.5.3.4 Use Github actions to orchestrate peer additions<br>1.1.5.3.5 Automate notification of peer going offline<br>1.1.5.3.6 Upgrade stream.routeviews.org to 3+ node Kafka cluster<br>1.1.5.3.7 Compare Open Source BGP stacks<br>1.1.5.3.8 Explore viability of goBMP as platform instead of OpenBMP (in C) to meeting monitoring needs. Need to do following development to evaluate feasibility:<br>a) Incorporate openBMP header and message formats into goBMP<br>b) implement *bmp_raw* message type<br>c) Add ability to introduce new BGP peers without restarting session<br>d) Leverage goBMP topics with group mapping<br>e) Enabled topic/group updates without interrupting goBMP process<br>f) Find similar use cases of goBMP, e.g. CodeBGP<br>g) Explore collaboration with goBMP authors<br>h) Write and present RV data processing pipeline<br>***1.1.5.4 Active data monitoring software***<br>1.1.5.4.1 Createa system for supporting authenticated remote use of Scamper deployment on the RouteViews platform, including TLS support.<br>1.1.5.4.1.1 Add authentication capability to the scamper client software, (regular TLS authentication, -- i.e., scamper (as client) presents a certificate to the server signed by a cert that the server trusts.<br>***1.1.5.5 DNS data monitoring software*** |

| | | |
|---|---|---|
| 1.1.6 | Prototype Monitors Deployment | Deploy prototype monitors in collaboration with R&E networks.<br>1.1.6.1 Mailing lists for collaborating partners created<br>1.1.6.2 Telescope data monitoring deployment<br>1.1.6.3 Two-way traffic data monitoring deployment<br>1.1.6.4 BGP data monitoring deployment<br>1.1.6.4.1 Explore possibility of VM at LINX (in conjunction with ILANDS)<br>1.1.6.4.2 Deploy collector in South Africa (in conjunction with ILANDS)<br>1.1.6.5 Active probing measurements data monitoring deployment<br>1.1.6.6 DNS data monitoring deployment<br>1.1.6.7 Ten nodes of at least one measurement deployed<br>1.1.6.8 Ten nodes of multiple measurements deployed<br>1.1.6.9 Additional 10 nodes of multiple measurements deployed |
| 1.1.7 | Data Acquisition Component Evaluation | Demonstrate operation of target measurements, use of platform for active probing, and deployment of a third-party experiment.<br>1.1.7.1 Software to support active probing measurements deployed<br>1.1.7.2 Third party experiment deployed<br>1.1.7.3 Evaluation report of Data Acquisition component published |
| 1.1.8 | Prototype Virtualization Capabilities | 1.1.8.1 Select preferred Network Function Virtualization framework. Package monitor software into this frame. Demonstrate deployment and operation.<br>1.1.8.1.1 Discuss Kubernetes functionalities<br>1.1.8.1.2 Explore pros and cons of packaging monitor software stacks into this frame<br>1.1.8.2 Study and experiment with Network Function Virtualization<br>1.1.8.3 Report on Virtualization capabilities<br>1.1.8.4 Put all RV collectors on VMs (in conjunction with ILANDS) |
| 1.1.9 | Combine "Internet infrastructure security vulnerabilities" (1.1.1) and Data Needs (1.1.2) reports into a single report for increasingly wide review | 1.1.9.1 Draft (1) post for internal feedback<br>1.1.9.2 Process internal feedback<br>1.1.9.3 Draft (2) post for internal feedback<br>1.1.9.4 Draft (3) publish for community feedback<br>1.1.9.5 Process community feedback<br>1.1.9.6 Draft (4) publish for public feedback<br>1.1.9.7 Process public feedback<br>1.1.9.8 Publish Final Report |
| 1.2 | Design Infrastructure for Data Management | Prototype data curation and sharing infrastructure. Design meta-data ontologies; standardize data exchange formats; design and prototype tools to support data curation; write documentation that includes guidance on valid usage and caveats, and techniques for efficient data sharing and dissemination. |

| | | |
|---|---|---|
| 1.2.1 | Data Storage Hardware Requirements | Based on 1.1.1 and 1.1.2, determine data rates from monitors, and requirements for curation and dissemination.<br>Weekly meetings with sysadmins about evolving data needs. Compile data storage estimates, collect feedbacks, document data storage hardware requirements<br>1.2.1.1 Data storage hardware requirement documented – draft posted for stakeholders' review<br>1.2.1.2 Community feedback incorporated into the document |
| 1.2.2 | Data Storage Systems Specifications | Based on results from 1.2.1 develop a system design specification for data curation and dissemination. Create DevOps Infrastructure workgroup, convene monthly meetings to plan hypervisor environments for researcher access. Document data storage systems specifications<br>1.2.2.1 Data Management Infrastructure (DevOps) group created<br>1.2.2.2 Data storage systems specifications documented; draft posted for stakeholders' review<br>1.2.2.3 Community feedback incorporated into the document |
| 1.2.3 | Data and Metadata Standards | Based on results of task 1.1.1, identify resulting data elements and necessary meta-data. Develop necessary standards.<br>1.2.3.1 Develop metadata schema for existing CAIDA datasets<br>1.2.3.1.01 Create metadata template for ASrank<br>1.2.3.1.02 Create metadata templates for other ongoing data sets<br>1.2.3.1.03 Create metadata templates for the datasets completed in the last 5 years<br>1.2.3.2 Research the state-of-the-art metadata approaches<br>1.2.3.2.1 Consult with Joel Sommers on his metadata research and software development<br>1.2.3.2.2 Discuss schema for putting BGP RouteViews data into BigQuery column store<br><br>1.2.3.3 Quarterly meetings to discuss proposed metadata<br>1.2.3.4 Data and metadata standards specifications updated annually<br>1.2.3.4.1 Year 1 update of data and metadata standards specifications<br>1.2.3.4.2 Year 2 update of data and metadata standards specifications<br>1.2.3.4.3 Year 3 update of data and metadata standards specifications |
| 1.2.4 | Tools for Data Curation and Documentation | Design tools for data anonymization and post-processing analytics. |

| | | |
|---|---|---|
| | | 1.2.4.1 Explore existing state-of-the-art anonymization tools starting with Cryptopan and ONTAS<br>1.2.4.2 Explore existing pcap on-the fly analysis tools starting with https://packettotal.com/ and https://dynamite.ai/<br>1.2.4.3 Specification of tools for data curation and documentation, report created<br>1.2.4.4 Specification of tools for data curation and documentation, final report created (Year 3) |
| 1.2.5 | Data and Metadata APIs | Design new data and metadata application programming interfaces for access to various data sets.<br>1.2.5.1 Design improved unified web interface to download heterogeneous datasets.<br>1.2.5.2 Increase the number of supported data sources, including non-CAIDA datasets<br>1.2.5.2.1 Incorporate datasets used by IYP<br>1.2.5.3 Improve existing APIs<br>1.2.5.3.1 Improve AS Rank API<br>1.2.5.3.1.1 Improve sort functionalities, add sort by relationship<br>1.2.5.3.1.2 Fix a bug causing request with number Addresses to produce a server error<br>1.2.5.3.1.3 Add IPv6 and country to asrank-tools<br>1.2.5.4 Document data and metadata APIs, update annually.<br>1.2.5.4.1 Document Data and metadata APIs<br>1.2.5.4.2 Year 2 Update of data and metadata APIs<br>1.2.5.4.3 Year 3 Update of data and metadata APIs |
| 1.2.6 | Software Development (SDK) Libraries | Design software development libraries that help synthesize and map structural information about the Internet with high- level phenomena investigated by domain researchers, e.g., geolocation, ownership of Internet address resources.<br>1.2.6.1 Expand, document, share CAIDA's software library libipmeta and use as an example for future tools.<br>1.2.6.1.1 Make *ipmeta* stable enough to *bgpsrv*<br>1.2.6.2 New libraries created<br>1.2.6.2.1 Develop architecture and design, implement, evaluate, and deploy software systems and libraries for automating the collection, curation, processing, and searching of Internet Topology Data Kits<br>1.2.6.3 SDK Libraries evaluated; report published |
| 1.2.7 | Tools for Additional Data Sources Integration | Based on results of 1.1.1 and 1.1.2, identify candidate external data sets to integrate into the infrastructure. Set targets for integration of each candidate.<br>1.2.7.1 Create instance of Internet Health Report platform<br>1.2.7.1.1 Become expert on AS hegemony code and BGPstream<br>1.2.7.1.2 Develop Spark expertise |

| | | |
|---|---|---|
| | | 1.2.7.1.3 Create hegemony code as a BGPstream consumer<br>1.2.7.1.4 Create functionality to annotate which AS paths were used to create a specific inference<br>1.2.7.1.5 Document how to add new sources into catalog |
| 1.2.8 | Tools for Dissemination | Leverage existing work on efficient distribution of scientific data, including the use of multicast to broadcast stream of research traffic to RE net- works. Investigate collaboration options with the National Science Data Fabric, and Pacific Research Platform projects<br>1.2.8.1 Learn about latest big data storage/management technologies |
| 1.3 | Design Infrastructure for Broad Usability | Prototype the infrastructure that makes the collected and curated data accessible and easy to use. |
| 1.3.1 | Data Discovery Tools | Design and prototype data discovery tools that address challenges of finding and evaluating usability of data from the platform. This will include the following sub-tasks<br>1.3.1.1 Bring together industry and academic stakeholders to exchange information on data availability, use and accessibility. Conduct 19 quarterly 2-hour zoom meetings.<br>1.3.1.2 Improve documentation of existing tools and datasets to lower participation barriers and improve user experience. Enhance metadata. Enable user provided recipes to be added. We will collaborate with authors of https://stat.ripe.net/about/ to design and prototype the GMI2S Science Gateway interface (based on existing CAIDA Science Gateway Portal ) that allows selection, joining, processing of subsets of different data sets.<br>1.3.1.3 Prototype GMI3S Data Catalog, starting with catalog.caida.org.<br>1.3.1.3.1 Design and run a cron job updating README files on disk based on the corresponding info in *md files<br>1.3.1.3.2 Add start and end dates for each dataset in catalog<br>1.3.1.3.3 Design and run a cron job that periodically updates all datasets sizes including databases<br>1.3.1.3.4 Mint DOIs for all CAIDA datasets and APIs<br>1.3.1.4 Conduct meetings to discuss approaches to integrating non-CAIDA datasets and tools into GMI3S catalog<br>1.3.1.5 Explore integration of automated meta-data/data citation creation into catalog<br>1.3.1.6 Create metadata databases to increase data |

| | | |
|---|---|---|
| | | accessibility |
| 1.3.2 | Software for Disclosure Controls | Design disclosure control approaches. Identify and evaluate potential software packages that provide privacy protections, such as differential privacy and secure multi-party computation, and their usability for cybersecurity research needs. This will include the following subtasks.<br>1.3.2.1 Understand gaps between privacy-preservation techniques and network and security research needs<br>1.3.2.2 Conduct workshops twice a year with the goal of bringing together diverse experts to find common grounds and develop solutions<br>1.3.2.3 Identify specific gaps that privacy techniques can support.<br>1.3.2.4 Create taxonomy of data, including proprietary data<br>1.3.2.5 Design and prototype repeatable practices to enable legitimate research access to various data types<br>1.3.2.6 Design and prototype authentication/authorization solution that supports both SSO and API/keys |
| 1.3.3 | Policy Tools for Disclosure Control | Investigate existing disclosure control policy frameworks. Identify candidate providers of potential sensitive data. Collaborate on development of sharing agreements. Collect and evaluate sharing agreements from other con- texts.<br>The subtasks include:<br>1.3.3.1 Work with industry partners to compile a list of existing data sharing policies and best practices (starting with Kentik, Farsight, DomainTools, Zvelo, IPinfo, Netacuity, Iconectiv, Censys, CommonCrawl, Telescope, Merit-Telescope)<br>1.3.3.2 Work with the academic researchers (esp. with other NSF-funded projects) to leverage labor and experience starting with NSF-funded projects RSOC (Research Security Operations Center) and CCRI CLASSNET (USC-ISI).<br>1.3.3.3 Work with DoD to compile their needs and requirements<br>1.3.3.4 Work with UCSD OCGA and Rev-Up (https://www.sdsc.edu/services/rev-up.html) to develop agreements to share sensitive data<br>1.3.3.5 Identify lessons learned from previous data-sharing efforts. Notably, the Menlo Report proposed a summary of principles to guide the identification and resolution of ethical issues in information technology |

| | | |
|---|---|---|
| | | research, and a companion report that applies these principles to real and synthetic case studies.<br>1.3.3.6 Investigate how Europe and other parts of the world are approaching this same issue<br>1.3.3.7 Pursue agreements with commercial data providers (Kentik, Farsight) on sharing practices that will not expose them to liability for privacy violations, and that embody the aspiration that what the research community learns from shared data can be valuable to the firm that shares it.<br>1.3.3.8 Design new data sharing agreements based on lessons learned above, and share with community |
| 1.3.4 | Case Studies on Extensibility | Demonstrate Extensibility of Policyframework<br>with case studies. Use first case study to test and refine a policy framework to support additional external data integration, including standardization and quality assurance policies, data use agreements, and legal disclosures.<br>Subtasks include:<br>1.3.4.1 Evaluate manrs_core.ipynb<br>1.3.4.2 Conduct meetings and compile data to create community-authored "State of the Internet report"<br>1.3.4.2 Lead the effort of compiling the "State of the DDoS attacks" report<br>1.3.4.3 Lead the effort of compiling the "State of the DDoS attacks" report<br>1.3.4.4 Case study comparing the darknet dataset use by three entities<br>1.3.4.5 Identify appropriate external datasets and tools, and organizations to include into the case studies<br>1.3.4.6 Conduct case studies, share with community |
| 1.4 | Design Infrastructure for Outreach | Design an infrastructure that enables collaboration and scales the STEM workforce training |
| 1.4.1 | Conduct Workshops | Organize, host, and document results of biannual workshops to develop consensus around the priorities on the data needs and solutions related to the research of the Internet vulnerabilities issues.<br>Conduct quarterly calls for each topic:<br>gmi-traffic (include telescope)<br>gmi-routing (bgp)<br>gmi-topology (ITDK)<br>gmi-dns<br>gmi-ddos<br>gmi-policy (mapping security)<br>gmi-economics<br>gmi-security: review security risks and question and map to data needed |

| | | |
|---|---|---|
| | | gmi-dataops<br>1.4.1.1 GMI-traffic workshops<br>1.4.1.2 GMI-BGP workshops<br>1.4.1.3 GMI-DNS workshops<br>1.4.1.4 GMI-topology workshops<br>1.4.1.5 GMI-Active DDOS workshops<br>1.4.1.6 GMI-Policy workshops<br>1.4.1.7 GMI-Economics Workshops<br>1.4.1.8 GMI-Security workshops<br>1.4.1.9 GMI-Dataops workshops |
| 1.4.2 | Virtual Collaboration Environment | Deploy and host a sustainable virtual collaboration environment that includes channel- based messaging platform (for each of 1.4.1 topics), townhall meetings, BOFs, and lunch-time presentations.<br>1.4.2.1 Virtual collaboration environment launched<br>1.4.2.2 Virtual collaboration environment evaluated and improved (1)<br>1.4.2.3 Virtual collaboration environment evaluated and improved (2)<br>1.4.2.4 Virtual collaboration environment evaluated and improved (3) |
| 1.4.3 | Modules to ScaleSTEM Workforce Development | Develop online course on Network Infrastructure Data Science (NIDS) that will promote the use of the datasets and analytics.<br>1.4.3.1 Conduct monthly calls with Dr. Fraenkel from Halıcıoğlu Data Science Institute.<br>1.4.3.2 Develop online course on Network Infrastructure Data Science (NIDS) that will promote the use of the datasets and analytics.<br>1.4.3.3 Create video tutorials to teach community/public how to deploy and manage nodes |
| 1.4.4 | Conduct quarterly calls with stakeholders | Conduct quarterly calls with all stakeholders |
| 1.4.5 | Present project outcomes at various meetings | Make presentations on various conferences and workshops, submit publications. |
| 1.5 | Project Management | Encompasses creating and maintaining the Project Execution Plan, updating and maintaining the schedule, and project web portal, reporting metrics and management of the Project Leadership Team, and Advisory Steering committee. |

| | | |
|---|---|---|
| 1.5.1 | Project Support | Ensure that all project support mechanisms are in place.<br>1.5.1.1 Design and maintain Jira/Confluence environment for project management<br>1.5.1.2 Periodically review PEP, add tasks into jira |
| 1.5.2 | Project Controls | Maintain and update schedule as milestones/tasks are completed. Develop and submit biannual reports.<br>1.5.2.1 Convene biweekly project status meetings (CAIDA/MIT/UO)<br>1.5.2.2 Convene RV/CAIDA weekly technical meetings<br>1.5.2.3 Document deliverables and responsibilities<br>1.5.2.4 Submit biannual reports to NSF |
| 1.5.3 | Quality Management | Ensure that quality expectations are met. Administer periodic stakeholders' surveys, communicate with NSF and collaborators.<br>1.5.3.1 Compile a list of advisory board and send invitations<br>1.5.3.2 Convene advisory board meetings |

## Updated Level 2 Milestones (PEP Table 11)

| ID | Milestone Description | Completion Month |
|---|---|---|
| 1.1 | Design Infrastructure for Data Acquisition | |
| **1.1.1** | **Preliminary report on Internet infrastructure security vulnerabilities that the GMI project will gather/manage/share data to address** | **M6 (A)** |
| **1.1.2.1** | **Slide deck created** | **M3 (A)** |
| 1.1.2.2 | Complete data needs draft report (based on 1.1.1) | M17 |
| **1.1.2.3** | **Preliminary Data catalog created** | **M9( A)** |
| 1.1.2.4 | Database of Peers (RV) created | M13 |
| 1.1.3 | Monitors Requirements documented, hardware and software needs assessed | M24 |
| **1.1.3.1** | **Inventory of CAIDA machines created** | **M11 (A)** |
| **1.1.3.2** | **Inventory of RV-related hardware created** | **M11 (A)** |
| 1.1.3.3 | Telescope data monitoring needs compiled | M20 |
| 1.1.3.4 | Two-way traffic data monitoring needs compiled | M20 |
| 1.1.3.5 | BGP data monitoring needs compiled | M24 |
| 1.1.3.6 | Active measurements needs compiled | M18 |
| 1.1.3.7 | DNS monitoring needs compiled | M19 |
| 1.1.4 | Monitor specification report | M24 |
| 1.1.4.1 | Draft (1) post monitor hardware specifications report for internal feedback | M18 |
| 1.1.4.2 | Draft (2) for community feedback | M20 |
| 1.1.4.3 | Post final document for public comments | M24 |
| 1.1.5 | Monitor software prototyped | M26 |
| 1.1.5.1 | Telescope data monitor software prototyped | M18 |
| 1.1.5.2 | Two-way traffic data monitor software prototyped | M16 |
| 1.1.5.3 | BGP data monitoring software prototyped | M17 |
| 1.1.5.4 | Active data monitoring software prototyped | M17 |
| 1.1.5.5 | DNS data monitoring software prototyped | M26 |
| 1.1.6 | Monitor deployment prototyped | M30 |
| **1.1.6.1** | **Mailing lists for collaborating partners created** | **M3 (A)** |
| 1.1.6.2 | Telescope data monitoring deployment | M21 |
| 1.1.6.3 | Two-way traffic data monitoring deployment | M16 |
| **1.1.6.4** | **BGP data monitoring deployment** | **M10 (A)** |
| 1.1.6.5 | Active probing measurements data monitoring deployment | M17 |
| 1.1.6.6 | DNS data monitoring deployment | M20 |
| 1.1.6.7 | 10 nodes of at least one measurement deployed | M18 |
| 1.1.6.8 | 10 nodes of multiple measurements deployed | M24 |
| 1.1.6.9 | additional 10 nodes of multiple measurements deployed | M30 |
| 1.1.7 | Evaluation report of Data Acquisition Component | M31 |
| 1.1.7.1 | Software to support active probing measurements deployed | M12 |
| 1.1.7.2 | Third-party experiment deployed | M25 |
| 1.1.7.3 | Evaluation report of Data Acquisition Component published | M30 |
| 1.1.8. | Prototype virtualization capabilities | M30 |
| 1.1.8.1 | Select preferred Network Function Virtualization framework. Package monitor software into this frame. Demonstrate deployment and operation. | M12 |
| 1.1.8.2 | Study and experiment with Network Function Virtualization | M13 |

| | | |
|---|---|---|
| | frameworks for collection nodes and analysis machines started | |
| 1.1.8.3 | Virtualization capabilities documented and evaluated; report published | M30 |
| **1.1.8.4** | **Put all RV collectors on VMs (in conjunction with ILANDS)** | **M11 (A)** |
| 1.1.9 | "Internet infrastructure security vulnerabilities" and "Data Needs" reports integrated into a single report | M24 |
| | | |
| 1.1.9 | "Internet infrastructure security vulnerabilities" and "Data Needs" reports integrated into a single report for increasingly wide review per the following calendar: | M17 |
| **1.1.9.1** | **Draft (1): post for internal feedback** | **M7 (A)** |
| **1.1.9.2** | **Process internal feedback** | **M9 (A)** |
| **1.1.9.3** | **Draft (2): post for internal feedback** | **M9 (A)** |
| 1.1.9.4 | Draft (3): publish for community feedback | M13 |
| 1.1.9.5 | Collect community feedback | M14 |
| | Process community feedback | M14 |
| 1.1.9.6 | Draft (4): publish for public feedback | M15 |
| 1.1.9.7 | Process public feedback | M16 |
| 1.1.9.8 | Publish Final report | M17 |
| 1.2 | Design Infrastructure for Data Management | |
| 1.2.1 | Data storage hardware requirement | M18 |
| 1.2.1.1 | Data storage hardware requirement documented – draft posted for stakeholders' review | M15 |
| 1.2.1.2 | Community feedback incorporated into the document | M18 |
| 1.2.2 | Data storage systems specifications published | M21 |
| **1.2.2.1** | **DevOps Infrastructure group created** | **M3 (A)** |
| 1.2.2.2 | Data storage systems specifications documented– draft posted for stakeholders' review | M17 |
| 1.2.2.3 | Community feedback incorporated into the document | M21 |
| 1.2.3.1 | Develop metadata schema for existing CAIDA datasets | M25 |
| **1.2.3.1.1** | **Create metadata template for ASrank** | **M10 (A)** |
| 1.2.3.1.2 | Create metadata templates for other ongoing data sets | M18 |
| 1.2.3.1.3 | Create metadata templates for the datasets completed in the last 5 years | M13 |
| 1.2.3.2 | Research the state-of-the-art metadata approaches | M13 |
| **1.2.3.3** | **Quarterly Workgroups meetings, reports posted 4 weeks after quarterly meetings** | **M7,10**(A),13,16,19,22,25,28,31,34 |
| 1.2.3.4 | Data and metadata standards specifications published (annual revisions) | M13, M24,M33 |
| **1.2.4.1** | **Report on existing state-of-the-art anonymization tools** | **M12 (A)** |
| 1.2.4.2 | Report on existing pcap on-the fly analysis tools | M21 |
| 1.2.4.3 | Specification of tools for data curation and documentation, report created, annually updated | M18, M30 |

| | | |
|---|---|---|
| 1.2.5.1 | Unified web interface to download heterogeneous datasets designed | M18 |
| 1.2.5.2 | Increase the number of supported data sources, including non-CAIDA datasets | M30 |
| 1.2.5.2.1 | Incorporate datasets used by IYP | M21 |
| 1.2.5.3 | Improve existing API | M17 |
| 1.2.5.4.1 | Data and metadata API documented | M18 |
| 1.2.5.4.2 | Year 2 Update of data and metadata APIs | M24 |
| 1.2.5.4.3 | Year 3 Update of data and metadata APIs | M36 |
| 1.2.6 | SDK libraries developed | M36 |
| 1.2.6.1 | Libipmeta expanded | M18 |
| 1.2.6.2 | New libraries created | M30 |
| 1.2.6.3 | SDK Libraries evaluated, report published | M36 |
| 1.2.7 | Tools for additional data sources integration created | M25 |
| 1.2.7.1 | Create instance of Internet Health Report platform | M25 |
| 1.2.8 | Approaches to dissemination designed, documentation created | M24 |
| 1.2.8.1 | Create report on the latest big data storage/management technologies | M18 |
| 1.3 | Design Infrastructure for Broad Usability | |
| 1.3.1 | Data discovery tools prototyped | M21 |
| **1.3.1.1** | **Quarterly meetings with stakeholders conducted, minutes published** | **M3,6,9,12**,15,18,21 |
| 1.3.1.2 | Documentations of the existing tools and datasets improved | M13 |
| 1.3.1.3 | GMI3S Data Catalog, starting with catalog.caida.org prototyped | M13 |
| 1.3.1.4 | Report on other non-CAIDA datasets and tools integration | M20 |
| 1.3.1.5 | Report on integration of automated meta-data/data citation creation into catalog | M21 |
| 1.3.1.6 | Metadata databases to increase data accessibility created | M16 |
| 1.3.2 | Software for disclosure control developed | M30 |
| 1.3.2.1 | Report on the gaps between privacy-preservation techniques and network and security research needs | M15 |
| 1.3.2.2 | Workshops conducted; notes shared with community | **M6**,14,18,26,30, |
| 1.3.2.3 | Gaps that privacy techniques can support identified, report created and shared | M16 |
| 1.3.2.4 | Report on taxonomy of data, including proprietary data | M15 |
| 1.3.2.5 | At least 2 practices prototyped and evaluated | M25 |
| 1.3.2.6 | Authentication/authorization solution that supports both SSO and API keys prototyped | M15 |
| 1.3.3 | Report on policy tools | M30 |
| 1.3.3.1 | A list of existing data sharing policies and best practices compiled and shared with community. This will include Kentik, Farsight, DomainTools, Zvelo, IPinfo, Netacuity, Iconectiv, Censys, CommonCrawl, Telescope, Merit-Telescope | M18 |
| **1.3.3.2** | **Biannual meetings with at least two NSF-funded projects** | **M6**,12,18,24,30 |
| 1.3.3.3 | List of the DoD needs and requirements compiled | M18 |
| **1.3.3.4** | **At least one agreement created and evaluated** | **M6** |
| 1.3.3.5 | Lessons learned identified and documented | M18 |
| 1.3.3.6 | Report on other countries' approaches | M18 |
| 1.3.3.7 | Agreements with at least two commercial data providers put in place | M21 |
| 1.3.3.8 | New agreements designed and shared | M25 |

| | | |
|---|---|---|
| 1.3.4 | Extensibility case studies documented and shared | M30 |
| 1.3.4.1 | manrs_core.ipynb evaluation report | M18 |
| 1.3.4.2 | Quarterly meetings conducted; minutes shared | M15,18,21,24 |
| 1.3.4.3 | State of Internet report created and shared | M24 |
| 1.3.4.4 | Case study comparing the darknet dataset use by three entities documented and shared | M18 |
| 1.3.4.5 | Appropriate external datasets and tools, and organizations to include in the extensibility case studies identified, documented | M21 |
| 1.3.4.6 | Case studies conducted, documented and shared | M30 |
| 1.4 | Infrastructure for Outreach | |
| 1.4 | Infrastructure for outreach created | M30 |
| **1.4.1** | **Meetings conducted. Minutes shared** | **M4,7,10**,13,16,19,22,25,28,31,34 |
| **1.4.2.1** | **Virtual collaboration environment launched** | **M3 (A)** |
| 1.4.2.2 | Virtual collaboration environment evaluated and improved | M12,24,36 |
| 1.4.3.1 | Modules to scale STEM workforce developed | M24 |
| 1.4.3.2 | Online course on NIDS developed | M24 |
| 1.4.3.3 | Video tutorials on nodes deployment and management created | M18 |
| **1.4.3.4** | **Quarterly calls conducted, minutes shared** | **M4,7,10**,13,16,19,22,25,28,31,34 |
| **1.4.3.5** | **At least 2 presentation each year** | **M12**,M24,M36 |