Semantic Unified Access to Current and Future Internet Measurement Infrastructures

Ángel Ferreiro[1], Thomas Fichtel[2], Jorge E. López de Vergara[3], Peter Mátray[4], Felix Strohmeier[2], Giuseppe Tropea[5], Udi Weinsberg[6]

[1] Telefónica Investigación y Desarrollo, Madrid, Spain, olivo@tid.es

[2] Salzburg Research Forschungsgesellschaft mbH, Salzburg, Austria

[3] Dept. Ingeniería Informática, Universidad Autónoma de Madrid, Madrid, Spain

[4] Department of Physics of Complex Systems, Eötvös University, Budapest, Hungary

[5] Consorzio Nazionale Interuniversitario per le Telecomunicazioni, Rome, Italy

[6] School of Electrical Engineering, Tel-Aviv University, Israel

Abstract. The deployment of various new network services generates a continuously increasing amount of traffic in IP networks. This process has alerted network operators and service providers and led them to search for tools to measure and control such IP traffic. No realistic support of Quality of Service (QoS) can be envisaged without an efficient traffic monitoring system. Until now numerous different projects have installed instruments to collect a vast amount of data about various Internet traffic characteristics. The data they capture and store are of great interest for network research and network management purposes, however no unified platform serves them so far. The sharing of network data in a standardized format under an integrated and user friendly interface would stimulate the field, and enable the emergence of more comprehensive studies. This paper describes the architecture and key features of a semantic unification layer which aims to bridge this gap and is being developed as the main component of the MOMENT project.

Key words: Future Internet, Network Monitoring, Unified Interface, Network Ontology, QoS.

1 Introduction

The rapid growth of the Internet, with new bandwidth demanding services that require dedicated quality (QoS), was confronted by increasing bandwidth, mainly in the access segment. In the Future Internet, this must be replaced by smart management of the existing resources. Sophisticated applications such as high-quality live video delivery, huge P2P overlays and distributed computation, partially using encrypted traffic, translate into strict requirements for network operators and service providers[7].

[7] NOBEL (http://www.ist-nobel.org/Nobel2/servlet/Nobel2.Main) and MUSE (http://www.ist-muse.org) projects have dealt with the proper classification of network services.

Since the Internet topology and traffic flows are not completely analyzed and understood, it is increasingly difficult for service providers to guarantee QoS and protect against even simple Denial of Service (DoS) attacks. Furthermore, network operators often lack the information required to optimize their networks [1] and to tune its behavior so that traffic engineering becomes effective and network control is pervasive, failures can be monitored or avoided, and Service Level Agreements are met.

To gain the needed insight, several organizations have deployed measurement infrastructures to track the Internet at various levels of granularity in an attempt to better understand its structure and dynamic behavior. Having different objectives, they employ different measurement techniques and publish data in distinct formats. Researchers, operators and other market agents using these data sources for combined queries face major difficulties due to this heterogeneity. If the distinct datasets could be cross-matched and analyzed together, it would be possible to propose more comprehensive studies, and the power of merging data that comes from different infrastructures deliver improved inputs for modeling the traffic and topology of the Internet. Currently, there is no unifying tool that provides researchers and commercial organizations the ability to simultaneously query, merge and analyze the results from more than a single infrastructure.

This paper presents a novel and modular architecture, based on the latest semantic technologies and currently developed within the MOMENT[8] project, providing a unified interface for *generic heterogeneous* Internet measurement infrastructures and data repositories. The introduction of a network measurement-oriented ontology allows the system to overcome the vast differences between infrastructures. The unified interface relies on a *Mediator* to which a user conveys semantic queries. The mediator requests all available data by browsing each repository and can transform, obfuscate and normalize data to common formats, prior to its release. Such capabilities help shifting the focus from the tedious task of data collection to effective data analysis.

2 State of the art

Standardizing network measurement protocols is a key point to provide a common understanding of measures across different networks and organizations. In this Section we focus on the work of IETF, OGF and CAIDA.

IETFs most flexible framework for exchange of network flow information is provided by the IPFIX [2] working group. It defines a protocol to transmit information about captured flows[9] to a collector. In the context of SNMP, the family of RMON MIBs [3] is another source of network traffic measurements and statistics, ranging from the data link to the application layers. Another IETF working group, PSAMP (Packet Sampling) [4], defines a standard set of capabilities for network elements to sample subsets of packets by statistical and other methods. Standardization of metrics is the goal of another IETF group,

[8] See `http://www.fp7-moment.eu`

[9] In the IPFIX context, a flow is defined as a set of packets having common properties

the IPPM (IP performance metrics) [5], that has developed a set of quantitative unbiased standard metrics applied to the quality, performance, and reliability of Internet data delivery services. Connectivity, one-way delay and loss, round-trip delay and loss, delay variation, loss patterns, packet reordering, bulk transport capacity and link bandwidth capacity are standardized metrics.

Open Grid Forum's Network Measurement Working Group (OGF NMWG) has developed a message format for the communication among different network measurement systems [6]. Basically, their work is focused on creating a common vocabulary used to provide information about different measuring tools. Actually, a set of XML Schemas have been defined for each of these tools using RELAX NG (Regular Language for XML Next Generation) [7], and the information is sent in XML code defined by those schemas. Related to OGF NMWG, the PerfSonar project [8], is developing a monitoring architecture designed to allow autonomous network operators to create measurement tool daemons, open to the world but governed by locally-defined policies and limits. It allows discovering measurement tool daemons along the end-to-end path, and facilitates the use of federated trust models. The design is service-oriented and decentralized, scalable and has minimum administrative overhead. The objective of such design is to enable users to perform network monitoring and measurements in a multi-domain environment.

In 2002 US-based CAIDA began to develop the Internet Measurement Data Catalog "DatCat" [9], a system which serves the global network research community by allowing anyone to find, annotate, and cite data contributed by others. DatCat does not store real data, only metadata, i.e. descriptions of the data and instructions for obtaining them. In fact, it is a mediator that does not dictate the terms of availability of the data, but helps you with the first step of finding data. In 2005 the MOME project launched the MOME database [10], an advanced measurement meta-repository for network monitoring tools and data.

The MOMENT project approaches the goal of interfacing users and measurement tools having in mind a unified ontology and using semantics in the core of the mediator engine.

3 Ontologies for Network Measurement Infrastructures

An ontology [11] provides a vocabulary of classes and relations to describe a domain, stressing knowledge sharing and knowledge representation. The advantages of using ontologies are manifold: the ontology can be downloaded from the web and read by anyone freely, the information is modeled in a more flexible way than using tables; their semantic definition of information enables a classification of knowledge (e.g. a tool that performs active measurement is an active tool) and inference (e.g. if a measurement is over a threshold then the network is overloaded); at the same time it is possible to query this knowledge (e.g. obtain all measurements with a given destination address).

Ontologies have been used in other information integration problems, including network management [12]. In that work, ontologies are proposed as a way to solve the heterogeneity of network management information models, following a methodology that merges all information in a single model, providing mappings from that new model to the old ones. The same solution is being applied in MOMENT, where several ontologies have been defined [13].

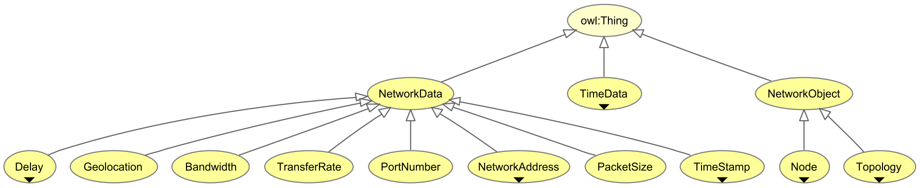

Fig. 1. A fragment of the ontology

First of all, it was necessary to have a state of the art of existing monitoring tools. For this, an ontology was created, which contains about one hundred classes. A central class, MonitoringTool, has been defined to describe the information that characterizes such tools. The rest are mostly used as a taxonomy of measurement, input/output and control data, communication paradigm, platform, license or filter. This taxonomy can be later used for the classification of tools. This ontology was populated with available information, obtaining a knowledge base of monitoring tools[10].

Then, other ontologies have also been defined to provide several functionalities. A metadata ontology has been specified to describe the repositories and measurements available to the mediator. It includes all necessary information to access to these repositories: where they are and what they contain. This ontology can be processed to query the information that a concrete user requests. For its specification, several sources have been used: existing metadata schemas, such as DatCat or MOME, and domain ontologies such as W3C Time or FOAF.

An upper ontology of measurements has also been defined. This upper ontology contains the concepts of the different measurements (see Figure 1 for the uppermost layers). This network data ontology has been obtained in a process in which, on the one hand, all existing data sources in the MOMENT project have been analyzed[11], taking their schemas to obtain a common one. On the other hand, existing common schemas such as those proposed by OGF NMWG have also been used to generate the vocabulary. It should be noted that this upper ontology does not have any measurement instances. Those instances are contained in existing repositories (implemented, for instance, as relational databases), simplifying the integration process. Joint with the upper ontology it is also necessary to have a mapping ontology. This ontology is needed to translate semantic queries

[10] This knowledge base is available at `http://www.fp7-moment.eu/images/tools_final.owl`

[11] The schemas describing the data sources are available at `http://www.fp7-moment.eu/index.php?option=com_content&view=article&id=81&Itemid=61`

into data queries, specifying which information in a concrete repository map to which classes and properties in the upper ontology.

Besides the ontology of measurements, a taxonomy of possible anonymization approaches has been conceived. The main effort was focused on the dependencies between the possible anonymization strategies and the role and purpose of data consumers that query data from the mediator. The design and identification of all possible values for the various anonymization components of the ontology was the most important task, because they represent the "common vocabulary" that shall fit all possible context. This is the minimum superset of concepts and/or functionalities that describes the "anonymization of network measurements" domain. As well known from the scientific literature ([14–16]), generic anonymization schemes are difficult to design, since different organizations have different needs. Anonymization requirements vary a great deal form environment to environment: in some contexts IP addresses are highly sensible while in other obscuring the type of traffic could be the major need.

4 Mediator

The main concern for an integrated interface to different monitoring and measurement infrastructures which store their data in arbitrary formats is to be flexible in order to provide solutions for three different problems of heterogeneity: 1) heterogeneous data semantics, 2) heterogeneous data access, and 3) heterogeneous data formats. First, the semantics of data from mediated infrastructures needs to be unified, as different measurement infrastructures use different names (e.g. delay, one-way-delay or owd) and units (e.g. second, millisecond, nanosecond) for the same things. Second, data access is provided by different means (e.g. direct SQL, FTP or Web Services). Finally the data format can differ between infrastructures and needs to be harmonized (e.g. XML, binary or CSV).

The MOMENT middleware design also takes into account the importance of preserving confidential data of Internet users. Controlling the processing of data prior to releasing them is a strong requirement for the mediation. Thus, aside from the obvious mechanism to characterize the mediator users and assigning them different permissions, a more sophisticated procedure is needed. This is necessary because user roles and their data usage-patterns are tightly connected to anonymization strategies that must be applied to the data. Being as open as possible for every kind of data consumer is the goal of the mediation layer, although without breaking the obfuscation rules suggested by the data issuer.

We present in [17] a list of use cases that the system architecture we design throughout the next sections is able to deal with. They detail the various system functional requirements as seen from the user-perspective, which the mediator covers by exploiting the modular design of its core middleware and its Unified Interface.

4.1 Architecture

To get acquainted with the mediator building blocks, Figure 2 provides the high-level architecture followed by short descriptions. The infrastructures that are indicated as “Mediated Infrastructures” are current or future measurement infrastructures, that are willing to share their data using the MOMENT mediator. The selected examples are typically used access interfaces to their data. Many of the studied measurement infrastructures store their data in SQL databases, some of them provide a web service interface using REST or SOAP, and others already have semantically enriched data and provide them via a SPARQL interface. As a final example measurement data accessible via FTP is mentioned. Goal of the MOMENT mediator is to allow the integration of all different kinds of data sources from very straightforward up to very sophisticated.

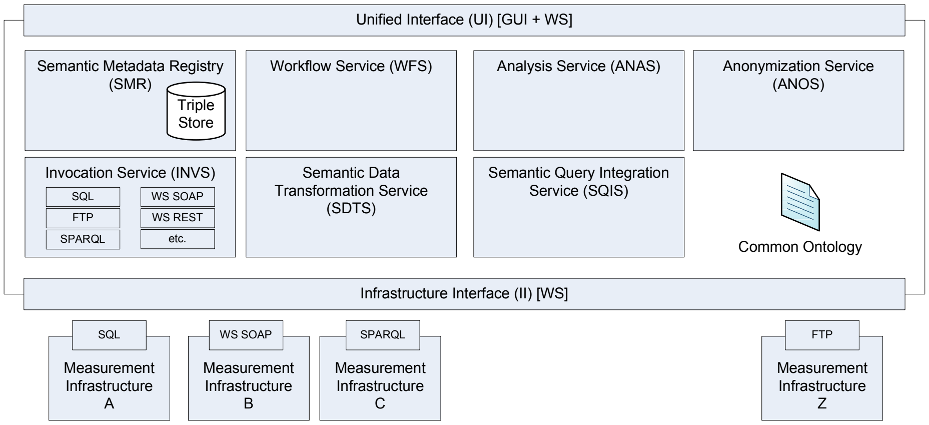
Unified Interface (UI) [GUI + WS]
Semantic Metadata Registry (SMR)
Triple Store
Workflow Service (WFS)
Analysis Service (ANAS)
Anonymization Service (ANOS)
Invocation Service (INVS)
SQL
WS SOAP
FTP
WS REST
SPARQL
etc.
Semantic Data Transformation Service (SDTS)
Semantic Query Integration Service (SQIS)
Common Ontology
Infrastructure Interface (II) [WS]
SQL
Measurement Infrastructure A
WS SOAP
Measurement Infrastructure B
SPARQL
Measurement Infrastructure C
FTP
Measurement Infrastructure Z

Fig. 2. Block diagram showing the main components of MOMENT mediator

The *Unified Interface (UI)* provides access to the *Mediator* for both, humans through a Graphical User Interface (GUI) and machines via web service (WS). The first one enables the user to browse or query the metadata registry, execute measurement data queries, define and execute workflows, set a data analysis method or define a data transformation; the second one allows users to invoke the individual mediator components to execute a workflow. Obviously, authentication and differentiated treatment of users are features supported by this component as well as the interfacing to other modules so as to compose the queries and present the results. It also presents the facilities to record the usage of the mediator, keep a repository of its functionality and help beginners to use it as well as enable fora about its results, namely monitoring the Internet.

The *Semantic Metadata Registry (SMR)* is a central component of the system. Its innovative approach is based on it since it stores the semantically enhanced information based on the ontology model, including measurement infrastructures description (metrics it deals with, data format, period of a measurement, etc.), data about workflows (its description, involved components, etc.) and data about analysis methods (objective of method, input and output format). All data in the metadata registry are based on a common ontology and therefore semantically interpretable. The metadata registry also stores information about the method to access measurement data in a mediated infrastructure. This information is passed on to the invocation service, e.g. the SQL URL, username, password and query.

The *Semantic Query Integration Service (SQIS)* is the module responsible for executing queries that access infrastructures directly, allowing a fine grained search for measurement data stored in query-enabled sources. These queries are requested by an application on top of the unified interface using semantic technologies such as SPARQL and the answer is integrated from existing sources, using a common vocabulary taken from the network data ontology.

The *Analysis Service (ANAS)* is responsible for further analysis of measurement results. This service can process data from one ore more infrastructure(s). Information about available analysis services is stored in the metadata registry.

The *Anonymization Service (ANOS)* provides methods to anonymize measurement data in order to preserve privacy. The service can be invoked if anonymization of measurement data is necessary and it can be also invoked, if needed, in those situations when data has already been anonymized, for internal policy reasons, directly at the source. This might be useful, and also needed, depending on the final user rights and privileges within the mediator. Measurement infrastructures that already have a well structured privacy policy will map their policies onto the anonymization ontology so that the correct strategy will be enforced over each mediated user. This module will also represent the interface with existing platforms that don't have encoded specific policies for data release, providing default mappings and policies based on semantic reasoning.

The *Workflow Service (WFS)* helps to automate recurring processes. In a workflow a user can define a sequence of MOMENT operations, e.g. get data from infrastructure A and B and analyze this data with method X.

The *Invocation Service (INVS)* is used when a user wants to retrieve specific measurement data from an infrastructure. To retrieve specific measurement data a user first has to query or browse the registry in order to identify the data he wants. For this data the mediator can than retrieve the reference to the data and pass it on to the invocation service. Depending on the type of the reference (SQL, WS, FTP etc.) the invocation service can instantiate an appropriate handler for the reference and retrieve the data from the mediated infrastructure.

The *Semantic Data Transformation Service (SDTS)* is able to transform the data in cases, where measurement data provided by an infrastructure is not exactly the data a user or application requires, e.g. transforming measurement units.

Finally, the *Infrastructure Interface (II)* infrastructure interface is mainly responsible for receiving semantically enhanced metadata information from the infrastructures. It provides a web service (WS) interface to push their metadata into the system in order to make their measurement data findable and accessible. Infrastructures can publish their metadata in RDF triples which are defined in the MOMENT ontology.

4.2 Functional description of the mediator

In this section we describe a selected service orchestration that can be performed by the MOMENT mediator. Figure 3 shows the process of requesting measurement data from several measurement infrastructures using the Unified Interface, and combines the results. This exploits two key features of the MOMENT mediator, which is first to hide complexity of a measurement infrastructure from the end user, and second to get value added information from multiple measurement infrastructures compared to a single one. In the given figure SPARQL is used as query language, but the same queries can be provided via user-friendly web interfaces.

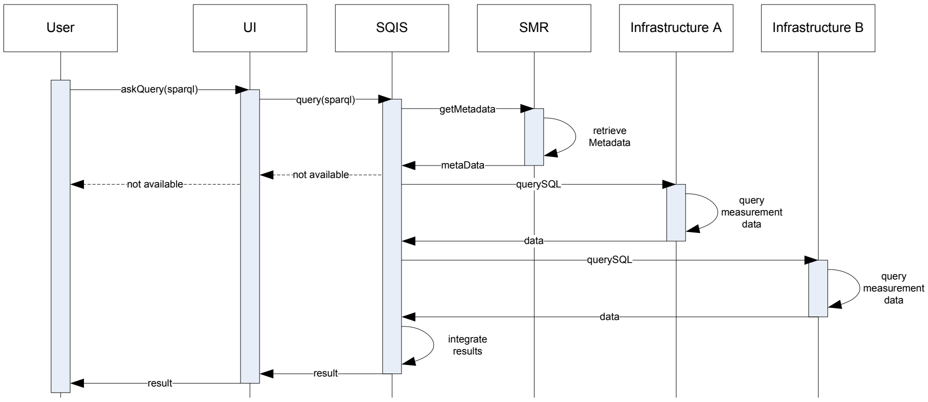

Fig. 3. Retrieving integrated measurement results from multiple infrastructures

Therefore the user queries the results he wants to retrieve from the UI, that forwards the request to the SQIS. Subsequently, the SQIS asks the SMR if according metadata is available. Depending on the metadata, the infrastructures hosting the parts of the requested measurement results are contacted (e.g. via SQL). The SQIS combines these replies into the integrated result, which is given back to the user. On several places during this process, optionally the results can be sent through the ANOS module, to be obfuscated. Upon each obfuscation, a skeletal Acceptable Use Policy document is generated by the ANOS to be shown to the user. It represents an informative document, although structured, about what the provider expects from the user regarding the usage of the data that the provider itself is willing to release. It can be accepted and signed by the user. Although this does not imply any technical enforcement by the mediator, it can be regarded as a kind of End User Legal Agreement.

5 Conclusions and Future Work

This paper presents a novel architecture for a semantic unifying interface, the MOMENT *Mediator*, to access Internet measurement infrastructures. To resolve the heterogeneity problem of the different data sources, a common ontology was designed, that includes several semantic building blocks for monitoring tools, measurement data, metadata, and anonymization issues. The unified access allows network operators and researchers to easily merge network data originating from different infrastructures, enabling complex analysis of available network data to gain a more complete view of the evolving Internet behavior. A sophisticated anonymization service is embedded into the architecture, that guarantees the appropriate respect to privacy laws.

The current design and implementation of the Mediator provides a basis for an even more advanced system, that enables not only unified access to data of various measurement infrastructures, but also a method for requesting new measurements for authorized entities. This establishes new possibilities for standard QoS references and mechanisms to govern the Future Internet. It is not only a powerful solution to access multiple data sources but allows to cross check different measuring infrastructures and control their behavior. Further work will be carried out in the MOMENT project to deploy such complex achievements.

Acknowledgments. The authors thank the partial support of the EU ICT MOMENT Collaborative Project (Grant Agreement No.215225). P. Mátray thanks the partial support of the National Office for Research and Technology (NAP 2005/ KCKHA005) and the ELTE eScience RET.

References

1. Mahajan, R., Wetherall, D., Anderson, T.: Understanding BGP misconfiguration. In: ACM SIGCOMM Computer Communications Review. (2002) 3–16
2. Claise, B.: Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of IP Traffic Flow Information. RFC 5101 (Proposed Standard) (January 2008)
3. Waldbusser, S., Cole, R., Kalbfleisch, C., Romascanu, D.: Introduction to the Remote Monitoring (RMON) Family of MIB Modules. RFC 3577 (Informational) (August 2003)
4. Zseby, T., Molina, M., Duffield, N., Niccolini, S., Raspall, F.: Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of IP Traffic Flow Information. draft-ietf-psamp-sample-tech-11.txt (July 2008)
5. Paxson, V., Almes, G., Mahdavi, J., Mathis, M.: Framework for IP Performance Metrics. RFC 2330 (Informational) (May 1998)
6. Zurawski, J., Swany, M., Gunter, D.: A scalable framework for representation and exchange of network measurements. In: Proc. 2nd International IEEE/Create-Net Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities, Tridentcom 2006. (2006)
7. van der Vlist, E.: RELAX NG. O'Reilly Media, Inc. (2003)

8. Hanemann, A., Boote, J., Boyd, E., Durand, J., Kudarimoti, L., Lapacz, R., Swany, M., Trocha, S., Zurawski, J.: Perfsonar: A service oriented architecture for multi-domain network monitoring. Lecture Notes in Computer Science **3826** (December 2005) 241–254
9. Shannon, C., Moore, D., Keys, K., Fomenkov, M., Huffaker, B., Claffy, K.: The internet measurement data catalog. ACM SIGCOMM Computer Communications Review (CCR) **35**(5) (October 2005) 97–100
10. Gutierrez, P.A.A., Bulanza, A., Dabrowski, M., Kaskina, B., Quittek, J., Schmoll, C., Strohmeier, F., Vidacs, A., Zsolt, K.S.: An Advanced Measurement Meta-Repository. In: Proceedings of 3rd International Workshop on Internet Performance, Simulation, Monitoring and Measurement; IPS-MoMe 2005, Warsaw, Poland (March 2005)
11. Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge Acquisition **5**(2) (1993) 199–220
12. López de Vergara, J.E., Villagrá, V.A., Asensio, J.I., Berrocal, J.: Ontologies: Giving semantics to network management models. IEEE Network **17**(3) (2003) 15–21
13. López de Vergara, J.E., Aracil, J., Martínez, J., Salvador, A., Hernández, J.A.: Application of ontologies for the integration of network monitoring platforms. In: Proc. 1st European Workshop on Mechanisms for Mastering Future Internet. (2008)
14. Alllman, M., Paxson, V.: Issues and etiquette concerning use of shared measurement data. In: IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, New York, NY, USA, ACM (2007) 135–140
15. Koukis, D., Antonatos, S., Antoniades, D., Markatos, E., Trimintzios, P.: A generic anonymization framework for network traffic. Communications, 2006. ICC '06. IEEE International Conference on **5** (June 2006) 2302–2309
16. Paxson, V.: Strategies for sound internet measurement. In: IMC '04: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement, New York, NY, USA, ACM (2004) 263–271
17. Aracil, J., Salvador, A., Martínez, J., López de Vergara, J., eds.: Requirements for MOMENT. http://www.fp7-moment.eu/images/FP7-MOMENT-WP2-D2.3.pdf (June 2008)