A SEMANTICALLY DISTRIBUTED APPROACH TO MAP IP TRAFFIC MEASUREMENTS TO A STANDARDIZED ONTOLOGY

Alfredo Salvador[1], Jorge E. López de Vergara[1], Giuseppe Tropea[2], Nicola Blefari-Melazzi[2], Ángel Ferreiro[3], Álvaro Katsu[1]

[1]High Performance Computing and Networking Research Group, Universidad Autónoma de Madrid, Madrid, Spain, {alfredo.salvador, jorge.lopez_vergara}@uam.es
[2]Consorzio Nazionale Interuniversitario per le Telecomunicazioni, Rome, Italy, giuseppe.tropea@cnit.it
[3]Telefónica Investigación y Desarrollo, Madrid, Spain, olivo@tid.es

ABSTRACT

Traffic monitoring in IP networks is a key issue for operators to guarantee Service Level Agreement both to their clients and with regards to other connectivity providers. Thus, having efficient solutions for traffic measurement and monitoring supports a good deal of their business and it is essential to fair development of Internet. However, even if service management is well recognized, QoS strategies must evolve from circuit switching technological framework towards next generation networks and convergent services concepts. Standardizing IP traffic measurement is a requirement for interoperable service aware management systems upon which future Internet business would be based.

A few projects have recently tackled the task of building rich infrastructures to provide IP traffic measurements. The European project MOMENT approach combines SOA and semantic search concepts: a mediator between clients and measurement tools has been designed in order to offer integrated access to the infrastructures, regardless their specific details, with the possibility of achieving complex queries. Pervasiveness of ontologies has been used for various purposes in the project. As such, one ontology deals traffic measurement data, another one describes metadata that is used instead of data for practical reasons, a third one focuses on anonymization required for ethical (and legal) restrictions and the last one describes general concepts from the field. This paper outlines the role of these ontologies and presents the process to achieve them from a set of traffic measurement databases as well as the integration of specific modules in the mediator to achieve the semantic queries.

KEYWORDS

Ontology, Network Management, Anonymization, Middleware, TMA, Semantic Searching, MOMENT project, Network measurements

1. INTRODUCTION

Service Level Agreements are defined in Quality of Service terms so efficient solutions for traffic measurement and monitoring available for clients and providers are desirable. However, network measurements heterogeneity, in terms of data representation, has been previously reported. Therefore, using a standardized common vocabulary would provide the basis for interoperable service management and network monitoring.

This paper presents the work done to develop an ontology of IP traffic measurements, within the project MOMENT [1], to implement a mediator, or unified interface, to heterogeneous measurement infrastructures and data repositories. Its design is based on a Service Oriented Architecture (SOA); the introduction of the traffic measurement ontologies enables the system

to overcome the differences between the accessed measurement tools and sets a taxonomy of possible anonymization treatments.

The advantages of using ontologies, which provide a vocabulary of classes and relations to describe a domain, stressing knowledge sharing and knowledge representation [2], are then manifold as the design of a mediator for traffic monitoring infrastructures concerns:

- The ontology can be downloaded from the web and read by anyone freely.
- The information is modelled in a more flexible way than using relational tables.
- The 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 for a given destination address). So, all functional requirements can be fully accomplished by a proper set up and design of ontologies. Besides, the ontologies have been used to enable internal tasks in the mediator. The mediator's architecture relies on the ontologies, not only for the semantic interface, but also, for example, for the anonymization of data. Figure 1 depicts the structure of the mediator where the role of common ontologies looks clear as well as its semantic nature [3]: a practical integration and correct functioning of modules such as SDTS (Semantic Data Transformation Service), SMR (Semantic Metadata Registry), SQIS (Semantic Query Integration Service), ANAS (Analysis Service), and ANOS (Anonymization Service) to perform complex requests (from more than a single measurement infrastructure) would not be possible without those common ontologies.

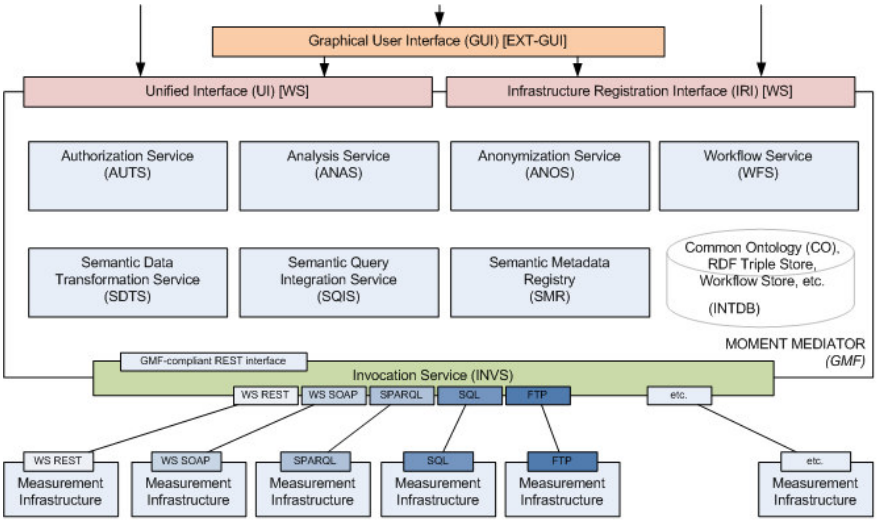


Figure 1. MOMENT mediator architecture.

MOMENT proposal for traffic measurement data and metadata ontologies takes into account previous works (eg. OGF-NMWG [4], W3C Time [5], NASA Units [6]), which were adapted to the mediator requirements. Thus, a core ontology has been built from them and from all databases schemas the mediator connects to. The metadata ontology describes the available repositories and measurements: where they are and what data they contain. It was derived mainly from metadata schemas, such as CAIDA's DatCat [7] or MOME [8] project. The data ontology contains the concepts of measurements. Common schemas, such as the one proposed by OGF NMWG/Perfsonar were used as a starting point, but an exhaustive analysis of all measurement tools reached by the MOMENT mediator was carried out to build it. On the other hand, for anonymization purposes, a new ontology had to be specifically created. 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 contexts. This is the minimum superset of concepts and/or functionalities that describes the "anonymization of network measurements" domain.

This paper intends to show the work accomplished by means of both heuristic and systematic methodology to merge all information in a single model, providing mappings from previous (heterogeneous) models to the new one. After considering the importance of standardizing the achieved ontology, chapter 3 describes the MOMENT ontologies in detail and chapter 4 is dedicated to describe the mapping process and illustrates the way ad hoc tools have been used. The relevance of the work described in this paper about mapping measurement tools to MOMENT ontologies is easy to understand because these ontologies will not be adopted if the developers of measurement tools have to make additional efforts to use the ontologies. The mapping procedures described in section 4 solve this potential problem.

2. STANDARDIZATION REQUIREMENT

Standardizing network measurement parameters, algorithms and protocols is the basis of fair SLA (Service Level Agreement), and hence future business in Telecommunications. Without a common understanding of measurements across different networks and organizations, QoS parameters could not be exchanged through different network domains. Hence, the problem of exploiting extended TMA (Traffic Monitoring and Analysis) infrastructures is not simply developing systems for conversion (values) but, in fact, agreeing on the meaning of measurements and protocols to obtain them. In fact, the task of reaching TMA standard tools has been tackled by different organizations:

The IETF IPFIX [9] working group defined a protocol to transmit information about captured flows; in the context of SNMP, the family of RMON MIBs [10] 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) [11], defined a standard set of capabilities for network elements to sample subsets of packets by statistics and other methods. Standardization of metrics is the goal of the IPPM (IP performance metrics) [12], that has developed a set of quantitative unbiased standard metrics applied to the quality, performance, and reliability of Internet data delivery services. The Open Grid Forum Network Measurement Working Group (OGF NMWG) has developed a message format for the communication among different network measurement systems [13]: a common vocabulary, used to provide information about different measuring tools, has been worked out; actually, a set of XML Schemas has been defined for each of these tools using RELAX NG (Regular Language for XML Next Generation) [14], and the information is sent in XML code defined by those schemas.

The work of TMA projects has also contributed to clarify the situation by practical implementations:

PerfSonar project [15] has developing a monitoring architecture designed to allow network operators the creation of measurement tool daemons; it is open to the world but governed by locally-defined policies and limits. CAIDA project has developed the Internet Measurement Data Catalog (DatCat), 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. Recently the MOME [16] project launched its database, a measurement meta-repository for network monitoring tools and data.

Under the MOMENT perspective, the standardization of its ontologies is considered basic not only for practical reasons but also to set a universal vocabulary, common formats and units to define interoperable TMA mechanisms that allow to use semantic tools and progress in QoS evaluation.

A preliminary ISG (Industrial Specification Group) in the European Telecommunications Standards Institute (ETSI), named MOI (Monitoring Ontology for IP traffic [17]), has been recently approved to start developing all the fundamental ontologies required to set up IP traffic measurement data exchange protocols and formats of data bases to be queried. As an ISG, MOI is supported by ETSI members but open to other participants as established in its Terms of Reference document. Currently, MOI work plan is defined by work items, namely:

- Report on information models for IP traffic measurement
- Requirements for IP traffic measurement ontologies development
- IP traffic measurement ontologies architecture.

To develop such work, MOI seeks for future cooperation with other standardization bodies like OGF (NMWG), IETF, (IPPM, IPFIX), ITU-T (Group 12) and TMForum (QoE groups) and intends to publish a first Group Specifications (deliverable) before 2010: The agenda of the ISG MOI is available in the ETSI portal [18]. As a pre-standardization body, MOI has to submit its Group Specifications to the ETSI technical board so that the interest of continuing the work in a standardization group can be understood. The future work of MOI will cover:

- A definition of the ontologies proposed to be standardized as they are presented in this paper
- A study of bridging metadata and data ontologies for IP traffic measurement to other standardization groups. Coming back to the first deliverable, this will analyze the possibilities of seamless spreading the defined ontologies thought adjacent fields in the development of Future Internet and next generation networks management. The tools presented in this paper will be presented for consideration to those fora.

The interest of this effort is clear: Once the traffic parameters be practically mapped to the ontology classes and their relationships, it should not be difficult to define complex KPI, QPI and BPI (Key, quality and business performance parameters, respectively). Then the development of interoperable systems for QoE-driven network management will arise for the development and benefit of (see Figure 2):

- Applications supervision systems, for service providers
- Network operation driven by QoS
- Machine learning instruments for correlating QoS to QoE

- Internet governance as far as multi-domain network supervision concerns
- Regulatory authorities since they cannot afford issuing fair rules upon unclear parameters definitions
- SLA between network operators (incumbent and virtual) and between them and service providers, for similar reasons

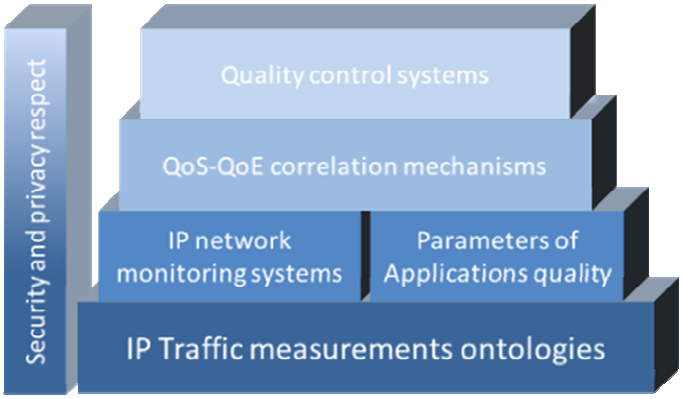


Figure 2. IP traffic ontology supporting QoS for SLA and network operation

3. THE MOMENT ONTOLOGIES FOR TMA

Traffic monitoring analysis can be performed by means of either, semantic and non-semantic solutions. Some non-semantic solutions to network measurement tools heterogeneity, such as MOME, DatCat and PerfSONAR, are focused on measurement data representation and measurement data storage. Although MOMENT approach consists of using semantic solutions, previous achievements by different projects have been taken into account: thus, PerfSONAR utilization of message communication protocol (NMWG schema) highlighting the difference between measurement data and measurement metadata was combined with the constraints of MOMENT project (to avoid direct access to data sources while retrieving information from all measurements at a time for very large data sources) motivated the division of the ontology into four differentiated parts: Upper Ontology, Data, Metadata and Anonymization.

MonONTO [19] is a similar semantic approach which also uses DatCat and NMWG as sources of inspiration. However MonONTO goal is creating a knowledge base for the networking area to advise the user about the performance of the network. MOMENT ontology goals also comprise the creation of a standard to integrate current and future measurements, allowing researchers to define new semantic queries and inference rules to extract more complex metrics from the network.

The ontology must take into account all the measurements from the data sources and provide an extensible and flexible framework for new applications to adopt it easily. Thus the ontology must describe very general concepts from network measurements such as delay or network segment but it also needs to be able to represent the particular measurements from the project partners such as statistical measurement analysis or GPS (Global Positioning System) logs.

A first version of the ontology [20] was developed in MOMENT but the hierarchy of network measurements proposed was found to be very strict and establishing a mapping to data sources was very complicated and resulted in complex associations. Also, it did not take into account some of the modules in the architecture which could benefit from the unified layer of information the common vocabulary provides.

The new version of the ontology [21] is structured in four parts, the Data ontology which describes the hierarchy of the different kind of measurements and their relations to the physical components of the network they measure; the Metadata ontology which describes all the information about how the measurements are stored, how and when they were measured, etc; Additionally all the common concepts from those ontology parts such as time, units and context are described in an Upper Ontology and finally, the Anonymization ontology describes dependencies between the possible anonymization strategies that data have to undergo, prior to being released to the user requesting them, and the role and purpose of such a user within the community of people interested in network measurements. [*]

3.1 The Upper Ontology

The Upper Ontology provides a collection of necessary concepts used in the Data and Metadata ontologies, such as units, time, location and network protocols.

The units ontology has replaced all the units from the physics field in the NASA version, with units from the computer sciences like bit Byte, bps, IP addresses, etc. Also bearing in mind the final use of the ontology which is mapping heterogeneous data-sources different units for different network address representations have been designed and also the transformation properties between them are no longer numerical factors like in the International System units but are regular expressions to match and perform the transformation of the unit.

Network protocols are defined for Transport and Application layers because the ontology is about IP measurements only.

Also the Upper Ontology describes the context concept for measurements in the class *Location*. This class represents a location, physical or logical, for any other object of the MOMENT Ontology. The class itself only contains two properties *description* and *nameID* to identify and describe the location, but it does not add any other information of the location, providing only the inheritance point for the Location definitions as shown in Figure 3.

Two classes derive from it, *LogicalLocation* and *PhysicalLocation*.*PhysicalLocation* is used to define where the measurement was done or where a Network object is placed (the city where a router is located, the geographical coordinates of some network node, etc) and *LogicalLocation* represents the location where measurements are stored and could be accessible.

[*] The defined ontologies are available at
http://www.fp7-moment.eu/index.php?option=com_content&view=article&id=103

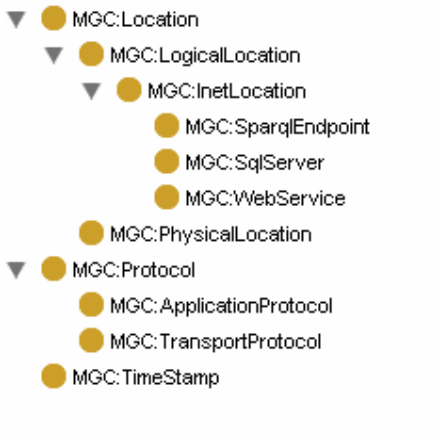


Figure 3. MOMENT upper ontology location hierarchy

3.2 The Metadata Ontology

This ontology describes the information about the measurements. It was designed, based on DatCat object types, around the *DataMetadata* class, which contains basic information about the measurement: human readable description of the measurement data, its size, a hash of the contents (for data integrity reasons) and a textual description of the creation process as well as other significant information of the data sources such as: *FileFormat* (for non SPARQL data-sources the system is capable of accessing them via the Invocation Service), *LogisticLocation* (which represents where is the measurement stored) and *Contributor* (the person or organization responsible of the data in the system).

For every measurement described in the Data ontology, or at least every collection of related measurements, one individual of *DataMetadata* class should exist in the Metadata ontology with all the information of the referenced measurement. Also metadata can describe external repositories but it will be impossible to treat semantically the measurement data referenced unless the data source is mapped to the ontology somehow.

Also information about the result set each data source returns is stored in the *DataMetadata* Object to let the user select which results are going to be shown and also to request data transformations (Semantic Data Transformation Service) or Analysis of the results (Analysis Service) before having the results, which is very important in the MOMENT scenario due the high volume of information contained in the data sources.

In the new version, the ontology has been better aligned with DatCat's object types allowing the Semantic Metadata Registry (which stores the information of the data sources expressed in terms of the Metadata ontology) to import semi-automatically all the information contained in DatCat providing MOMENT will have at least the same information as DatCat.

3.3 The Data Ontology

The Data Ontology has been designed based on the input from MOMENT partners. The previous version of the ontology was designed having a strict hierarchy for network measurements, imposing high requirements for a data source, and at the end complex mappings where being produced which not always represented the data sources very well. Also some modules of the MOMENT architecture which were previously omitted from the analysis of requirements of the ontology were willingly to use the layer of unified information the ontology provides; such modules as the Graphical user interface and the Workflow Service are now benefiting of the new version of the ontology.

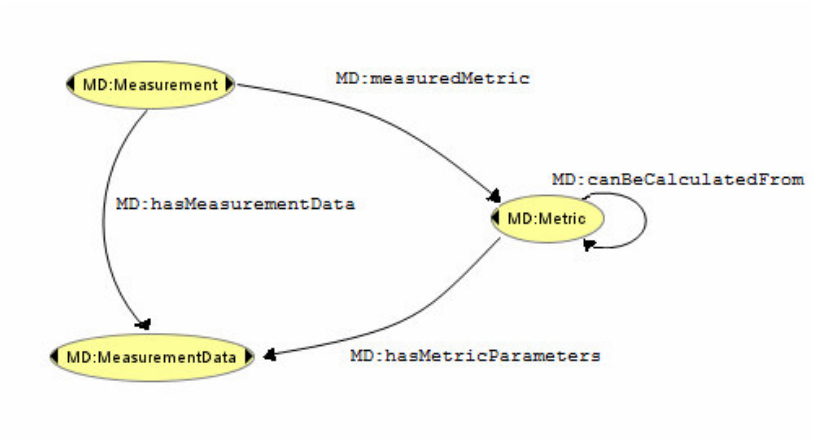


Figure 4. MOMENT Data ontology structure

Now the ontology is designed giving more importance to what information the measurement has and not on how measurements are categorized under a fixed hierarchy. A *Measurement* instance only contains an Id (Not the rdf:id, the id which identifies the measurement in the data source) and a time stamp about when the measurement was performed.

All the additional information and values the measurement has are modelled through the *hasMeasurementData* property and instances of *MeasurementData* subclasses. There is a subclass of MeasurementData (Figure 5.) for every possible measurement value: source node, delay value, number of hops, tool configuration parameter, scenario configuration parameter, etc; organized based on the concept being measured: parameters of the measurement, network physical information, statistical information or more general information.

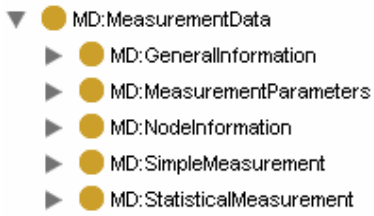


Figure 5. *MeasurementData* basic hierarchy

High level concepts from network measurements such as *Route*, *Capacity*, etc. which cannot be determined with single values are represented with the Metric class. Each instance of *Metric*

will define the required *MeasurementData* a measurement needs to have in order to be measuring that concept. *Measurement* instances can explicitly state which metrics are measuring but it can also be inferred from the *MeasurementData* instances they have. This is extremely useful to allow researchers to query MOMENT architecture for *Measurements* of a certain *Metric* without actually knowing which values the query will return. Also if a data source has more information than the expected (specified with the *hasMetricAttributes* property) it will not be lost.

Also, an important subset of network measurements are from well-known measurement tools or techniques, like ping or traceroute; and the MOMENT ontology should be able to represent them. Those measurements are described as subclasses of *Measurement* but not adding particular properties for each of the specific results they return, just adding constraints about the *MeasurementData* a measurement should have (i.e :Traceroute is described in Figure 6. with its set of constraints).

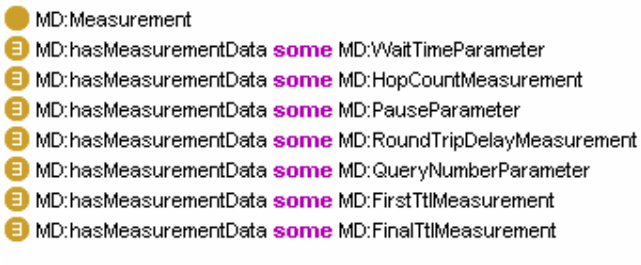


Figure 6. Example of property constrain: MD:Traceroute

3.4 The Anonymization Ontology

This model represents the "common vocabulary" that shall fit all possible anonymization contexts, i.e. the minimum superset of concepts and/or functionalities that describes the "anonymization of network measurements" domain. While designing the Anonymization Ontology we have tried to maintain an high level of generality, at various different levels: it is well known that generic anonymization schemes are difficult to design, since different organizations have different needs. Thus we have investigated the common ideas, and found that always including the purpose the data will be used for, and the user role, into the obfuscation process is a key concept of all desirable anonymization schemes. We came up with a model where the PolicyObject is the cornerstone of MOMENT's Anonymization Ontology. It is an N-ary relation that associates a number of *UserRoles*, a number of *UsagePurposes*, applied to a number of *PrivacyScopes*. Moreover, the *PolicyObject* specifies one well-defined *AnonymizationStrategy* and an associated *AcceptableUsePolicy*. The *AnonymizationStrategy* is composed of a group of *AnonymizationTargets* and an external *AnonymizationBackend* to support and implement that strategy.

Another quite general concept we have introduced is modeled by the *DataAge* value partition class, which has been designed as a general characteristic of the *PrivacyScope*, and represents the mean aging (OLD, RECENT, NEW) of a single specific measurement or measurement set. This enables to apply looser obfuscation algorithms to data which becomes less sensitive as time passes by. Practically this is achieved by means of fuzzy-logic-based mapping techniques. A fuzzy membership function is designed, and tuned, in order to map a numerical value coming from the measurement database (representing the exact date at which the data has been taken)

into a vague concept represented by the linguistic labels NEW, OLD, RECENT. This mapping is quite flexible, and although in principle it could also depend on the context of the specific measurement database, we tried to capture the "absolute" meaning of, say, "recent measurement" (thus only depending on actual date and absolute timescale) as it is perceived by the human observer. This helps maintaining an abstract common model for obfuscating data differently, based on its ageing, according to the law and not to the specific requirements of a specific data owner.

Once those general concepts and inner workings of the model are established, the Anonymization Ontology needs to be fully integrated with the Data Ontology. This leads us to a different level of generality. The thing to avoid is to embed into the Anonymization Ontology a direct reference to the specific Data Ontology classes. This design choice is needed to avoid rewriting the Anonymization model each time the Data model changes, and to achieve a perfect decoupling between the description of how IP traffic measurements have to be obfuscated and the description of IP traffic measurements themselves.

This design goal has lead to a complex and flexible mapping between the *AnonymizationTargets* and the fields of the *Measurement* class. This mapping has been designed and implemented both in OWL and by means of supporting Java code that extends the Ontology with a dynamic behavior, so that the correct fields can be handled by the external *AnonymizationBackend* which performs the obfuscation task. Basically the Anonymization Ontology has been designed so that it does not contain explicit references to the *Data Ontology* classes and their names. Those references are contained in the Java code that imports the *Data Ontology* and does the matching.

To clarify this approach, the following Figure is useful. It represents a selection of the vocabulary which is private to the Anonymization Ontology. The sub-classes of the *AnonymizationTarget* class describe the generic Anonymization Targets as they are modeled from the point of view of a typical privacy-protection framework.

Figure 7. Anonymization Ontology

Based on the input parameter (i.e. an instance of the Data Ontology, a specific Measurement) of the obfuscation function, a specific target is in that anonymization sub-tree has to be selected. This selection is based on a rather flexible string matching, which would be cumbersome or even impossible to hardwire in the anonymization model itself, and has thus been implemented via Java support functions.

In the end, this effort has been highly rewarded when we had to face a substantial change in the model of the Data Ontology, and that change was seamlessly absorbed by the adjustment of some matching functions in the Java code, rather than by a painful revisiting of the whole anonymization properties.

Another indirection step is achieved by noting that the Anonymization Targets can be differently interpreted in the context of the different Anonymization Backends, namely in the context of a specific external anonymization framework or APIs.

As an example, the AAPI anonymization framework [22] has been chosen as a first candidate able to provide the necessary external data-obfuscation functions upon which implementation of a real proof-of-concept Anonymization Backend is possible. A Java version of this library has recently been produced by its authors and allows smoother integration with the overall architecture of the modules developed in the present work.

4. ONTOLOGY SET UP

To provide the homogeneous interface over several measurement databases in MOMENT, a semantic infrastructure was conceived in [23]. To implement it we have used D2R-server [24], but applying the designed ontology instead of the automatically generated by the schema inspector described below (section 4.1).

D2R-server is an existing Java application to publish relational databases and allow HTML and RDF browsers to navigate its contents. Although RDB2RDF (Relational Database to RDF) applications are in an early stage of development and query optimization for SPARQL (RDF Query language) is an open field for research, the advantages those tools give removing the necessity of ad-hoc interfaces for data-sources makes very easy for any data source maintainer to publish their information via RDF easily and more willingly to participate in the Semantic Web.

W3C RDB2RDF Incubator Group performed a survey [25] on existing approaches to the problem and a comparison of the main advantages of each tool was presented. From that comparison and the fact that D2R is the most active project of them all with periodical releases of bug-fixes, open source community and an active mailing list we decided to use it in the MOMENT architecture.

4.1 D2R Generated Ontology

D2R schema inspector automatically builds the RDF description of the database generating a Notation3 (N3) file with the individuals of d2r-map ontology [26] representing the tables and columns. This automatic application assumes that each table represents a particular rdf:Class, where columns represent the stated properties of the Class. It assigns names for the created classes and properties using the tables and columns ones, generating the default vocabulary of the database as RDF. This vocabulary has to be overridden, as shown in Figure 8, by the database maintainer, matching the vocabulary with the MOMENT Ontologies; this process, called the Mapping Process, is explained below.

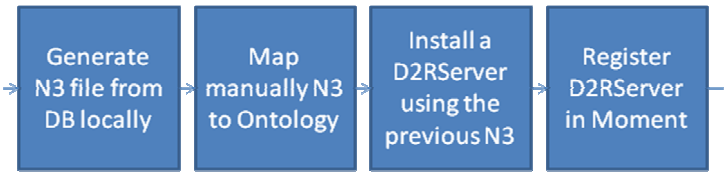


Figure 8. Mapping process

4.2 The Mapping Process

The mapping between the database and the MOMENT ontology (for now on MO) is the process of matching database tables and columns to the MO by the substitution of the vocabulary generated automatically by D2R. More precisely:

Given the **MO** is complete (assuming it could describe every measurement possible in a network) the process of mapping the Database vocabulary (from now on **DV**) is a function **M** from the **DV** properties to the **MO** properties where for each property $\mathbf{v} \in \mathbf{DV}$, it assigns a **MO** property $\mathbf{m_1}$ where the semantics of **v** are a specialization of $\mathbf{m_1}$, and it does not exists another property $\mathbf{m_2} \in \mathbf{MO}$ such that its specialization to **DV** property adds less semantic content than $\mathbf{m_1}$.

Notice that this does not entails necessarily that it exists a function $\mathbf{M^c}$ between **DV** classes and **MO** classes. **DV** classes are defined by the properties they satisfy according to the mapping **M** so there will be cases where it exists a **MO** class with the same property restrictions and such a map between **DV** class and **MO** class will exits but there will be times when the properties from the **DV** class maps to properties in the intersection of multiple ontology classes.

The new structure of the ontology makes easier the mapping process because each table in the database will represented as a *Measurement* instance and then every column the measurement has is mapped to a certain subclass of *MeasurementData* which represents the concept of the column value. Units of the properties are defined by default in the *MeasurementData* class but this can be overridden if the database has a different unit.

A mapping GUI was developed to help in the mapping process. It presents the database vocabulary and the MOMENT ontology as a pair of trees to simplify the process of mapping, hiding the representation structure (XML or N3) of the vocabularies. This application enables matching and replacing D2R vocabulary names clicking on the desired concept from the MOMENT ontology.

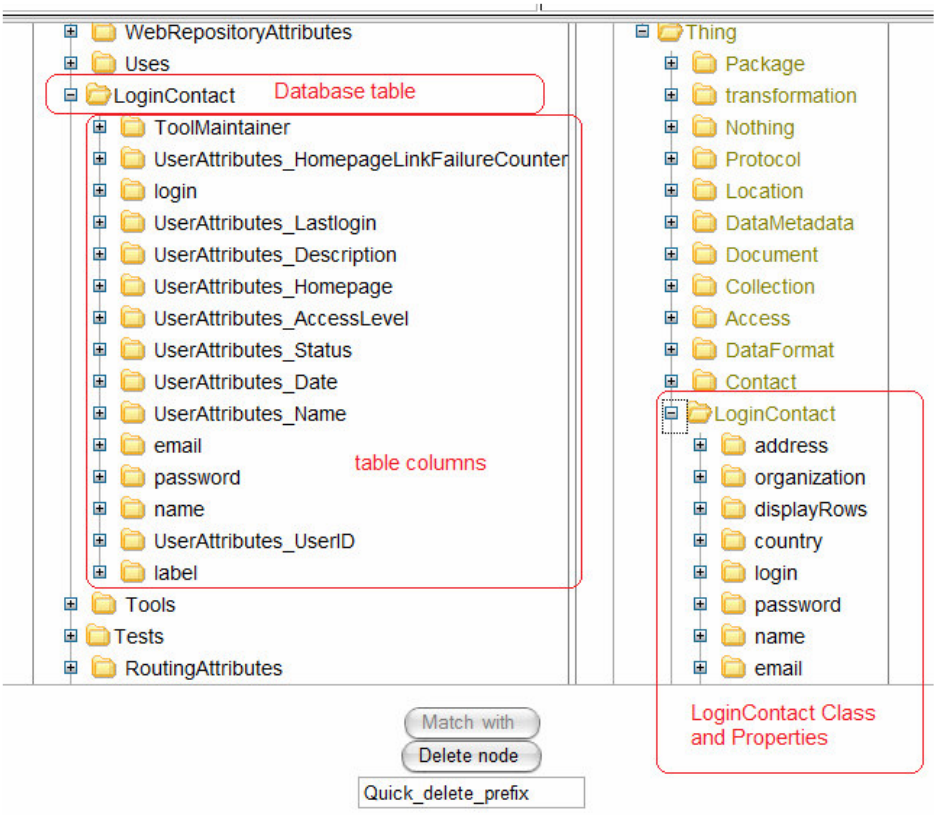


Figure 9. GUI for the Mapping process, representing the *LoginContact* class and a database partially mapped to the class properties.

4.3 Configuration with multiple databases.

Finally the functionality of D2R has been extended to enable a distributed architecture where multiple D2R servers combine their results. When multiple databases are involved (and MOMENT objective is to handle several network measurement databases) multiple mapping files are generated so it is necessary to provide methods to merge all this information together.

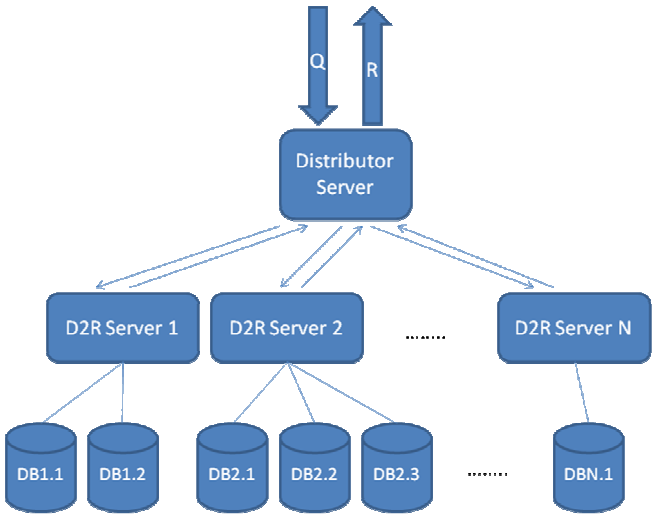


Figure 10. Distributed Architecture

The hierarchical architecture proposed relies on a central server which distributes the queries among the registered D2R-servers and then merges the results from all of them as shown in Figure 10. Also, D2R provides mechanisms to configure it with multiple databases so a first unification of the information is done in each node of the distributed architecture (D2R Server *i*, in Figure 10).

SPARQL query language uses graph patterns which are evaluated against a RDF Graph in a datasheet. Those graph patterns can be expressed in terms of simpler ones, triple patterns, which are similar to RDF statements but with variables. A SPARQL query matches a subgraph of the datasheet and returns the bindings for the selected variables if all the triple patterns in the query match. If one of the triple patterns does not match, the whole query fails.

This is particularly dangerous in the MOMENT scenario because it imposes a restriction on the structure of the queries because the information is partitioned horizontally (An instance of a class can have its properties stored in several databases) and vertically (instances of the same class can be stored in different databases) so the SQIS should ask each D2R server only the information it can answer and store the results in a temporal knowledge base which will be queried at last, joining all the information.

Previous work has been done in this field by other projects like DARQ [27] and SemWIQ [28] but both are in an early stage of development giving few warranties of performance with the large datasheets MOMENT project needs to handle.

DARQ generates the sub-queries based on a configuration file which states the statements stored in each endpoint and an estimate of the result size to perform query optimization. Some limitations with DARQ are:

- The query optimizer does not perform very well nor support many endpoint and triples which is essential in the MOMENT architecture as the project is targeted to handle as many data-sources as possible.

- The initial estimate of number of results cannot be changed, and in network measurement databases, new measurements are added very frequently, being impossible to give a fixed estimate for the result size.

- Query rewriting and optimization algorithm is based on predicates which means that predicates must be bound in each triple pattern, limiting the expressiveness of SPARQL. Also blank nodes are not supported avoiding the usage of collections in RDF.

SemWIQ tries to solve the limitations of DARQ, removing the configuration file and asking each data-source the information contained in the knowledge base. The current version of SemWIQ queries for all the instances of classes and properties in the endpoint to generate statistics of each one and to be able to generate the sub-queries and the optimization plan. MOMENT databases contain millions of rows which translated using D2R-server gives dozens of millions of triples; querying for them periodically is almost impossible, given the analyzed response times of D2R, as shown below.

Performance measurements were done using the free tool Firebug (a Firefox plug-in for web-developers which computes web page load times) under Windows XP on an Intel Pentium D 2,8GHz with 1GByte of RAM. D2R-server was configured in both versions as a stand-alone server in an Intel Dual Core server with 2GByte of RAM assigned to the Java Virtual Machine and the mapping file describing only one datasheet of a remote database. The new version of D2R was executed with the "*fast*" flag to enable all possible optimizations. To determine the response-time function of each D2R-server version, the test consisted on limiting the maximum result size which D2R-server delivers to the user using a configuration parameter. These measurements provide an approximate idea of the performance of D2R-server, but in any case their exactitude depends on many conditions such as server bandwidth to data-sources and user bandwidth to D2R-server, which could not be measured but are assumed to be similar in both tested versions.

As it can be observed in Figure 11, the improvement of the "*fast*" flag in d2rserver 0.7 is significant but in the same order of magnitude as the previous version. The response-time function behaves in a similar way in both cases, but it is important to notice that is not convenient to request queries without imposing a limit under 75,000 rows due to the high response times. Assuming the results follow a linear behavior, each query that SemWIQ does to retrieve each knowledge base will last a day for each server, which is not feasible for MOMENT system.

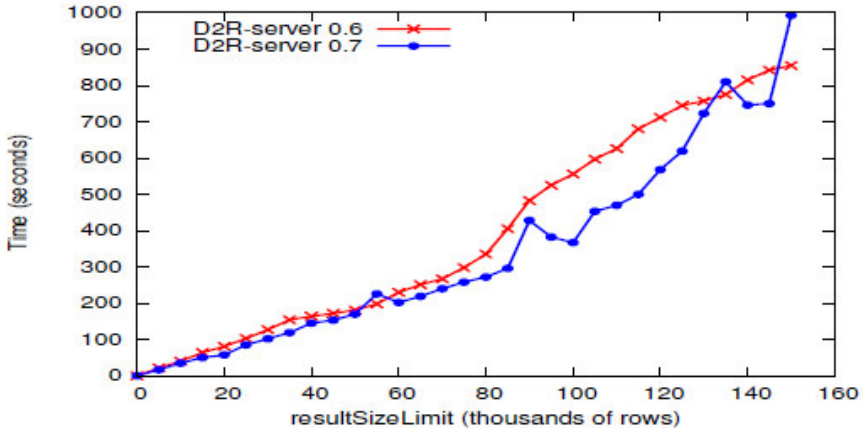


Figure 11. D2R-Server performance comparison between versions

The idea behind the SQIS is to use the information we have from the mapping files, which states what classes and properties are stored in each D2R SPARQL endpoint in a similar way as DARQ does, but also be able to distribute the query in a more general way as SemWIQ does.

When a new D2R server is registered in the architecture, it sends its N3 information file to the SQIS, which parses the file, storing the *propertyBridges* and *classMap* mapped URI's in a Database. This way, when a new query is received, the SQIS rewrites the query for each D2R server asking only for the information available at each endpoint. When it receives the individual results it inserts all the bindings as RDF statements (using the original query) into a temporal JENA [29] knowledge base and at the end, the original query is requested to the temporal knowledge base and the results are delivered.

Also this architecture provides solutions to different access policies and security problems such as restricting the access to the data, because each D2R-server can be configured to ignore some tables of the database, publishing to the system only the desired measurements. Also, the connection to D2R-server is over HTTP so it can be configured over HTTPS to deny external entities to alter the results.

5. CONCLUSIONS

An ontology that comprises all aspects of the IP measurement domain has been developed: it includes a Data ontology as well as a Metadata ontology, Upper ontology and an Anonymization ontology. Their application to build a mediator, to access multiple and heterogeneous measurement databases and tools, is being tested in the MOMENT project design and implementation activities.

The semantic approach to enable complex queries about IP traffic has become then easier. The final purpose of defining the so-called MOMENT Ontologies has needed intensive work: software tools, like D2R, and mapping algorithms were used to complete them from previous information models underlying traffic monitoring infrastructures already available. The procedure to map eventual new data sources or measurement tools has been described so that the established ontology can be easily adopted by other developers.

Further work needs to be done in query execution planning for SPARQL as our algorithm only rewrites the query for each data source based on the configuration file and not in previous query results which would be able to give an estimation about the knowledge base size. Also D2R-server memory and performance limitations should be improved using the built-in capabilities of SPARQL to split results using LIMIT and OFFSET keywords.

ACKNOWLEDGEMENTS

The authors thank the partial support of the EU ICT MOMENT Collaborative Project (Grant Agreement No.215225) and all partners in the project for their comments and useful support.

REFERENCES

[1] MOMENT: Monitoring and Measurement in the Next Generation Technologies (http://www. fp7-moment.eu)

[2] Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge Acquisition 5(2) (1993), pp. 199-220.

[3] Ferreiro, A., Fichtel, T., López de Vergara, J., Mátray, P., Strohmeier, F., Tropea, G., Weinsberg, U.: "Semantic Unified Access to Traffic Measurement systems for Internet Monitoring Service", ICTMobileSummit2009, Santander (Spain), June 10-12, 2009

[4] OGF Network Measurement Working group (NMWG): http://nmwg.internet2.edu/

[5] W3C Time Ontology in OWL: http://www.w3.org/TR/owl-time/

[6] Units Ontology: http://sweet.jpl.nasa.gov/ontology/units.owl

[7] 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

[8] 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)

[9] Claise, B.: Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of IP Traffic Flow Information. RFC 5101 (January 2008).

[10] Waldbusser, S., Cole, R., Kalbeisch, C., Romascanu, D.: Introduction to the Remote Monitoring (RMON) Family of MIB Modules. RFC 3577 (Informational) (August 2003)

[11] Zseby, T., Molina, M., Du_eld, N., Niccolini, S., Raspall, F.: Speci_cation 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).

[12] Paxson, V., Almes, G., Mahdavi, J., Mathis, M.: Framework for IP Performance Metrics. RFC 2330 (Informational) (May 1998)

[13] 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)

[14] van der Vlist, E.: RELAX NG. O'Reilly Media, Inc. (2003)

[15] Hanemann, A., Boote, J., Boyd, E., Durand, J., Kudarimoti, L., Lapacz, R., Swany, M., Trocha, S., Zurawski, J.: Perfsonar: A service oriented architecture for multidomain network monitoring. Lecture Notes in Computer Science 3826 (December 2005) 241-254

[16] Dabrowski, M., Sliwinski, J., Burakowski, W.: The MOME Measurement Database, Krajowe Sympozjum Telekomunikacji, 2005

[17] Monitoring Ontology for IP traffic ETSI working group : http://portal.etsi.org/moi

[18] http://portal.etsi.org/Portal_Common/home.asp

[19] Moraes, P.S.; Sampaio, L.N.; Monteiro, J.A.S.; Portnoi, M., "MonONTO: A Domain Ontology for Network Monitoring and Recommendation for Advanced Internet Applications Users," *Network Operations and Management Symposium Workshops, 2008. NOMS Workshops 2008. IEEE* , vol., no., pp.116-123, 7-11 April 2008

[20] Salvador, A., López de Vergara, J.E., Tropea, G., Blefari, N., Ferreiro, A.: Ontology design and implementation for IP networks monitorning First International Workshop on Web & Semantic Technology

[21] Salvador, A., Katsu, A., López de Vergara, J.E., Aracil, J.: Designing a network measurement ontology for a semantically driven architecture. Workshop Future Internet Design: Aspects of network monitoring, privacy and security within the 4th FP7-FP6 concertion meeting

[22] Koukis, D. and Antonatos, S. and Antoniades, D. and Markatos, E.P. and Trimintzios, P.: A Generic Anonymization Framework for Network Traffic. In: Proc. IEEE International Conference on Communications, 2006. ICC '06.

[23] 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, 10-11 July 2008, Salzburg, Austria

[24] D2R Server: http://www4.wiwiss.fu-berlin.de/bizer/d2r-server

[25] Satya S. Sahoo, Wolfgang Halb, Sebastian Hellmann, Kingsley Idehen, Ted Thibodeau Jr, Sören Auer, Juan Sequeda, Ahmed Ezzat: A Survey of Current Approaches for Mapping of Relational Databases to RDF , 2009-01-31.

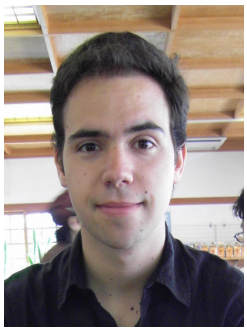
[26] D2R Map ontology: http://www4.wiwiss.fu-berlin.de/bizer/pub/www2003-D2R-Map.pdf

[27] Quilitz, B., Leser, U.: Querying Distributed RDF Data Sources with SPARQL, European Semantic Web Conference ESWC2008

[28] Langegger, A.,Wöß, W., Blöchl, M.: SemWIQ A Semantic Web Middleware for Virtual Data Integration on the Web, European Semantic Web Conference ESWC2008

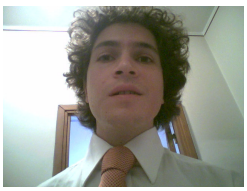
[29] McBride B. : JENA: Implementing the RDF Model and Syntax Specification, http://www.hpl.hp.com/semweb/publications.htm

Authors



Alfredo Salvador Gonzalez is currently an Ph.D student at Universidad Autónoma de Madrid. He received his M. Sc. Degree in september 2009 but has already published three articles in several workshops from the results of his work in Excellence Scholarships and Collaboration Scholarships during his degree. He is participating in the EU FP7 MOMENT project as an assistant researcher and developer of the semantic core of the application. His current interest, other than network monitoring and managing are: Information models, Logic Programming and distributed datasources integration.

Jorge E. López de Vergara is currently an associate professor in the Computer Science Department of the Universidad Autónoma de Madrid. He received his M.Sc. Degree in telecommunications from the Technical University of Madrid in 1998 and finished his Ph.D. in telematics engineering at the same university in 2003, where he held a research grant. He has participated in several Spanish and EU research projects, and has authored more than 50 papers in international conferences and journals. His current research topics include network and service management and monitoring, with special focus on management information models.



Giuseppe Tropea received his Laurea Degree in Informatics Engineering (spec. Computer Systems) from University of Catania, Italy, in October 2002. Since 2006 he is with Consorzio Nazionale Inter-universitario per le Telecomunicazioni (CNIT), inside the DISCREET and MOMENT European

projects, dealing with privacy protection in digital services and obfuscation of sensible IP traffic data by means of semantic, ontology-bases approaches. He is active in the MOI (Monitoring Ontology for IP traffic) ETSI Industry Specification Group. Ha has also been technical head for the FuLL (Fuzzy Logic and Language) Italian research project involving software companies and CNR's Institute of Computational Linguistics and Knowledge Discovery and Delivery Laboratory. Project's goal was a new technology for Natural Language Human Interaction with databases and structured data, based on spatially and temporally enhanced semantic Ontologies. His research interests include neural networks, software modelling, semantic ontologies and natural language parsers, as well as data mining and GIS applications.

Nicola Blefari-Melazzi earned his Ph.D. in Information and Communication Engineering in 1994 at the University of Roma, "La Sapienza", Italy. From November 2003 to November 2008 he has been the director of the PhD program in "Telecommunications and Microelectronic Engineering". Dr. Blefari-Melazzi has been the coordinator of two cooperative projects funded by the European Union and has played the role of evaluator for many research proposals and of reviewer for numerous EU and ITEA sponsored projects. Dr. Blefari-Melazzi has been a member of the Technical Program Committee for several IEEE Conferences. He has served as reviewer and session chair for IEEE Conferences and as reviewer and guest editor for IEEE Journals. He is also conducting research on multimedia traffic modelling, mobile and personal communications, quality of service in the Internet, ubiquitous computing, reconfigurable systems and networks, service personalization, autonomic computing.

Ángel Ferreiro was graduated in Physics in 1981 (Complutense University, Madrid) and has been teaching since then at different Spanish universities (UPM, UC3M, UAM) while working in several high technology companies like IBM and AT&T. In 1991 he joined Telefonica I+D after spending four years in the LETI (Grenoble, France) to develop his thesis (PhD degree obtained in 1989). Dr. Ferreiro has cooperated with several research groups in Spain and abroad, has published over twenty scientific papers: In the *Journal of Magnetism & Magnetic Materials, Journal of Physics E, International Journal of Mathematics and Education Science Technology, American Journal of Physics and Journal of Lightwave Technology.* He has also presented works in more than twenty international meetings and workshops and recently he has contributed, as author, to the book "Core and Metro Networks" (John Willey & Sons, 2009). Contributions to seminars, a couple of patents and a few technical notes complete his CV. Dr. Ferreiro has been working in EU-funded projects since more than 25 years; currently he is working in research projects about IP traffic measurement and QoS monitoring, he is chairman of the ETSI ISG "Measurement Ontology for IP traffic" and also cooperates actively in the MANA group of Future Internet Assembly.